Foundations of Multisensory Artificial Intelligence

Paul Pu Liang

May 2024
CMU-ML-24-103

Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:
Louis-Philippe Morency, Co-chair
Ruslan Salakhutdinov, Co-chair
Manuel Blum
Lenore Blum (UC Berkeley)
Trevor Darrell (UC Berkeley)

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

This research was funded by: National Science Foundation awards IIS1722822 and IIS1750439; National Institutes of Health awards R01MH096951 and U01MH116923; graduate fellowships from Meta Platforms and Siebel Scholars; and grants from Meta Platforms, Nippon Telegraph and Telephone Corporation, Oculus VR, and Samsung Electronics.

Keywords: Multimodal Machine Learning, Multisensory Artificial Intelligence, Deep Learning, Information Theory, Quantification, Generalization, Affective Computing, AI and Healthcare

Abstract

Building multisensory artificial intelligence systems that learn from multiple sensory inputs such as text, speech, video, real-world sensors, wearable devices, and medical data holds great promise for impact in many scientific areas with practical benefits, such as in supporting human health and well-being, enabling multimedia content processing, and enhancing real-world autonomous agents.

However, the breadth of progress in multimodal research has made it difficult to identify the common themes and open questions in the field. By synthesizing a range of theoretical frameworks and application domains, this thesis aims to advance the foundations of multimodal machine learning. We start by defining three key principles of modality *heterogeneity*, *connections*, and *interactions* often present in multimodal problems [371]. Using these principles as a foundation, we propose a taxonomy of six core challenges in multimodal research: *representation*, *alignment*, *reasoning*, *generation*, *transference*, and *quantification*. Recent technical achievements will be presented through this taxonomy, allowing researchers to understand the similarities and differences across approaches, and identifying open problems for future research.

The bulk of the thesis covers our recent progress towards tackling two key problems in multimodal learning: the machine learning foundations of multimodal interactions, as well as practical methods for building multisensory foundation models that generalize to many modalities and tasks in the real world.

In the first part, we study the foundations of multimodal interactions: the basic principle of how modalities combine to give rise to new information for a task. We present a theoretical framework formalizing how *modalities interact* with each other to give rise to new information for a task, such as sarcasm identified from the incongruity between spoken words and vocal expressions [372]. Using this theoretical framework, we propose two practical estimators to quantify the interactions in real-world datasets. Quantifying the types of interactions a multimodal task requires enables researchers to decide which modality to collect [376], design suitable approaches to learn these interactions [374], and analyze whether their model has succeeded in learning [375].

In the second part, we study the design of practical multimodal foundation models that generalize over many modalities and tasks, which presents a step toward grounding large language models to real-world sensory modalities. We first introduce MULTIBENCH, a unified large-scale benchmark across a wide range of modalities, tasks, and research areas [367]. We will also present the *cross-modal attention* [101, 359] and *multimodal transformer* [613] architectures that now underpin many of today's multimodal foundation models. Scaling these architectures on MULTIBENCH enables the creation of general-purpose multimodal multitask models across a variety of tasks, and we have collaborated broadly with practitioners to apply these models for real-world impact on affective computing, mental health, and cancer prognosis.

We conclude this thesis by discussing how future work can leverage these ideas toward more general, interactive, and safe multimodal artificial intelligence.

Acknowledgments

I owe my greatest acknowledgments to my advisors, mentors, and thesis committee members for their invaluable guidance during my PhD. To Louis-Philippe Morency and Ruslan Salakutdinov, who have closely guided my research and personal development at every stage over the past 5 years. LP has mentored me closely in all aspects of research - brainstorming ideas, idea execution, and written and oral presentations. Some of the best memories I've had during my PhD have been whiteboard brainstorming sessions, coming up with good names for problems and models, and collaboratively figuring out the best ways to visually depict technical concepts. Russ's incredibly sharp insight and keen eye for impactful problems has shaped my thinking and forced me to work on problems that matter in practice, and I've thoroughly enjoyed our recent push towards interactive multimodal agents with other folks in the group and at CMU. Thank you LP and Russ for additionally giving me the opportunity to co-instruct and guest lecture CMU courses multimodal ML, deep learning, and socially intelligent AI. I also had the pleasure of working closely with Manuel Blum and Lenore Blum during the senior years of my PhD. I have learned a lot from our discussions at the intersection of artificial intelligence, consciousness, and neuroscience, which have changed how I look at long-term problems and approach them. Manuel and Lenore have also inspired me to think big and make broad impact across CS and beyond, giving me many opportunities to communicate my ideas and contributions to a wide audience in neuroscience, psychology, and more. Finally, Trevor Darrell has been a source of inspiration as a senior faculty and has given me great advice for my PhD research and broadly for my research career. Some of his early works in multimodal machine learning and multimodal interaction are still some of my favorite works in this space.

Beyond my committee members, I would like to acknowledge other CMU faculty and students with whom I've had fruitful collaborations, discussions, and received helpful feedback on ideas, paper drafts, and presentations. Fantastic students in LP and Russ's research groups: Hubert Tsai, Amir Zadeh, Chaitanya Ahuja, Volkan Cirik, Torsten Wortwein, Martin Ma, Alex Wilf, Leena Mathur, Victoria Lin, Yousouf Kebe, Devendra Chaplot, Bhuwan Dhingra, Lisa Lee, Shrimai Prabhumoye, Ben Eysenbach, Jing Yu Koh, Minji Yoon, Brandon Trabucco, Murtaza Dalal, Yue Wu, and Kelly He. In addition, Hai Pham, Shaojie Bai, and their advisors Barnabas Poczos, and Zico Kolter with whom I did some early work in multimodal representation learning. Yonatan Bisk, Daniel Fried, Albert Gu, Zack Lipton, Tom Mitchell, Graham Neubig, Mayank Goel, and Haiyi Zhu who have given me a lot of advice (both personal and professional) over the years. Roni Rosenfeld, Ryan Tibshirani, and Tai Sing Lee who mentored me on undergraduate research projects at CMU. Finally, CMU Machine Learning Department and Language Technologies Institute are some of the best places to do AI research, and this could not be possible without the fantastic support from staff like Diane Stidle, Dorothy Holland-Minkley, and John Friday.

Some folks outside CMU I would like to thank include: Faisal Mahmood's group at Harvard Medical School, especially students Richard Chen, Guillaume Jaume, and Anurag Vadiya for a series of fruitful collaborations regarding multimodal computational pathology; David Brent at UPMC, Nicholas Allen at University of Oregon, and Randy Auerbach at Columbia University for collaborations on daily mood assessment, markers of suicide ideation, and mobile health; Liangqiong Qu, Yuyin Zhou, Daniel Rubin, and James Zou at Stanford for investigations into multimodal and federated learning for biomedical applications; and most recently Jack Hessel, Yejin Choi, and Jae Sung Park at University of Washington/AI2 for many discussions regarding research and projects on vision-language commonsense reasoning.

I was also lucky to be mentored by several fantastic researchers in industry labs during my internships. To Manzil Zaheer at Google, you have made me a more mature researcher by reminding me to focus deeply on problems rather than jumping around during my junior researcher days. Our close collaborations have also strengthened my expertise in both fundamental and practical machine learning. To Yuke Zhu, Anima Anandkumar, and Sanja Fidler at Nvidia, you have done a great job setting up a vibrant and flexible research environment at Nvidia and I have learned a lot about the latest progress in multisensor robotics, AI for science, and vision-language models from our collaborations. To Makoto Yamada and Qibin Zhao at Riken AIP, where I learned more about tensors and kernels for multimodal learning. To Brandon Amos, Tim Rocktäschel, and Ed Grefenstette at Facebook AI, where I learned a lot about optimization, control, and reinforcement learning. And finally, to Dani Yogatama, Lisa Anne Hendricks and Aida Nematzadeh at DeepMind, where I gained practical experience training large-scale multimodal foundation models.

The most personally rewarding part of my PhD was definitely the many undergraduate, masters, and PhD students I have had the pleasure of advising - both at CMU and around the world: Adejuwon Fasanya, Akshay Goindani, Aviv Bick, Arav Agarwal, Chengfeng Mao, Chiyu Wu, Dong Won Lee, Edmund Tong, Gunjan Chhablani, Haofei Yu, Haoli Yin, Holmes Wu, Irene Li, Jiewen Hu, Jingyi Zhang, Jivat Neet, Katrina Jiao, Marian Qian, Peter Wu, Rana Shahroz, Richard Zhu, Rohan Pandey, Rulin Shao, Samuael Adnew, Samuel Yu, Seong Hyeon Park, Shentong Mo, Siyuan Wu, Talha Chafekar, Terrance Liu, Xiang Fan, Xiangru Tang, Yao Chong Lim, Ying Shen, Yiwei Lyu, Yudong Liu, Yun Cheng, Yuxin Xiao, Zhun Liu, Zihao Deng, Ziyin Liu. All of you have taught me so much and become experts in your own fields. I'm delighted to see all of you make great strides in PhD studies and industry, and look forward to hearing about your successes in the future.

Finally, I could not have done all this without the close support of my family and friends, especially from my mom, dad, sister, and grandparents, Jane, Truffle, and Tigger, Jane's family, close friends Chun Kai Ling, Yue Niu, Raahul Sriram, Dylan Sam, Rattana Pukdee, Jennifer Hsia, Clara Na, Cindy Wu, Pratyush Maini, Ananye Agarwal, Yiding Jiang, Sam Sokota, Alex Wilf, Leena Mathur, Yiwei Lyu, Chirag Gupta, Tom Yan, Helen Zhou, Manzil Zaheer, and many more.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Multimodal artificial intelligence is a vibrant multi-disciplinary research field that aims to design computer agents that can perceive, reason, and interact through multiple communicative modalities, including linguistic, acoustic, visual, tactile, sensory, and physiological messages [46, 371]. Multimodal AI systems can bring great impact in many scientific areas with practical benefits, such as in supporting human health and well-being [360, 427, 715], enabling multimedia content processing [11, 485, 513], and enhancing real-world autonomous agents [63, 93, 334, 522, 545].

However, the breadth of progress in multimodal research has made it difficult to identify the common themes and open questions in the field. By synthesizing a broad range of theoretical frameworks and application domains from both historical and recent perspectives, this thesis is designed to advance the theoretical and computational foundations of multimodal machine learning. We start by defining three key principles of modality *heterogeneity*, *connections*, and *interactions* often present in multimodal problems which brings unique challenges to machine learning. The heterogeneity of multimodal data makes learning challenging, for example, language is often seen as symbolic while audio and video are represented as continuous signals. At the same time, these modalities contain overlapping connected information, and interact to give rise to new information relevant for a task. It is crucial to learn these connections and interactions for systems to perform well. Using these principles as a foundation, we propose a taxonomy of six core challenges in multimodal research: *representation*, *alignment*, *reasoning*, *generation*, *transference*, and *quantification*. Recent technical achievements will be presented through the lens of this taxonomy, allowing researchers to understand the similarities and differences across new approaches, and enabling us to identify key open problems for future research.

Using our taxonomy for multimodal machine learning, we highlight two key challenges that are important for progress in multimodal learning: (1) building the **foundations** of multimodal interactions so we can quantify the interactions present in datasets and model these interactions correctly using machine learning methods, and (2) constructing multimodal models and datasets that enable **generalization** across a large number of modalities and tasks for real-world societal impact (Figure 1.1).

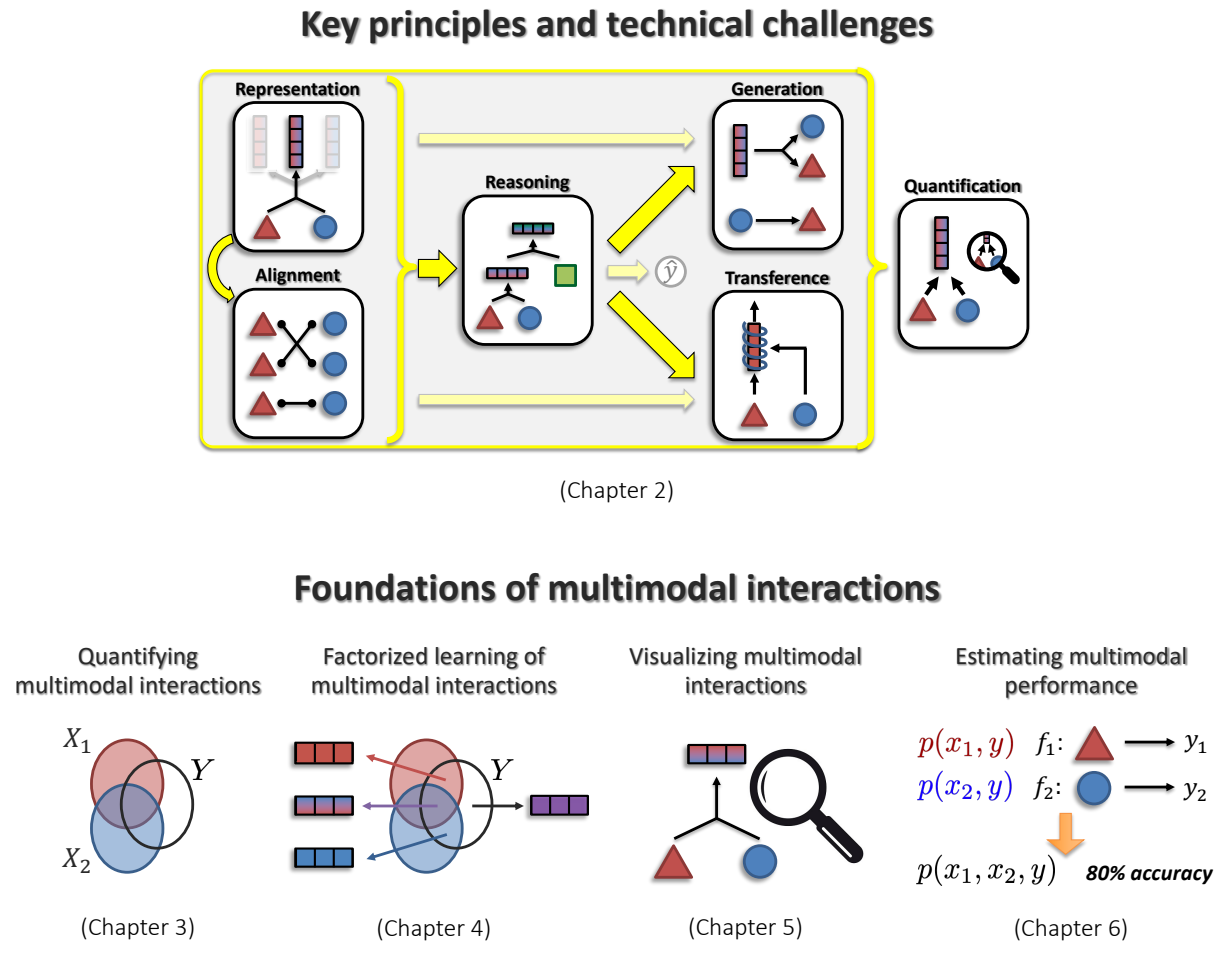


Figure 1.1: This thesis is designed to advance the theoretical and computational foundations of multimodal machine learning, and enable the creation of next-generation multimodal technologies. It starts by identifying the common themes and open questions in the field, through a taxonomy of six **core challenges** in multimodal research: representation, alignment, reasoning, generation, transference, and quantification. The bulk of the thesis studies two core challenges in multimodal learning: (1) building a **foundation for multimodal interactions** that enables the quantification of multimodal interactions in data and their principled modeling using machine learning methods, and (2) the data requirements and model building blocks enabling **generalization** of knowledge between modalities, tasks, and their representations.

1.1 Foundations of Multimodal Interactions

Multimodal interactions can be categorized into redundancy, uniqueness, and synergy: *redundancy* quantifies information shared between modalities, such as smiling while telling an overtly humorous joke; *uniqueness* quantifies the information present in only one, such as each medical sensor designed to provide new information; and *synergy* quantifies the emergence of new information using both, such as conveying sarcasm through disagreeing verbal and nonverbal cues [371]. These interactions are the basic principles of how modalities combine to give rise to new information for a task, which is present in all multimodal problems. While there have been intuitive definitions of these multimodal interactions, we still lack a formal foundation and systematic understanding of how to learn these interactions from data. As a result, there remain basic open questions like:

What interactions are in my data?
What interactions are learned by different models?
What models are suitable for my data?

To answer these questions, the first part of the thesis presents a theoretical framework formalizing the *useful information* in each modality and how *modalities interact* with each other to give rise to new information for a task [372]. Based on this theoretical framework, we propose two practical estimators to quantify the interactions in high-dimensional datasets, which can also be used more broadly for estimating information-theoretic quantities in real-world distributions. These estimators allow us to understand the information and interactions in multimodal datasets, and design the right models that provably learn the desired interactions in data.

We further show several broader implications that quantifying multimodal interactions can have on practitioners. Firstly, we operationalize the learning of multimodal interactions through a new approach called Factorized Contrastive Learning to capture both shared and unique information across modalities [374]. Secondly, a formal definition of multimodal interactions also enables us to analyze through qualitative visualizations whether a trained model has succeeded in learning the desired interactions from data [375]. Finally, we show how to use this information-theoretic framework to estimate the performance of optimal multimodal models given only unimodal data, which can inform practitioners which modalities to collect, and whether multimodal modeling is worth it for maximum increase in performance [376]. We release all code for quantifying multimodal interaction (both exact and approximate), and their implications on understanding datasets and models at `https://github.com/pliang279/PID`, code for Factorized Contrastive Learning at `https://github.com/pliang279/FactorCL`, and code for visualizing and debugging multimodal models at `https://github.com/pliang279/MultiViz`, which can help practitioners navigate the multimodal modeling pipeline.

1.2 Multisensory Foundation Models

There has been substantial impact of foundation models (e.g., large language models) trained on vast amounts of unlabeled data to obtain general-purpose capabilities over many prediction tasks. The future will lie in multisensory foundation models that are *grounded in the world*: being able to simultaneously process a large number of modalities beyond language, to

vision, audio [11, 360, 381, 502], and leveraging advances in sensing technologies such as cellphones [366], wearable devices [218], autonomous vehicles [697], healthcare technologies [287], and robots [53, 304]. The large number of heterogeneous modalities creates challenges in building multisensory foundation models. For example, the healthcare domain typically collects tabular data and high-frequency sensors [287], and it remains an open question how to best combine large language models with tabular data and sensors [546]. In the second part of this thesis, we take steps towards both data and modeling requirements to build the next generation of multisensory foundation models:

What data sources do we need to train foundation models over many heterogeneous modalities?

What modeling architectures are suitable for scaling to many heterogeneous modalities?

To answer the first question, we introduce MULTIBENCH, the largest and most comprehensive multimodal benchmark enabling the training of multisensory foundation models. MULTIBENCH collects and standardizes 15 realistic datasets across 10 diverse modalities, 20 prediction tasks, and 6 research areas from multimedia, affective computing, robotics, HCI, finance, and healthcare. MULTIBENCH is publicly available at `https://github.com/pliang279/MultiBench`, and has been broadly used in the community to train and evaluate multimodal architectures.

On the modeling side, prior work on multimodal learning has focused on a fixed set of modalities (e.g., image and text), without tackling generalization to many heterogeneous modalities and tasks necessary for truly multisensory models. To tackle the heterogeneity across many different modalities, we treat modalities in their most general form as sequences of elements, and present the *cross-modal attention* [101, 359] and *multimodal transformer* [613] architectures to learn interactions between all sequences of elements. These multimodal transformers are scalable and achieve strong results over a wide range of modalities, and we show their applications to image, text, video, sensors, and medical data. Finally, using MULTIBENCH, we scale multimodal transformers to the high-modality setting, resulting in a single model architecture with the same set of parameters that can function across a large number of modalities partially observed for different tasks [370] (e.g., image and text on the internet, video and audio in human communication, video and sensors in robotics, and so on). This represents the most realistic setting of how humans process the multisensory world, and we believe that general-purpose AI systems will also need to be trained in the high-modality setting. Our collection of high-modality models, available at `https://github.com/pliang279/HighMMT`, has already been extended for learning over many modalities in the medical, internet-of-things, and affective computing domains.

We end the thesis by discussing our collaborative efforts in applying these multisensory models for real-world impact on affective computing, mental health, and cancer prognosis.

1.3 Summary of Contributions

In this section, we provide a highlight of our main thesis contributions.

1. **Literature survey and taxonomy of multimodal challenges (Chapter 2)**
   (a) **Three key principles**: We begin by defining three key principles that have driven technical challenges and innovations: (1) modalities are *heterogeneous* because the information present often shows diverse qualities, structures, and representations, (2)

modalities are *connected* since they are often related and share commonalities, and (3) modalities *interact* to give rise to new information when used for task inference.

(b) **Six technical challenges**: Building upon these principles, we propose a new taxonomy of six core challenges in multimodal learning: (1) *Representation* studies how to summarize multimodal data to reflect the heterogeneity and interconnections between individual modality elements, before (2) *alignment* captures the connections and interactions between multiple local elements according to their structure. After representation and alignment comes (3) *reasoning*, which aims to combine the information from multimodal evidence in a principled way that respects the structure of the problem to give more robust and interpretable predictions. While most systems aim to predict the label y, there are also cases where the goal is (4) *generation*, to learn a generative process to produce raw modalities that reflect cross-modal interactions, structure, and coherence, or (5) *transference*, to transfer information from high-resource modalities to low-resource ones and their representations. Finally, (6) *quantification* revisits the previous challenges to give deeper empirical and theoretical understanding of modality heterogeneity, interconnections, and the learning process.

(c) **Current work and open directions**: For each challenge, we create a taxonomy of subchallenges and categorize recent advances in the field. This new taxonomy will enable researchers to better understand the state of research, and we identify several key directions for future work.

2. **Foundations of multimodal interactions (Chapter 3)**

(a) **Multimodal interactions** can be categorized into redundancy, uniqueness, and synergy: *redundancy* quantifies information shared between modalities, such as smiling while telling an overtly humorous joke; *uniqueness* quantifies the information present in only one, such as each medical sensor designed to provide new information; and *synergy* quantifies the emergence of new information using both, such as conveying sarcasm through disagreeing verbal and nonverbal cues [371].

(b) **Formal framework and estimation**: By introducing a new connection between information theory and multimodal interactions [372], I designed *scalable estimators to quantify the interactions in large-scale multimodal datasets and those learned by multimodal models*. These estimators are based on max-entropy convex optimization and a scalable end-to-end estimator suitable for high-dimensional continuous data.

(c) **Model selection**: We show that quantifying the interactions enables practitioners to analyze their datasets and select the most appropriate model that captures the right interactions in the data. We implemented these methods in two real-world case studies in mental health assessment [366] and cancer prognosis [372] from multimodal data. Domain experts appreciated the transparency that these methods convey as opposed to black-box neural networks, resulting in trust and adoption in real-world practice.

3. **Learning multimodal interactions using self-supervised learning (Chapter 4)**

(a) **From estimation to learning**: Naturally, a formal definition of multimodal interactions also translates to new training objectives to learn these interactions using neural networks. We show how to better learn task-relevant *unique information* [374, 614] using self-supervised learning, going beyond shared information between modalities.

(b) **Factorized learning of each interaction**: FACTORCL is built from three new con-

tributions: (1) factorizing task-relevant information into shared and unique representations, (2) capturing task-relevant information via maximizing MI lower bounds and removing task-irrelevant information via minimizing MI upper bounds, and (3) multimodal data augmentations to approximate task relevance without labels.

(c) **Real-world settings with unique information**: On large-scale real-world datasets, FACTORCL captures both shared and unique information and achieves state-of-the-art results on six benchmarks, including tasks involving medical sensors or robotics with force sensors that provide unique information, or cartoon images and figurative captions (i.e., not literal but metaphoric or idiomatic descriptions of the images).

4. **Visualizing multimodal interactions in trained models (Chapter 5)**
   (a) **Interpreting multimodal models**: MULTIVIZ is a framework for visualizing and understanding multimodal models across multiple stages: (1) modality importance, (2) multimodal interactions, and (3) multimodal reasoning. It includes tools to visualize what the model has learned about each stage of the prediction process.
   (b) **Model simulation**: To evaluate the fidelity of MULTIVIZ visualizations, we worked with real-world stakeholders to judge the accuracy of explanations at each fine-grained stage to determine if it helps users gain a deeper understanding of model behavior.
   (c) **Model debugging**: Furthermore, we ran user studies to show MULTIVIZ as a tool to highlight errors made by models and help users debug multimodal models for real-world deployment.
5. **Estimating multimodal performance for modality selection (Chapter 6)**
   (a) **Modality selection**: We extended our analysis to quantify interactions in a semi-supervised setting with only labeled unimodal data $(x_1, y), (x_2, y)$ and naturally co-occurring multimodal data (x_1, x_2) (e.g., unlabeled images and captions, video and corresponding audio) but when labeling them is time-consuming [376]. We show how to approximately estimate the multimodal interactions in the unseen full distribution (x_1, x_2, y), which enables practitioners to prioritize collecting data for modalities that has the most synergy with existing ones.
   (b) **Estimating performance**: Our approximation is based on lower and upper bounds for synergy: a lower bound based on the *disagreement* between modality predictors, and an upper bound based on a connection to *min-entropy couplings*. Lower and upper bounds on synergistic information translate to bounds on multimodal performance.
   (c) **On disagreement**: Finally, we show that disagreement is a critical quality that can result in synergy between modalities, and propose a learning algorithm that captures disagreement between modalities beyond agreement that is typically done.
6. **MULTIBENCH: A benchmark for real-world generalization (Chapter 7)**
   (a) **Real-world benchmarks**: We describe MULTIBENCH, the largest unified benchmark for multimodal representation learning [367]. MULTIBENCH provides an end-to-end machine learning pipeline that simplifies and standardizes data loading, experimental setup, and model evaluation, while ensuring reproducibility and ease of use.
   (b) **Standardized building blocks**: To accompany this benchmark, we also provide a standardized implementation of 20 core approaches in multimodal learning spanning innovations in fusion paradigms, optimization objectives, and training approaches.
   (c) **Benefits of standardization**: We find that standardizing and sharing methods pro-

posed in different research areas can improve performance on several datasets. MULTIBENCH also provides a better understanding of the capabilities and limitations of multimodal models.

7. **Learning multimodal interactions across time (Chapter 8)**
   (a) **Temporal interactions**: To tackle heterogeneity across many different modalities, we treat modalities in their most general form as a sequence of elements, such as words in a sentence, patches in an image, frames in a video, and time steps in time-series data. This introduces a critical challenge of learning multimodal interactions across sequences, such as relating a word with a facial expression within a long video.
   (b) **Recurrent cross-modal attention**: While prior work summarized temporal modalities into a single static feature before fusion, I developed a new method for *fine-grained temporal fusion* to learn interactions between all elements across the sequence, such as between individual words, gestures, and vocal expressions [101, 359]. We call this module recurrent cross-modal attention, by using attention weights to recursively learn interactions based on the current input and previous signals.
   (c) **Multimodal transformers**: We extended recurrent attention into multimodal transformers that learn all interactions across sequences in parallel [613]. The multimodal transformer learns a cross-modal attention matrix to highlight related signals across time (e.g., rolling eyes and sighing). This matrix is used to learn a new representation for each modality fused with other modalities in parallel over the entire sequence, which provides huge efficiency gains when trained on modern GPUs.
8. **Multimodal and multitask foundation models (Chapter 9)**
   (a) **High-modality learning**: Chapter 9 builds upon the diverse modalities and tasks provided by MULTIBENCH by designing methods for high-modality learning: where there are a large number of modalities partially observed for different tasks [370]. This represents the most realistic setting of how humans process the multisensory world, and we believe that general-purpose AI systems will also need to be multisensory.
   (b) **A single model for many modalities and tasks**: We propose HIGHMMT, a single shared high-modality model that achieves generalization over more than 10 modalities and 15 tasks, and transfers to new modalities and tasks.
   (c) **Tackling extreme heterogeneity**: We've seen two extremes - full parameter sharing across everything, and no sharing at all across modality and task-specific models. A key idea in HIGHMMT is to find the optimal amount of parameter sharing balancing performance and efficiency. We do this by defining a new measure of which modalities are similar, and which modality pairs interact similarly, to inform parameter sharing.

1.4 Other Contributions

I have also pursued the following selected research directions during my Ph.D. studies, which are excluded from this thesis. The first major direction lies in datasets and methods for learning representations from a fixed set of input modalities (i.e., without modeling generalization). To that end, I have contributed core resources and models for multimodal representation learning, especially in the application domain of modeling human communication. I have also engaged in

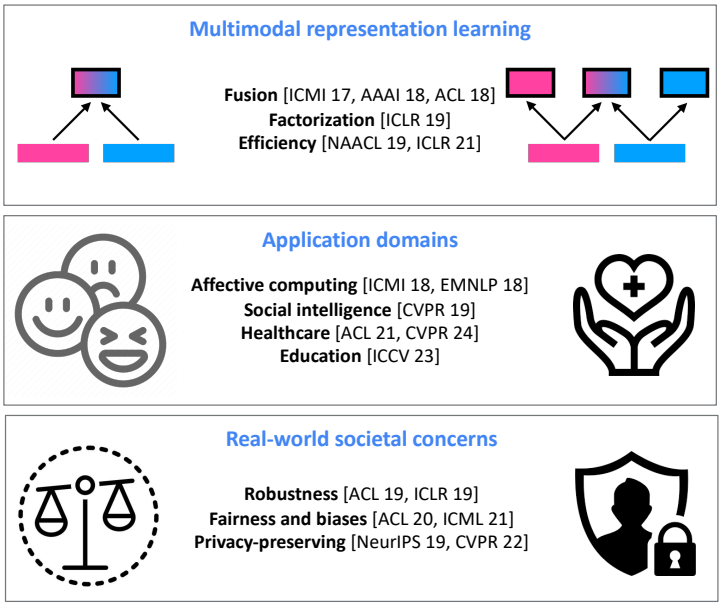


Figure 1.2: I have also pursued the following directions during my Ph.D. studies: (1) new machine learning and deep learning models to learn multimodal representations (without modeling generalization), (2) collaborating with real-world stakeholders to apply these methods in affective computing, socially intelligent AI, healthcare, and education, and (3) mitigating real-world issues of deploying multimodal models in the face of real-world noise topologies, dataset biases, and privacy concerns.

collaborations with real-world stakeholders particularly in the healthcare and affective computing space where multimodal learning paradigms offer opportunities to learn from high-dimensional multimodal data. Finally, I have also worked on addressing the real-world societal concerns these models, such as improving their robustness, fairness, and privacy.

1.4.1 Multimodal representation learning

Computational modeling of human multimodal language: From a computational perspective, the modeling of human communication across both verbal and nonverbal behaviors enables real-world tasks such as multimodal sentiment analysis [427], emotion recognition [74], and personality traits recognition [467]. To comprehend human communication, there is a need for 1) large multimodal resources with diversity in training samples, topics, speakers, and annotations, as well as 2) powerful models for multimodal communication.

As a first step, we have worked towards addressing the lack of multimodal resources by collecting and releasing the largest dataset of multimodal sentiment and emotion recognition enabling generalizable studies of human communication. CMU-MOSEI contains $23,500$ annotated video segments from $1,000$ distinct speakers and 250 topics. The diversity in topics, speakers, annotations, and modalities allows for generalizable studies of speaker and topic-independent features. The multimodal dataset and a general multimodal data loading framework are provided to the scientific community to encourage valuable research in human communication analysis [360, 717]. Since then, the dataset has also been the subject of two workshop challenges in modeling human multimodal language at ACL 2018 and ACL 2020, and has been a standard benchmark dataset for

the multimodal machine learning community.

Multimodal gated fusion: With the increasing popularity of video sharing websites such as YouTube and Facebook, multimodal sentiment analysis has received increasing attention from the scientific community [427, 477, 663]. We develop a novel deep architecture for multimodal sentiment analysis that performs modality fusion at the word level [101]. We proposed the GME-LSTM model that is composed of 2 modules. The Gated Multimodal Embedding alleviates the difficulties of fusion when there are noisy modalities. The LSTM with Temporal Attention performs word level fusion at a finer fusion resolution between input modalities and attends to the most important time steps. As a result, the GME-LSTM is able to better model the multimodal structure of speech through time and perform better sentiment comprehension. We demonstrate the effectiveness of this approach by achieving state-of-the-art sentiment classification and regression results. Qualitative analysis on our model emphasizes the importance of the Temporal Attention Layer in sentiment prediction because the additional acoustic and visual modalities are noisy. We also demonstrate the effectiveness of the Gated Multimodal Embedding layer in selectively filtering these noisy modalities out. Our results and analysis open new areas in the study of sentiment analysis in human communication and provide new models for multimodal fusion.

Factorized multimodal representations: Using MULTIBENCH and other related multimodal benchmarks enables us a deeper study of the desiderata for multimodal representations beyond discriminative performance [614]. While the two main pillars of research in multimodal representation learning have considered discriminative [88, 101, 181, 553, 712] and generative [442, 482, 555, 565, 585] objectives individually, we demonstrate that factorizing multimodal representations into multimodal discriminative and modality-specific generative factors marries the strengths of discriminative learning of joint features across modalities that achieves state-of-the-art performance for affect analysis with controllable generation of human language based on individual factors, robustness to partially missing modalities, and interpretable local contributions from each modality during prediction. Our resulting Multimodal Factorization Model (MFM) defines a flexible latent variable framework balancing prediction with robustness and understandability for real-world human multimodal language.

Efficient statistical baselines: The constraints of real-world edge devices have created a demand for data and compute-efficient multimodal learning via simple yet strong models [569]. We proposed an approach based on stronger statistical baselines rather than black-box neural networks. By assuming a fully-factorized probabilistic generative model of multimodal data from a latent representation, careful model design allows us to capture expressive unimodal, bimodal, and trimodal interactions while at the same time retaining simplicity and efficiency during learning and inference [362]. These models show strong performance on both supervised and semi-supervised multimodal prediction, as well as significant (10 times) speedups over neural models during inference.

1.4.2 Applications in affective computing, social intelligence, and healthcare

Improving the generalization and quantification of multimodal models enables a step towards real-world models capturing the benefits of multimodal data while mitigating its risks. However, tangible real-world impact requires direct collaboration with real-world stakeholders to determine their precise computational needs. During my PhD, I have had the pleasure of collaborating on

the following real-world applications:

Multimodal affective computing: As an application-specific instantiation of multimodal learning, we studied the problem of continuous-time human affect analysis and proposed a new perspective by modeling both person-independent and person-dependent signals through insights from human psychology [361]. Some emotional expressions are almost universal person-independent behaviors and can be recognized directly from a video [145, 305]. For example, an open mouth with raised eyebrows and a loud voice is likely to be associated with surprise. However, emotions are also expressed in a person-dependent fashion with idiosyncratic behaviors where it may not be possible to directly estimate absolute emotion intensities. Instead, it would be easier to compare two video segments of the same person and judge whether there was a relative change in emotion intensities [163, 419, 567]. For example, a person could have naturally furrowed eyebrows and we should not always interpret this as a display of anger, but rather compare two video segments to determine relative changes in anger. By designing a model combining both signals, we are able to achieve state-of-the-art audio-visual emotion recognition performance and allow for fine-grained investigation of person-independent and person-dependent behaviors.

Social intelligence question-answering: As intelligent systems increasingly blend into our everyday life, artificial social intelligence becomes a prominent area of research. Intelligent systems must be socially intelligent in order to comprehend human intents and maintain a rich level of interaction with humans [268, 302, 600, 636]. Human language offers a unique unconstrained approach to probe *through questions* and reason *through answers* about social situations [11, 343]. This unconstrained approach extends previous attempts to model social intelligence through numeric supervision (e.g. sentiment and emotions labels). We introduced the Social-IQ dataset [715], an unconstrained benchmark specifically designed to train and evaluate socially intelligent technologies. By providing a rich source of open-ended questions and answers, Social-IQ opens the door to explainable social intelligence. The dataset contains rigorously annotated and validated videos, questions and answers, as well as annotations for the complexity level of each question and answer. Social-IQ contains $1,250$ natural in-the-wild social situations, $7,500$ questions and $52,500$ correct and incorrect answers. Although humans can reason about social situations with very high accuracy (95.08%), existing state-of-the-art computational models struggle on this task. As a result, Social-IQ brings novel challenges that will spark future research in social intelligence modeling, visual reasoning, and multimodal question answering (QA).

Privacy-preserving mood prediction from mobile data: Mental health conditions remain underdiagnosed even in countries with common access to advanced medical care [178, 328]. The ability to accurately and efficiently predict mood from easily collectible data has several important implications for the early detection, intervention, and treatment of mental health disorders [200, 433]. One promising data source to help monitor human behavior is daily smartphone usage [492]. However, care must be taken to summarize behaviors without identifying the user through personal (e.g., personally identifiable information) or protected (e.g., race, gender) attributes [313, 365, 539, 580]. Through data collected via a collaboration with psychiatrists and psychologists at the University of Oregon, Columbia University, and the University of Pittsburgh, we study behavioral markers of daily mood using a recent dataset of mobile behaviors from adolescent populations at high risk of suicidal behaviors [366]. Using computational models, we find that language and multimodal representations of mobile *typed text* (spanning typed characters,

words, keystroke timings, and app usage) are predictive of daily mood. However, we find that models trained to predict mood often also capture private user identities in their intermediate representations. To tackle this problem, we evaluate approaches that obfuscate user identity while remaining predictive. By combining multimodal representations with privacy-preserving learning, we are able to push forward the performance-privacy frontier.

1.4.3 Real-world robustness, fairness, and privacy

Finally, the third major direction studies the real-world concerns of deploying multimodal models in the face of real-world noise topologies, dataset biases, and privacy concerns.

Robustness to noisy modalities: Different modalities often display different noise topologies, and real-world multimodal signals possibly suffer from missing or noisy data in at least one of the modalities [46, 150, 336]. Human-centric data is also often imperfect due to personal idiosyncrasies which affect the contribution of certain modalities during social interactions [204, 524]. For example, multimodal dialogue systems trained on acted TV shows are susceptible to poor performance when deployed in the real world where users might be less expressive in using facial gestures. This calls for robust models that can still make accurate predictions despite only having access to a (possibly noisy) subset of signals.

As a step towards robustness, we propose a tensor representation learning method to deal with noisy modalities in time-series data (e.g., text, videos, audio) [364]. This method is based on the observation that multimodal time series data often exhibits correlations across time and modalities which lead to low-rank multimodal representations [237, 327, 690]. However, the presence of noise or incomplete values breaks these correlations and results in tensor representations of higher rank. Regularizing the rank of tensor representations therefore provides a denoising effect which achieves strong results across various levels of imperfection. We show how to integrate an upper-bound of tensor rank minimization as a simple regularizer for training in the presence of imperfect data, thereby combining the strength of temporal non-linear transformations of multimodal data with principled regularization on tensor structures. Through experiments on multimodal video data, our results back up our intuitions that imperfect data increases tensor rank and demonstrates strong results across various levels of imperfection.

Learning fair sentence representations: To safely deploy human-centric multimodal models in real-world scenarios such as healthcare, legal systems, and social science, it is also necessary to recognize the role they play in shaping social biases and stereotypes. Previous work has revealed the presence of *representational biases* in widely used word embeddings - harmful biases resulting from stereotyping that propagate negative generalizations involving gender, race, religion, and other social constructs [64, 193, 233, 432, 518, 542]. While some methods were proposed to debias these word-level embeddings [67, 405], there is a need to perform debiasing at the sentence-level given the recent shift towards new contextualized sentence representations such as ELMo [479] and BERT [144] which have become core components in both real-world language [19, 257, 656] and multimodal prediction systems [348, 390]. We investigated the presence of social biases in sentence-level representations and proposed a new method, SENT-DEBIAS, to reduce these biases [365]. We show that SENT-DEBIAS is effective in reducing biases from the geometry of contextual representation spaces, and at the same time, preserves performance on sentence-level downstream NLP tasks such as sentiment analysis, linguistic

acceptability, and natural language understanding.

Mitigating social biases in language models: In addition to sentence representations deployed primarily for discriminative tasks, large-scale pretrained language models (LMs) have also become widely-deployed for generative applications such as text generation [496], dialog systems [730], recommendation systems [533], and search engines [43, 455]. Recent work has found that these language models can potentially generate text propagating negative generalizations about particular social groups [432], language that is denigrating to particular social groups [542], and toxic speech [193], while at the same time also being unable to reason about human-aligned values such as ethics [233], social bias implications [518], and allocational harms across social groups [386]. As a step towards improving the fairness of LMs, we carefully defined several sources of representational biases before proposing new benchmarks and metrics to measure them [368]. With these tools, we propose A-INLP, an approach towards post-hoc debiasing of large pretrained LMs. The key to our approach lies in dynamically finding bias-sensitive tokens rather than relying on a predefined set of bias-sensitive words that are common in existing literature [67]. Our empirical results and human evaluation on large language models such as GPT-2 demonstrate effectiveness in mitigating bias while retaining crucial context information for high-fidelity text generation, thereby pushing forward the performance-fairness Pareto frontier. These steps are critical towards improving the safety of language and multimodal models.

Privacy-preserving federated learning: More broadly, federated learning is a method of training models on private data distributed over multiple devices [68, 356, 413, 552]. To keep device data private, a single global model is trained by only communicating parameters and updates which poses scalability challenges for large models [446]. Furthermore, current approaches use the same model architecture across all local models and the global aggregated model, which causes federated learning to struggle with data heterogeneity across devices [248, 356, 731]. This is made worse when each device contains multimodal data sources that are used unequally across users [426]. To this end, we propose a new federated learning algorithm, Local Global Federated Averaging (LG-FEDAVG), that jointly learns compact *local representations* on each device and a global model across all devices [363]. As a result, the global model can be smaller since it only operates on local representations, reducing the number of communicated parameters. Furthermore, well-designed local models enable learning of personalized representations for user-specific behavior modeling while enjoying the benefit of global model learning across many users' data. Theoretically, we provide a generalization analysis which shows that a combination of local and global models reduces both variance in the data as well as variance across device distributions. Empirically, we demonstrate that local models enable communication-efficient training while retaining performance. We also evaluate on the task of personalized mood prediction from real-world mobile data where privacy is key. Finally, we show that local models handle heterogeneous data from new devices, and learn fair representations that obfuscate protected attributes such as race, age, and gender [67].

Chapter 2

Literature Survey and Taxonomy of Multimodal Challenges

2.1 Introduction

It has always been a grand goal of artificial intelligence to develop computer agents with intelligent capabilities such as understanding, reasoning, and learning through multimodal experiences and data, similar to how humans perceive and interact with our world using multiple sensory modalities. With recent advances in embodied autonomous agents [69, 522], self-driving cars [674], image and video understanding [18, 576], image and video generation [502, 550], and multisensor fusion in application domain such as robotics [335, 408] and healthcare [287, 367], we are now closer than ever to intelligent agents that can integrate and learn from many sensory modalities. This vibrant multi-disciplinary research field of multimodal machine learning brings unique challenges given the heterogeneity of the data and the interconnections often found between modalities, and has widespread applications in multimedia [436], affective computing [489], robotics [308, 335], human-computer interaction [450, 538], and healthcare [76, 429].

However, the rate of progress in multimodal research has made it difficult to identify the common themes underlying historical and recent work, as well as the key open questions in the field. By synthesizing a broad range of research, this paper is designed to provide an overview of the methodological, computational, and theoretical foundations of multimodal machine learning. We begin by defining (in §2.2) three key principles that have driven technical challenges and innovations: (1) modalities are *heterogeneous* because the information present often shows diverse qualities, structures, and representations, (2) modalities are *connected* since they are often related and share commonalities, and (3) modalities *interact* to give rise to new information when used for task inference. Building upon these definitions, we propose a new taxonomy of six core challenges in multimodal learning: *representation*, *alignment*, *reasoning*, *generation*, *transference*, and *quantification* (see Figure 2.1). These core multimodal challenges are understudied in conventional unimodal machine learning and need to be tackled in order to progress the field forward:

1. **Representation (§2.3):** Can we learn representations that reflect heterogeneity and interconnections between modality elements? We will cover approaches for (1) *representation fusion*: integrating information from two or more modalities to capture cross-modal interactions, (2)

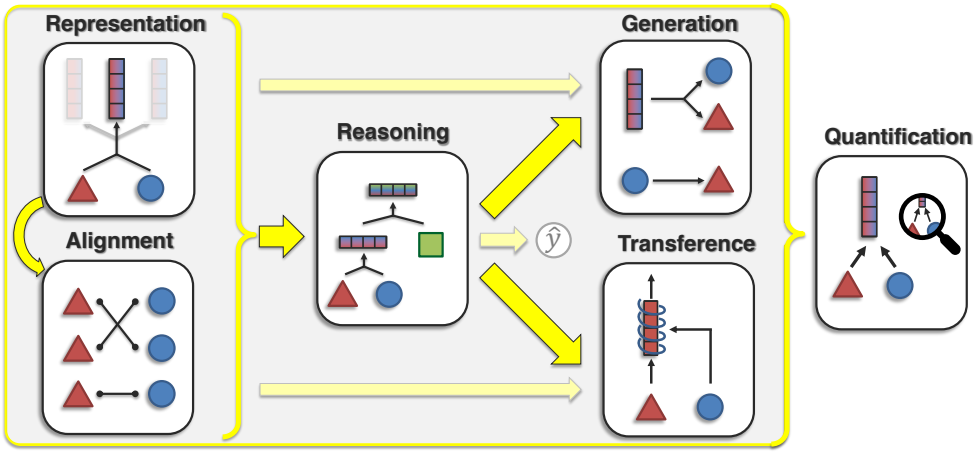


Figure 2.1: Core research challenges in multimodal learning: Every multimodal problem typically requires tackling representation and alignment: (1) *Representation* studies how to summarize multimodal data to reflect the heterogeneity and interconnections between individual modality elements, before (2) *alignment* captures the connections and interactions between multiple local elements according to their structure. After representation and alignment comes (3) *reasoning*, which aims to combine the information from multimodal evidence in a principled way that respects the structure of the problem to give more robust and interpretable predictions. While most systems aim to predict the label y, there are also cases where the goal is (4) *generation*, to learn a generative process to produce raw modalities that reflect cross-modal interactions, structure, and coherence, or (5) *transference*, to transfer information from high-resource modalities to low-resource ones and their representations. Finally, (6) *quantification* revisits the previous challenges to give deeper empirical and theoretical understanding of modality heterogeneity, interconnections, and the learning process.

representation coordination: interchanging cross-modal information to keep the same number of representations but improve multimodal contextualization, and (3) *representation fission*: creating a larger set of disjoint representations that reflects knowledge about internal structure such as data clustering or factorization.

2. **Alignment (§2.4):** How can we identify the connections and interactions between modality elements? Alignment is challenging since it may depend on long-range dependencies, involves ambiguous segmentation (e.g., words or utterances), and could be either one-to-one, many-to-many, or not exist at all. We cover (1) *discrete alignment*: identifying connections between discrete elements across modalities, (2) *continuous alignment*: modeling alignment between continuous modality signals with ambiguous segmentation, and (3) *contextualized representations*: learning better representations by capturing cross-modal interactions between elements.
3. **Reasoning (§2.5)** is defined as composing knowledge, usually through multiple inferential steps, that exploits the problem structure for a specific task. Reasoning involves (1) *modeling the structure* over which composition occurs, (2) the *intermediate concepts* in the composition process, (3) understanding the *inference paradigm* of more abstract concepts, and (4) leveraging large-scale *external knowledge* in the study of structure, concepts, and inference.
4. **Generation (§2.6)** involves learning a generative process to produce raw modalities. We categorize its subchallenges into (1) *summarization*: summarizing multimodal data to reduce

information content while highlighting the most salient parts of the input, (2) *translation*: translating from one modality to another and keeping information content while being consistent with cross-modal connections, and (3) *creation*: simultaneously generating multiple modalities to increase information content while maintaining coherence within and across modalities.

5. **Transference (§2.7)** aims to transfer knowledge between modalities, usually to help the target modality, which may be noisy or with limited resources. Transference is exemplified by (1) *cross-modal transfer*: adapting models to tasks involving the primary modality, (2) *co-learning*: transferring information from secondary to primary modalities by sharing representation spaces between both modalities, and (3) *model induction*: keeping individual unimodal models separate but transferring information across these models.
6. **Quantification (§2.8):** The sixth and final challenge involves empirical and theoretical studies to better understand (1) the dimensions of *heterogeneity* in multimodal datasets and how they subsequently influence modeling and learning, (2) the presence and type of modality *connections and interactions* in multimodal datasets and captured by trained models, and (3) the *learning* and optimization challenges involved with heterogeneous data.

Finally, we conclude this paper with a long-term perspective on multimodal learning by motivating open research questions identified by this taxonomy. This survey was also presented by the authors in a visual medium through tutorials at CVPR 2022 and NAACL 2022, as well as courses 11-777 Multimodal Machine Learning and 11-877 Advanced Topics in Multimodal Machine Learning at CMU. The reader is encouraged to refer to these public video recordings, additional readings, and discussion probes for more mathematical depth on certain topics, visual intuitions and explanations, and more open research questions in multimodal learning.

This paper is designed to complement other surveys that belong broadly to the study of multiple modalities or views: multi-view learning [443, 577, 685] is concerned with settings where different views (e.g., camera views) typically provide overlapping (redundant) information but not the other core challenges we cover, surveys on multimodal foundation models [157, 185] go into detail on tackling representation, fusion, and alignment using large-scale pretraining but do not cover other core challenges, and several application-oriented surveys in vision-language models [627], language and reinforcement learning [394], multimedia analysis [34], and multimodal human-computer interaction [277] discuss specific multimodal challenges faced in these applications. This survey presents a telescoping overview suitable as a starting point for researchers who can then diver deeper into methodology or application-specific research areas.

2.1.1 Key modalities and application domains

In this subsection, we first contextualize our subsequent discussion of multimodal machine learning by listing some key modalities of interest, standard multimodal datasets and toolkits, and major applications of multimodal learning in the real world.

Affective computing studies the perception of human affective states such as emotions, sentiment, and personalities from multimodal human communication: spoken language, facial expressions and gestures, body language, vocal expressions, and prosody [483]. Some commonly studied tasks involve predicting sentiment [556, 710], emotions [717], humor [225], and sarcasm [83] from multimodal videos of social interactions.

Healthcare: Machine learning can help integrate complementary medical signals from lab

tests, imaging reports, patient-doctor conversations, and multi-omics data to assist doctors in the clinical process [5, 21, 383]. Multimodal physiological signals recorded regularly from smartphones and wearable devices can also provide non-invasive health monitoring [134, 189, 366]. Public datasets include MIMIC [287] with patient tabular data, medical reports, and medical sensor readings, question answering on pathology [230] and radiology [329] images, and multi-omics data integration [610].

Robotics systems are often equipped with multiple sensors to aid in robust decision-making for real-world physical tasks such as grasping, cleaning, and delivery. These sensors can include vision (RGB and depth), force, and proprioception [335]. These multi-sensor robots have been successfully applied in haptic [459, 530] and surgical robots [4, 60]. More generally, language [394] and audio [135] have also emerged as useful signals for robot learning.

Interactive agents in the virtual world can assist humans in multimedia web tasks and computer tasks [183] as well as in the social world through virtual agents [471, 472]. These agents need to understand human commands and behaviors, process various forms of visual, tabular, and multimedia content, use external web tools and APIs, and interact in multi-step decision-making tasks. Webshop [695] and WebAreana [735] are recent environments testing the capabilities of AI agents in navigating image and text content to solve web tasks.

Multimedia data spanning text, images, videos, audio, and music is abundant on the internet and has fueled a significant body of multimodal research [34], such as classification [728], retrieval [504], and recommendation [423, 513, 732] of multimedia content, image and video question answering [11, 317, 339] and captioning [156, 640]), multimedia and entertainment content description [532] (including movies [32], memes [301, 537], and cartoons [236]), and more recently in automatic creation of text [726], images [511], videos [667], music [9], and more.

Human-computer interaction has sought to endow computers with multimodal capabilities to provide more natural, powerful, and compelling interactive user experiences [624]. These systems have leveraged speech, touch, vision, gestures, affective states [462] and affordable wearable and mobile sensors [277, 457, 624]. Public datasets have enabled the study of multimodal user interfaces [340, 644], speech and gesture interactions [166], and human sensing [89, 139, 526].

Science and environment: Deepening our knowledge of the natural sciences and physical environments can bring about impactful changes in scientific discovery, sustainability, and conservation. This requires processing modalities such as chemical molecules [570], protein structures [725], satellite images [109, 693], remote sensing [243, 346], wildlife movement [389], scientific diagrams and texts [392], and various physical sensors [424].

Education: AI can broaden access to educational content by digitizing lecture slides and videos, creating personalized tutors, and designing interactive learning curricula. It introduces challenges in processing recorded lecture slides and videos [333], and modeling student learning via asked questions, spoken feedback and non-verbal gestures [87, 574, 679].

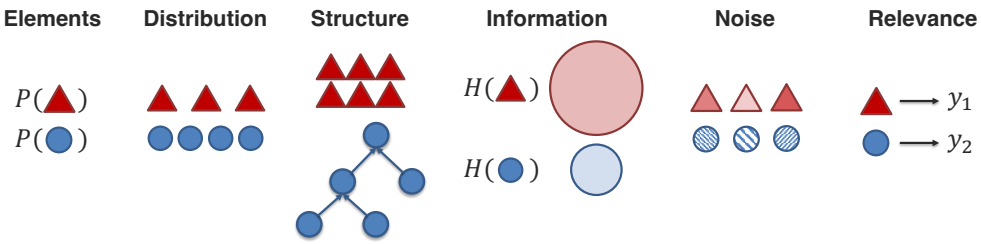


Figure 2.2: The information present in different modalities will often show diverse qualities, structures, and representations. **Dimensions of heterogeneity** can be measured via differences in individual elements and their distribution, the structure of elements, as well as modality information, noise, and task relevance.

2.2 Foundational Principles in Multimodal Research

A *modality* refers to a way in which a natural phenomenon is perceived or expressed. For example, modalities include speech and audio recorded through microphones, images and videos captured via cameras, and force and vibrations captured via haptic sensors. Modalities can be placed along a spectrum from *raw* to *abstract*: raw modalities are those more closely detected from a sensor, such as speech recordings from a microphone or images captured by a camera. Abstract modalities are those farther away from sensors, such as language extracted from speech recordings, objects detected from images, or even abstract concepts like sentiment intensity and object categories.

Multimodal refers to situations where multiple modalities are involved. From a research perspective, multimodal entails the computational study of *heterogeneous* and *interconnected* modalities. Firstly, modalities are *heterogeneous* because the information present in different modalities will often show diverse qualities, structures, and representations. Secondly, these modalities are not independent entities but rather share *connections* due to complementary information. Thirdly, modalities *interact* in different ways when they are integrated for a task. We expand on these three foundational principles of multimodal research in the following subsections.

2.2.1 Principle 1: Modalities are heterogeneous

The principle of heterogeneity reflects the observation that the information present in different modalities will often show diverse qualities, structures, and representations. Heterogeneity should be seen as a spectrum: two images from the same camera that capture the same view modulo camera wear and tear are closer to homogeneous, two different languages that capture the same meaning but from different language families are slightly heterogeneous, language and vision are even more heterogeneous, and so on. In this section, we present a non-exhaustive list of dimensions of heterogeneity (see Figure 2.2 for an illustration). These dimensions are complementary and may overlap; each multimodal problem likely involves heterogeneity in multiple dimensions.

1. **Element representation**: Each modality is typically comprised of a set of elements - the most basic unit of data which cannot (or rather, the user chooses to not) be broken down into further units [49, 358]. For example, typed text is recorded via a set of characters, videos are recorded via a set of frames, and graphs are recorded via a set of nodes and edges. What are the basic elements present in each modality, and how can we represent them? Formally, this dimension measures heterogeneity in the sample space or representation space of modality elements.
2. **Distribution** refers to the frequency and likelihood of modality elements. Elements typically

follow a unique distribution, with words in a linguistic corpus following Zipf's Law [743] as an example. Distribution heterogeneity refers to the differences in frequencies and likelihoods of elements, such as different frequencies in recorded signals and the density of elements.

3. **Structure**: Natural data exhibits structure in the way individual elements are composed to form entire modalities [71]. For example, images exhibit spatial structure across objects, language is hierarchically composed of words, and signals exhibit temporal structure across time. Structure heterogeneity refers to differences in this underlying structure.
4. **Information** measures the total information content present in each modality. Subsequently, information heterogeneity measures the differences in information content across modalities, which could be formally measured by information theoretic metrics [535].
5. **Noise**: Noise can be introduced at several levels across naturally occurring data and also during the data recording process. Natural data noise includes occlusions, imperfections in human-generated data (e.g., imperfect keyboard typing or unclear speech), or data ambiguity due to sensor failures [367]. Noise heterogeneity measures differences in noise distributions across modalities, as well as differences in signal-to-noise ratio.
6. **Relevance**: Finally, each modality shows different relevance toward specific tasks and contexts - certain modalities may be more useful for certain tasks than others [192]. Task relevance describes how modalities can be used for inference, while context relevance describes how modalities are contextualized with other modalities.

It is useful to take these dimensions of heterogeneity into account when studying both unimodal and multimodal data. In the unimodal case, specialized encoders are typically designed to capture these unique characteristics in each modality [71]. In the multimodal case, modeling heterogeneity is useful when learning representations and capturing alignment [718], and is a key subchallenge in quantifying multimodal models [370].

2.2.2 Principle 2: Modalities are connected

Although modalities are heterogeneous, they are often connected due to shared complementary information. The presence of *shared* information is often in contrast to *unique* information that exists solely in a single modality [662]. Modality connections describe the extent and dimensions to which information can be shared across modalities. When reasoning about the connections in multimodal data, it is helpful to think about both bottom-up (statistical) and top-down (semantic) approaches (see Figure 2.3). From a statistical data-driven perspective, connections are identified from distributional patterns in multimodal data, while semantic approaches define connections based on our domain knowledge about how modalities share and contain unique information.

1. **Statistical association** exists when the values of one variable relate to the values of another. For example, two elements may co-occur with each other, resulting in a higher frequency of both occurring at the same time. Statistically, this could lead to correlation - the degree to which elements are linearly related, or other non-linear associations. From a data-driven perspective, discovering which elements are associated with each other is important for modeling the joint distributions across modalities during multimodal representation and alignment [605].
2. **Statistical dependence** goes deeper than association and requires an understanding of the exact type of statistical dependency between two elements. For example, is there a causal dependency from one element to another, or an underlying confounder causing both elements

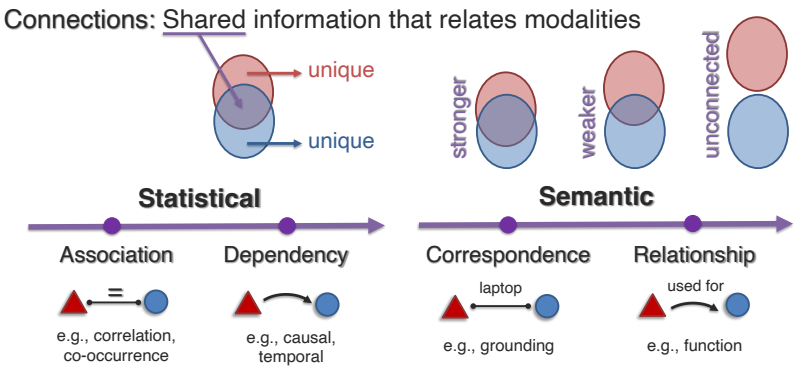


Figure 2.3: Modality connections describe how modalities are related and share commonalities, such as correspondences between the same concept in language and images or dependencies across spatial and temporal dimensions. Connections can be studied through both statistical and semantic perspectives.

to be present at the same time? Other forms of dependencies could be spatial or temporal: one element occurring above the other, or after the other. Typically, while statistical association can be estimated purely from data, understanding the nature of statistical dependence requires some knowledge of the elements and their underlying relationships [445, 625].

3. **Semantic correspondence** can be seen as the problem of ascertaining which elements in one modality share the same semantic meaning as elements in another modality [456]. Identifying correspondences is fundamental in many problems related to language grounding [86], translation and retrieval [485], and cross-modal alignment [588].
4. **Semantic relations**: Finally, semantic relations generalize semantic correspondences: instead of modality elements sharing the same exact meaning, semantic relations include an attribute describing the exact nature of the relationship between two modality elements, such as semantic, logical, causal, or functional relations. Identifying these semantically related connections is important for higher-order reasoning [49, 410].

2.2.3 Principle 3: Modalities interact

Modality interactions study how modality elements interact to give rise to new information when integrated together for task *inference*. We note an important difference between modality connections and interactions: connections exist within multimodal data itself, whereas interactions only arise when modalities are integrated and processed together to bring a new response. In Figure 2.4, we provide a high-level illustration of some dimensions of interactions that can exist.

1. **Interaction information** investigates the type of connected information that is involved in an interaction. When an interaction involves shared information common to both modalities, the interaction is *redundant*, while a *non-redundant* interaction is one that does not solely rely on shared information, and instead relies on different ratios of shared, unique, or possibly even synergistic information [372, 662].
2. **Interaction mechanics** are the functional operators involved when integrating modality elements for task inference. For example, interactions can be expressed as statistically additive, non-additive, and non-linear forms [283], as well as from a semantic perspective where two elements interact through a logical, causal, or temporal operation [626].
3. **Interaction response** studies how the inferred response changes in the presence of multiple

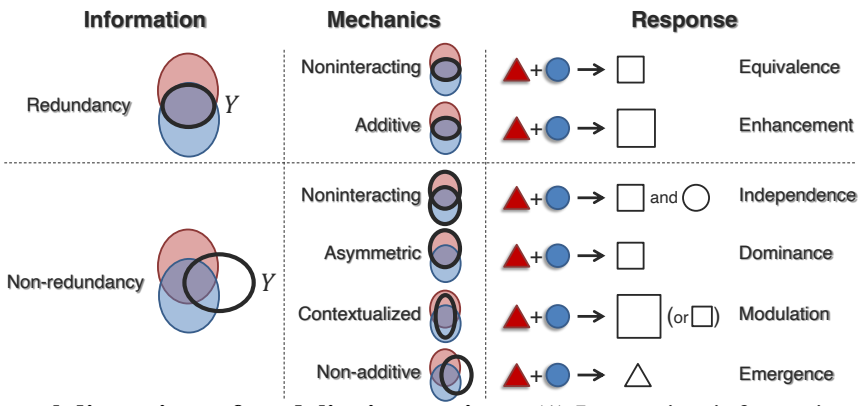


Figure 2.4: Several dimensions of modality interactions: (1) Interaction information studies whether common redundant information or unique non-redundant information is involved in interactions; (2) interaction mechanics study the manner in which interaction occurs, and (3) interaction response studies how the inferred task changes in the presence of multiple modalities.

modalities. For example, through sub-dividing redundant interactions, we can say that two modalities create an equivalence response if the multimodal response is the same as responses from either modality, or enhancement if the multimodal response displays higher confidence. On the other hand, non-redundant interactions such as modulation or emergence happen when there exist different multimodal versus unimodal responses [468].

2.2.4 Core technical challenges

Building on these three core principles and our detailed review of recent work, we propose a new taxonomy to characterize the core technical challenges in multimodal research: representation, alignment, reasoning, generation, transference, and quantification. In Table 2.1 we summarize our full taxonomy of these six core challenges, their subchallenges, categories of corresponding approaches, and recent examples in each category. In the following sections, we describe our new taxonomy in detail and also revisit the principles of heterogeneity, connections, and interactions to see how they pose research questions and inspire research in each of these six challenges.

2.3 Challenge 1: Representation

The first fundamental challenge is to learn representations that reflect cross-modal interactions between individual elements across different modalities. This challenge can be seen as learning a 'local' representation between elements, or a representation using holistic features. This section covers (1) *representation fusion*: integrating information from 2 or more modalities, effectively reducing the number of separate representations, (2) *representation coordination*: interchanging cross-modal information by keeping the same number of representations but improving multimodal contextualization, and (3) *representation fission*: creating a new decoupled set of representations, usually larger number than the input set, that reflects knowledge about internal structure such as data clustering or factorization (Figure 2.5).

Table 2.1: This table summarizes our taxonomy of 6 core challenges in multimodal machine learning, their subchallenges, categories of corresponding approaches, and representative examples. We believe that this taxonomy can help to catalog rapid progress in this field and better identify the open research questions.

Challenge	Subchallenge	Approaches & key examples
Representation (2.3)	Fusion (2.3.1)	Abstract [283, 712] & raw [47, 501] fusion
	Coordination (2.3.2)	Strong [181, 497] & partial [634, 727] coordination
	Fission (2.3.3)	Modality-level [235, 614] & fine-grained [1, 92] fission
Alignment (2.4)	Discrete connections (2.4.1)	Local [121, 247] & global [351] alignment
	Continuous alignment (2.4.2)	Warping [224, 252] & segmentation [576]
	Contextualization (2.4.3)	Joint [348], cross-modal [232, 390] & graphical [687]
Reasoning (2.5)	Structure modeling (2.5.1)	Hierarchical [26], temporal [676], interactive [394], discovery [478]
	Intermediate concepts (2.5.2)	Attention [680], discrete symbols [22, 632], language [265, 722]
	Inference paradigm (2.5.3)	Logical [203, 586] & causal [8, 448, 698]
	External knowledge (2.5.4)	Knowledge graphs [213, 739] & commonsense [466, 719]
Generation (2.6)	Summarization (2.6.1)	Extractive [96, 628] & abstractive [345, 461]
	Translation (2.6.2)	Exemplar-based [294, 331] & generative [13, 281, 502]
	Creation (2.6.3)	Conditional decoding [142, 452, 737]
Transference (2.7)	Cross-modal transfer (2.7.1)	Tuning [500, 622], multitask [370, 551] & transfer [391]
	Co-learning (2.7.2)	Representation [285, 716] & generation [482, 589]
	Model Induction (2.7.3)	Co-training [65, 159] & co-regularization [563, 692]
Quantification (2.8)	Heterogenity (2.8.1)	Importance [192, 465], bias [231, 473] & noise [398]
	Interconnections (2.8.2)	Connections [7, 79, 601] & interactions [235, 375, 654]
	Learning (2.8.3)	Generalization [370, 505], optimization [651, 670], tradeoffs [367]

2.3.1 Subchallenge 1a: Representation fusion

Representation fusion aims to learn a joint representation that models cross-modal interactions between individual elements of different modalities, effectively *reducing* the number of separate representations. We categorize these approaches into *fusion with abstract modalities* and *fusion with raw modalities* (Figure 2.6). In fusion with abstract modalities, suitable unimodal encoders are first applied to capture a holistic representation of each element (or modality entirely), after which several building blocks for representation fusion are used to learn a joint representation. As a result, fusion happens at the abstract representation level. On the other hand, fusion with raw modalities entails representation fusion at very early stages with minimal preprocessing, perhaps even involving raw modalities themselves.

Fusion with abstract modalities: We begin our treatment of representation fusion of abstract representations with *additive and multiplicative interactions*. These operators can be seen as differentiable building blocks combining information from two streams of data that can be flexibly inserted into almost any unimodal machine learning pipeline. Given unimodal data or features $\mathbf{x}_1$ and $\mathbf{x}_2$, additive fusion can be seen as learning a new joint representation $\mathbf{z}_{\text{mm}} = w_0 + w_1\mathbf{x}_1 + w_2\mathbf{x}_2 + \epsilon$, where w_1 and w_2 are the weights learned for additive fusion of $\mathbf{x}_1$ and $\mathbf{x}_2$, w_0 the bias term, and ϵ the error term. If the joint representation $\mathbf{z}_{\text{mm}}$ is directly taken as a prediction $\hat{y}$, then additive fusion resembles late or ensemble fusion $\hat{y} = f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2)$ with unimodal predictors f_1 and f_2 [180]. Otherwise, the additive representation $\mathbf{z}_{\text{mm}}$ can also undergo subsequent unimodal or multimodal processing [46]. Multiplicative interactions extend additive interactions to include a cross term $w_3(\mathbf{x}_1 \times \mathbf{x}_2)$. These models have been used extensively in statistics, where it can

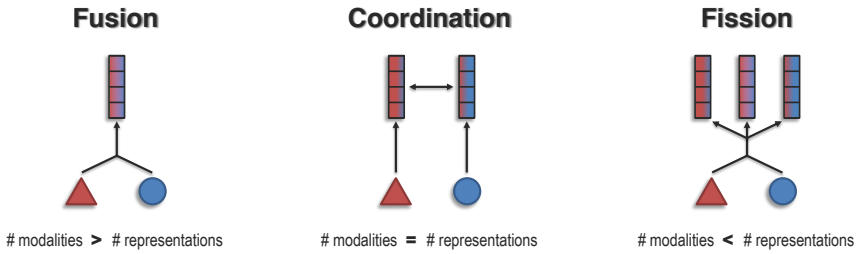


Figure 2.5: Challenge 1 aims to learn **representations** that reflect cross-modal interactions between individual modality elements, through (1) *fusion*: integrating information to reduce the number of separate representations, (2) *coordination*: interchanging cross-modal information by keeping the same number of representations but improving multimodal contextualization, and (3) *fission*: creating a larger set of decoupled representations that reflects knowledge about internal structure.

Figure 2.6: We categorize **representation fusion** approaches into (1) *fusion with abstract modalities*, where unimodal encoders first capture a holistic representation of each element before fusion at relatively homogeneous representations, and (2) *fusion with raw modalities* which entails representation fusion at very early stages, perhaps directly involving heterogeneous raw modalities.

be interpreted as a *moderation* effect of $\mathbf{x}_1$ affecting the linear relationship between $\mathbf{x}_2$ and y [48]. Overall, purely additive interactions $\mathbf{z}_{\text{mm}} = w_0 + w_1\mathbf{x}_1 + w_2\mathbf{x}_2$ can be seen as a first-order polynomial between input modalities $\mathbf{x}_1$ and $\mathbf{x}_2$, combining additive and multiplicative $\mathbf{z}_{\text{mm}} = w_0 + w_1\mathbf{x}_1 + w_2\mathbf{x}_2 + w_3(\mathbf{x}_1 \times \mathbf{x}_2)$ captures a second-order polynomial.

To further go beyond first and second-order interactions, *tensors* are specifically designed to explicitly capture higher-order interactions across modalities [712]. Given unimodal data $\mathbf{x}_1, \mathbf{x}_2$, tensors are defined as $\mathbf{z}_{\text{mm}} = \mathbf{x}_1 \otimes \mathbf{x}_2$ where $\otimes$ denotes an outer product [55, 182]. Tensor products of higher order represent polynomial interactions of higher order between elements [245]. However, computing tensor products is expensive since their dimension scales exponentially with the number of modalities, so several efficient approximations based on low-rank decomposition have been proposed [245, 388]. Finally, *Multiplicative Interactions (MI)* generalize additive and multiplicative operators to include learnable parameters that capture second-order interactions [283]. In its most general form, MI defines a bilinear product $\mathbf{z}_{\text{mm}} = \mathbf{x}_1\mathbb{W}\mathbf{x}_2 + \mathbf{x}_1^\intercal\mathbf{U} + \mathbf{V}\mathbf{x}_2 + \mathbf{b}$ where $\mathbb{W}, \mathbf{U}, \mathbf{Z}$, and $\mathbf{b}$ are trainable parameters.

Multimodal gated units/attention units learn representations that dynamically change for every input [88, 651]. Its general form can be written as $\mathbf{z}_{\text{mm}} = \mathbf{x}_1 \odot h(\mathbf{x}_2)$, where h represents a function with sigmoid activation and $\odot$ denotes element-wise product. $h(\mathbf{x}_2)$ is commonly referred to as 'attention weights' learned from $\mathbf{x}_2$ to attend on $\mathbf{x}_1$. Recent work has explored more expressive forms of learning attention weights such as using Query-Key-Value mechanisms [613], fully-connected neural network layers [32, 88], or even hard gated units for sharper attention [101].

Fusion with raw modalities entails representation fusion at very early stages, perhaps even involving raw modalities themselves. These approaches typically bear resemblance to early

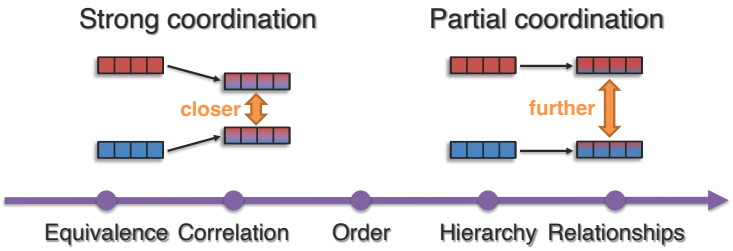


Figure 2.7: There is a spectrum of **representation coordination** functions: *strong coordination* aims to enforce strong equivalence in all dimensions, whereas in *partial coordination* only certain dimensions may be coordinated to capture more general connections such as correlation, order, hierarchies, or relationships.

fusion [46], which performs concatenation of input data before applying a prediction model (i.e., $\mathbf{z}_{\text{mm}} = [\mathbf{x}_1, \mathbf{x}_2]$). Fusing at the raw modality level is more challenging since raw modalities are likely to exhibit more dimensions of heterogeneity. Nevertheless, Barnum et al. [47] demonstrated robustness benefits of fusion at early stages, while Gadzicki et al. [184] also found that complex early fusion can outperform abstract fusion. To account for the greater heterogeneity during complex early fusion, many approaches rely on generic encoders that are applicable to both modalities, such as convolutional layers [47, 184] and Transformers [370, 378]. However, do these complex non-additive fusion models actually learn non-additive interactions between modality elements? Not necessarily, according to Hessel and Lee [235]. We cover these fundamental analysis questions and more in the quantification challenge (§2.8).

2.3.2 Subchallenge 1b: Representation coordination

Representation coordination aims to learn multimodal contextualized representations that are coordinated through their interconnections (Figure 2.7). In contrast to representation fusion, coordination keeps the same number of representations but improves multimodal contextualization. We start our discussion with *strong coordination* that enforces strong equivalence between modality elements, before moving on to *partial coordination* that captures more general connections such as correlation, order, hierarchies, or relationships beyond similarity.

Strong coordination aims to bring semantically corresponding modalities close together in a coordinated space, thereby enforcing strong *equivalence* between modality elements. For example, these models would encourage the representation of the word 'dog' and an image of a dog to be close (i.e., semantically positive pairs), while the distance between the word 'dog' and an image of a car to be far apart (i.e., semantically negative pairs) [181]. The coordination distance is typically cosine distance [414] or max-margin losses [250]. Recent work has explored large-scale representation coordination by scaling up contrastive learning of image and text pairs [497], and also found that contrastive learning provably captures redundant information across the two views [604, 608] (but not non-redundant information). In addition to contrastive learning, several approaches instead learn a coordinated space by mapping corresponding data from one modality to another [162]. For example, Socher et al. [553] maps image embeddings into word embedding spaces for zero-shot image classification. Similar ideas were used to learn coordinated representations between text, video, and audio [482], as well as between pretrained language models and image features [589].

Partial coordination: Instead of capturing strong equivalences, partial coordination captures

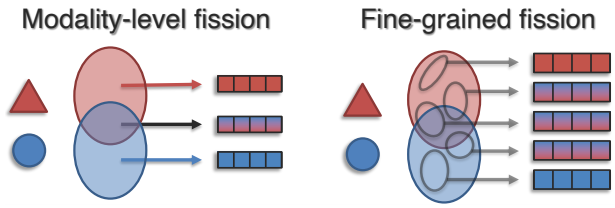


Figure 2.8: Representation fission creates a larger set of decoupled representations that reflects knowledge about internal structure. (1) *Modality-level fission* factorizes into modality-specific information primarily in each modality, and multimodal information redundant in both modalities, while (2) *fine-grained fission* attempts to further break multimodal data down into individual subspaces.

more general modality connections such as correlation, order, hierarchies, or relationships. Partially coordinated models enforce different types of constraints on the representation space beyond semantic similarity, and perhaps only on certain dimensions of the representation.

Canonical correlation analysis (CCA) computes a linear projection that maximizes the correlation between two random variables while enforcing each dimension in a new representation to be orthogonal to each other [599]. CCA models have been used extensively for cross-modal retrieval [504] audio-visual signal analysis [521], and emotion recognition [439]. To increase the expressiveness of CCA, several nonlinear extensions have been proposed including Kernel CCA [325], Deep CCA [27], and CCA Autoencoders [650].

Ordered and hierarchical spaces: Another example of representation coordination comes from order-embeddings of images and language [634], which aims to capture a partial order on the language and image embeddings to enforce a hierarchy in the coordinated space. A similar model using denotation graphs was also proposed by Young et al. [702] where denotation graphs are used to induce such a partial ordering hierarchy.

Relationship coordination: In order to learn a coordinated space that captures semantic relationships between elements beyond correspondences, Zhang et al. [727] use structured representations of text and images to create multimodal concept taxonomies. Delaherche and Chetouani [138] learn coordinated representations capturing hierarchical relationships, while Alviar et al. [20] apply multiscale coordination of speech and music using partial correlation measures. Finally, Xu et al. [678] learn coordinated representations using a Cauchy loss to strengthen robustness to outliers.

2.3.3 Subchallenge 1c: Representation fission

Finally, representation fission aims to create a new decoupled set of representations (usually a larger number than the input representation set) that reflects knowledge about internal multimodal structure such as data clustering, independent factors of variation, or modality-specific information. In comparison with joint and coordinated representations, representation fission enables careful interpretation and fine-grained controllability. Depending on the granularity of decoupled factors, methods can be categorized into *modality-level* and *fine-grained* fission (Figure 2.8).

Modality-level fission aims to factorize into modality-specific information primarily in each modality and multimodal information redundant in both modalities [249, 374, 614]. *Disentangled representation learning* aims to learn mutually independent latent variables that each explain a particular variation of the data [57, 238], and has been useful for modality-level fission by

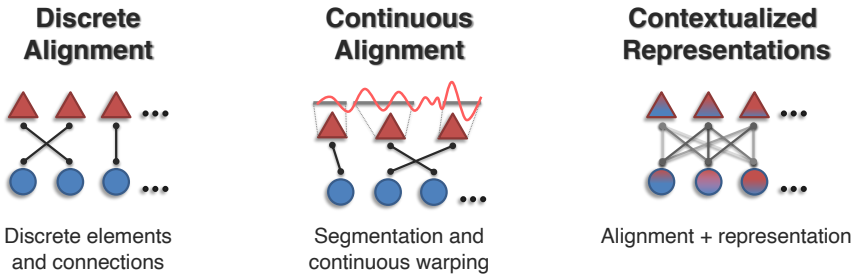


Figure 2.9: Alignment aims to identify cross-modal connections and interactions between modality elements. Recent work has involved (1) *discrete alignment* to identify connections among discrete elements, (2) *continuous alignment* of continuous signals with ambiguous segmentation, and (3) *contextualized representation* learning to capture these cross-modal interactions between connected elements.

enforcing independence constraints on modality-specific and multimodal latent variables [249, 614]. Tsai et al. [614] and Hsu and Glass [249] study factorized multimodal representations and demonstrate the importance of modality-specific and multimodal factors towards generation and prediction. Shi et al. [544] study modality-level fission in multimodal variational autoencoders using a mixture-of-experts layer, while Wu and Goodman [668] instead use a product-of-experts layer.

Post-hoc representation disentanglement is suitable when it is difficult to retrain a disentangled model, especially for large pretrained multimodal models. Empirical multimodally-additive function projection (EMAP) [235] is an approach for post-hoc disentanglement of the effects of unimodal (additive) contributions from cross-modal interactions in multimodal tasks, which works for arbitrary multimodal models and tasks. EMAP is also closely related to the use of Shapley values for feature disentanglement and interpretation [417], which can also be used for post-hoc representation disentanglement in general models.

Fine-grained fission: Beyond factorizing only into individual modality representations, fine-grained fission attempts to further break multimodal data down into the individual subspaces covered by the modalities [637]. *Clustering* approaches that group data based on semantic similarity [402] have been integrated with multimodal networks for end-to-end representation fission and prediction. For example, Hu et al. [250] combine k-means clustering in representations with unsupervised audiovisual learning. Chen et al. [92] combine k-means clustering with self-supervised contrastive learning on videos. Subspace clustering [1, 299], manifold learning [354] approximate graph Laplacians [298], conjugate mixture models [297], and dictionary learning [306] have also been integrated with multimodal models. *Matrix factorization* techniques have also seen several applications in multimodal fission for prediction [16] and cross-modal retrieval [77].

2.4 Challenge 2: Alignment

A second challenge is to identify cross-modal connections and interactions between elements of multiple modalities. For example, when analyzing the speech and gestures of a human subject, how can we align specific gestures with spoken words or utterances? Alignment between modalities is challenging since it may depend on long-range dependencies, involves ambiguous segmentation (e.g., words or utterances), and could be either one-to-one, many-to-many, or not

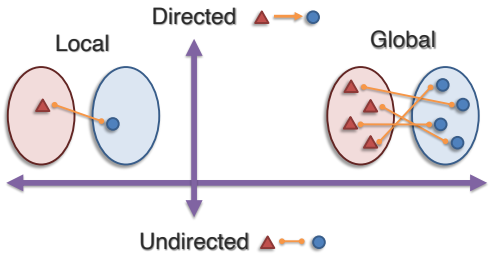


Figure 2.10: Discrete alignment identifies connections between discrete elements, spanning (1) *local alignment* to discover connections given matching pairs, and (2) *global alignment* where alignment must be performed globally to learn both the connections and matchings between modality elements.

exist at all. This section covers recent work in multimodal alignment involving (1) *discrete alignment*: identifying connections between discrete elements across modalities, (2) *continuous alignment*: modeling alignment between continuous modality signals with ambiguous segmentation, and (3) *contextualized representations*: learning better multimodal representations by capturing cross-modal interactions between elements (Figure 2.9).

2.4.1 Subchallenge 2a: Discrete alignment

The first subchallenge aims to identify connections between discrete elements of multiple modalities. We describe recent work in (1) *local alignment* to discover connections between a given matching pair of modality elements, and (2) *global alignment* where alignment must be performed globally to learn both the connections and matchings (Figure 2.10).

Local alignment between connected elements is particularly suitable for multimodal tasks where there is clear segmentation into discrete elements such as words in text or object bounding boxes in images or videos (e.g., tasks such as visual coreference resolution [315], visual referring expression recognition [120, 122], and cross-modal retrieval [181, 485]). When we have supervised data in the form of connected modality pairs, *contrastive learning* is a popular approach where the goal is to match representations of the same concept expressed in different modalities [46]. Several objective functions for learning aligned spaces from varying quantities of paired [80, 260] and unpaired [207] data have been proposed. Many of the ideas that enforce strong [181, 369] or partial [27, 634, 727] representation coordination (§2.3.2) are also applicable for local alignment. Several examples include aligning books with their corresponding movies/scripts [740], matching referring expressions to visual objects [407], and finding similarities between image regions and their descriptions [254]. Methods for local alignment have also enabled the learning of shared semantic concepts not purely based on language but also on additional modalities such as vision [260], sound [121, 553], and multimedia [740] that are useful for downstream tasks.

Global alignment: When the ground-truth modality pairings are not available, alignment must be performed globally between all elements across both modalities. Optimal transport (OT)-based approaches [639] (which belong to a broader set of matching algorithms) are a potential solution since they jointly optimize the coordination function and optimal coupling between modality elements by posing alignment as a divergence minimization problem. These approaches are useful for aligning multimodal representation spaces [351, 491]. To alleviate computational issues, several recent advances have integrated them with neural networks [99], approximated

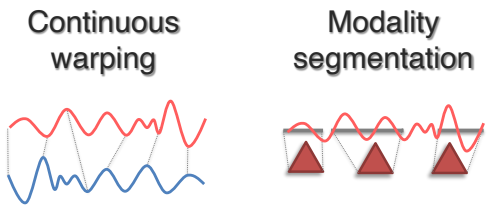


Figure 2.11: Continuous alignment tackles the difficulty of aligning continuous signals where element segmentation is not readily available. We cover related work in (1) *continuous warping* of representation spaces and (2) *modality segmentation* of continuous signals into discrete elements at an appropriate granularity.

optimal transport with entropy regularization [659], and formulated convex relaxations for efficient learning [207].

2.4.2 Subchallenge 2b: Continuous alignment

So far, one important assumption we have made is that modality elements are already segmented and discretized. While certain modalities display clear segmentation (e.g., words/phrases in a sentence or object regions in an image), there are many cases where the segmentation is not readily provided, such as in continuous signals (e.g, financial or medical time-series), spatiotemporal data (e.g., satellite or weather images), or data without clear semantic boundaries (e.g., MRI images). In these settings, methods based on warping and segmentation have been recently proposed:

Continuous warping aims to align two sets of modality elements by representing them as continuous representation spaces and forming a bridge between these representation spaces, such as aligning continuous audio and video data [187, 602, 603]. *Adversarial training* is a popular approach to warp one representation space into another. Initially used in domain adaptation [54], adversarial training learns a domain-invariant representation across domains where a domain classifier is unable to identify which domain a feature came from [14]. These ideas have been extended to align multimodal spaces [247, 252, 431]. Hsu et al. [247] use adversarial training to align images and medical reports, Hu et al. [252] design an adversarial network for cross-modal retrieval, and Munro and Damen [431] design both self-supervised alignment and adversarial alignment objectives for multimodal action recognition. *Dynamic time warping (DTW)* [321] segments and aligns multi-view time-series data by maximizing their similarity via time warping (inserting frames) such that they are aligned across time. For multimodal tasks, it is necessary to design similarity metrics between modalities [28, 593], such as combining DTW with CCA or other coordination functions [611].

Modality segmentation involves dividing high-dimensional data into elements with semantically meaningful boundaries. A common problem involves *temporal segmentation*, where the goal is to discover the temporal boundaries across sequential data. Several approaches for temporal segmentation include forced alignment, a popular approach to align discrete speech units with individual words in a transcript [708]. Malmaud et al. [404] explore multimodal alignment using a factored hidden Markov model to align ASR transcripts to the ground truth. *Clustering* approaches have also been used to group continuous data based on semantic similarity [402]. Clustering-based discretization has recently emerged as an important preprocessing step for generalizing language-based pretraining (with clear word/byte pair segmentation boundaries and

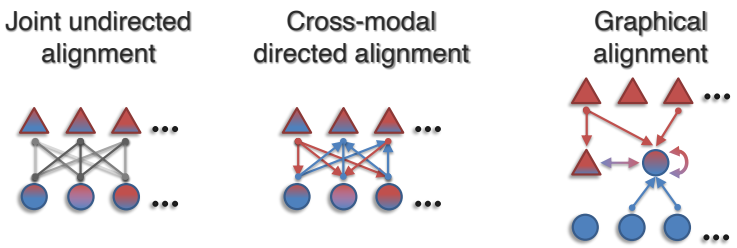


Figure 2.12: Contextualized representation learning aims to model modality connections to learn better representations. Recent directions include (1) *joint undirected alignment* that captures undirected symmetric connections, (2) *cross-modal directed alignment* that models asymmetric connections in a directed manner, and (3) *graphical alignment* that generalizes the sequential pattern into arbitrary graph structures.

discrete elements) to video or audio-based pretraining (without clear segmentation boundaries and continuous elements). By clustering raw video or audio features into a discrete set, approaches such as VideoBERT [576] perform masked pretraining on raw video and audio data. Similarly, approaches such as DALL.E [502], VQ-VAE [629], and CMCM [384] also utilize discretized intermediate layers obtained via vector quantization and showed benefits in modality alignment.

2.4.3 Subchallenge 2c: Contextualized representations

Finally, contextualized representation learning aims to model all modality connections and interactions to learn better representations. Contextualized representations have been used as an intermediate (often latent) step enabling better performance on a number of downstream tasks including speech recognition, machine translation, media description, and visual question-answering. We categorize work in contextualized representations into (1) *joint undirected alignment*, (2) *cross-modal directed alignment*, and (3) *alignment with graph networks* (Figure 2.12).

Joint undirected alignment aims to capture undirected connections across pairs of modalities, where the connections are symmetric in either direction. This is commonly referred to in the literature as unimodal, bimodal, trimodal interactions, and so on [401]. Joint undirected alignment is typically captured by parameterizing models with alignment layers and training end-to-end for a multimodal task. These alignment layers can include attention weights [88], tensor products [388, 712], and multiplicative interactions [283]. More recently, transformer models [631] have emerged as powerful encoders for sequential data by automatically aligning and capturing complementary features at different time steps. Building upon the initial text-based transformer model, multimodal transformers have been proposed that perform joint alignment using a full self-attention over modality elements concatenated across the sequence dimension (i.e., early fusion) [348, 576]. As a result, all modality elements become jointly connected to all other modality elements similarly (i.e., modeling all connections using dot-product similarity kernels).

Cross-modal directed alignment relates elements from a source modality in a directed manner to a target modality, which can model asymmetric connections. For example, *temporal attention models* use alignment as a latent step to improve many sequence-based tasks [676, 724]. These attention mechanisms are typically directed from the output to the input so that the resulting weights reflect a soft alignment distribution over the input. *Multimodal transformers* perform directed alignment using query-key-value attention mechanisms to attend from one modality's sequence to another, before repeating in a bidirectional manner. This results in two sets of

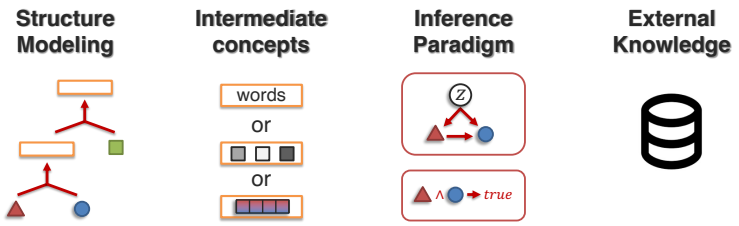


Figure 2.13: Reasoning aims to combine knowledge, usually through multiple inferential steps, exploiting the problem structure. Reasoning involves (1) *structure modeling*: defining or learning the relationships over which reasoning occurs, (2) the *intermediate concepts* used in reasoning, (3) *inference* of increasingly abstract concepts from evidence, and (4) leveraging *external knowledge* in the study of structure, concepts, and inference.

asymmetric contextualized representations to account for the possibly asymmetric connections between modalities [390, 588, 613]. These methods are useful for sequential data by automatically aligning and capturing complementary features at different time-steps [613].

Large vision-language foundation models have emerged as powerful models capable of learning contextualized representations for multiple tasks involving natural language, images, video, and audio [185, 370, 454, 497, 505]. These models typically build on top of pretrained language models [496], pretrained visual encoders [154] combined with an alignment layer. Alignment can be done via end-to-end training with multimodal transformers [681] (e.g., Flamingo [18], OpenFlamingo [37], Kosmos [474]), or keeping the language and vision parts frozen and only training a post-hoc alignment layer (e.g., MiniGPT-4 [736], BLIP-2 [347], InstructBLIP [128], LLaMA-Adapter V2 [186]). Self-supervised pretraining has emerged as an effective way to train these architectures to learn general-purpose representations from larger-scale unlabeled multimodal data before transferring to specific downstream tasks via supervised fine-tuning [155, 348, 736]. Pretraining objectives typically consist of unimodal language modeling [496, 498], image-to-text or text-to-image alignment [232, 736], and multimodal instruction tuning [128, 385, 393]. We refer the reader to recent survey papers discussing these large vision-language models in more detail [157, 185].

Graphical alignment generalizes the sequential pattern seen in undirected or directed alignment into arbitrary graph structures between elements. This has several benefits since it does not require all elements to be connected, and allows the user to choose different edge functions for different connections. Graph neural networks [633] can be used to recursively learn element representations contextualized with the elements in locally connected neighborhoods [523, 633], such as in MTAG [687] and F2F-CL [661] for multimodal and multi-speaker videos.

2.5 Challenge 3: Reasoning

Reasoning is defined as combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and the problem structure. We categorize work towards multimodal reasoning into 4 subchallenges of structure modeling, intermediate concepts, inference paradigm, and external knowledge (Figure 2.13). (1) *Structure modeling* involves defining or learning the relationships over which reasoning occurs, (2) *intermediate concepts* studies the parameterization of individual multimodal concepts in the reasoning process, (3) *inference paradigm* learns how

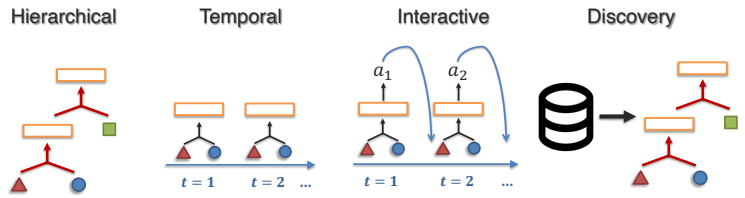


Figure 2.14: Structure modeling aims to define the relationship over which composition occurs, which can be (1) *hierarchical* (i.e., more abstract concepts are defined as a function of less abstract ones), (2) *temporal* (i.e., organized across time), (3) *interactive* (i.e., where the state changes depending on each step's decision), and (4) *discovered* when the latent structure is unknown and instead directly inferred from data and optimization.

increasingly abstract concepts are inferred from individual multimodal evidence, and (4) *external knowledge* aims to leverage large-scale databases in the study of structure, concepts, and inference.

2.5.1 Subchallenge 3a: Structure modeling

Structure modeling aims to capture the hierarchical relationship over which composition occurs, usually via a data structure parameterizing atoms, relations, and the reasoning process. Commonly used data structures include trees [244], graphs [706], or neural modules [26]. We cover recent work in modeling latent *hierarchical*, *temporal*, and *interactive* structure, as well as *structure discovery* when the latent structure is unknown (Figure 2.14).

Hierarchical structure defines a system of organization where abstract concepts are defined as a function of less abstract ones. Hierarchical structure is present in many tasks involving language syntax, visual syntax, or higher-order reasoning. These approaches typically construct a graph based on predefined node and edge categories before using (heterogeneous variants of) graph neural networks to capture a representation of structure [543], such as using language syntactic structure to guide visual modules that discover specific information in images [26, 120]. Graph-based reasoning approaches have been applied for visual commonsense reasoning [380], visual question answering [519], machine translation [699], recommendation systems [592], web image search [647], and social media analysis [525].

Temporal structure extends the notion of compositionality to elements across time, which is necessary when modalities contain temporal information, such as in video, audio, or time-series data. Explicit memory mechanisms have emerged as a popular choice to accumulate multimodal information across time so that long-range cross-modal interactions can be captured through storage and retrieval from memory. Rajagopalan et al. [501] explore various memory representations including multimodal fusion, coordination, and factorization. Insights from key-value memory [676] and attention-based memory [713] have also been successfully applied to applications including question answering, video captioning, emotion recognition, and sentiment analysis.

Interactive structure extends the challenge of reasoning to interactive settings, where the state of the reasoning agent changes depending on the local decisions made at every step. Typically formalized by the sequential decision-making framework, the challenge lies in maximizing long-term cumulative reward despite only interacting with the environment through short-term actions [582]. To tackle the challenges of interactive reasoning, the growing research field of mul-

timodal reinforcement learning (RL) has emerged from the intersection of language understanding, embodiment in the visual world, deep reinforcement learning, and robotics. We refer the reader to the extensive survey paper by Luketina et al. [394] and the position paper by Bisk et al. [62] for a full review of this field. Luketina et al. [394] separate the literature into multimodal-conditional RL (in which multimodal interaction is necessitated by the problem formulation itself, such as instruction following [88, 653]) and language-assisted RL (in which multimodal data is optionally used to facilitate learning, such as reading instruction manuals [437]).

Structure discovery: It may be challenging to define the structure of multimodal composition without some domain knowledge of the given task. As an alternative approach, recent work has also explored using differentiable strategies to automatically search for the structure in a fully data-driven manner. To do so, one first needs to define a candidate set of reasoning atoms and relationships, before using a 'meta' approach such as architecture search to automatically search for the ideal sequence of compositions for a given task [478, 682]. These approaches can benefit from optimization tricks often used in the neural architecture search literature. Memory, Attention, and Composition (MAC) similarly search for a series of attention-based reasoning steps from data in an end-to-end approach [266]. Finally, Hu et al. [255] extend the predefined reasoning structure obtained through language parsing in Andreas et al. [26] by instead using policy gradients to automatically optimize a compositional structure over a discrete set of neural modules.

2.5.2 Subchallenge 3b: Intermediate concepts

The second subchallenge studies how we can parameterize individual multimodal concepts within the reasoning process. While intermediate concepts are usually dense vector representations in standard neural architectures, there has also been substantial work towards interpretable attention maps, discrete symbols, and language as an intermediate medium for reasoning.

Attention maps are a popular choice for intermediate concepts since they are, to a certain extent, human-interpretable, while retaining differentiability. For example, Andreas et al. [26] design individual modules such as 'attend', 'combine', 'count', and 'measure' that are each parametrized by attention operations on the input image for visual question answering. Xu et al. [680] explore both soft and hard attention mechanisms for reasoning in image captioning generation. Related work has also used attention maps through dual attention architectures [434] or stacked latent attention architectures [169] for multimodal reasoning. These are typically applied for problems involving complex reasoning steps such as CLEVR [289] or VQA [729].

Discrete symbols: A further level of discretization beyond attention maps involves using discrete symbols to represent intermediate concepts. Recent work in neuro-symbolic learning aims to integrate these discrete symbols as intermediate steps in multimodal reasoning in tasks such as visual question answering [26, 406, 632] or referring expression recognition [120]. A core challenge in this approach lies in maintaining the differentiability of discrete symbols, which has been tackled via logic-based differentiable reasoning [22, 531].

Language as a medium: Finally, perhaps the most human-understandable form of intermediate concepts is natural language (through discrete words or phrases) as a medium. Recently, Zeng et al. [722] explored using language as an intermediate medium to coordinate multiple separate pretrained models in a zero-shot manner. Several approaches also used language phrases obtained from external knowledge graphs to facilitate interpretable reasoning [213, 739]. Hudson and

Manning [265] designed a neural state machine to simulate the execution of a question being asked about an image, while using discrete words as intermediate concepts.

2.5.3 Subchallenge 3c: Inference paradigms

The third subchallenge in multimodal reasoning defines how increasingly abstract concepts are inferred from individual multimodal evidence. While advances in local representation fusion (such as additive, multiplicative, tensor-based, attention-based, and sequential fusion, see §2.3.1 for a full review) are also generally applicable here, the goal of reasoning is to be more interpretable in the inference process through domain knowledge about the multimodal problem. To that end, we cover recent directions in explicitly modeling the inference process via logical and causal operators as examples of recent trends in this direction.

Logical inference: Logic-based differentiable reasoning has been widely used to represent knowledge in neural networks [22, 531]. Many of these approaches use differentiable fuzzy logic [630] which provides a probabilistic interpretation of logical predicates, functions, and constants to ensure differentiability. These logical operators have been applied for visual question answering [203] and visual reasoning [22]. Among the greatest benefits of logical reasoning lies in its ability to perform interpretable and compositional multi-step reasoning [267]. Logical frameworks have also been useful for visual-textual entailment [586] and geometric numerical reasoning [94], fields where logical inductive biases are crucial toward strong performance.

Causal inference extends the associational level of reasoning to interventional and counterfactual levels [470], which requires extensive knowledge of the world to imagine counterfactual effects. For example, Yi et al. [698] propose the CLEVRER benchmark focusing on four specific elements of reasoning on videos: descriptive (e.g., 'what color'), explanatory ('what's responsible for'), predictive ('what will happen next'), and counterfactual ('what if'). Beyond CLEVRER, recent work has also proposed Causal VQA [8] and Counterfactual VQA [448] to measure the robustness of VQA models under controlled interventions to the question as a step towards mitigating language bias in VQA models. Methods inspired by integrating causal reasoning capabilities into neural network models have also been shown to improve robustness and reduce biases [649].

2.5.4 Subchallenge 3d: External knowledge

The final subchallenge studies the derivation of knowledge in the study of defining composition and structure. Knowledge can refer to any data source that is complementary to the limited supervised training data that models typically see, which encapsulates larger banks of unlabeled internet data (e.g., textbooks, Wikipedia, videos), curated knowledge graphs and knowledge bases, and expert domain knowledge for specific tasks such as healthcare and robotics.

Multimodal knowledge graphs extend classic work in language and symbolic knowledge graphs (e.g., Freebase [66], DBpedia [35], YAGO [572], WordNet [420]) to semantic networks containing multimodal concepts as nodes and multimodal relationships as edges [738]. Multimodal knowledge graphs are important because they enable the grounding of structured information in the visual and physical world. For example, Liu et al. [387] constructs multimodal knowledge graphs containing both numerical features and images for entities. Visual Genome is another example containing dense annotations of objects, attributes, and relationships in images

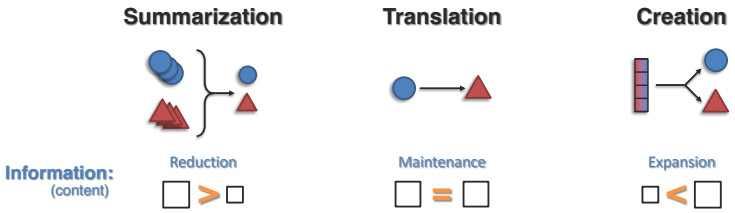


Figure 2.15: How can we learn a generative process to produce raw modalities that reflect cross-modal interactions, structure, and coherence? **Generation** involves (1) *summarizing* multimodal data to highlight the most salient parts, (2) *translating* from one modality to another while being consistent with modality connections, and (3) *creating* multiple modalities simultaneously while maintaining coherence.

and text [317]. These multimodal knowledge bases have been shown to benefit visual question answering [671, 739], knowledge base completion [480], and image captioning [415]. Gui et al. [213] integrates knowledge into vision-and-language transformers for automatic reasoning over both knowledge sources. Another promising approach is multimodal knowledge expansion [495, 672, 683] using knowledge distillation to expand knowledge from unimodal data to multimodal settings. We refer the reader to a comprehensive survey by Zhu et al. [738] for additional references.

Multimodal commonsense reasoning requires deeper real-world knowledge potentially spanning logical, causal, and temporal relationships between concepts. For example, elements of causal reasoning are required to answer the questions regarding images in VCR [719] and VisualCOMET [466], while other works have also introduced datasets with video and text inputs to test for temporal reasoning (e.g., MovieQA [594], MovieFIB [403], TVQA [339]). Benchmarks for multimodal commonsense typically require leveraging external knowledge from knowledge bases [558] or pretraining paradigms on large-scale datasets [390, 720].

2.6 Challenge 4: Generation

The fourth challenge involves learning a generative process to produce raw modalities that reflect cross-modal interactions, structure, and coherence, through *summarization*, *translation*, and *creation* (Figure 9.3). These three categories are distinguished based on the information change from input to output modalities, following categorizations in text generation [141]. We will cover recent advances as well as the evaluation of generated content.

2.6.1 Subchallenge 4a: Summarization

Summarization aims to compress data to create an abstract that represents the most important or relevant information within the original content. Recent work has explored various input modalities to guide text summarization, such as images [95], video [352], and audio [167, 282, 345]. Recent trends in multimodal summarization include *extractive* and *abstractive* approaches. Extractive approaches aim to filter words, phrases, and other unimodal elements from the input to create a summary [96, 282, 345]. Beyond text as output, video summarization is the task of producing a compact version of the video (visual summary) by encapsulating the most informative parts [515]. Li et al. [345] collected a dataset of news videos and articles paired with manually annotated summaries as a benchmark towards multimodal summarization. Finally, UzZaman et al.

[628] aim to simplify complex sentences by extracting multimodal summaries for accessibility. On the other hand, abstractive approaches define a generative model to generate the summary at multiple levels of granularity [95, 350]. Although most approaches only focus on generating a textual summary from multimodal data [461], several directions have also explored generating summarized images to supplement the generated textual summary [95, 352].

2.6.2 Subchallenge 4b: Translation

Translation aims to map one modality to another while respecting semantic connections and information content [640]. For example, generating a descriptive caption of an image can help improve the accessibility of visual content for blind people [215]. Multimodal translation brings about new difficulties involving the generation of high-dimensional structured data as well as their evaluation. Recent approaches can be classified as *exemplar-based*, which are limited to retrieving from training instances to translate between modalities but guarantee fidelity [171], and *generative* models which can translate into arbitrary instances interpolating beyond the data but face challenges in quality, diversity, and evaluation [310, 502, 622]. Despite these challenges, recent progress in large-scale generative models has yielded impressive results in text-to-image [502, 511], text-to-video [550], audio-to-image [281], text-to-speech [507], speech-to-gesture [13], speaker-to-listener [441], language to pose [12], and speech and music generation [9, 124, 452].

2.6.3 Subchallenge 4c: Creation

Creation aims to generate novel high-dimensional data (which could span text, images, audio, video, and other modalities) from small initial examples or latent conditional variables. This *conditional decoding* process is extremely challenging since it needs to be (1) conditional: preserve semantically meaningful mappings from the initial seed to a series of long-range parallel modalities, (2) synchronized: semantically coherent across modalities, (3) stochastic: capture many possible future generations given a particular state, and (4) auto-regressive across possibly long ranges. Many modalities have been considered as targets for creation. Language generation has been explored for a long time [496], and recent work has explored high-resolution speech and sound generation using neural networks [452]. Photorealistic image generation has also recently become possible due to advances in large-scale generative modeling [295]. Furthermore, there have been a number of attempts at generating abstract scenes [587], computer graphics [418], and talking heads [737]. While there has been some progress toward video generation [550], complete synchronized generation of realistic video, text, and audio remains a challenge.

Finally, one of the biggest challenges facing multimodal generation is difficulty in evaluating generated content, especially when there exist serious ethical issues when fake news [56], hate speech [3, 193], deepfakes [222], and lip-syncing videos [583] can be easily generated. While the ideal way to evaluate generated content is through user studies, it is time-consuming, costly, and can potentially introduce subjectivity bias [195]. Several automatic proxy metrics have been proposed [25, 98] by none are universally robust across many generation tasks.

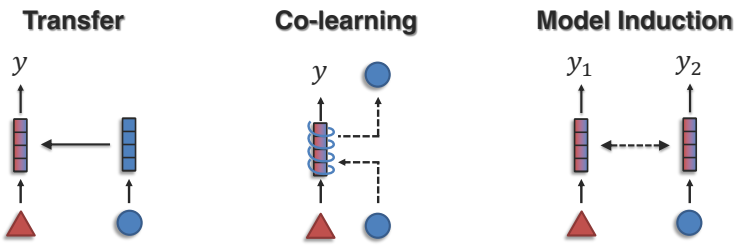


Figure 2.16: Transference studies the transfer of knowledge between modalities, usually to help a noisy or limited primary modality, via (1) *cross-modal transfer* from models trained with abundant data in the secondary modality, (2) *multimodal co-learning* to share information across modalities by sharing representations, and (3) *model induction* that keeps individual unimodal models separate but induces behavior in separate models.

2.7 Challenge 5: Transference

Transference aims to transfer knowledge between modalities and their representations, and is often used when there is a *primary modality* that we care about making predictions on but suffers from limited resources - a lack of annotated data, noisy inputs, or unreliable labels, and a *secondary modality* with more abundant or reliable data. How can knowledge learned from a secondary modality (e.g., predicted labels or representation) help a model trained on a primary modality? We call this challenge transference, since the transfer of information from the secondary modality gives rise to new behaviors previously unseen in the primary modality. We identify three types of approaches: (1) *cross-modal transfer*, (2) *multimodal co-learning*, and (3) *model induction* (Figure 2.16).

2.7.1 Subchallenge 5a: Cross-modal transfer

In most settings, it may be easier to collect either labeled or unlabeled data in the secondary modality and train strong supervised or pretrained models. These models can then be conditioned or fine-tuned for a downstream task involving the primary modality. In other words, this line of research extends unimodal transfer and fine-tuning to cross-modal settings.

Tuning: Inspired by prior work in NLP involving prefix tuning [357] and prompt tuning [342], recent work has also studied the tuning of pretrained language models to condition on visual and other modalities. For example, Tsimpoukelli et al. [622] quickly conditions a pretrained, frozen language model on images for image captioning. Related work has also adapted prefix tuning for image captioning [97], multimodal fusion [226], and summarization [705]. While prefix tuning is simple and efficient, it provides the user with only limited control over how information is transferred. Representation tuning goes a level deeper by modifying the inner representations of the language model via contextualization with other modalities. For example, Ziegler et al. [741] includes additional self-attention layers between language model layers and external modalities. Rahman et al. [500] design a shifting gate to adapt language model layers with audio and visual information.

Multitask learning aims to use multiple large-scale tasks to improve performance as compared to learning on individual tasks. Several models such as Perceiver [276], MultiModel [291], ViT-BERT [353], and PolyViT [378] have explored the possibility of using the same unimodal

encoder architecture for different inputs across unimodal tasks (i.e., language, image, video, or audio-only). The Transformer architecture has emerged as a popular choice due to its suitability for serialized inputs such as text (sequence of tokens) [144], images (sequence of patches) [154], video (sequence of images) [576], and other time-series data (sequence of timesteps) [379]. There have also been several attempts to build a single model that works well on a suite of multimodal tasks, including both not limited to HighMMT [370], VATT [15], FLAVA [551], and Gato [505].

Transfer learning: While more research has focused on transfer within the same modality with external information [553, 675, 716], Liang et al. [369] studies transfer to new modalities using small amounts of paired but unlabeled data. Lu et al. [391] found that Transformers pretrained on language transfer to other sequential modalities as well. Liang et al. [370] builds a single multimodal model capable of transferring to completely new modalities and tasks. Recently, there has also been a line of work investigating the transfer of pretrained language models for planning [259], interactive decision-making [355], and robotics [70].

2.7.2 Subchallenge 5b: Multimodal co-learning

Multimodal co-learning aims to transfer information learned through secondary modalities to target tasks involving the primary modality by sharing intermediate representation spaces between both modalities. These approaches essentially result in a single joint model across all modalities.

Co-learning via representation aims to learn a joint or coordinated representation space using both modalities as input. Typically, this involves adding secondary modalities during the training process, designing a suitable representation space, and investigating how the multimodal model transfers to the primary modality during testing. For example, DeViSE learns a coordinated space between image and text to improve image classification [181]. Marino et al. [409] use knowledge graphs for image classification via a graph-based joint representation. Jia et al. [285] improve image classifiers with contrastive learning between images and noisy captions. Finally, Zadeh et al. [716] showed that implicit co-learning is also possible without explicit co-learning objectives.

Co-learning via generation instead learns a translation model from the primary to secondary modality, resulting in enriched representations of the primary modality that can predict both the label and 'hallucinate' secondary modalities containing shared information. Classic examples in this category includes language modeling by mapping contextualized text embeddings into images [589], image classification by projecting image embeddings into word embeddings [553], and language sentiment analysis by translating language into video and audio [482].

2.7.3 Subchallenge 5c: Model induction

In contrast to co-learning, model induction approaches keep individual unimodal models across primary and secondary modalities separate but transfer information across them. There are two general ways of doing so. The first is co-training, where each unimodal model's predictions on their own modality are used to pseudo-label new unlabeled examples in the other modality, thereby enlarging the training set of the other modality [65]. The second is co-regularization [549, 563], in which the predictions from separate unimodal classifiers are regularized to be similar, thereby encouraging both classifiers to share information (i.e., redundancy). Therefore, information is transferred across modalities through model predictions instead of shared representation spaces.

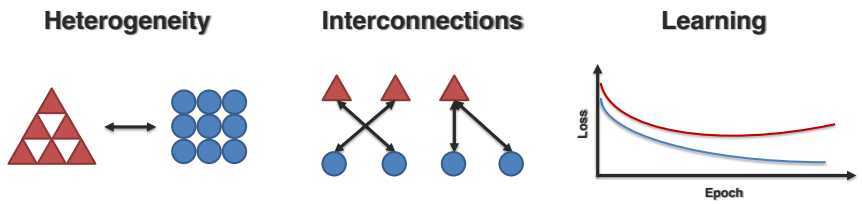


Figure 2.17: Quantification: what are the empirical and theoretical studies we can design to better understand (1) the dimensions of *heterogeneity*, (2) the presence and type of *interconnections*, and (3) the *learning* and optimization challenges?

Multimodal co-training extends co-training by jointly learning classifiers for multiple modalities [239]. Guillaumin et al. [214] use a classifier on both image and text to pseudo-label unlabeled images before training a final classifier on both labeled and unlabeled images. Cheng et al. [111] performs semi-supervised multimodal learning using a diversity-preserving co-training algorithm. Finally, Dunnmon et al. [159] applies ideas from data programming to the problem of cross-modal weak supervision, where weak labels derived from a secondary modality (e.g., text) are used to train models over the primary modality (e.g., images).

Co-regularization methods employs a regularizer that penalizes functions from either modality that disagree with each other. These methods are useful in controlling model complexity by preferring hypothesis classes with redundancy across the two modalities [549]. Sridharan and Kakade [563] provide guarantees for these approaches using an information-theoretic framework. More recently, similar co-regularization approaches have also been applied for multimodal feature selection [246], semi-supervised multimodal learning [692], and video summarization [428].

2.8 Challenge 6: Quantification

Quantification aims to provide a deeper empirical and theoretical study of multimodal models to gain insights and improve their robustness, interpretability, and reliability in real-world applications. We break down quantification into 3 sub-challenges: (1) quantifying the *dimensions of heterogeneity* and how they subsequently influence modeling and learning, (2) quantifying the presence and type of *connections and interactions* in multimodal datasets and trained models, and (3) characterizing the *learning and optimization* challenges involved when learning from heterogeneous data (Figure 2.17).

2.8.1 Subchallenge 6a: Dimensions of heterogeneity

This subchallenge aims to understand the dimensions of heterogeneity commonly encountered in multimodal research, and how they subsequently influence modeling and learning (Figure 2.18).

Modality information: Understanding the information of modalities and their constituents is important for determining which parts contributed to subsequent modeling. Recent work can be categorized into (1) interpretable methods that explicitly model how each modality is used [465, 615, 717] or (2) post-hoc explanations of black-box models [85, 205]. In the former, methods such as Concept Bottleneck Models [311] and fitting sparse linear layers [666] or decision trees [641] on top of deep feature representations have emerged as promising choices. In the latter, gradient-based visualizations [205, 529, 548]) and feature attributions (e.g., modality

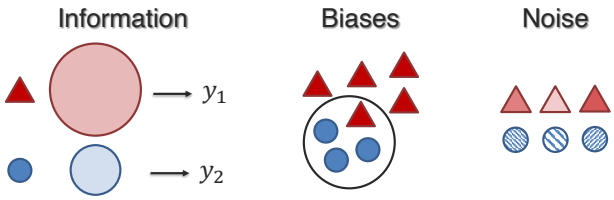


Figure 2.18: The subchallenge of **heterogeneity** quantification aims to understand the dimensions of heterogeneity commonly encountered in multimodal research, such as (1) different quantities and usages of *modality information*, (2) the presence of *modality biases*, and (3) quantifying and mitigating *modality noise*.

contribution [192], LIME [509], and Shapley values [417]) have been used to highlight regions of modality importance.

Modality biases are unintended correlations between input and outputs that could be introduced during data collection [61, 67], modeling [194], or during human annotation [143]. Modality biases can lead to unexpectedly poor performance in the real world [516], or even more dangerously, potential for harm towards underrepresented groups [231, 473]. For example, Goyal et al. [206] found *unimodal biases* in the language modality of VQA tasks, resulting in mistakes due to ignoring visual information [10]. Subsequent work has developed carefully curated diagnostic benchmarks to mitigate data collection biases, like VQA 2.0 [206], GQA [267], and NLVR2 [573]. Recent work has also found compounding *social biases* in multimodal systems [114, 512, 564] stemming from gender bias in both language and visual modalities [73, 542], which may cause danger when deployed [473].

Modality noise topologies and robustness: The study of modality noise topologies aims to benchmark and improve how multimodal models perform in the presence of real-world data imperfections. Each modality has a unique noise topology, which determines the distribution of noise and imperfections that it commonly encounters. For example, images are susceptible to blurs and shifts, typed text is susceptible to typos following keyboard positions, and multimodal time-series data is susceptible to correlated imperfections across synchronized time steps. Liang et al. [367] collect a comprehensive set of targeted noisy distributions unique to each modality. In addition to natural noise topologies [338, 399], related work has also explored adversarial attacks [149] and distribution shifts [176] in multimodal systems. Finally, there have been recent efforts on incomplete multimodal learning [370, 398, 648, 691] to account for noisy or missing modalities, such as modality imputation using probabilistic models [398], autoencoders [609], translation models [482], low-rank approximations [364], or knowledge distillation [648], or training general models with a wide range of modalities so they can still operate on partial subsets [370, 505]. However, they may run the risk of possible error compounding and require knowing which modalities are imperfect beforehand.

2.8.2 Subchallenge 6b: Modality interconnections

Modality connections and interactions are an essential component of multimodal models, which has inspired an important line of work in visualizing and understanding the nature of modality interconnections in datasets and trained models. We divide recent work into quantifying (1) *connections*: how modalities are related and share commonality, and (2) *interactions*: how

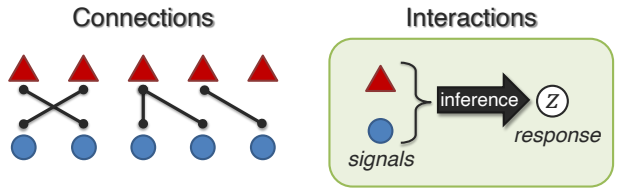


Figure 2.19: Quantifying **modality interconnections** studies (1) *connections*: can we discover what modality elements are related to each other and why, and (2) *interactions*: can we understand how modality elements interact during inference?

modality elements interact during inference (Figure 2.19).

Connections: Recent work has explored the quantification of modality connections through visualization tools on joint representation spaces [271] or attention maps [7]. Perturbation-based analysis perturbs the input and observes changes in the output to understand internal connections [375, 448]. Finally, specifically curated diagnostic datasets are also useful in understanding semantic connections: Winoground [601] probes vision and language models for visio-linguistic compositionality, and PaintSkills [114] measures the connections necessary for visual reasoning.

Interactions: One common categorization of interactions involves redundancy, uniqueness, and synergy [662]. Redundancy describes task-relevant information shared among features, uniqueness studies the task-relevant information present in only one of the features, and synergy investigates the emergence of new information when both features are present. From a statistical perspective, measures of redundancy include mutual information [44, 65] and contrastive learning estimators [608, 616]. Other approaches have studied these measures in isolation, such as redundancy via distance between prediction logits using either feature [411], statistical distribution tests on input features [36], or via human annotations [514]. From the semantic view, recent work in Causal VQA [8] and Counterfactual VQA [448] seek to understand the interactions captured by trained models by measuring their robustness under controlled semantic edits to the question or image. Finally, recent work has formalized definitions of non-additive interactions to quantify their presence in trained models [562, 619, 684]. Parallel research such as EMAP [235], DIME [396], M2Lens [654], and MultiViz [375] take a more visual approach to visualize the interactions in real-world multimodal datasets and models through higher-order gradient activations of learned representations. Despite this, accurately visualizing multimodal information and interactions remains a challenge due to the brittleness of interpretation methods [197], difficulty in evaluation [318], and challenges in extending visualization methods to applications such as biomedical data integration, imaging, intelligent systems and user interfaces.

2.8.3 Subchallenge 6c: Multimodal learning process

Finally, there is a need to characterize the learning and optimization challenges involved when learning from heterogeneous data. This section covers recent work in (1) *generalization* across modalities and tasks, (2) better *optimization* for balanced and efficient training, and (3) balancing the *tradeoffs* between performance, robustness, and complexity in real-world deployment (Figure 2.20).

Generalization: With advances in sensing technologies, many real-world platforms such as cellphones, smart devices, self-driving cars, healthcare technologies, and robots now integrate a

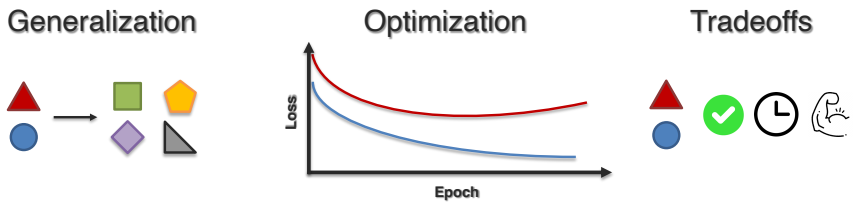


Figure 2.20: Studying the multimodal **learning process** involves understanding (1) *generalization* across modalities and tasks, (2) *optimization* for balanced and efficient training, and (3) *tradeoffs* between performance, robustness, and complexity in the real-world deployment of multimodal models.

much larger number of sensors beyond the prototypical text, video, and audio modalities [263]. Recent work has studied generalization across paired modality inputs [369, 497] and in unpaired scenarios where each task is defined over only a small subset of all modalities [370, 391, 505].

Optimization challenges: Related work has also explored the optimization challenges of multimodal learning, where multimodal networks are often prone to overfitting due to increased capacity, and different modalities overfit and generalize at different rates so training them jointly with a single optimization strategy is sub-optimal [651]. Subsequent work has studied why joint training of multimodal networks may be difficult and proposed methods to improve the optimization process via weighting approaches [670], adaptive learning [261, 262], or contrastive learning [377].

Modality Tradeoffs: In real-world deployment, a balance between performance, robustness, and complexity is often required. Therefore, one often needs to balance the utility of additional modalities with the additional complexity in data collection and modeling [367] as well as increased susceptibility to noise and imperfection in the additional modality [482]. How can we formally quantify the utility and risks of each input modality, while balancing these tradeoffs for reliable real-world usage? There have been several attempts toward formalizing the semantics of a multimodal representation and how these benefits can transfer to downstream tasks [358, 598, 616], while information-theoretic arguments have also provided useful insights [65, 563].

Chapter 3

Machine Learning Foundations of Multimodal Interactions

A core challenge in machine learning lies in capturing the interactions between multiple input modalities. Learning different types of multimodal interactions is often quoted as motivation for many successful multimodal modeling paradigms, such as contrastive learning to capture redundancy [307, 497], modality-specific representations to retain unique information [614], as well as tensors and multiplicative interactions to learn higher-order interactions [283, 364, 712]. However, several fundamental research questions remain: *How can we quantify the interactions that are necessary to solve a multimodal task? Subsequently, what are the most suitable multimodal models to capture these interactions?* This paper aims to formalize these questions by proposing an approach to quantify the *nature* (i.e., which type) and *degree* (i.e., the amount) of modality interactions, a fundamental principle underpinning our understanding of multimodal datasets and models [371].

By bringing together two previously disjoint research fields of Partial Information Decomposition (PID) in information theory [59, 210, 662] and multimodal machine learning [46, 371], we provide precise definitions categorizing interactions into *redundancy*, *uniqueness*, and *synergy*. Redundancy quantifies information shared between modalities, uniqueness quantifies the information present in only one of the modalities, and synergy quantifies the emergence of new information not previously present in either modality. A key aspect of these four measures is that they not only quantify interactions between modalities, but also how they relate to a downstream task. Figure 3.1 shows a depiction of these four measures, which we refer to as PID statistics. Leveraging insights from neural representation learning, we propose two new estimators for PID statistics that can scale to high-dimensional multimodal datasets and models. The first estimator is exact, based on convex optimization, and is able to scale to features with discrete support, while the second estimator is an approximation based on sampling, which enables us to handle features with large discrete or even continuous supports. We validate our estimation of PID in 2 ways: (1) on synthetic datasets where PID statistics are known from the nature of data generation, and (2) on real-world data where PID is compared with human annotation. Finally, we demonstrate that estimated PID statistics can help in multimodal applications involving:

1. **Dataset quantification**: We apply PID to quantify large-scale multimodal datasets, showing that these estimates match common intuition for interpretable modalities (e.g., language, vision,

and audio) and yield new insights in other domains (e.g, healthcare, HCI, and robotics).

2. **Model quantification**: Across a suite of models, we apply PID to interpret model predictions and find consistent patterns of interactions that different models capture.
3. **Model selection**: Given our findings from dataset and model quantification, a new research question naturally arises: *given a new multimodal task, can we quantify its* PID *values to infer (a priori) what type of models are most suitable?* Our experiments show success in model selection for both existing benchmarks and completely new case studies engaging with domain experts in computational pathology, mood prediction, and robotics to select the best multimodal model.

Finally, we release a suite of trained models across 10 model families and 30 datasets to accelerate future analysis of multimodal interactions at `https://github.com/pliang279/PID`.

3.1 Background and Related Work

Let $\mathcal{X}_i$ and $\mathcal{Y}$ be sample spaces for features and labels. Define Δ to be the set of joint distributions over $(\mathcal{X}_1, \mathcal{X}_2, \mathcal{Y})$. We are concerned with features X_1, X_2 (with support $\mathcal{X}_i$) and labels Y (with support $\mathcal{Y}$) drawn from some distribution $p \in \Delta$. We denote the probability mass (or density) function by $p(x_1, x_2, y)$, where omitted parameters imply marginalization. Key to our work is defining estimators that given p or samples $\{(x_1, x_2, y) : \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y}\}$ thereof (i.e., dataset or model predictions), estimates the amount of redundant, unique, and synergistic interactions.

3.1.1 Partial information decomposition

Information theory formalizes the amount of information that one variable provides about another [535]. However, its extension to 3 variables is an open question [190, 412, 596, 658]. In particular, the natural three-way mutual information $I(X_1; X_2; Y) = I(X_1; X_2) - I(X_1; X_2|Y)$ [412, 596] can be both positive and negative, which makes it difficult to interpret. In response, Partial information decomposition (PID) [662] generalizes information theory to multiple variables by decomposing $I_p(X_1, X_2; Y)$, the total information 2 variables X_1, X_2 provide about a task Y into 4 quantities (see Figure 3.1): redundancy R between X_1 and X_2, uniqueness U_1 in X_1 and U_2 in X_2, and synergy S that only emerges when both X_1 and X_2 are present. We adopt the PID definition proposed by Bertschinger et al. [59]:

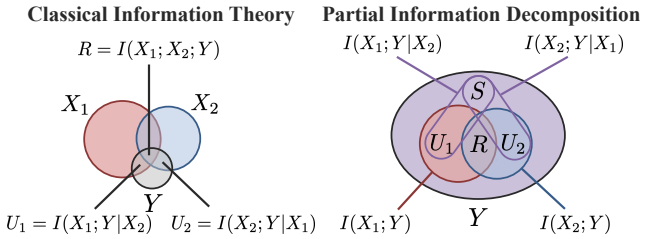


Figure 3.1: PID decomposes $I(X_1, X_2; Y)$ into redundancy R between X_1 and X_2, uniqueness U_1 in X_1 and U_2 in X_2, and synergy S in both X_1 and X_2.

$$R = \max_{q \in \Delta_p} I_q(X_1; X_2; Y), \tag{3.1}$$

$$U_1 = \min_{q \in \Delta_p} I_q(X_1; Y|X_2), \quad U_2 = \min_{q \in \Delta_p} I_q(X_2; Y|X_1), \tag{3.2}$$

$$S = I_p(X_1, X_2; Y) - \min_{q \in \Delta_p} I_q(X_1, X_2; Y), \tag{3.3}$$

where $\Delta_p = \{q \in \Delta : q(x_i, y) = p(x_i, y)\ \forall y \in \mathcal{Y}, x_i \in \mathcal{X}_i, i \in [2]\}$ and the notation $I_p(\cdot)$ and $I_q(\cdot)$ disambiguates mutual information under p and q respectively. The key lies in optimizing $q \in \Delta_p$ to satisfy the marginals $q(x_i, y) = p(x_i, y)$, but relaxing the coupling between x_1 and x_2: $q(x_1, x_2)$ need not be equal to $p(x_1, x_2)$. The intuition behind this is that one should be able to infer redundancy and uniqueness given only access to $p(x_1, y)$ and $p(x_2, y)$, and therefore they should only depend on $q \in \Delta_p$. Synergy is the only term that should depend on the coupling $p(x_1, x_2)$, and this is reflected in (6.2) depending on the full p distribution. This definition enjoys several useful properties in line with intuition, as we will see in comparison with related frameworks for interactions below [59].

3.1.2 Related frameworks for feature interactions

Information-theoretic definitions: Perhaps the first measure of redundancy in machine learning is co-training [44, 65, 116], where 2 variables are redundant if they are conditionally independent given the task: $I(X_1; X_2|Y) = 0$. As a result, redundancy can be measured by $I(X_1; X_2; Y)$. The same definition of redundancy is used in multi-view learning [563, 605, 608] which further define $I(X_1; Y|X_2)$ and $I(X_2; Y|X_1)$ as unique information in X_1, X_2. However, $I(X_1; X_2; Y)$ can be both positive and negative [280]. PID resolves this by separating R and S such that $R - S = I(X_1; X_2; Y)$, identifying that prior measures confound redundancy and synergy. This crucially provides an explanation for the distinction between *mediation*, where one feature conveys the information already in another (i.e., $R > S$), versus *moderation*, where one feature affects the relationship of other features (i.e., $S > R$) [48, 196]. Furthermore, if $I(X_1; X_2; Y) = 0$ then existing frameworks are unable to distinguish between positive R and S canceling each other out.

Statistical measures: Other approaches have studied interaction measures via statistical measures, such as redundancy via distance between prediction logits using either feature [411], statistical distribution tests on input features [36, 703], or via human annotations [514]. However, it is unclear how to extend these definitions to uniqueness and synergy while remaining on the same standardized scale like PID provides. Also of interest are notions of redundant and synergistic interactions in human and animal communication [175, 468, 469, 514], which we aim to formalize.

Model-based methods: Prior research has formalized definitions of non-additive interactions [180] to quantify their presence [235, 562, 619, 620] in trained models, or used Shapley values on trained features to measure interactions [272]. Parallel research has also focused on qualitative visualizations of real-world multimodal datasets and models, such as DIME [396], M2Lens [654], and MultiViz [375].

3.2 Scalable Estimators for PID

PID as a framework for multimodality: Our core insight is that PID provides a formal framework to understand both the *nature* and *degree* of interactions involved when two features X_1 and X_2 are used for task Y. The nature of interactions is afforded by a precise decomposition into redundant, unique, and synergistic interactions, and the degree of interactions is afforded by a standardized unit of measure (bits). However, computing PID is a considerable challenge, since it involves optimization over Δ_p and estimating information-theoretic measures. Up to now, analytic

approximations of these quantities were only possible for discrete and small support [59, 210, 664] or continuous but low-dimensional variables [460, 493, 665]. Leveraging ideas in representation learning, Sections 3.2.1 and 3.2.2 are our first technical contributions enabling scalable estimation of PID for high-dimensional distributions. The first, CVX, is exact, based on convex optimization, and is able to scale to problems where $|\mathcal{X}_i|$ and $|\mathcal{Y}|$ are around 100. The second, BATCH, is an approximation based on sampling, which enables us to handle large or even continuous supports for X_i and Y. Applying these estimators in Section 3.3, we show that PID provides a path towards understanding the nature of interactions in datasets and those learned by different models, and principled approaches for model selection.

3.2.1 CVX: Dataset-level optimization

Our first estimator, CVX, directly compute PID from its definitions using convex programming. Crucially, Bertschinger et al. [59] show that the solution to the max-entropy optimization problem: $q^\star = \arg\max_{q \in \Delta_p} H_q(Y|X_1, X_2)$ equivalently solves (3.1)-(6.2). When $\mathcal{X}_i$ and $\mathcal{Y}$ are small and discrete, we can represent all valid distributions $q(x_1, x_2, y)$ as a set of tensors Q of shape $|\mathcal{X}_1| \times |\mathcal{X}_2| \times |\mathcal{Y}|$ with each entry representing $Q[i, j, k] = p(X_1 = i, X_2 = j, Y = k)$. The problem then boils down to optimizing over valid tensors $Q \in \Delta_p$ that match the marginals $p(x_i, y)$.

Given a tensor Q representing q, our objective is the concave function $H_q(Y|X_1, X_2)$. While Bertschinger et al. [59] report that direct optimization is numerically difficult as routines such as Mathematica's FINDMINIMUM do not exploit convexity, we overcome this by rewriting conditional entropy as a KL-divergence [201], $H_q(Y|X_1, X_2) = \log|\mathcal{Y}| - KL(q\|\tilde{q})$, where $\tilde{q}$ is an auxiliary product density of $q(x_1, x_2) \cdot \frac{1}{|\mathcal{Y}|}$ enforced using linear constraints: $\tilde{q}(x_1, x_2, y) = q(x_1, x_2)/|\mathcal{Y}|$. Finally, optimizing over $Q \in \Delta_p$ that match the marginals can also be enforced through linear constraints: the 3D-tensor Q summed over the second dimension gives $q(x_1, y)$ and summed over the first dimension gives $q(x_2, y)$, yielding the final optimization problem:

$$\arg\max_{Q, \tilde{Q}} KL(Q\|\tilde{Q}), \quad \text{s.t.} \quad \tilde{Q}(x_1, x_2, y) = Q(x_1, x_2)/|\mathcal{Y}|, \tag{3.4}$$

$$\sum_{x_2} Q = p(x_1, y), \sum_{x_1} Q = p(x_2, y), Q \geq 0, \sum_{x_1, x_2, y} Q = 1. \tag{3.5}$$

The KL-divergence objective is recognized as convex, allowing the use of conic solvers such as SCS [451], ECOS [152], and MOSEK [30]. Plugging $q^\star$ into (3.1)-(6.2) yields the desired PID.

Pre-processing via feature binning: In practice, X_1 and X_2 often take continuous rather than discrete values. Thus, Q is no longer a finite dimensional polytope. We work around this by histogramming each X_i, thereby estimating the continuous joint density by discrete distributions with finite support. To make our discretization as data-independent as possible, we focus on a prespecified number of fixed-width bins (except for the first and last). For example, it is known that with a fixed number of samples, making the width of bins arbitrarily small will cause KL estimates to diverge. It is known that the number of bins should grow sub-linearly with the number of samples. For example, Rice [510] suggest setting the number of bins to be the cubed-root of number of samples.

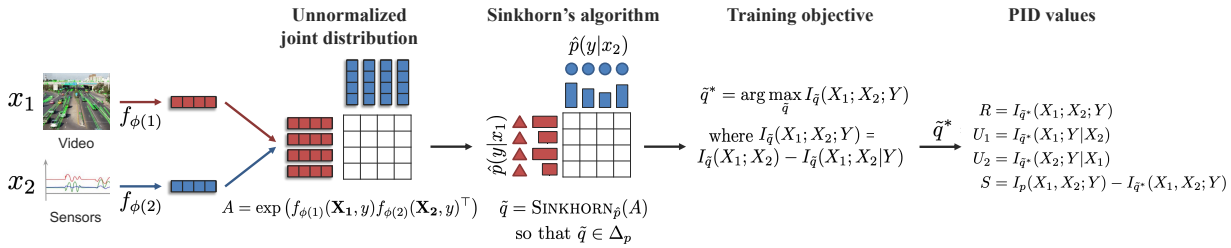


Figure 3.2: We propose BATCH, a scalable estimator for PID over high-dimensional continuous distributions. BATCH parameterizes $\tilde{q}$ using a matrix A learned by neural networks such that mutual information objectives over $\tilde{q}$ can be optimized via gradient-based approaches over minibatches. Marginal constraints $\tilde{q} \in \Delta_p$ are enforced through a variant of the Sinkhorn-Knopp algorithm on A.

3.2.2 BATCH: Batch-level amortization

We now present BATCH, our next estimator that is suitable for large datasets where $\mathcal{X}_i$ is high-dimensional and continuous ($|\mathcal{Y}|$ remains finite). To estimate PID given a sampled dataset $\mathcal{D} = \{(x_1^{(j)}, x_2^{(j)}, y^{(j)})\}$ of size n, we propose an end-to-end model parameterizing marginal-matching joint distributions in Δ_p and a training objective whose solution returns approximate PID values.

Simplified algorithm sketch: Our goal, loosely speaking, is to optimize $\tilde{q} \in \Delta_p$ for objective (3.1) through an approximation using neural networks instead of exact optimization. We show an overview in Figure 3.2. To explain our approach, we first describe (1) how we parameterize $\tilde{q}$ using neural networks such that it can be learned via gradient-based approaches, (2) how we ensure the marginal constraints $\tilde{q} \in \Delta_p$ through a variant of the Sinkhorn-Knopp algorithm, and finally (3) how to scale this up over small subsampled batches from large multimodal datasets.

Parameterization using neural networks: The space of joint distributions Δ is often too large to explicitly specify. To tackle this, we implicitly parameterize each distribution $\tilde{q} \in \Delta$ using a neural network f_ϕ that takes in batches of modalities $\mathbf{X_1} \in \widetilde{\mathcal{X}}_1^n, \mathbf{X_2} \in \widetilde{\mathcal{X}}_2^n$ and the label $\mathbf{Y} \in \mathcal{Y}^n$ before returning a matrix $A \in \mathbb{R}^{n \times n \times |\mathcal{Y}|}$ representing an (unnormalized) joint distribution $\tilde{q}$, i.e., we want $A[i][j][y] = \tilde{q}(\mathbf{X_1}[i], \mathbf{X_2}[j], y)$ for each $y \in \mathcal{Y}$. In practice, f_ϕ is implemented via a pair of encoders $f_{\phi(1)}$ and $f_{\phi(2)}$ that learn modality representations, before an outer product to learn joint relationships $A_y = \exp(f_{\phi(1)}(\mathbf{X_1}, y) f_{\phi(2)}(\mathbf{X_2}, y)^\top)$ for each y, yielding the desired $n \times n \times |\mathcal{Y}|$ joint distribution. As a result, optimizing over $\tilde{q}$ can be performed via optimizing over parameters ϕ.

Respecting the marginal constraints: How do we make sure the $\tilde{q}$'s learned by the network satisfies the marginal constraints (i.e., $\tilde{q} \in \Delta_p$)? We use an unrolled version of Sinkhorn's algorithm [127] which projects A onto Δ_p by iteratively normalizing A's rows and columns to sum to 1 and rescaling to satisfy the marginals $p(x_i, y)$. However, $p(x_i, y)$ is not easy to estimate for high-dimensional continuous x_i's. In response, we first expand $p(x_i, y)$ into $p(y|x_i)$ and $p(x_i)$ using Bayes' rule. Since A was constructed by samples x_i from the dataset, the rows and columns of A are already distributed according to $p(x_1)$ and $p(x_2)$ respectively. This means that it suffices to approximate $p(y|x_i)$ with unimodal classifiers $\hat{p}(y|x_i)$ parameterized by neural networks and trained separately, before using Sinkhorn's algorithm to normalize each row to $\hat{p}(y|x_1)$ and each column to $\hat{p}(y|x_2)$.

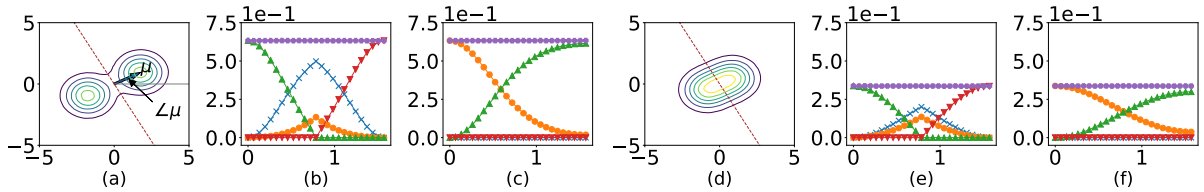


Figure 3.3: Left to right: (a) Contour plots of the GMM's density for $||\mu||_2 = 2.0$. Red line denotes the optimal linear classifier. (b) PID (Cartesian) computed for varying $\angle\mu$ with respect to the x axis. (c) PID (Polar) for varying $\angle\mu$, with U_1 and U_2 corresponding to unique information from (r, θ). Plots (d)-(f) are similar to (a)-(c), but repeated for $||\mu||_2 = 1.0$. Legend: ✖(R), ▲(U_1), ▼(U_2), ●(S), ✚(Sum). Observe how PID changes with the change of variable from Cartesian (b and e) to Polar (c and f), as well as how a change in $||\mu||_2$ can lead to a disproportionate change across PID (b vs e).

Objective: We choose the objective $I_q(X_1; X_2; Y)$, which equivalently solves the optimization problems in the other PID terms [59]. Given matrix A representing $\tilde{q}(x_1, x_2, y)$, the objective can be computed in closed form through appropriate summation across dimensions in A to obtain $\tilde{q}(x_i)$, $\tilde{q}(x_1, x_2)$, $\tilde{q}(x_i|y)$, and $\tilde{q}(x_1, x_2|y)$ and plugging into $I_{\tilde{q}}(X_1; X_2; Y) = I_{\tilde{q}}(X_1; X_2) - I_{\tilde{q}}(X_1; X_2|Y)$. We maximize $I_{\tilde{q}}(X_1; X_2; Y)$ by updating parameters ϕ via gradient-based methods. Overall, each gradient step involves computing $\tilde{q} = \text{SINKHORN}_{\hat{p}}(A)$, and updating ϕ to maximize (3.1) under $\tilde{q}$. Since Sinkhorn's algorithm is differentiable, gradients can be backpropagated end-to-end.

Approximation with small subsampled batches: Finally, to scale this up to large multimodal datasets where the full $\tilde{q}$ may be too large to store, we approximate $\tilde{q}$ with small subsampled batches: for each gradient iteration t, the network f_ϕ now takes in a batch of $m \ll n$ datapoints sampled from $\mathcal{D}$ and returns $A \in \mathbb{R}^{m \times m \times |\mathcal{Y}|}$ for the subsampled points. We perform Sinkhorn's algorithm on A and a gradient step on ϕ as above, *as if* $\mathcal{D}_t$ was the full dataset (i.e., mini-batch gradient descent). While it is challenging to obtain full-batch gradients since computing the full A is intractable, we found our approach to work in practice for large m. Our approach can also be informally viewed as performing amortized optimization [23] by using ϕ to implicitly share information about the full batch using subsampled batches. Upon convergence of ϕ, we extract PID by plugging $\tilde{q}$ into (3.1)-(6.2).

Implementation details such as the network architecture of f, approximation of objective (3.1) via sampling from $\tilde{q}$, and estimation of $I_{\tilde{q}}(\{X_1, X_2\}; Y)$ from learned $\tilde{q}$ are included in the full version of the paper [372].

3.3 Evaluation and Applications of PID in Multimodal Learning

We design experiments to (1) understand PID on synthetic data, (2) quantify real-world multimodal benchmarks, (3) understand the interactions captured by multimodal models, (4) perform model selection across different model families, and (5) applications on novel real-world tasks.

3.3.1 Validating PID estimates on synthetic data

Our first goal is to evaluate the accuracy of our proposed estimators with respect to the ground truth (if it can be computed) or human judgment (for cases where the ground truth cannot be readily obtained). We start with a suite of datasets spanning both synthetic and real-world distributions.

Synthetic bitwise features: We sample from a binary bitwise distribution: $x_1, x_2 \sim \{0, 1\}, y = x_1 \wedge x_2, y = x_1 \vee x_2, y = x_1 \oplus x_2$,. Each bitwise operator's PID can be solved exactly when the x_i's and labels are discrete and low-dimensional [59]. Compared to the ground truth in Bertschinger et al. [59], both our estimators exactly recover the correct PID values (Table 3.1).

Table 3.1: Results on estimating PID on synthetic bitwise datasets. Both our estimators exactly recover the correct PID values as reported in Bertschinger et al. [59].

Task	OR				AND				XOR			
PID	R	U_1	U_2	S	R	U_1	U_2	S	R	U_1	U_2	S
Exact	0.31	0	0	0.5	0.31	0	0	0.5	0	0	0	1
CVX	0.31	0	0	0.5	0.31	0	0	0.5	0	0	0	1
BATCH	0.31	0	0	0.5	0.31	0	0	0.5	0	0	0	1

Gaussian Mixture Models (GMM): Consider a GMM, where X_1, $X_2 \in \mathbb{R}$ and the label $Y \in \{-1, +1\}$, comprising two equally weighted standard multivariate Gaussians centered at $\pm\mu$, where $\mu \in \mathbb{R}^2$, i.e., $Y \sim \text{Bernoulli}(1/2)$, $(X_1, X_2)|Y = y \sim \mathcal{N}(y \cdot \mu, I)$. PID was estimated by sampling $1e6$ points, histogramming them into 50 bins spanning $[-5, +5]$ to give p, and then applying the CVX estimator. We term this PID-Cartesian. We also compute PID-Polar, which are PID computed using *polar coordinates*, (r, θ). We use a variant where the angle θ is given by the arctangent with principal values $[0, \pi]$ and the length $r \in \mathbb{R}$ could be negative. θ specifies a line (through the origin), and r tells us where along the line the datapoint lies on.

Results: We consider $\|\mu\|_2 \in \{1.0, 2.0\}$, where for each $\|\mu\|_2$, we vary the angle $\measuredangle\mu$ that μ makes with the horizontal axis. Our computed PID is presented in Figure 3.3. Overall, we find that the PID matches what we expect from intuition. For Cartesian, unique information dominates when the angle goes to 0 or $\pi/2$ — if centroids share a coordinate, then observing that coordinate yields no information about y. Conversely, synergy and redundancy peak at $\pi/4$. Interestingly, synergy seems to be independent of $\|\mu\|_2$. For Polar, redundancy is 0. Furthermore, θ contains no unique information, since θ shows nothing about y unless we know r (in particular, its sign). When the angle goes to $\pi/2$, almost all information is unique in r. The distinctions between Cartesian and Polar highlight how different representations of data can exhibit wildly different PID values, even if total information is the same.

Synthetic generative model: We begin with a set of latent vectors $z_1, z_2, z_c \sim \mathcal{N}(0_d, \Sigma_d^2), d = 50$ representing information unique to X_1, X_2 and common to both respectively. $[z_1, z_c]$ is transformed into high-dimensional x_1 using a fixed transformation T_1 and likewise $[z_2, z_c]$ to x_2 via T_2. The label y is generated as a function of (1) only z_c, in which case we expect complete redundancy, (2) only z_1 or z_2 which represents complete uniqueness, (3) a combination of z_1 and z_2 representing complete synergy, or (4) arbitrary ratios of each of the above with z_i^* representing half of the dimensions from z_i and therefore half of each interaction. In total, Table 3.2 shows the 10 synthetic datasets we generated: 4 specialized datasets $\mathcal{D}_I$, $I \in \{R, U_1, U_2, S\}$ where y only depends on one interaction, and 6 mixed datasets with varying interaction ratios. We also report the ground-truth interactions as defined by the label-generating process and the total capturable information using the bound in Feder and Merhav [172], which relates the accuracy of the best model on these datasets with the mutual information between the inputs to the label. Since the test

Table 3.2: Estimating PID on synthetic generative model datasets. Both CVX and BATCH measures agree with each other on relative values and are consistent with ground truth interactions.

Task	$\mathcal{D}_R$				$\mathcal{D}_{U_1}$				$\mathcal{D}_{U_2}$				$\mathcal{D}_S$				$y = f(z_1^*, z_2^*, z_c^*)$			
PID	R	U_1	U_2	S	R	U_1	U_2	S	R	U_1	U_2	S	R	U_1	U_2	S	R	U_1	U_2	S
CVX	**0.16**	0	0	0.05	0	**0.16**	0	0.05	0	0	**0.17**	0.05	0.07	0	0.01	**0.14**	**0.04**	0.01	0	**0.07**
BATCH	**0.29**	0.02	0.02	0	0	**0.30**	0	0	0	0	**0.30**	0	0.11	0.02	0.02	**0.15**	**0.06**	0.01	0.01	**0.06**
Truth	0.58	0	0	0	0	0.56	0	0	0	0	0.54	0	0	0	0	0.56	0.13	0	0	0.27

Task	$y = f(z_1, z_2^*, z_c^*)$				$y = f(z_1, z_2, z_c^*)$				$y = f(z_1^*, z_2^*, z_c)$				$y = f(z_2^*, z_c^*)$				$y = f(z_2^*, z_c)$			
PID	R	U_1	U_2	S	R	U_1	U_2	S	R	U_1	U_2	S	R	U_1	U_2	S	R	U_1	U_2	S
CVX	0.04	**0.06**	0	**0.07**	**0.07**	0	0	**0.12**	**0.1**	0	0.01	**0.07**	**0.03**	0	**0.04**	0.05	**0.1**	0	0.04	0.05
BATCH	0.04	**0.09**	0	**0.06**	**0.11**	0.02	0.02	**0.10**	**0.11**	0.02	0.02	**0.05**	**0.07**	0	**0.06**	0	**0.19**	0	0.06	0
Truth	0	0.25	0	0.25	0.18	0	0	0.36	0.22	0	0	0.22	0.21	0	0.21	0	0.34	0	0.17	0

accuracies for Table 3.2 datasets range from 67-75%, this corresponds to total MI of $0.42 - 0.59$ bits.

Results: From Table 3.2, both CVX and BATCH agree in relative PID values, correctly assigning the predominant interaction type and interactions with minimal presence consistent with the ground-truth based on data generation. For example, $\mathcal{D}_R$ has the highest R value, and when the ratio of z_1 increases, U_1 increases from 0.01 on $y = f(z_1^*, z_2^*, z_c^*)$ to 0.06 on $y = f(z_1, z_2^*, z_c^*)$. We also note some interesting observations due to the random noise in label generation, such as the non-zero synergy measure of datasets such as $\mathcal{D}_R, \mathcal{D}_{U_1}, \mathcal{D}_{U_2}$ whose labels do not depend on synergy.

3.3.2 Quantifying real-world multimodal benchmarks

We now apply these estimators to quantify the interactions in real-world multimodal datasets.

Real-world multimodal data setup: We use a large collection of real-world datasets in MultiBench [367] which test *multimodal fusion* of different input signals (including images, video, audio, text, time-series, sets, and tables) for different tasks (predicting humor, sentiment, emotions, mortality rate, ICD-9 codes, image-captions, human activities, digits, and design interfaces). We also include experiments on *question-answering* (Visual Question Answering 2.0 [29, 206] and CLEVR [289]) which test grounding of language into the visual domain. For the 4 datasets (top row of Table 3.3) involving images and text where modality features are available and readily clustered, we apply the CVX estimator on top of discrete clusters. For the remaining 4 datasets (bottom row of Table 3.3) with video, audio, and medical time-series modalities, clustering is not easy, so we use the end-to-end BATCH estimator.

Human judgment of interactions: Real-world multimodal datasets do not have reference PID values, and exact PID computation is impossible due to continuous data. We therefore use human judgment as a reference. We design a new annotation scheme where we show both modalities and the label and ask each annotator to annotate the degree of redundancy, uniqueness, and synergy on a scale of 0-5, alongside their confidence in their answers on a scale of 0-5. We give 50 datapoints from each dataset (except MIMIC and ENRICO which require specialized knowledge) to 3 annotators each. We show a sample user interface and annotation procedures in the full paper [372], and also provide an in-depth study of how humans annotate multimodal

Table 3.3: Estimating PID on real-world MultiBench [367] datasets. Many of the estimated interactions align well with human judgement as well as unimodal performance.

Task	AV-MNIST				ENRICO				VQA 2.0				CLEVR			
PID	R	U_1	U_2	S	R	U_1	U_2	S	R	U_1	U_2	S	R	U_1	U_2	S
CVX	0.10	**0.97**	0.03	0.08	**0.73**	0.38	0.53	0.34	0.79	0.87	0	**4.92**	0.55	0.48	0	**5.16**
Human	**0.57**	**0.61**	0	0	-	-	-	-	0	0	0	**6.58**	0	0	0	**6.19**

Task	MOSEI				UR-FUNNY				MUSTARD				MIMIC			
PID	R	U_1	U_2	S	R	U_1	U_2	S	R	U_1	U_2	S	R	U_1	U_2	S
BATCH	**0.26**	**0.49**	0.03	0.04	0.03	0.04	0.01	**0.08**	0.14	0.01	0.01	**0.30**	0.05	**0.17**	0	0.01
Human	**0.32**	**0.20**	0.15	0.15	0.04	**0.05**	0.03	**0.04**	0.13	**0.17**	0.04	**0.16**	-	-	-	-

interactions in a subsequent follow-up work [373].

Results on multimodal fusion: From Table 3.3, we find that different datasets do require different interactions. Some interesting observations: (1) all pairs of modalities on MUSTARD sarcasm detection show high synergy values, which aligns with intuition since sarcasm is often due to a contradiction between what is expressed in language and speech, (2) uniqueness values are strongly correlated with unimodal performance (e.g., modality 1 in AV-MNIST and MIMIC), (3) datasets with high synergy do indeed benefit from interaction modeling as also seen in prior work (e.g., MUSTARD, UR-FUNNY) [83, 225], and (4) conversely datasets with low synergy are those where unimodal performance is relatively strong (e.g., MIMIC) [367].

Results on QA: We observe very high synergy values as shown in Table 3.3 consistent with prior work studying how these datasets were balanced (e.g., VQA 2.0 having different images for the same question such that the answer can only be obtained through synergy) [206] and that models trained on these datasets require non-additive interactions [235]. CLEVR has a higher proportion of synergy than VQA 2.0 (83% versus 75%): indeed, CLEVR is a more balanced dataset where the answer strictly depends on both the question and image with a lower likelihood of unimodal biases.

Comparisons with human judgment: For human judgment, we cannot ask humans to give a score in bits, so it is on a completely different scale (0-5 scale). To put them on the same scale, we normalize the human ratings such that the sum of human interactions is equal to the sum of PID estimates. The resulting comparisons are in Table 3.3, and we find that the human-annotated interactions overall align with estimated PID: the highest values are the same for 4 datasets: both explain highest synergy on VQA and CLEVR, image (U_1) being the dominant modality in AV-MNIST, and language (U_1) being the dominant modality in MOSEI. Overall, the Krippendorff's alpha for inter-annotator agreement is high (0.72 for R, 0.68 for U_1, 0.70 for U_2, 0.72 for S) and the average confidence scores are also high ($4.36/5$ for R, $4.35/5$ for U_1, $4.27/5$ for U_2, $4.46/5$ for S), indicating that the human-annotated results are reliable. For the remaining two datasets (UR-FUNNY and MUSTARD), estimated PID matches the second-highest human-annotated interaction. We believe this is because there is some annotator subjectivity in interpreting whether sentiment, humor, and sarcasm are present in language only (U_1) or when contextualizing both language and video (S), resulting in cases of low annotator agreement in U_1 and S: -0.14, -0.03 for UR-FUNNY and -0.08, -0.04 for MUSTARD.

Comparisons with other interaction measures: Our framework allows for easy general-

Table 3.4: Average interactions ($R/U/S$) learned by models alongside their average performance on interaction-specialized datasets ($\mathcal{D}_R/\mathcal{D}_U/\mathcal{D}_S$). Synergy is the hardest to capture and redundancy is relatively easier to capture by existing models.

Model	EF	ADDITIVE	AGREE	ALIGN	ELEM	TENSOR	MI	MULT	LOWER	REC	AVERAGE
R	0.35	**0.48**	**0.44**	**0.47**	0.27	**0.55**	0.20	0.40	**0.47**	**0.53**	**0.41 ± 0.11**
Acc($\mathcal{D}_R$)	0.71	**0.74**	**0.73**	**0.74**	0.70	**0.75**	0.67	0.73	**0.74**	**0.75**	0.73 ± 0.02
U	0.29	0.31	0.19	0.44	0.20	0.52	0.18	0.45	**0.55**	**0.55**	**0.37 ± 0.14**
Acc($\mathcal{D}_U$)	0.66	0.55	0.60	0.73	0.66	0.73	0.66	0.72	**0.73**	**0.73**	0.68 ± 0.06
S	0.13	0.09	0.08	0.29	0.14	**0.33**	0.12	**0.29**	**0.31**	**0.32**	**0.21 ± 0.10**
Acc($\mathcal{D}_S$)	0.56	0.66	0.63	0.72	0.66	**0.74**	0.65	**0.72**	**0.73**	**0.74**	0.68 ± 0.06

ization to other interaction definitions: we also implemented 3 information theoretic measures **I-min** [662], **WMS** [91], and **CI** [447]. These results are included in the full paper [372], where we explain the limitations of these methods as compared to PID, such as over- and under-estimation, and potential negative estimation [210]. These are critical problems with the application of information theory for shared $I(X_1; X_2; Y)$ and unique information $I(X_1; Y|X_2)$, $I(X_2; Y|X_1)$ often quoted in the co-training [44, 65] and multi-view learning [563, 605, 608] literature. We also tried 3 non-info theory measures: Shapley values [395], Integrated gradients (IG) [579], and CCA [27], which are based on quantifying interactions captured by a multimodal model. Our work is fundamentally different in that interactions are properties of data before training any models.

3.3.3 Quantifying multimodal model predictions

We now shift our focus to quantifying multimodal models. *Do different multimodal models learn different interactions?* A better understanding of the types of interactions that our current models struggle to capture can provide new insights into improving these models.

Setup: For each dataset, we train a suite of models on the train set $\mathcal{D}_{\text{train}}$ and apply it to the validation set $\mathcal{D}_{\text{val}}$, yielding a predicted dataset $\mathcal{D}_{\text{pred}} = \{(x_1, x_2, \hat{y}) \in \mathcal{D}_{\text{val}}\}$. Running PID on $\mathcal{D}_{\text{pred}}$ summarizes the interactions that the model captures. We categorize and implement a comprehensive suite of models (spanning representation fusion at different feature levels, types of interaction inductive biases, and training objectives) that have been previously motivated to capture redundant, unique, and synergistic interactions.

Results: We show results in Table 3.4 and highlight the following observations:

General observations: We first observe that model PID values are consistently higher than dataset PID. The sum of model PID is also a good indicator of test performance, which agrees with their formal definition since their sum is equal to $I(\{X_1, X_2\}; Y)$, the total task-relevant information.

On redundancy: Several methods succeed in capturing redundancy, with an overall average of $R = 0.41 \pm 0.11$ and accuracy of $73.0 \pm 2.0\%$ on redundancy-specialized datasets. Additive, agreement, and alignment-based methods are particularly strong, and we do expect them to capture redundant shared information [147, 497]. Methods based on tensor fusion (synergy-based), including lower-order interactions, and adding reconstruction objectives (unique-based) also capture redundancy.

On uniqueness: Uniqueness is harder to capture than redundancy, with an average of $U = 0.37 \pm 0.14$. Redundancy-based methods like additive and agreement do poorly on uniqueness, while those designed for uniqueness (lower-order interactions [712] and modality reconstruction objectives [614]) do well, with on average $U = 0.55$ and 73.0% accuracy on uniqueness datasets.

On synergy: Synergy is the hardest to capture, with an average score of only $S = 0.21 \pm 0.10$. Some of the strong methods are tensor fusion [182], tensors with lower-order interactions [712], modality reconstruction [614], and multimodal transformer [681], which achieve around $S = 0.30$, acc $= 73.0\%$. Additive, agreement, and element-wise interactions do not seem to capture synergy well.

On robustness: Finally, we also show connections between PID and model performance in the presence of missing modalities. We find high correlation ($\rho = 0.8$) between the performance drop when X_i is missing and the model's U_i value. Inspecting Figure 3.4, we find that the implication only holds in one direction: high U_i coincides with large performance drops (in red), but low U_i can also lead to performance drops (in green). The latter can be further explained by the presence of large S values: when X_i is missing, synergy can no longer be learned which affects performance. For the subset of points when $U_i \leq 0.05$, the correlation between S and performance drop is $\rho = 0.73$ (in contrast, the correlation for R is $\rho = 0.01$).

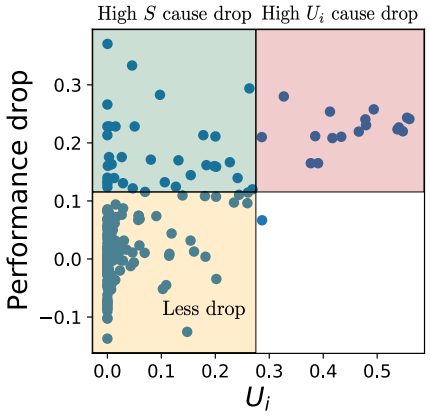


Figure 3.4: We find high correlation ($\rho = 0.8$) between the performance drop when X_i is missing and the model's U_i value: high U_i coincides with large performance drops (red), but low U_i can also lead to performance drops. The latter can be further explained by large S so X_i is necessary (green).

3.3.4 PID agreement and model selection

Now that we have quantified datasets and models individually, the natural next question unifies both: *what does the agreement between dataset and model* PID *measures tell us about model performance?* We hypothesize that models able to capture the interactions necessary in a given dataset should also achieve high performance. Given estimated interactions on dataset $\mathcal{D}$ and model $f(\mathcal{D})$ trained on $\mathcal{D}$, we define the agreement for each interaction $I \in \{R, U_1, U_2, S\}$ as:

$$\alpha_I(f, \mathcal{D}) = \hat{I}_{\mathcal{D}} I_{f(\mathcal{D})}, \quad \hat{I}_{\mathcal{D}} = \frac{I_{\mathcal{D}}}{\sum_{I' \in \{R,U_1,U_2,S\}} I'_{\mathcal{D}}}, \tag{3.6}$$

which summarizes the quantity of an interaction captured by a model ($I_{f(\mathcal{D})}$) weighted by its normalized importance in the dataset ($\hat{I}_{\mathcal{D}}$). The total agreement sums over $\alpha(f, \mathcal{D}) = \sum_I \alpha_I(f, \mathcal{D})$.

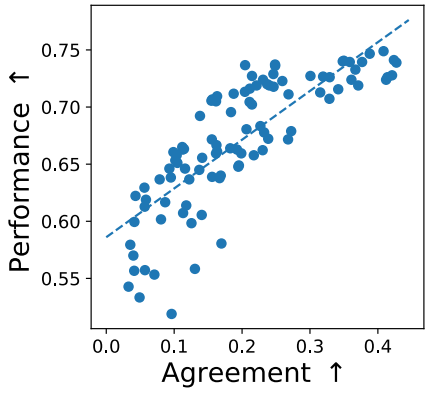


Figure 3.5: PID agreement $\alpha(f, \mathcal{D})$ between datasets and models strongly correlate with model accuracy ($\rho = 0.81$).

Table 3.5: Model selection results on unseen synthetic and real-world datasets. Given a new dataset $\mathcal{D}$, finding the closest synthetic dataset $\mathcal{D}'$ with similar PID values and recommending the best models on $\mathcal{D}'$ consistently achieves $95\% - 100\%$ of the best-performing model on $\mathcal{D}$.

Dataset	5 Synthetic Datasets	MIMIC	ENRICO	UR-FUNNY	MOSEI	MUSTARD	MAPS
% Performance	99.91%	99.78%	100%	98.58%	99.35%	95.15%	100%

Results: Our key finding is that PID agreement scores $\alpha(f, \mathcal{D})$ correlate (ρ = 0.81) with model accuracy across all 10 synthetic datasets as illustrated in Figure 3.5. This shows that PID agreement can be a useful proxy for model performance. For the specialized datasets, we find that the correlation between α_I and $\mathcal{D}_I$ is 0.96 for R, 0.86 for U, and 0.91 for S, and negatively correlated with other specialized datasets. For mixed datasets with roughly equal ratios of each interaction, the measures that correlate most with performance are α_R (ρ = 0.82) and α_S (ρ = 0.89); datasets with relatively higher redundancy see ρ = 0.89 for α_R; those with higher uniqueness have α_{U_1} and α_{U_2} correlate ρ = 0.92 and ρ = 0.85; those with higher synergy increases the correlation of α_S to ρ = 0.97.

Using these observations, our final experiment is model selection: *can we choose the most appropriate model to tackle the interactions required for a dataset?*

Setup: Given a new dataset $\mathcal{D}$, we first compute its difference in normalized PID values with respect to $\mathcal{D}'$ among our suite of 10 synthetic datasets, $s(\mathcal{D}, \mathcal{D}') = \sum_{I \in \{R, U_1, U_2, S\}} |\hat{I}_{\mathcal{D}} - \hat{I}_{\mathcal{D}'}|$, to rank the dataset $\mathcal{D}^*$ with the most similar interactions, and return the top-3 performing models on $\mathcal{D}^*$. In other words, we select models that best capture interactions that are of similar nature and degree as those in $\mathcal{D}$. We emphasize that even though we restrict dataset and model search to *synthetic datasets*, we evaluate model selection on real-world datasets and find that it *generalizes to the real world*.

Results: We test our selected models on 5 new synthetic datasets with different PID ratios and 6 real-world datasets, summarizing results in Table 3.5. We find that the top 3 chosen models achieve $95\% - 100\%$ of the best-performing model accuracy, and > 98.5% for all datasets except 95.2% on MUSTARD. For example, UR-FUNNY and MUSTARD have the highest synergy (S = 0.13, S = 0.3) and indeed transformers and higher-order interactions are helpful (MULT: 65%, MI: 61%, TENSOR: 60%). ENRICO has the highest R = 0.73 and U_2 = 0.53, and methods for redundant and unique interactions perform best (LOWER: 52%, ALIGN: 52%, AGREE: 51%). MIMIC has the highest U_1 = 0.17, and unimodal models are mostly sufficient [367].

3.3.5 Real-world applications

Finally, we apply PID to 3 real-world case studies: pathology, mental health, and robotic perception.

Case Study 1: Computational pathology. Cancer prognostication is a challenging task in anatomic pathology that requires integration of whole-slide imaging (WSI) and molecular features for patient stratification [92, 383, 425]. We use The Cancer Genome Atlas (TCGA), a large public data consortium of paired WSI, molecular, and survival information [607, 660], including modalities: (1) pre-extracted histology image features from diagnostic WSIs and (2) bulk gene mutation status, copy number variation, and RNA-Seq abundance values. We evaluate on two cancer datasets in TCGA, lower-grade glioma (LGG [440], n = 479) and pancreatic

adenocarcinoma (PAAD [503], n = 209).

Results: In TCGA-LGG, most PID measures were near-zero except U_2 = 0.06 for genomic features, which indicates that genomics is the only modality containing task-relevant information. This conclusion corroborates with the high performance of unimodal-genomic and multimodal models in Chen et al. [102], while unimodal-pathology performance was low. In TCGA-PAAD, the uniqueness in pathology and genomic features was less than synergy (U_1 = 0.06, and U_2 = 0.08 and S = 0.15), which also match the improvement of using multimodal models that capture synergy.

Case Study 2: Mental health. Suicide is the second leading cause of death among adolescents [84]. Intensive monitoring of behaviors via adolescents' frequent use of smartphones may shed new light on the early risk of suicidal ideations [200, 433], since smartphones provide rich behavioral markers [366]. We used a dataset, MAPS, of mobile behaviors from high-risk consenting adolescent populations (approved by IRB). Passive sensing data is collected from each participant's smartphone across 6 months. The modalities include (1) *text* entered by the user represented as a bag of top 1000 words, (2) *keystrokes* that record the exact timing and duration of each typed character, and (3) *mobile applications* used per day as a bag of 137 apps. Every morning, users self-report their daily mood, which we discretized into $-1, 0, +1$. In total, MAPS has 844 samples from 17 participants.

Results: We first experiment with MAPS$_{T,K}$ using text and keystroke features. PID measures show that MAPS$_{T,K}$ has high synergy (0.40), some redundancy (0.12), and low uniqueness (0.04). We found the purely synergistic dataset $\mathcal{D}_S$ has the most similar interactions and the suggested models LOWER, REC, and TENSOR that work best on $\mathcal{D}_S$ were indeed the top 3 best-performing models on MAPS$_{T,K}$, indicating that model selection is effective. Model selection also retrieves the best-performing model on MAPS$_{T,A}$ using text and app usage features.

Case Study 3: Robotic Perception. MuJoCo PUSH [334] is a contact-rich planar pushing task in MuJoCo [606], where a 7-DoF Panda Franka robot is pushing a circular puck with its end-effector in simulation. The dataset consists of 1000 trajectories with 250 steps sampled at 10Hertz. The multimodal inputs are gray-scaled images from an RGB camera, force and binary contact information from a force/torque sensor, and the 3D position of the robot end-effector. We estimate the 2D position of the unknown object on a table surface while the robot intermittently interacts with it.

Results: We find that BATCH predicts U_1 = 1.79 as the highest PID value, which aligns with our observation that image is the best unimodal predictor. Comparing both estimators, CVX underestimates U_1 and R since the high-dimensional time-series modality cannot be easily described by clusters without losing information. In addition, both estimators predict a low U_2 value but attribute high R, implying that a multimodal model with higher-order interactions would not be much better than unimodal models. Indeed, we observe no difference in performance between these two.

3.4 Conclusion

Our work aims to quantify the nature and degree of feature interactions by proposing scalable estimators for redundancy, uniqueness, and synergy suitable for high-dimensional heterogeneous

datasets. Through comprehensive experiments and real-world applications, we demonstrate the utility of our proposed framework in dataset quantification, model quantification, and model selection. We are aware of some potential **limitations**:

1. These estimators only approximate real interactions due to cluster preprocessing or unimodal models, which naturally introduce optimization and generalization errors. We expect progress in density estimators, generative models, and unimodal classifiers to address these problems.
2. It is harder to quantify interactions for certain datasets, such as ENRICO which displays all interactions which makes it difficult to distinguish between R and S or U and S.
3. Finally, there exist challenges in quantifying interactions since the data generation process is never known for real-world datasets, so we have to resort to human judgment, other automatic measures, and downstream tasks such as estimating model performance and model selection.

Future work can leverage PID for targeted dataset creation, representation learning optimized for PID values, and applications of information theory to higher-dimensional data. More broadly, there are several exciting directions in investigating more applications of multivariate information theory in modeling feature interactions, predicting multimodal performance, and other tasks involving feature interactions such as privacy-preserving and fair representation learning from high-dimensional data [161, 219]. Being able to provide guarantees for fairness and privacy-preserving learning can be particularly impactful.

Chapter 4

Factorized Learning of Multimodal Interactions

4.1 Introduction

Using the mathematical foundation of multimodal interactions that we just presented, we now seek to learn representations from multimodal data that are suitable in capturing each of these interactions. Learning representations from different modalities is a central paradigm in machine learning [371]. Today, a popular learning method is to first pre-train general representations on unlabeled multimodal data before fine-tuning on task-specific labels [72, 370, 371, 390]. These current multimodal pre-training approaches have largely been inherited from prior work in multi-view learning [103, 453] that exploit a critical assumption of *multi-view redundancy*: the property that shared information between modalities is almost exactly what is relevant for downstream tasks [563, 608, 616]. When this assumption holds, approaches based on contrastive pre-training to capture shared information [103, 300, 497, 605], followed by fine-tuning to keep task-relevant shared information [616], have seen successful applications in learning from images and captions [497], video and audio [31], speech and transcribed text [453], and instructions and actions [168]. However, our paper studies two fundamental limitations in the application of contrastive learning (CL) to learn multimodal interactions in real-world settings

1. **Low *shared* information** relevant to tasks: There exists a wide range of multimodal tasks involving small amounts of shared information, such as between cartoon images and figurative captions (i.e., not literal but metaphoric or idiomatic descriptions of the images [410, 700]). In these situations, standard multimodal CL will only receive a small percentage of information from the learned representations and struggle to learn the desired task-relevant information.
2. **High *unique* information** relevant to tasks: Many real-world modalities can provide unique information not present in other modalities. Examples include healthcare with medical sensors or robotics with force sensors [367, 372]. Standard CL will discard task-relevant unique information, leading to poor downstream performance.

We refer the reader to Figure 9.1 for a visual depiction and experimental results showing the performance drop of CL in these two settings of low shared information and high unique information.

In light of these limitations, how can we design suitable multimodal learning objectives that

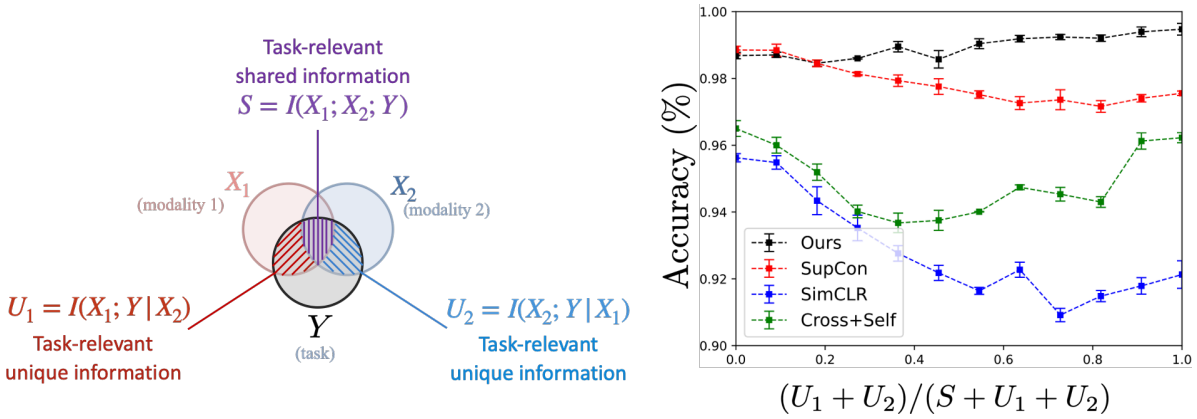


Figure 4.1: Left: We define $S = I(X_1; X_2; Y)$ as task-relevant shared information and $U_1 = I(X_1; Y|X_2)$, $U_2 = I(X_2; Y|X_1)$ as task-relevant unique information. **Right**: On controllable datasets with varying ratios of S, U_1, and U_2, standard CL captures S but struggles when there is more U_1 and U_2. Our FACTORCL approach maintains best performance, whereas SimCLR [103] and SupCon [300] see performance drops as unique information increases, and Cross+Self [258, 278, 337, 709] recovers in fully unique settings but suffers at other ratios.

work beyond multi-view redundancy? In this paper, starting from the first principles in information theory, we provide formal definitions of shared and unique information via conditional mutual information and propose an approach, FACTORIZED CONTRASTIVE LEARNING (FACTORCL for short), to learn these multimodal representations beyond multi-view redundancy using three key ideas. The first idea is to explicitly *factorize* shared and unique representations. The second idea is to *capture task-relevant* information via maximizing lower bounds on MI and *remove task-irrelevant* information via minimizing upper bounds on MI, resulting in representations with sufficient and necessary information content. Finally, a notion of task relevance without explicit labels in the self-supervised setting is achieved by leveraging *multimodal augmentations*. Experimentally, we evaluate the effectiveness of FACTORCL on a suite of synthetic datasets and large-scale real-world multimodal benchmarks involving images and figurative language [700], prediction of human sentiment [710], emotions [717], humor [225], and sarcasm [83], as well as patient disease and mortality prediction from health indicators and sensor readings [286], achieving new state-of-the-art performance on six datasets. Overall, we summarize our key technical contributions here:

1. A new analysis of contrastive learning performance showing that standard multimodal CL fails to capture task-relevant unique information under low shared or high unique information cases.
2. A new contrastive learning algorithm called FACTORCL:
   (a) FACTORCL factorizes task-relevant information into shared and unique information, expanding contrastive learning to better handle low shared or high unique information.
   (b) FACTORCL optimizes shared and unique information separately, by removing task-irrelevant information via MI upper bounds and capturing task-relevant information via lower bounds, yielding optimal task-relevant representations.
   (c) FACTORCL leverages multimodal augmentations to approximate task-relevant information, enabling self-supervised learning from our proposed FACTORCL.

4.2 Analysis of Multi-view Contrastive Learning

We begin by formalizing definitions of four types of information: shared, unique, task-relevant, and task-irrelevant information in multimodal data. To formalize the learning setting, we assume there exist two modalities expressed as random variables X_1 and X_2 with outcomes x_1 and x_2, and a task with the random variable Y and outcome y. We denote X_{-i} as the other modality where appropriate.

Shared and unique information: We formalize shared and unique information by decomposing the total multimodal information $I(X_1, X_2; Y)$ into three conditional mutual information (MI) terms:

$$I(X_1, X_2; Y) = \underbrace{I(X_1; X_2; Y)}_{S \text{ = shared}} + \underbrace{I(X_1; Y|X_2)}_{U_1 \text{ = uniqueness in } X_1} + \underbrace{I(X_2; Y|X_1)}_{U_2 \text{ = uniqueness in } X_2}, \tag{4.1}$$

where $I(X_1, X_2; Y) = \int p(x_1, x_2, y) \log \frac{p(x_1,x_2,y)}{p(x_1,x_2)p(y)} dx_1 dx_2 dy$ is the total MI between the joint random variable X_1, X_2 and the task Y, $S = I(X_1; X_2; Y) = I(X_1; X_2) - I(X_1; X_2|Y) = \int p(x_1, x_2) \log \frac{p(x_1,x_2)}{p(x_1)p(x_2)} dx_1 dx_2 - I(X_1; X_2|Y)$ is the task-relevant shared information, $I(X_1; X_2|Y) = \int p(x_1, x_2, y) \log \frac{p(x_1,x_2|y)}{p(x_1|y)p(x_2|y)} dx_1 dx_2 dy$ is the task-irrelevant shared information, and $U_1 = I(X_1; Y|X_2)$, $U_2 = I(X_2; Y|X_1)$ denote unique task-relevant information.

Limitations of CL: Current approaches for CL maximize mutual information $I(X_1; X_2)$ (and subsequently task-relevant shared information $I(X_1; X_2; Y)$ during supervised fine-tuning), without modeling unique information. These methods generally learn a pair of representations [608, 616],

$$Z_1 = \underset{Z_1 := f_\theta(X_1)}{\arg\max} I(Z_1; X_2), \quad Z_2 = \underset{Z_2 := f_\theta(X_2)}{\arg\max} I(X_1; Z_2). \tag{4.2}$$

For example, Z_1 could encode images X_1 and Z_2 encodes text X_2 via maximizing a lower bound on $I(X_1; X_2)$ using the NCE objective [453]. The NCE objective falls into a broader class of contrastive learning methods [103, 107, 229, 300, 497] that model the ratio between joint densities $p(x_1, x_2)$ and product of marginal densities $p(x_1)p(x_2)$ using positive and negative samples [444, 458, 487, 621, 669] or probabilistic classifiers [430, 617]. It has been shown that contrastive learning works well under the assumption of multi-view redundancy [39, 240, 563, 604]:

Definition 1. *(Multi-view redundancy)* $\exists \epsilon > 0$ *such that* $I(X_1; Y|X_2) \leq \epsilon$ *and* $I(X_2; Y|X_1) \leq \epsilon$.

In other words, the task-relevant information in data is mostly shared across both views and the unique information is at most a small ϵ. From a representation perspective, Tian et al. [605] further introduces the assumption that the optimal representation is minimal and sufficient, where all learned task-relevant information is shared information: $I(Z_1; Y|X_2) = I(Z_2; Y|X_1) = 0$. While the multi-view redundancy is certainly true for particular types of multimodal distributions, it crucially ignores settings that display *multi-view non-redundancy* and unique information can be important, such as when health indicators, medical sensors, and robotic visual or force sensors each provide unique information not present in other modalities [367, 372].

Definition 2. *(Multi-view non-redundancy)* $\exists \epsilon > 0$ *such that* $I(X_1; Y|X_2) > \epsilon$ *or* $I(X_2; Y|X_1) > \epsilon$.

Under multi-view non-redundancy, we show that standard CL only receives a weak training signal since it can only maximize a lower bound on shared information $I(X_1; X_2)$, and struggles to learn task-relevant unique information. We formalize this intuition with the following statement:

Theorem 1. *(Suboptimality of standard CL) When there is multi-view non-redundancy as in Definition 2, given optimal representations $\{Z_1, Z_2\}$ that satisfy Eq.(4.2 and $I(Z_1; Y|X_2) = I(Z_2; Y|X_1) = 0$ [605], we have that*

$$I(Z_1, Z_2; Y) = I(X_1, X_2; Y) - I(X_1; Y|X_2) - I(X_2; Y|X_1) = I(X_1; X_2) - I(X_1; X_2|Y) < I(X_1, X_2; Y). \tag{4.3}$$

Correspondingly, the Bayes error rate $P_e(Z_1, Z_2) := 1 - \mathbb{E}_{p(z_1,z_2)}\left[\max_{y \in Y} P\left(\hat{Y} = y \mid z_1, z_2\right)\right]$ of contrastive representations $\{Z_1, Z_2\}$ for a downstream task Y is given by:

$$P_e \leq 1 - \exp\left[I(X_1, X_2; Y) - I(X_1; Y|X_2) - I(X_2; Y|X_1) - H(Y)\right] \tag{4.4}$$

$$= 1 - \exp\left[I(X_1; X_2; Y) - H(Y)\right] \tag{4.5}$$

We include proofs and a detailed discussion of the assumptions in the full paper [374]. Based on Eq.(4.3), $I(Z_1, Z_2; Y)$ decreases with higher task-relevant unique information $I(X_1; Y|X_2)$ and $I(X_2; Y|X_1)$; we call this the difference $I(X_1, X_2; Y) - I(Z_1, Z_2; Y)$ the *uniqueness gap*. The uniqueness gap measures the loss in task-relevant information between the input and encoded representation: as task-relevant unique information grows, the uniqueness gap increases. In addition, $I(Z_1, Z_2; Y)$ also drops with lower $I(X_1; X_2)$ (i.e., two modalities sharing little information to begin with), or with higher $I(X_1; X_2|Y)$ (i.e., when the shared information is mostly task-irrelevant). Similarly, in Eq.(4.5), the Bayes error rate of using $\{Z_1, Z_2\}$ for prediction is directly related to the task-relevant information in $\{Z_1, Z_2\}$: error on the downstream task increases with higher unique information and lower shared information.

4.3 Factorized Contrastive Learning

We now present a suite of new CL objectives that alleviate the challenges above and work at all ranges of shared and unique information. At a high level, we aim to learn a set of factorized representations $Z_{S_1}, Z_{S_2}, Z_{U_1}, Z_{U_2}$ representing task-relevant information in X_1 shared with X_2, in X_2 shared with X_1, unique to X_1, and unique to X_2 respectively. As common in practice [497, 605], we define neural networks f_θ with trainable parameters θ to extract representations from inputs X_1 and X_2. Learning these parameters requires optimizing differentiable and scalable training objectives to capture task-relevant shared and unique information (see overview in Figure 4.2):

$$Z_{S_1} = \underset{Z_1 = f_\theta(X_1)}{\arg\max} I(Z_1; X_2; Y), \qquad Z_{S_2} = \underset{Z_2 = f_\theta(X_2)}{\arg\max} I(Z_2; X_1; Y), \tag{4.6}$$

$$Z_{U_1} = \underset{Z_1 = f_\theta(X_1)}{\arg\max} I(Z_1; Y|X_2), \qquad Z_{U_2} = \underset{Z_2 = f_\theta(X_2)}{\arg\max} I(Z_2; Y|X_1). \tag{4.7}$$

where $I(Z_1; X_2; Y) = I(Z_1; X_2) - I(Z_1; X_2|Y)$ is the shared information and $I(Z_2; X_1; Y) = I(Z_2; X_2) - I(Z_2; X_1|Y)$ is the unique information. One important characteristic of our framework

is that when unique information is zero: $I(X_1; Y|X_2) = 0$ and $I(X_2; Y|X_1) = 0$, or all shared information is task-relevant: $I(X_1; X_2; Y) = I(X_1; X_2)$, our framework recovers standard CL as in Eq.(4.2). However, as we have previously indicated and will show empirically, these assumptions can easily be violated, and our framework enlarges Eq.(4.2) to cases where unique information is present.

The learned Zs can then be used as input to a linear classifier and fine-tuned to predict the label for multimodal classification or retrieval tasks. However, the shared and unique MI terms above are often intractable in practice. In the next section, we will build up our method step by step, eventually showing that each term in Eqs.(4.6- 4.7) can be approximated as follows:

$$S = I(X_1; X_2; Y) \geq I_{\text{NCE}}(X_1; X_2) - I_{\text{NCE-CLUB}}(X_1; X_2|X_1', X_2') \tag{4.8}$$

$$U_i = I(X_i; Y|X_{-i}) \geq I_{\text{NCE}}(X_i; X_i') - I_{\text{NCE-CLUB}}(X_1; X_2) + I_{\text{NCE}}(X_1; X_2|X_1', X_2') \tag{4.9}$$

where I_{NCE} and $I_{\text{NCE-CLUB}}$ are scalable contrastive estimators (Section 4.3.1) and X_1', X_2' are suitable data augmentations (Section 4.3.2) on each modality. Overall, these equations can be interpreted as both positive and negative signals to learn representations for S and U. For shared information S, the estimator maximizes task-relevant shared information via $I_{\text{NCE}}(X_1; X_2)$ while removing task-irrelevant shared information via a novel upper bound $-I_{\text{NCE-CLUB}}(X_1; X_2|X_1', X_2')$. For unique information U_i, we capture task-relevant uniqueness via $+I_{\text{NCE}}(X_i; X_i')$ while non-unique information is removed via $-(I_{\text{NCE-CLUB}}(X_1; X_2) - I_{\text{NCE}}(X_1; X_2|X_1', X_2'))$. In the following sections, we derive this final objective step-by-step: (1) approximating the MI objectives in S and U with CL estimators, (2) relaxing the dependence on labels Y with self-supervised data augmentations, finally (3) discussing overall training and implementation details of end-to-end self-supervised learning.

4.3.1 Supervised FACTORCL with shared and unique information

To capture shared and unique information via an objective function, we will need to maximize lower bounds for all terms with a positive sign in Eq.(4.8) and (4.9) $(I(X_1; X_2), I(X_i; Y), I(X_1; X_2|Y))$ and minimize upper bounds for all terms with a negative sign $(I(X_1; X_2), I(X_1; X_2|Y))$. Our first theorem derives general lower and upper bounds for MI terms as variants of contrastive estimation:

Theorem 2. *(Contrastive estimators for $I(X_1; X_2)$) Defining the NCE and NCE-CLUB estimators,*

$$I_{\text{NCE}}(X_1; X_2) = \mathbb{E}_{\substack{x_1, x_2^+ \sim p(x_1, x_2) \\ x_2^- \sim p(x_2)}} \left[\log \frac{\exp f(x_1, x_2^+)}{\sum_k \exp f(x_1, x_2^-)} \right] \tag{4.10}$$

$$I_{\text{NCE-CLUB}}(X_1; X_2) = \mathbb{E}_{x_1, x_2^+ \sim p(x_1, x_2)} \left[f^*(x_1, x_2^+) \right] - \mathbb{E}_{\substack{x_1 \sim p(x_1) \\ x_2^- \sim p(x_2)}} \left[f^*(x_1, x_2^-) \right] \tag{4.11}$$

where $f^*(x_1, x_2)$ is the optimal critic from I_{NCE} plugged into the I_{CLUB} objective [110]. We call the proposed plug-in objective Eq.(4.11) $I_{\text{NCE-CLUB}}$, and obtain lower and upper bounds on $I(X_1; X_2)$:

$$I_{\text{NCE}}(X_1; X_2) \leq I(X_1; X_2) \leq I_{\text{NCE-CLUB}}(X_1; X_2). \tag{4.12}$$

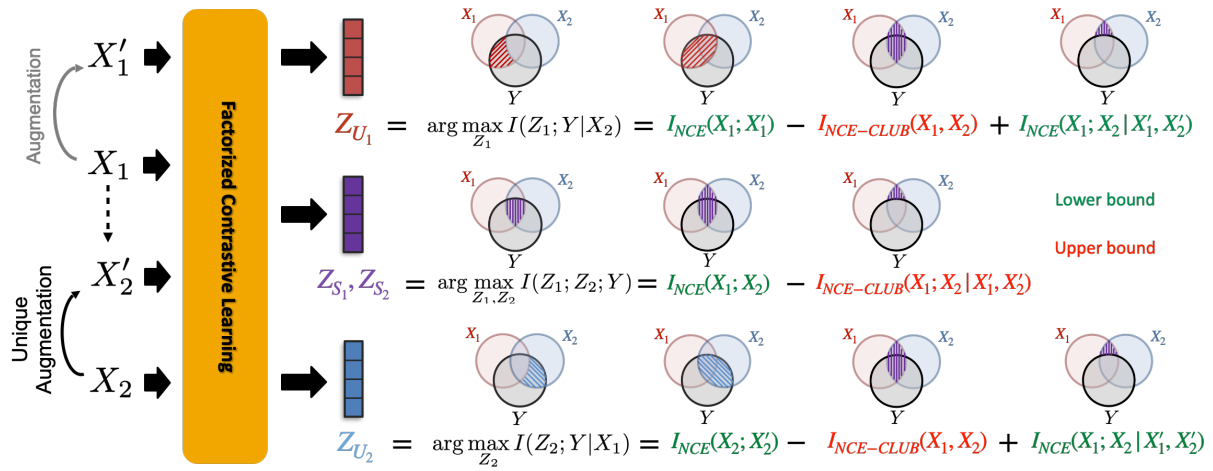


Figure 4.2: FACTORCL: We propose a self-supervised CL method to learn *factorized* representations Z_{S_1}, Z_{S_2}, Z_{U_1}, and Z_{U_2} to capture task-relevant information shared in both X_1 and X_2, unique to X_1, and unique to X_2. By starting with information-theoretic first principles of shared and unique information, we design contrastive estimators to both *capture task-relevant* and *remove task-irrelevant* information, where a notion of task-relevance without explicit labels is afforded by a new definition of *multimodal augmentations* X_1', X_2'. Lower bounds are in green and upper bounds are in red.

Proof. The lower bound $I_{\text{NCE}}(X_1;X_2) \leq I(X_1;X_2)$ follows from Oord et al. [453]: optimizing the objective leads to an optimal critic [487] $f^* = \log p(x_1|x_2) + c(x_1)$, with a deterministic function $c(\cdot)$. Plugging optimal critic f^* into $I_{\text{NCE-CLUB}}(X_1;X_2)$ cancels out the $c(x_1)$ term and yields $I_{\text{NCE-CLUB}}(X_1;X_2)$ and $I(X_1;X_2) \leq I_{\text{NCE-CLUB}}$. We include a detailed proof in the full paper [374]. □

$I_{\text{NCE-CLUB}}(X_1;X_2)$ gives a desired upper bound of $I(X_1;X_2)$ "for free" while avoiding separately optimizing lower bound and upper bounds. In Figure 4.3, we show these two bounds in practice across two Gaussian distributions X_1 and X_2 with varying amounts of MI $I(X_1;X_2)$. We use the second formulation of I_{CLUB} [110], which assumes $p(x_1|x_2)$ to be unknown. Our upper bound is empirically tighter (see Figure 4.3) and comes for "free" via jointly maximizing the lower bound I_{NCE}. These lower and upper bounds can be seen as new contrastive objectives over positive and negative (x_1, x_2) pairs, enabling a close integration with existing pre-training paradigms. Finally, we can similarly obtain bounds for the conditional MI $I_{\text{NCE}}(X_1;X_2|Y) \leq I(X_1;X_2|Y) \leq I_{\text{NCE-CLUB}}(X_1;X_2|Y)$:

$$I_{\text{NCE}}(X_1;X_2|Y) = \mathbb{E}_{p(y)}\left[\mathbb{E}_{\substack{x_1,x_2^+\sim p(x_1,x_2|y)\\ x_2^-\sim p(x_2|y)}}\left[\log \frac{\exp f(x_1,x_2^+,y)}{\sum_k \exp f(x_1,x_2^-,y)}\right]\right] \tag{4.13}$$

$$I_{\text{NCE-CLUB}}(X_1;X_2|Y) = \mathbb{E}_{p(y)}\left[\mathbb{E}_{x_1,x_2^+\sim p(x_1,x_2|y)}\left[f^*(x_1,x_2^+,y)\right] - \mathbb{E}_{\substack{x_1\sim p(x_1|y)\\ x_2^-\sim p(x_2|y)}}\left[f^*(x_1,x_2^-,y)\right]\right] \tag{4.14}$$

These two bounds result in *conditional CL* objectives [397, 612, 618] - they differ critically from standard CL methods since they capture task-irrelevant shared information that remains

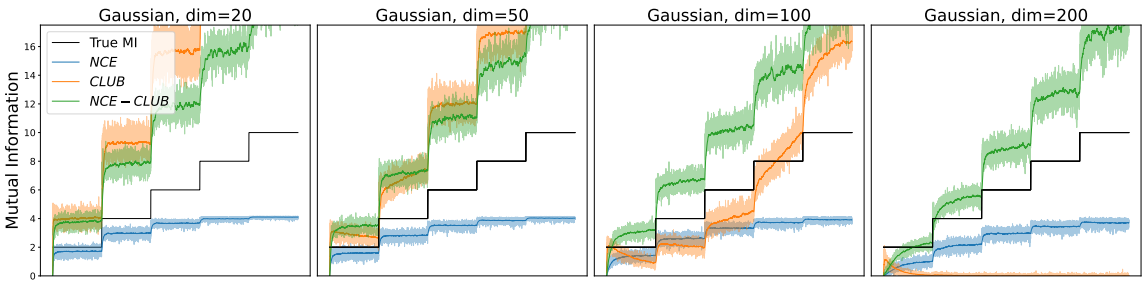


Figure 4.3: Estimated I_{NCE} lower bound [453] and our proposed upper bound $I_{\text{NCE-CLUB}}$ on sample distributions with changing mutual information: our upper bound is tighter, more accurate, and more stable than I_{CLUB} upper bound [110], and also comes for 'free' via jointly estimating both lower and upper bounds simultaneously. We find that as dimension increases, the I_{CLUB} estimator collapses to zero and no longer tracks true MI.

between X_1 and X_2 after observing Y. This task-irrelevant shared information is removed by minimizing its upper bound. Note that $f(x_1, x_2, y)$ here denotes a different function from $f(x_1, x_2)$ in Eq.(4.10), as the general forms are different (taking in x_1, x_2 versus x_1, x_2, y). $f(x_1, x_2, y)$ can be implemented in different ways, e.g., $g([x_1, y])^T h(x_2)$ where $g(), h()$ are trainable encoders and $[x_1, y]$ denotes concatenation [561].

4.3.2 Self-supervised FACTORCL via multimodal augmentations

The derivations above bring about supervised CL objectives with access to Y [300]. For unsupervised CL [453, 605], we derive similar objectives without access to Y by leveraging semantic augmentations on each modality. Denote X' as some augmentation of X (e.g., rotating, shifting, or cropping). Under the *optimal augmentation* assumption from Tian et al. [605] (restated below), replacing Y with X' in our formulations enables learning of task-relevant information without access to labels:

Definition 3. *(Optimal unimodal augmentation) [605] X_1' is an optimal unimodal augmentation for X_1 when $I(X; X') = I(X; Y)$, which implies that the only information shared between X and X' is task-relevant with no irrelevant noise.*

This assumption is satisfied when all information shared between X and X' is task-relevant, which implies that the augmentation keeps task-relevant information constant while changing task-irrelevant information. In the case of image classification, task-relevant information is the object in the picture, while task-irrelevant information is the background. By performing two separate unimodal augmentations giving X_1' and X_2', we can substitute contrastive estimators in Eqs.(4.13) and (4.14), by replacing $I(X_i; Y)$ terms with $I(X_i; X_i')$ and replacing $I(X_1; X_2|Y)$ terms with $I(X_1; X_2|X_1', X_2')$:

$$I_{\text{NCE}}(X_1; X_2|X_1', X_2') = \mathbb{E}_{p(x_1', x_2')}\left[\mathbb{E}_{\substack{x_1, x_2^+ \sim p(x_1, x_2|x_1', x_2') \\ x_2^- \sim p(x_2|x_1', x_2')}}\left[\log \frac{\exp f(x_1, x_2^+, x_1', x_2')}{\sum_k \exp f(x_1, x_2^-, x_1', x_2')}\right]\right] \tag{4.15}$$

$$I_{\text{NCE-CLUB}}(X_1; X_2|X_1', X_2') = \mathbb{E}_{p(x_1', x_2')}\Big[\mathbb{E}_{x_1, x_2^+ \sim p(x_1, x_2|x_1', x_2')}[f^*(x_1, x_2^+, x_1', x_2')] - \mathbb{E}_{\substack{x_1 \sim p(x_1|x_1', x_2') \\ x_2^- \sim p(x_2|x_1', x_2')}}[f^*(x_1, x_2^-, x_1', x_2')]\Big] \tag{4.16}$$

The objectives can be seen as conditional contrastive learning on augmentations (X_1', X_2'). Here again $f(x_1, x_2, x_1', x_2')$ is different from the critics in Eqs.(4.13 because of the different general forms. We implement $f()$ here as $g([x_1, x_1'])^T h([x_2, x_2'])$ where $g(), h()$ are trainable encoders specific for each modality and $[x_1, x_1']$ denotes concatenation. This concatenation is justified by the CMI estimators in Sordoni et al. [561], who show that concatenating the conditioning variable with the input in the critic $f(x_1, x_2, x_1', x_2')$ yields a Conditional InfoNCE estimator (Eq.(4.15)) that is a lower bound for CMI. However, the exact Conditional InfoNCE estimator learns a different conditional distribution $p(x_1, x_2|x_1', x_2')$ for each augmented pair x_1', x_2', which can be prohibitively expensive. We could approximate this by creating multiple augmentations of a single paired x_1, x_2. Our code uses one augmented pair x_1', x_2' for each x_1, x_2 but could be extended to multiple pairs, and we find this simple approach yields consistent CMI lower and upper bounds that are empirically comparable to existing CMI estimators [430, 561]. We include full comparisons and implementation details in the full paper [374].

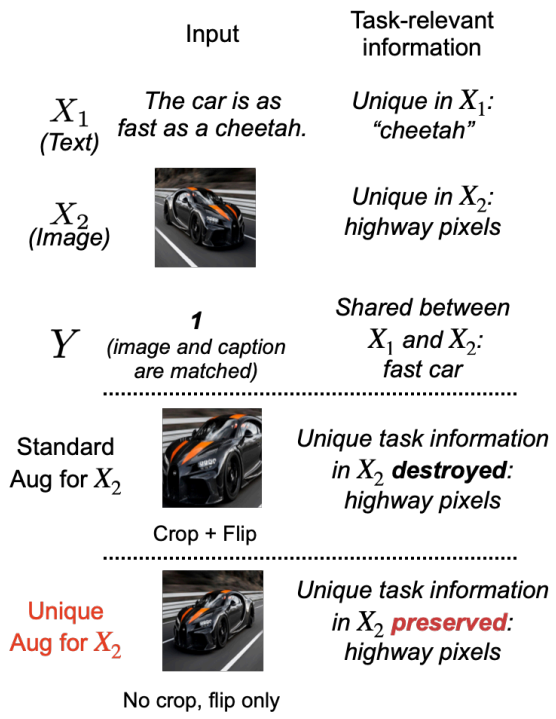


Figure 4.4: Standard vs. unique augmentations for the figurative language [700] dataset. After augmenting text modality X_1 independently (same for both augmentation types), we illustrate their differences for image augmentation: unique augmentation on images should avoid removing information referred to by X_1 (the text). The text mentions that the car is fast so unique augmentation for images should *not* remove the highway pixels of the image which can suggest the car is fast.

Although we find this method to work well in practice, a more careful analysis reveals that 2 separate unimodal augmentations X_1' and X_2' each satisfying $I(X_i; X_i') = I(X_i; Y)$ do not together satisfy $I(X_1; X_2|Y) = I(X_1; X_2|X_1', X_2')$ needed for the substitution in Eqs.(4.15) and (4.16) to hold with equality. To satisfy this property exactly, we define optimal multimodal augmentations:

Definition 4. *(Optimal multimodal augmentation) X_1' and X_2' are optimal multimodal augmentation for X_1 and X_2 when $I(X_1, X_2; X_1', X_2') = I(X_1, X_2; Y)$, which implies that the only information shared between X_1, X_2 and X_1', X_2' is task-relevant with no irrelevant noise.*

We satisfy $I(X_1, X_2; X_1', X_2') = I(X_1, X_2; Y)$ using two steps:

$$\textit{Unimodal aug: } X_1' \text{ s.t. } I(X_1; X_1') = I(X_1; Y), \tag{4.17}$$

$$\textit{Unique aug: } X_2' \text{ s.t. } I(X_2; X_2'|X_1) = I(X_2; Y|X_1). \tag{4.18}$$

We call the second step *unique augmentation*: after observing X_1, we create augmented X_2' from X_2 to keep task-relevant information not already in X_1. To empirically satisfy optimal multimodal augmentations, we avoid augmentations in one modality that will remove or strongly destroy information shared with the other modality. For example, in image captioning, we should avoid image augmentations such as cropping that destroy information

Algorithm 1 Standard multimodal CL.

Require: Multimodal dataset $\{\mathbf{X_1}, \mathbf{X_2}\}$.

Initialize networks $f(\cdot)$.
while not converged **do**
  for sampled batch $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ **do**
    Estimate $I_{\text{NCE}}(X_1; X_2)$ from Eq. 4.10
    $\mathcal{L} = -I_{\text{NCE}}(X_1; X_2)$
    Update $f(\cdot)$ to minimize $\mathcal{L}$
  end for
end while
return $f(\cdot)$

Algorithm 2 FACTORCL.

Require: Multimodal dataset $\{\mathbf{X_1}, \mathbf{X_2}\}$.

Initialize networks $f(\cdot)$.
while not converged **do**
  for sampled batch $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ **do**
    $\boldsymbol{x}_1' \leftarrow$ **Augment**$(\boldsymbol{x}_1)$
    $\boldsymbol{x}_2' \leftarrow$ **Unique-Augment**$(\boldsymbol{x}_2|\boldsymbol{x}_1)$
    Plug $\boldsymbol{x}_1'$ and $\boldsymbol{x}_2'$ into Eq. 4.15 and 4.16
    Estimate $\mathbf{S}, \mathbf{U_1}, \mathbf{U_2}$ from Eq. 4.8 and 4.9
    $\mathcal{L} = -(\mathbf{S} + \mathbf{U_1} + \mathbf{U_2})$
    Update $f(\cdot)$ to minimize $\mathcal{L}$
  end for
end while
return $f(\cdot)$

from the caption (e.g., cropping object parts referred to by the caption), and instead, only augment images via flipping or color jittering which retains all caption information. Figure 4.4 shows an example of unique augmentation that satisfies these conditions. In our experiments, we will show that our augmentations consistently perform better than standard augmentations (Table 4.3), suggesting that approximately satisfying Eqs.(4.17) and (4.18) can be empirically sufficient, which is simple and straightforward to implement on real-world datasets.

4.3.3 Overall method and implementation

The final algorithm sketch is in Algorithm 2, which we compare against standard CL in Algorithm 1. It can be shown that FACTORCL learns all the task-relevant information from both modalities:

Theorem 3. *(Optimality of* FACTORCL*) If $Z_{S_1}, Z_{S_2}, Z_{U_1}, Z_{U_2}$ perfectly maximize Eqs.(4.6-4.7) and the estimations in Eqs.(4.8) and (4.9) are tight, we obtain $I(X_1, X_2; Y) = I(Z_{S_1}; Z_{S_2}; Y) + I(Z_{U_1}; Y|Z_{S_2}) + I(Z_{U_2}; Y|Z_{S_1})$, suggesting that* FACTORCL *learns both shared and unique task-relevant information.*

We include the full proof in the full paper [374]. In practice, while we do not expect perfect estimation of MI quantities and maximization with respect to MI objectives, we show that our method still improves empirical performance on several real-world datasets.

Complexity: Compared to heuristic combinations of cross-modal and single-modality CL [258, 278, 337, 534, 646, 688, 709], our approach does not significantly increase complexity: (1) upper bounds on MI can be estimated "for free" by directly plugging in the optimal critic from I_{NCE}, (2) removal of task-irrelevant information via $I(X_1; X_2|X_1', X_2')$ shares encoders with I_{NCE}, and (3) separate unimodal augmentations perform empirically well.

Table 4.1: We probe whether contrastive representations learned by classic CL methods and FACTORCL contain shared w_s or unique w_1, w_2 information. FACTORCL captures the most unique information.

Model	SimCLR		Cross+self		SupCon		FACTORCL			
Representations	Z_1	Z_2	Z_1	Z_2	Z_1	Z_2	Z_{U_1}	Z_{U_2}	Z_{S_1}	Z_{S_2}
$I(Z;w_1)$	4.45	0.16	4.39	0.14	5.17	0.19	**7.83**	0.03	6.25	0.04
$I(Z;w_2)$	0.17	3.92	0.13	4.26	0.23	5.17	0.06	**7.17**	0.05	5.79
$I(Z;w_s)$	12.61	12.06	11.30	11.47	7.48	7.17	9.47	9.89	10.13	9.40

4.4 Experiments

We run comprehensive experiments on a suite of synthetic and large-scale real-world datasets with varying requirements of shared and unique task-relevant information, comparing our FACTORCL method to key baselines:

1. SimCLR [103]: the straightforward method of cross-modal (X_1, X_2) contrastive learning.
2. Cross+Self [258, 278, 337, 534, 688, 709]: captures a range of methods combining cross-modal (X_1, X_2) CL with additional unimodal (X_i, X_i') CL objectives. This category also includes other ways of preserving unique information, such as through (variational) autoencoder reconstructions [646].
3. Cross+Self+Fact [689, 709]: A factorized extension of Cross+Self, which is approximately done in prior work that adds separate (typically pre-trained) unimodal encoders for each modality.
4. SupCon [300], which learns $I(X_1; X_2|Y)$ using CL conditioned on Y from labeled data.

We also carefully ablate each component of our method and investigate factors, including training data size and choice of augmentations. The intermediate ablations that emerge include:

1. FACTORCL-SUP: The supervised CL version which uses labels Y in Eqs.(4.13) and (4.14).
2. FACTORCL-SSL: The fully self-supervised version of our approach replacing Y with multimodal augmentations X_1' and X_2' to approximate the task.
3. OurCL-SUP: FACTORCL-SUP but removing the factorization so only two features Z_1 is optimized for both $I(X_1; X_2; Y)$ and $I(X_1; Y|X_2)$, Z_2 optimized for both $I(X_1; X_2; Y)$ and $I(X_2; Y|X_1)$.
4. OurCL-SSL: FACTORCL-SSL but also removing the factorization in the self-supervised setting.

The formulation of each ablation and implementation can be found in the full paper [374].

4.4.1 Controlled experiments on synthetic datasets

Synthetic data generation: We begin by generating data with controllable ratios of task-relevant shared and unique information. Starting with a set of latent vectors $w_1, w_2, w_s \sim \mathcal{N}(0_d, \Sigma_d^2), d = 50$ representing information unique to X_1, X_2 and common to both respectively, the concatenated vector $[w_1, w_s]$ is transformed into high-dimensional x_1 using a fixed transformation T_1 and likewise $[w_2, w_s]$ to x_2 via T_2. The label y is generated as a function (with nonlinearity and noise) of varying ratios of w_s, w_1, and w_2 to represent shared and unique task-relevant information.

Results: In Figure 9.1, we show our main result on synthetic data comparing FACTORCL with existing CL baselines. FACTORCL consistently maintains the best performance, whereas SimCLR [103] and SupCon [300] see performance drops as unique information increases. Cross+Self [258, 278, 337, 709] recovers in fully unique settings (x-axis= 1.0) but suffers at other ratios.

Representation probing information: We run a probing experiment to compute how well different contrastive representations capture shared and unique information. In Table 4.1, for the Z_i's learned by each method, we approximately compute $I(Z_i; w_1)$, $I(Z_i; w_2)$, and $I(Z_i; w_s)$ with respect to ground truth generative variables w_s, w_1, and w_2. As expected, existing methods such as SimCLR capture smaller amounts of unique information (roughly 4 bits in $I(Z_i; w_1)$ and $I(Z_i; w_2)$), focusing instead on learning $I(Z_i; w_s)$ (12 bits). Cross+self captures slightly larger $I(Z_i; w_2) = 4.26$, and SupCon with labeled data captures up to 5 bits of unique information. Our FACTORCL approach captures 7 bits of unique information and maintains 10 bits of shared information, with total information captured higher than the other approaches. Furthermore, $\{Z_{S_1}, Z_{S_2}\}$ capture more information about w_s, Z_{U_1} about w_1, and Z_{U_2} about w_2, indicating that factorization in our approach is successful.

4.4.2 Self-supervised learning with low redundancy and high uniqueness

Multimodal fusion datasets: We use a large collection of real-world datasets provided in MultiBench [367], where we expect varying ratios of shared and unique information important for the task, to compare FACTORCL with other CL baselines:

1. MIMIC [286]: mortality and disease prediction from $36,212$ medical records (tabular patient data and medical time-series sensors from ICU).
2. MOSEI [717]: multimodal sentiment and emotion benchmark with $23,000$ monologue videos.
3. MOSI [710]: multimodal sentiment analysis from $2,199$ YouTube videos.
4. UR-FUNNY [225]: a dataset of humor detection from more than $16,000$ TED talk videos.
5. MUSTARD [83]: a corpus of 690 videos for research in sarcasm detection from TV shows.
6. IRFL [700]: $6,697$ matching images and figurative captions (rather than literal captions).

Together, these datasets cover seven different modalities from the healthcare, affective computing, and multimedia research areas and total more than $84,000$ data points. For MIMIC with tabular and medical sensor inputs, we train self-supervised CL models on top of raw modality inputs. For IRFL with image and caption inputs, we start with a pretrained CLIP model [497] and perform continued pre-training to update CLIP weights with our FACTORCL objectives, before linear classifier testing. For the remaining four video datasets, we train self-supervised CL models starting from standard pre-extracted text, video, and audio features [367]. Please refer to the full paper [374] for experimental details. We release our code and models at `https://github.com/pliang279/FactorCL`.

Multimodal fusion results: From Table 4.2, FACTORCL significantly outperforms the baselines that do not capture both shared and unique information in both supervised and self-supervised settings, particularly on MUSTARD (where unique information expresses sarcasm, such as sardonic facial expressions or ironic tone of voice), and on MIMIC (with unique health indicators and sensor readings). In Table 4.3, we also show that FACTORCL substantially improves the state-of-the-art in classifying images and figurative captions which are not literally descriptive of the image on IRFL, outperforming zero-shot and fine-tuned CLIP [497] as well as continued pre-training baselines on top of CLIP.

Modeling ablations: In Table 4.2, we also carefully ablate each component in our method and indicate either existing baselines or newly-run ablation models.

1. **Factorized representations**: In comparing FACTORCL-SSL with OurCL-SSL, and also FAC-

Table 4.2: Results on MultiBench [367] datasets with varying shared and unique information: FACTORCL achieves strong results vs self-supervised (top 5 rows) and supervised (bottom 3 rows) baselines that do not have unique representations, factorization, upper-bounds to remove irrelevant information, and multimodal augmentations.

Model	$(X_1; X_2)$	$(X_i; X_i')$	$(X_1; X_2 \vert Y)$	(X_2'')	Fact	MIMIC	MOSEI	MOSI	UR-FUNNY	MUSTARD
SimCLR [103]	✓	✗	✗	✗	✗	66.67%	71.03%	46.21%	50.09%	53.48%
Cross+Self [646]	✓	✓	✗	✗	✗	65.20%	71.04%	46.92%	56.52%	53.91%
Cross+Self+Fact [709]	✓	✓	✗	✗	✓	65.49%	71.07%	52.37%	59.91%	53.91%
OurCL-SSL	✓	✓	✓	✓	✗	65.22%	71.16%	48.98%	58.79%	53.98%
FACTORCL-SSL	✓	✓	✓	✓	✓	**67.34%**	**74.88%**	**52.91%**	**60.50%**	**55.80%**
SupCon [300]	✗	✗	✓	✗	✗	67.37%	72.71%	47.23%	50.98%	52.75%
OurCL-SUP	✓	✓	✓	✗	✗	68.16%	71.15%	65.32%	58.32%	65.05%
FACTORCL-SUP	✓	✓	✓	✗	✓	**76.79%**	**77.34%**	**70.69%**	**63.52%**	**69.86%**

TORCL-SUP with OurCL-SUP, we find that factorization is critical: without it, performance drops on average 6.1%, with performance drop as high as 8.6% for MIMIC.

2. **Information removal via upper bound**: By comparing FACTORCL with SimCLR, Cross+Self, and Cross+Self+Fact, and SupCon that only seek to capture task-relevant information via contrastive lower bounds on MI, we find that separately modeling the task-relevant information (to be captured) and task-irrelevant information (to be removed) is helpful. Without removing task-irrelevant information via the upper-bound objective, performance drops on average 13.6%, with performance drops as high as 23.5% for the MOSI dataset. We also found that training was more difficult without this objective, which is expected due to overwhelming superfluous information from the dataset [717].
3. **Multimodal augmentations**: Finally, we investigate the differences between separate unimodal augmentations (FACTORCL-IndAug in Table 4.3) versus a joint multimodal augmentation (FACTORCL-SSL) on the IRFL dataset. We choose this dataset since its images and captions are the easiest to visualize (see Figure 4.4 for augmentations from both strategies). In the self-supervised setting, we find that multimodal augmentations achieve 95% performance, higher than the 92% for separate unimodal augmentations, and both outperform baselines SimCLR and Cross+Self.

Ablations on S, U_1 and U_2: In Table 4.4, we also test FACTORCL when training linear classifiers on top of only shared $\{Z_{S_1}, Z_{S_2}\}$ and unique Z_{U_1}, Z_{U_2} separately. We call these models FACTORCL-S, FACTORCL-U_1, and FACTORCL-U_2. Immediately, we observe that performance drops as compared to the full FACTORCL model, indicating that both shared and unique information are critical in real-world multimodal tasks. As expected, the best-performing submodel is the one that captures the region with the largest amount of task-relevant information: MOSEI and MOSI are known to include a lot of redundancy and unique information since language is very important for detecting sentiment [717], so FACTORCL-S and FACTORCL-U_2 perform best. For sarcasm detection on MUSTARD, video information is most important with FACTORCL-U_1 performing best (59.4%), and ablation models are also the furthest away from full multimodal performance (69.9%). This is aligned with intuition where sarcasm

Table 4.3: Continued pre-training on CLIP with our FACTORCL objectives on classifying images and figurative language.

Task	IRFL
Zero-shot CLIP [497]	89.15%
SimCLR [103]	91.57%
Cross+Self [646, 709]	95.18%
FACTORCL-IndAug	92.77%
FACTORCL-SSL	**95.18%**
Fine-tuned CLIP [497]	96.39%
SupCon [300]	89.16%
FACTORCL-SUP	**98.80%**

Table 4.4: We ablate using only shared representations $\{Z_{S_1}, Z_{S_2}\}$, unique representation Z_{U_1}, and Z_{U_2} separately for prediction. Both shared and unique information are critical in real-world multimodal tasks.

Model	MIMIC	MOSEI	MOSI	UR-FUNNY	MUSTARD
FACTORCL-S	63.77%	77.17%	70.12%	63.42%	57.25%
FACTORCL-U_1	55.90%	77.06%	70.11%	62.00%	59.42%
FACTORCL-U_2	69.08%	71.01%	52.33%	54.35%	53.62%
FACTORCL-SUP	**76.79%**	**77.34%**	**70.69%**	**63.52%**	**69.86%**

is expressed through tone of voice and visual gestures (high U_1), as well as from contradictions between language and video (higher multimodal performance).

4.5 Related Work

Contrastive learning is a successful self-supervised learning paradigm for computer vision [82, 103, 106, 211, 229, 453], natural language [188, 416, 438], speech [42, 453, 527], and multimodal tasks [15, 285, 497]. Its foundational underpinnings are inspired by work in multiview information theory [173, 300, 563, 605, 616] studying the shared information between two views and whether they are necessary or sufficient in predicting the label. Recently, Wang et al. [646] and Kahana and Hoshen [290] discuss the limitations of assuming multiview redundancy and propose autoencoder reconstruction or unimodal contrastive learning to retain unique information, which resembles the Cross+self baselines in our experiments. We refer the reader to Shwartz-Ziv and LeCun [547] for a comprehensive review on multiview and contrastive learning. Our work also relates to conditional contrastive learning [112, 397, 618, 696], where positive or negative pairs are supposed to sample from conditional distributions.

Multimodal contrastive learning aims to align related data from different modalities, typically provided as positive pairs. This could be done via optimizing a contrastive objective for inter-modality pairs [15, 17, 285, 497], or both intra- and inter-modality data pairs [258, 278, 303, 337, 709]. Our work also relates to factorized representation learning, which primarily studies how to capture modality-specific information primarily in each modality and multimodal information redundant in both modalities [249, 614]. Prior work has used disentangled latent variable models [57, 238, 249, 614], mixture-of-experts [544], or product-of-experts [668] layer to explain factors in multimodal data.

Information theory [125, 535] has been used to study several phenomena in multimodal learning, including co-learning [499, 716] and multi-view learning [261, 616]. Due to its theoretical importance, several lower and upper bounds have been proposed for practical estimation [453, 458, 487, 669]. We build on the CLUB upper bound [110] to create a more accurate and stable bound. Our characterizations of shared and unique information are also related to partial information decomposition [662], co-information [51, 635], interaction information [412], and cross-domain disentanglement [269] research.

4.6 Conclusion

This paper studied how standard CL methods suffer when task-relevant information lies in regions unique to each modality, which is extremely common in real-world applications such as sensor

placement, medical testing, and multimodal interaction. In response, we proposed FACTORCL, a new method expanding CL techniques through the use of factorized representations, removing task-irrelevant information via upper bounds on MI, and multimodal data augmentations suitable for approximating the unobserved task. Based on FACTORCL's strong performance, there are several exciting directions in extending these ideas for masked and non-contrastive pre-training.

Chapter 5

Quantifying Multimodal Interactions in Trained Models

5.1 Introduction

Using our foundation of multimodal interactions, we now present our work in *model quantification*: visualizing and understanding the internal modeling of multimodal interactions in trained models. As multimodal models are increasingly deployed in real-world applications, it has become increasingly important to quantify and understand their internal mechanics [205, 371, 465] as a step towards accurately benchmarking their limitations for more reliable deployment [231, 274]. However, modern multimodal models are typically black-box neural networks, such as pretrained transformers [348, 390], which makes understanding what interactions they learn difficult.

As a step in interpreting multimodal models, this paper introduces an analysis and visualization method called MULTIVIZ (see Figure 5.1). To tackle the challenges of visualizing model behavior, we scaffold the problem of interpretability into 4 stages: (1) *unimodal importance*: identifying the contributions of each modality towards downstream modeling and prediction, (2) *cross-modal interactions*: uncovering the various ways in which different modalities can relate with each other and the types of new information possibly discovered as a result of these relationships, (3) *multimodal representations*: how unimodal and cross-modal interactions are represented in decision-level features, and (4) *multimodal prediction*: how decision-level features are composed to make a prediction for a given task. In addition to including current approaches for unimodal importance [205, 417, 508] and cross-modal interactions [235, 396], we additionally propose new methods for interpreting cross-modal interactions, multimodal representations, and prediction to complete these stages in MULTIVIZ. By viewing multimodal interpretability through the lens of these 4 stages, MULTIVIZ contributes a *modular* and *human-in-the-loop* visualization toolkit for the community to visualize popular multimodal datasets and models as well as compare with other interpretation perspectives, and for stakeholders to understand multimodal models in their research domains.

MULTIVIZ is designed to support many modality inputs while also operating on diverse modalities, models, tasks, and research areas. Through experiments on 6 real-world multimodal tasks (spanning fusion, retrieval, and question-answering), 6 modalities, and 8 models, we show that MULTIVIZ helps users gain a deeper understanding of model behavior as measured via a

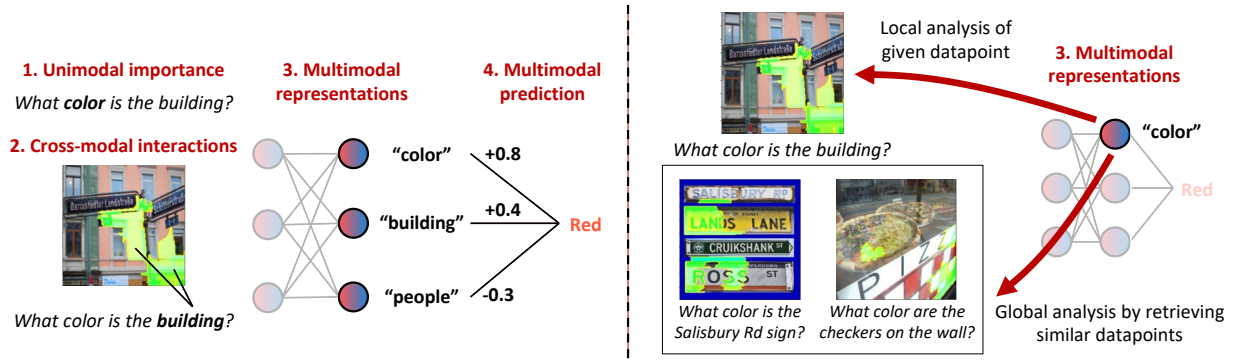


Figure 5.1: Left: We scaffold the problem of multimodal interpretability and propose MULTIVIZ, a comprehensive analysis method encompassing a set of fine-grained analysis stages: (1) **unimodal importance** identifies the contributions of each modality, (2) **cross-modal interactions** uncover how different modalities relate with each other and the types of new information possibly discovered as a result of these relationships, (3) **multimodal representations** study how unimodal and cross-modal interactions are represented in decision-level features, and (4) **multimodal prediction** studies how these features are composed to make a prediction. **Right**: We visualize multimodal representations through local and global analysis. Given an input datapoint, **local analysis** visualizes the unimodal and cross-modal interactions that activate a feature. **Global analysis** informs the user of similar datapoints that also maximally activate that feature, and is useful in assigning human-interpretable concepts to features by looking at similarly activated input regions (e.g., the concept of color).

proxy task of model simulation. We further demonstrate that MULTIVIZ helps human users assign interpretable language concepts to previously uninterpretable features and perform error analysis on model misclassifications. Finally, using takeaways from error analysis, we present a case study of human-in-the-loop model debugging. Overall, MULTIVIZ provides a practical toolkit for interpreting multimodal models for human understanding and debugging. MULTIVIZ datasets, models, and code are at `https://github.com/pliang279/MultiViz`.

5.2 MULTIVIZ: Visualizing & Understanding Multimodal Models

This section presents MULTIVIZ, our proposed analysis framework for analyzing the behavior of multimodal models. As a general setup, we assume multimodal datasets take the form $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{x}_2, y)_{i=1}^n\} = \{(x_1^{(1)}, x_1^{(2)}, ..., x_2^{(1)}, x_2^{(2)}, ..., y)_{i=1}^n\}$, with boldface $\mathbf{x}$ denoting the entire modality, each x_1, x_2 indicating modality atoms (i.e., fine-grained sub-parts of modalities that we would like to analyze, such as individual words in a sentence, object regions in an image, or time-steps in time-series data), and y denoting the label. These datasets enable us to train a multimodal model $\hat{y} = f(\mathbf{x}_1, \mathbf{x}_2; \theta)$ which we are interested in visualizing.

Modern parameterizations of multimodal models f are typically black-box neural networks, such as multimodal transformers [232, 613] and pretrained models [348, 390]. How can we visualize and understand the internal modeling of multimodal information and interactions in these models? Having an accurate understanding of their decision-making process would enable us to benchmark their opportunities and limitations for more reliable real-world deployment. However, interpreting f is difficult. In many multimodal problems, it is useful to first scaffold the problem

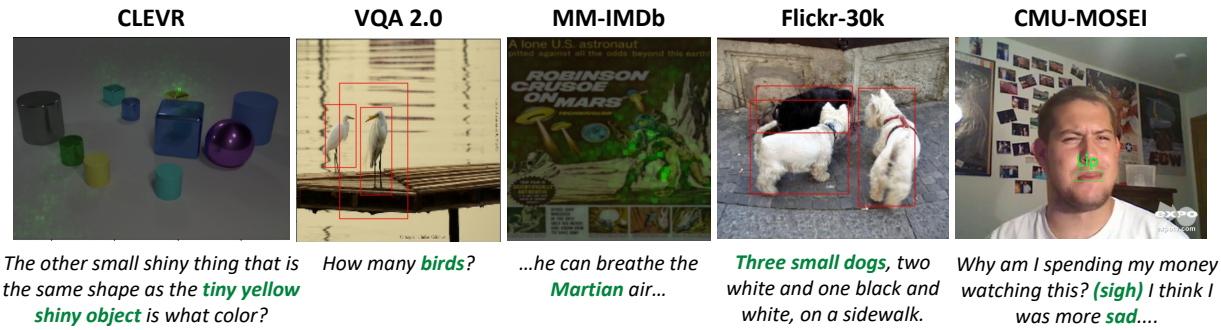


Figure 5.2: Examples of cross-modal interactions discovered by our proposed second-order gradient approach: first taking a gradient of model f with respect to an input word (e.g., x_1 = *birds*), before taking a second-order gradient with respect to all image pixels (highlighted in green) or bounding boxes (in red boxes) $\mathbf{x}_2$ indeed results in all birds in the image being highlighted.

of interpreting f into several intermediate stages from low-level unimodal inputs to high-level predictions, spanning *unimodal importance*, *cross-modal interactions*, *multimodal representations*, and *multimodal prediction*. Each of these stages provides complementary information on the decision-making process (see Figure 5.1). We now describe each step in detail and propose methods to analyze each step.

5.2.1 Unimodal importance (U)

Unimodal importance aims to understand the contributions of each modality towards modeling and prediction. It builds upon ideas of gradients [40, 165, 548] and feature attributions (e.g., LIME [508], Shapley values [417]). We implement unimodal feature attribution methods as a module $\text{UNI}(f_\theta, y, \mathbf{x})$ taking in a trained model f_θ, an output/feature y which analysis is performed with respect to, and the modality of interest $\mathbf{x}$. UNI returns importance weights across atoms x of modality $\mathbf{x}$.

5.2.2 Cross-modal interactions (C)

Cross-modal interactions describe various ways in which atoms from different modalities can relate with each other and the types of new information possibly discovered as a result of these relationships. Recent work [235, 396] has formalized a definition of cross-modal interactions by building upon literature in statistical non-additive interactions:

Definition 1 (Statistical Non-Additive Interaction [180, 562, 619, 620]). A function f learns a feature interaction $\mathcal{I}$ between 2 unimodal atoms x_1 and x_2 if and only if f cannot be decomposed into a sum of unimodal subfunctions g_1, g_2 such that $f(x_1, x_2) = g_1(x_1) + g_2(x_2)$.

This definition of non-additive interactions is general enough to include different ways that interactions can happen, including multiplicative interactions from complementary views of the data (i.e., an interaction term $x_1 \mathbb{W} x_2$ [283]), or cooperative interactions from equivalent views (i.e., an interaction term $\text{majority}(f(x_1), f(x_2))$ [146]). Using this definition, MULTIVIZ first includes two recently proposed methods for understanding cross-modal interactions: EMAP [235] decomposes $f(x_1, x_2) = g_1(x_1) + g_2(x_2) + g_{12}(x_1, x_2)$ into strictly unimodal representations g_1, g_2, and cross-modal representation $g_{12} = f - \mathbb{E}_{x_1}(f) - \mathbb{E}_{x_2}(f) + \mathbb{E}_{x_1,x_2}(f)$ to quantify the degree of global cross-modal interactions across an entire dataset. DIME [396] further extends EMAP

using feature visualization on each disentangled representation locally (per datapoint). However, these approaches require approximating expectations over modality subsets, which may not scale beyond 2 modalities. To fill this gap, we propose an efficient approach for visualizing these cross-modal interactions by observing that the following gradient definition directly follows from Definition 1:

Definition 2 (Gradient definition of statistical non-additive interaction). A function f exhibits non-additive interactions among 2 unimodal atoms x_1 and x_2 if $\mathbf{E}_{x_1,x_2}\left[\frac{\partial^2 f(x_1,x_2)}{\partial x_1 \partial x_2}\right]^2 > 0$.

Taking a second-order gradient of f zeros out the unimodal terms $g_1(x_1)$ and $g_2(x_2)$ and isolates the interaction $g_{12}(x_1, x_2)$. Theoretically, second-order gradients are necessary and sufficient to recover cross-modal interactions: purely additive models will have strictly 0 second-order gradients so $\mathbf{E}_{x_1,x_2}\left[\frac{\partial^2 f(x_1,x_2)}{\partial x_1 \partial x_2}\right]^2 = 0$, and any non-linear interaction term $g_{12}(x_1, x_2)$ has non-zero second-order gradients since g cannot be a constant or unimodal function, so $\mathbf{E}_{x_1,x_2}\left[\frac{\partial^2 f(x_1,x_2)}{\partial x_1 \partial x_2}\right]^2 > 0$.

Definition 2 inspires us to extend first-order gradient and perturbation-based approaches [221, 508, 701] to the second order. Our implementation first computes a gradient of f with respect to a modality atom which the user is interested in querying cross-modal interactions for (e.g., x_1 = *birds*), which results in a vector $\nabla_1 = \frac{\partial f}{\partial x_1}$ of the same dimension as x_1 (i.e., token embedding dimension). We aggregate the vector components of ∇_1 via summation to produce a single scalar $\|\nabla_1\|$, before taking a second-order gradient with respect to all atoms of the second modality $x_2 \in \mathbf{x}_2$ (e.g., all image pixels), which results in a vector $\nabla_{12} = \left[\frac{\partial^2 f}{\partial x_1 \partial x_2^{(1)}}, ..., \frac{\partial^2 f}{\partial x_1 \partial x_2^{(|\mathbf{x}_2|)}}\right]$ of the same dimension as $\mathbf{x}_2$ (i.e., total number of pixels). Each scalar entry in ∇_{12} highlights atoms x_2 that have non-linear interactions with the original atom x_1, and we choose the x_2's with the largest magnitude of interactions with x_1 (i.e., which highlights the birds in the image, see Figure 5.2 for examples on real datasets). We implement a general module $\text{CM}(f_\theta, y, x_1, \mathbf{x}_2)$ for cross-modal visualizations, taking in a trained model f_θ, an output/feature y, the first modality's atom of interest x_1, and the entire second modality of interest $\mathbf{x}_2$, before returning importance weights across atoms x_2 of modality $\mathbf{x}_2$.

5.2.3 Multimodal representations

Given these highlighted unimodal and cross-modal interactions at the input level, the next stage aims to understand how these interactions are represented at the feature representation level. Specifically, given a trained multimodal model f, define the matrix $M_z \in \mathbb{R}^{N \times d}$ as the penultimate layer of f representing (uninterpretable) deep feature representations implicitly containing information from both unimodal and cross-modal interactions. For the ith datapoint, $z = M_z(i)$ collects a set of individual feature representations $z_1, z_2, ..., z_d \in \mathbb{R}$. We aim to interpret these feature representations through both local and global analysis (see Figure 5.1 (right) for an example):

Local representation analysis ($\mathbf{R}_\ell$) informs the user on parts of the original datapoint that activate feature z_j. To do so, we run unimodal and cross-modal visualization methods with respect to feature z_j (i.e., $\text{UNI}(f_\theta, z_j, \mathbf{x})$, $\text{CM}(f_\theta, z_j, x_1, \mathbf{x}_2)$) in order to explain the input unimodal and cross-modal interactions represented in feature z_j. Local analysis is useful in explaining model predictions on the original datapoint by studying the input regions activating feature z_j.

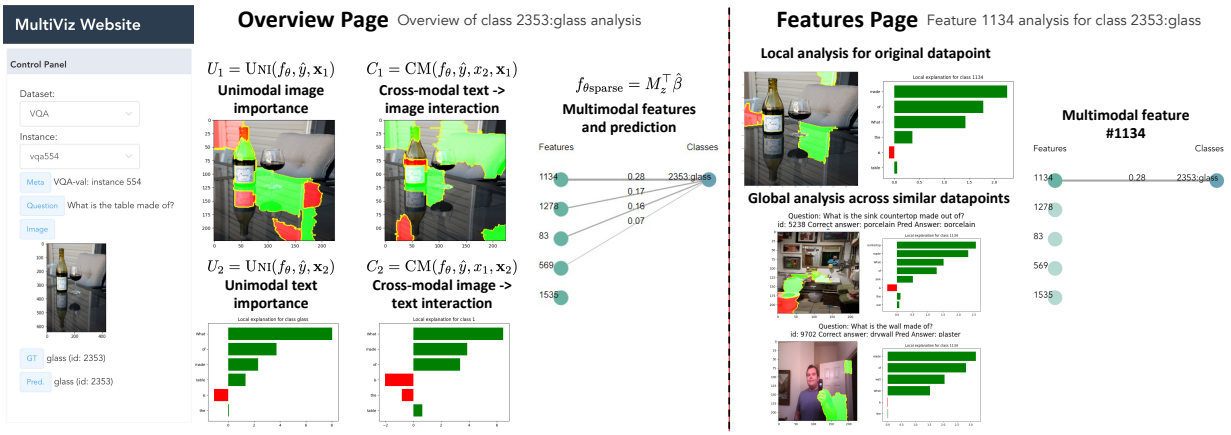


Figure 5.3: MULTIVIZ provides an interactive visualization API across multimodal datasets and models. The overview page shows general unimodal importance, cross-modal interactions, and prediction weights, while the features page enables local and global analysis of specific user-selected features.

Global representation analysis ($\mathbf{R}_g$) provides the user with the top k datapoints $\mathcal{D}_k(z_j) = \{(\mathbf{x}_1, \mathbf{x}_2, y)_{i=(1)}^k\}$ that also maximally activate feature z_j. By further unimodal and cross-modal visualizations on datapoints in $\mathcal{D}_k(z_j)$, global analysis is especially useful in helping humans assign interpretable language concepts to each feature by looking at similarly activated input regions across datapoints (e.g., the concept of color in Figure 5.1, right). Global analysis can also help to find related datapoints the model also struggles with for error analysis.

5.2.4 Multimodal prediction (P)

Finally, the prediction step takes the set of feature representations $z_1, z_2, ..., z_d$ and composes them to form higher-level abstract concepts suitable for a task. We approximate the prediction process with a linear combination of penultimate layer features by integrating a sparse linear prediction model with neural network features [666]. Given the penultimate layer $M_z \in \mathbb{R}^{N \times d}$, we fit a linear model $\mathbb{E}(Y|X = x) = M_z^\top \beta$ (bias β_0 omitted for simplicity) and solve for sparsity using:

$$\hat{\beta} = \arg\min_\beta \frac{1}{2N} \|M_z^\top \beta - y\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2. \tag{5.1}$$

The resulting understanding starts from the set of learned weights with the highest non-zero coefficients $\beta_{\text{top}} = \{\beta_{(1)}, \beta_{(2)}, ...\}$ and corresponding ranked features $z_{\text{top}} = \{z_{(1)}, z_{(2)}, ...\}$. β_{top} tells the user how features z_{top} are composed to make a prediction, and z_{top} can then be visualized with respect to unimodal and cross-modal interactions using the representation stage (Section 5.2.3).

5.2.5 Putting everything together

We summarize these proposed approaches for understanding each step of the multimodal process and show the overall MULTIVIZ user interface in Figure 5.3. This interactive API enables users to choose multimodal datasets and models and be presented with a set of visualizations at each stage, with an **overview page** for general unimodal importance, cross-modal interactions, and prediction weights, as well as a **feature page** for local and global analysis of user-selected features (see full paper [375] for details).

Table 5.1: MULTIVIZ enables fine-grained analysis across 6 datasets spanning 3 research areas, 6 input modalities (ℓ: language, i: image, v: video, a: audio, t: time-series, ta: tabular), and 8 models.

Area	Dataset	Model	Modalities	# Samples	Prediction task
Fusion	CMU-MOSEI	MULT	$\{\ell, v, a\} \rightarrow y$	$22,777$	sentiment, emotions
	MM-IMDB	LRTF	$\{\ell, i\} \rightarrow y$	$25,959$	movie genre classification
	MIMIC	LF	$\{t, ta\} \rightarrow y$	$36,212$	mortality, ICD-9 codes
Retrieval	FLICKR-30K	VILT	$\ell \leftrightarrow i$	$158,000$	image-caption retrieval
	FLICKR-30K	CLIP	$\ell \leftrightarrow i$	$158,000$	image-caption retrieval
QA	CLEVR	CNN-LSTM-SA	$\{i, \ell\} \rightarrow y$	$853,554$	QA
	CLEVR	MDETR	$\{i, \ell\} \rightarrow y$	$853,554$	QA
	VQA 2.0	LXMERT	$\{i, \ell\} \rightarrow y$	$1,100,000$	QA

5.3 Experiments

Our experiments are designed to verify the usefulness and complementarity of the 4 MULTIVIZ stages. We start with a model simulation experiment to test the utility of each stage towards overall model understanding (Section 5.3.1). We then dive deeper into the individual stages by testing how well MULTIVIZ enables representation interpretation (Section 5.3.2) and error analysis (Section 5.3.3), before presenting a case study of model debugging from error analysis insights (Section 5.3.4). We showcase the following selected experiments and defer results on other datasets to the full paper [375].

Setup: We use a large suite of datasets from MultiBench [367] which span real-world fusion [32, 287, 717], retrieval [485], and QA [206, 289] tasks. For each dataset, we test a corresponding state-of-the-art model: MULT [613], LRTF [388], LF [46], VILT [307], CLIP [497], CNN-LSTM-SA [289], MDETR [292], and LXMERT [588]. These cover models both pre-trained and trained from scratch. We summarize all 6 datasets and 8 models tested in Table 5.1, and provide more details in the full paper [375].

5.3.1 Model simulation

We first design a model simulation experiment to determine if MULTIVIZ helps users of multimodal models gain a deeper understanding of model behavior. If MULTIVIZ indeed generates human-understandable explanations, humans should be able to accurately simulate model predictions given these explanations only, as measured by correctness with respect to actual model predictions and annotator agreement (Krippendorff's alpha [316]). To investigate the utility of each stage in MULTIVIZ, we design a human study to see how accurately 21 humans users (3 users for each of the following 7 local ablation settings) can simulate model predictions:

(1) **U**: Users are only shown the unimodal importance (U) of each modality towards label y.

(2) **U + C**: Users are also shown cross-modal interactions (C) highlighted towards label y.

(3) **U + C + $\mathbf{R}_\ell$**: Users are also shown local analysis (R_ℓ) of unimodal and cross-modal interactions of top features $z_{\text{top}} = \{z_{(1)}, z_{(2)}, ...\}$ maximally activating label y.

(4) **U + C + $\mathbf{R}_\ell$ + $\mathbf{R}_g$**: Users are additionally shown global analysis (R_g) through similar datapoints that also maximally activate top features z_{top} for label y.

(5) **MULTIVIZ (U + C + $\mathbf{R}_\ell$ + $\mathbf{R}_g$ + P)**: The entire MULTIVIZ method by further including visualizations of the final prediction (P) stage: sorting top ranked feature neurons

Table 5.2: Model simulation: We tasked 15 humans users (3 users for each of the following local ablation settings) to simulate model predictions based on visualized evidences from MULTIVIZ. Human annotators who have access to all stages visualized in MULTIVIZ are able to accurately and consistently simulate model predictions (regardless of whether the model made the correct prediction) with high accuracy and annotator agreement, representing a step towards model understanding.

Research area	QA		Fusion		Fusion	
Dataset	VQA 2.0		MM-IMDB		CMU-MOSEI	
Model	LXMERT		LRTF		MULT	
Metric	Correctness	Agreement	Correctness	Agreement	Correctness	Agreement
U	55.0 ± 0.0	0.39	50.0 ± 13.2	0.34	71.7 ± 17.6	0.39
U + C	65.0 ± 5.0	0.50	53.7 ± 7.6	0.51	76.7 ± 10.4	0.45
U + C + R_ℓ	61.7 ± 7.6	0.57	56.7 ± 7.6	0.59	78.3 ± 2.9	0.42
U + C + R_ℓ + R_g	71.7 ± 15.3	0.61	61.7 ± 7.6	0.43	**100.0 ± 0.0**	**1.00**
MULTIVIZ	**81.7 ± 2.9**	**0.86**	**65.0 ± 5.0**	**0.60**	**100.0 ± 0.0**	**1.00**

$z_{\text{top}} = \{z_{(1)}, z_{(2)}, ...\}$ with respect to their coefficients $\beta_{\text{top}} = \{\beta_{(1)}, \beta_{(2)}, ...\}$ and showing these coefficients to the user.

Using 20 datapoints per setting, these experiments with 15 users on 3 datasets and 3 models involve 35 total hours of users interacting with MULTIVIZ, which is a significantly larger-scale study of model simulation compared to prior work [7, 396, 654].

Quantitative results: We show these results in Table 5.2 and find that having access to all stages in MULTIVIZ leads to significantly highest accuracy of model simulation on VQA 2.0, along with lowest variance and most consistent agreement between annotators. On fusion tasks with MM-IMDB and CMU-MOSEI, we also find that including each visualization stage consistently leads to higher correctness and agreement, despite the fact that fusion models may not require cross-modal interactions to solve the task [235]. More importantly, humans are able to simulate model predictions, regardless of whether the model made the correct prediction or not.

To test additional intermediate ablations, we conducted user studies on (6) $\mathbf{R}_\ell$ **+ P** (local analysis on final-layer features along with their prediction weights) and (7) $\mathbf{R}_g$ **+ P** (global analysis on final-layer features along with their prediction weights), to ablate the effect of overall analysis (**U** and **C**) and feature analysis ($\mathbf{R}_\ell$ or $\mathbf{R}_g$ in isolation). $\mathbf{R}_\ell$ **+ P** results in an accuracy of 51.7 ± 12.6 with 0.40 agreement, while $\mathbf{R}_g$ **+ P** gives 71.7 ± 7.6 with 0.53 agreement. Indeed, these underperform as compared to including overall analysis (**U** and **C**) and feature analysis ($\mathbf{R}_\ell$ **+** $\mathbf{R}_g$).

Finally, we also scaled to 100 datapoints on VQA 2.0, representing upwards of 10 hours of user interaction (for the full MULTIVIZ setting), and obtain an overall correctness of 80%, reliably within the range of model simulation using 20 points (81.7 ± 2.9). Therefore, the sample size of 20 points that makes all experiments feasible is still a reliable sample.

We also conducted **qualitative interviews** to determine what users found useful in MULTIVIZ:

(1) Users reported that they found local and global representation analysis particularly useful: global analysis with other datapoints that also maximally activate feature representations were important for identifying similar concepts and assigning them to multimodal features.

(2) Between Overview (**U + C**) and Feature ($\mathbf{R}_\ell$ **+** $\mathbf{R}_g$ **+ P**) visualizations, users found Feature visualizations more useful in 31.7%, 61.7%, and 80.0% of the time under settings (3), (4), and (5) respectively, and found Overview more useful in the remaining points. This means that for each

Table 5.3: Left: Across 15 human users (5 users for each of the following 3 settings), we find that users are able to consistently assign concepts to previously uninterpretable multimodal features using both local and global representation analysis. **Right**: Across 10 human users (5 users for each of the following 2 settings), we find that users are also able to categorize model errors into one of 3 stages they occur in when given full MULTIVIZ visualizations.

Research area	QA	
Dataset	VQA 2.0	
Model	LXMERT	
Metric	Confidence	Agree.
R_ℓ	1.74 ± 0.52	0.18
$R_\ell + R_g$ (no viz)	3.67 ± 0.45	0.60
$R_\ell + R_g$	$\mathbf{4.50 \pm 0.43}$	**0.69**

Research area	QA		QA	
Dataset	CLEVR		VQA 2.0	
Model	CNN-LSTM-SA		LXMERT	
Metric	Confidence	Agree.	Confidence	Agree.
No viz	2.72 ± 0.15	0.05	2.15 ± 0.70	0.14
MULTIVIZ	$\mathbf{4.12 \pm 0.45}$	**0.67**	$\mathbf{4.21 \pm 0.62}$	**0.60**

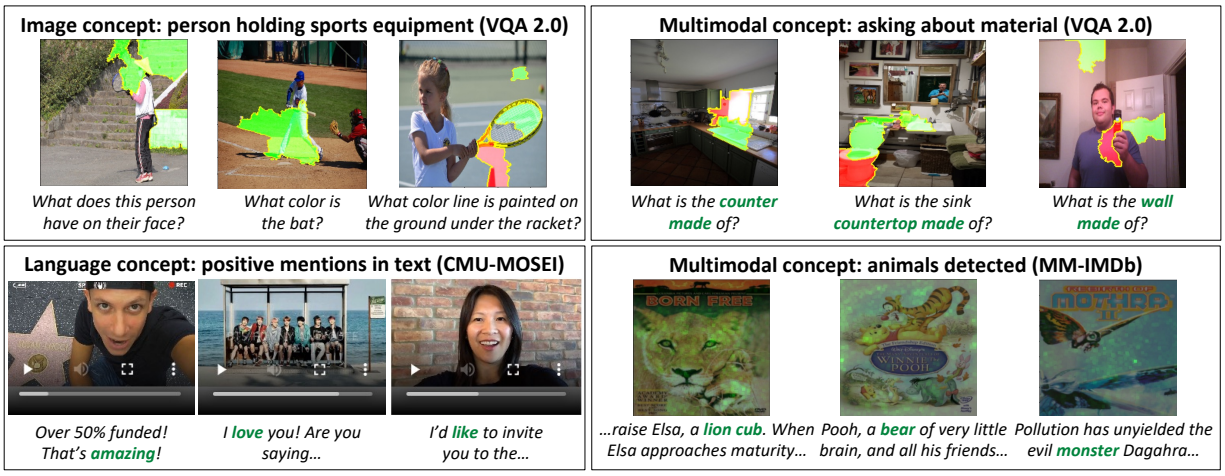


Figure 5.4: Examples of human-annotated **concepts** using MULTIVIZ on feature representations. We find that the features separately capture image-only, language-only, and multimodal concepts.

stage, there exists a significant fraction of data points where that stage is most needed.

(3) While it may be possible to determine the prediction of the model with a subset of stages, having more stages that confirm the same prediction makes them a lot more confident about their prediction, which is quantitatively substantiated by the higher accuracy, lower variance, and higher agreement in human predictions. We also include additional experiments in the full paper [375].

5.3.2 Representation interpretation

We now take a deeper look to check that MULTIVIZ generates accurate explanations of multimodal representations. Using local and global representation visualizations, can humans consistently assign interpretable concepts in natural language to previously uninterpretable features? We study this question by tasking 15 human users (5 users for each of the following 3 settings) to assign concepts to each feature z when given access to visualizations of (1) $\mathbf{R}_\ell$ (local analysis of unimodal and cross-modal interactions in z), (2) $\mathbf{R}_\ell + \mathbf{R}_g$ **(no viz)** (including global analysis through similar datapoints that also maximally activate feature z), and (3) $\mathbf{R}_\ell + \mathbf{R}_g$ (adding highlighted unimodal and cross-modal interactions of global datapoints). Using 20 datapoints per setting, these experiments with 15 users involve roughly 10 total hours of users interacting with

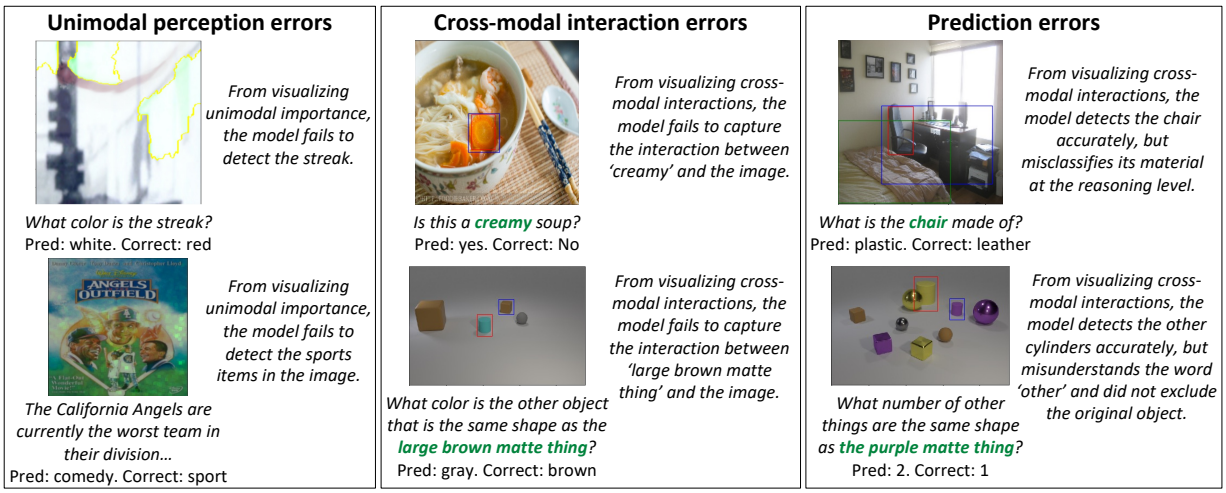


Figure 5.5: Examples of human-annotated **error analysis** using MULTIVIZ on multimodal models. Using all stages provided in MULTIVIZ enables fine-grained classification of model errors (e.g., errors in unimodal processing, cross-modal interactions, and predictions) for targeted debugging.

MULTIVIZ.

Quantitative results: Since there are no ground-truth labels for feature concepts, we rely on annotator confidence (1-5 scale) and annotator agreement [316] as a proxy for accuracy. From Table 5.3 (left), we find that having access to both local and global visualizations are crucial towards interpreting multimodal features, as measured by higher confidence with low variance in confidence, as well as higher agreement among users.

Qualitative interviews: We show examples of human-assigned concepts in Figure 5.4. Note that the 3 images in each box of Figure 5.4 (even without feature highlighting) does constitute a visualization generated by MULTIVIZ, as they belong to data instances that maximize the value of the feature neuron (i.e. R_g in stage 3 multimodal representations). Without MULTIVIZ, it would not be possible to perform feature interpretation without combing through the entire dataset. Participants also noted that feature visualizations make the decision a lot more confident if its highlights match the concept. Taking as example Figure 5.4 top left, the visualizations serve to highlight what the model's feature neuron is learning (i.e., highlighting the person holding sports equipment), rather than what category of datapoint it is. If the visualization was different, such as highlighting the ground, then users would have to conclude that the feature neuron is capturing '*outdoor ground*' rather than '*sports equipment*'. Similarly, for text highlights (Figure 5.4 top right), without using MULTIVIZ to highlight '*counter*', '*countertop*', and '*wall*', along with the image crossmodal interactions corresponding to these entities, one would not be able to deduce that the feature asks about material - it could also represent '*what*' questions, or '*household objects*', and so on. Therefore, these conclusions can only be reliably deduced with all MultiViz stages.

5.3.3 Error analysis

We examine a case study of error analysis on trained models. We task 10 human users (5 users for each of the following 2 settings) to use MULTIVIZ and highlight the errors that a model exhibits

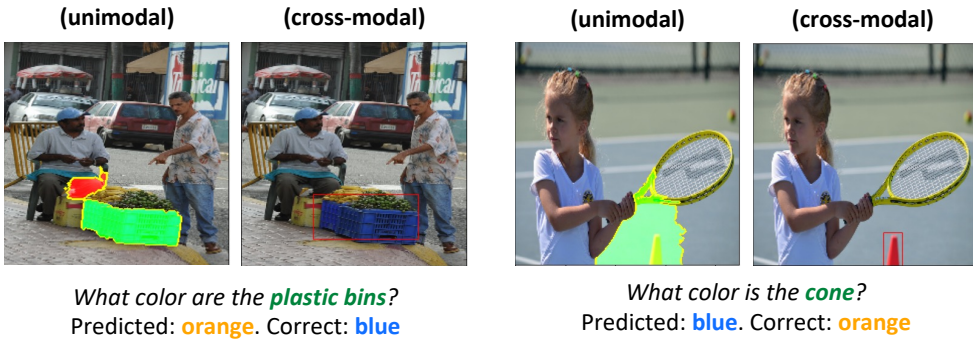


Figure 5.6: A case study on **model debugging**: we task 3 human users to use MULTIVIZ visualizations and highlight the errors that a pretrained LXMERT model fine-tuned on VQA 2.0 exhibits, and find 2 penultimate-layer neurons highlighting the model's failure to identify color (especially blue). Targeted localization of the error to this specific stage (prediction) and representation concept (blue) via MULTIVIZ enabled us to identify a bug in the popular Hugging Face LXMERT repository.

by categorizing these errors into one of 3 stages: failures in (1) unimodal perception, (2) capturing cross-modal interaction, and (3) prediction with perceived unimodal and cross-modal information. Again, we rely on annotator confidence (1-5 scale) and agreement due to lack of ground-truth error categorization, and compare (1) **MULTIVIZ** with (2) **No viz**, a baseline that does not provide any model visualizations to the user. Using 20 datapoints per setting, these experiments with 10 users on 2 datasets and 2 models involve roughly 15 total hours of users interacting with MULTIVIZ. From Table 5.3 (right), we find that MULTIVIZ enables humans to consistently categorize model errors into one of 3 stages. We show examples that human annotators classified into unimodal perception, cross-modal interaction, and prediction errors in Figure 5.5.

5.3.4 A case study in model debugging

Following error analysis, we take a deeper investigation into one of the errors on a pretrained LXMERT model fine-tuned on VQA 2.0. Specifically, we first found the top 5 penultimate-layer neurons that are most activated on erroneous datapoints. Inspecting these neurons carefully through MULTIVIZ local and global representation analysis, human annotators found that 2 of the 5 neurons were consistently related to questions asking about color, which highlighted the model's failure to identify color correctly (especially blue). The model has an accuracy of only 5.5% amongst all blue-related points (i.e., either have blue as correct answer or predicted answer), and these failures account for 8.8% of all model errors. We show examples of such datapoints and their MULTIVIZ visualizations in Figure 5.6. Observe that the model is often able to capture unimodal and cross-modal interactions perfectly, but fails to identify color at prediction.

Curious as to the source of this error, we looked deeper into the source code for the entire pipeline of LXMERT, including that of its image encoder, Faster R-CNN [506][1]. We in fact uncovered a bug in data preprocessing for Faster R-CNN in the popular Hugging Face repository that swapped the image data storage format from RGB to BGR formats responsible for these errors. This presents a concrete use case of MULTIVIZ: through visualizing each stage, we

[1]we used the popular Hugging Face implementation at `https://huggingface.co/unc-nlp/lxmert-vqa-uncased`

were able to (1) isolate the source of the bug (at prediction and not unimodal perception or cross-modal interactions), and (2) use representation analysis to localize the bug to the specific color concept. In our full paper [375], we further detail our initial attempt at tackling this error by using MULTIVIZ analysis to select additional targeted datapoints in an active learning scenario, which proved to be much more effective (higher improvement with fewer data) as compared to baselines that add data randomly or via uncertainty sampling [344], which may be of independent interest.

5.3.5 Additional experiments and takeaways messages

New models: We included results on VILT [307], CLIP [497], and MDETR [292] in the full paper [375], showing that MULTIVIZ is a general approach that can be quickly applied to new models. We also study the correlation between performance and cross-modal interactions across several older and recent models, and find that the ability to capture cross-modal alignment, as judged by MULTIVIZ, correlates strongly with final task performance.

Sanity checks: In our full paper [375], we show that MULTIVIZ passes the data randomization and model randomization sanity checks for interpretation approaches [6].

Intermediate-layer features: Finally, we show that MULTIVIZ can be extended to visualize any intermediate layer, not just the final layer of multimodal models. We showcase a few examples of $\mathbf{R}_\ell$ and $\mathbf{R}_g$ on intermediate-layer neurons and discuss several tradeoffs: while they reveal new visualization opportunities, they run the risk of overwhelming the user with the number of images they have to see multiplied by d^L (d: dimension of each layer, L: number of layers).

5.4 Related Work

Interpretable ML aims to further our understanding and trust of ML models, enable model debugging, and use these insights for joint decision-making between stakeholders and AI [104, 198]. Interpretable ML is a critical area of research straddling machine learning [6], language [597], vision [548], and HCI [117]. We categorize related work in interpreting multimodal models into:

Unimodal importance: Several approaches have focused on building interpretable components for unimodal importance through soft [465] and hard attention mechanisms [101]. When aiming to explain black-box multimodal models, related work rely primarily on gradient-based visualizations [40, 165, 548] and feature attributions (e.g., LIME [508], Shapley values [417]) to highlight regions of the image which the model attends to.

Cross-modal interactions: Recent work investigates the activation patterns of pretrained transformers [79, 349], performs diagnostic experiments through specially curated inputs [177, 320, 463, 601], or trains auxiliary explanation modules [293, 465]. Particularly related to our work is EMAP [235] for disentangling the effects of unimodal (additive) contributions from cross-modal interactions in multimodal tasks, as well as M2Lens [654], an interactive visual analytics system to visualize multimodal models for sentiment analysis through both unimodal and cross-modal contributions.

Multimodal representation and prediction: Existing approaches have used language syntax (e.g., the question in VQA) for compositionality into higher-level features [22, 26, 632]. Similarly,

logical statements have been integrated with neural networks for interpretable logical reasoning [203, 586]. However, these are typically restricted to certain modalities or tasks. Finally, visualizations have also uncovered several biases in models and datasets (e.g., unimodal biases in VQA questions [24, 75] or gender biases in image captioning [231]). We believe that MULTIVIZ will enable the identification of biases across a wider range of modalities and tasks.

5.5 Conclusion

This paper proposes MULTIVIZ for analyzing and visualizing multimodal models. MULTIVIZ scaffolds the interpretation problem into unimodal importance, cross-modal interactions, multimodal representations, and multimodal prediction, before providing existing and newly proposed analysis tools in each stage. MULTIVIZ is designed to be *modular* (encompassing existing analysis tools and encouraging research towards understudied stages), *general* (supporting diverse modalities, models, and tasks), and *human-in-the-loop* (providing a visualization tool for human model interpretation, error analysis, and debugging), qualities which we strive to upkeep by ensuring its public access and regular updates from community feedback.

Chapter 6

Estimating Multimodal Performance and Modality Selection

6.1 Introduction

To conclude the first part of this thesis, we provide a guideline for researchers to decide which modalities to collect that will lead to improved multimodal performance [376]. Specifically, we study how to quantify interactions in a semi-supervised setting where there is only *unlabeled multimodal data* $\mathcal{D}_M = \{(x_1, x_2)\}$ and some *labeled unimodal data* $\mathcal{D}_i = \{(x_i, y)\}$ collected separately for each modality. This multimodal semi-supervised paradigm is reminiscent of many real-world settings with separate unimodal datasets like visual recognition [140] and text classification [642], as well as naturally co-occurring multimodal data (e.g., news images and captions or video and audio), but when labeling them is time-consuming [247, 250] or impossible due to partially observed modalities [370] or privacy concerns [90]. We want to understand how the modalities can share, exchange, and create information to inform practitioners whether it is worth collecting multimodal data and trying multimodal models [283, 372, 712].

Using a precise information-theoretic definition of interactions [59], our key contributions are the derivations of lower and upper bounds to quantify multimodal interactions in this semi-supervised setting with only $\mathcal{D}_i$ and $\mathcal{D}_M$. We propose two lower bounds: the first relates interactions with the amount of *shared information* between modalities, and the second is based on the *disagreement* of classifiers trained separately on each modality. Finally, we propose an upper bound through connections to approximate algorithms for *min-entropy couplings* [118]. To validate our bounds, we experiment on both synthetic and large real-world datasets with varying amounts of interactions. In addition, these theoretical results naturally yield new guarantees regarding the performance of multimodal models. By analyzing the relationship between interaction estimates and downstream task performance assuming optimal multimodal classifiers are trained on labeled multimodal data, we can *closely predict multimodal model performance, before even training the model itself*. These performance estimates also help develop new guidelines for deciding when to *collect additional modality data* and *select the appropriate multimodal fusion models*. We believe these results shed light on the intriguing connections between multimodal interactions, modality disagreement, and model performance, and release our code and models at `https://github.com/pliang279/PID`.

6.2 Related Work and Technical Background

6.2.1 Semi-supervised multimodal learning

Let $\mathcal{X}_i$ and $\mathcal{Y}$ be finite sample spaces for features and labels. Define Δ to be the set of joint distributions over $(\mathcal{X}_1, \mathcal{X}_2, \mathcal{Y})$. We are concerned with features X_1, X_2 (with support $\mathcal{X}_i$) and labels Y (with support $\mathcal{Y}$) drawn from some distribution $p \in \Delta$. We denote the probability mass function by $p(x_1, x_2, y)$, where omitted parameters imply marginalization. Many real-world applications such as multimedia and healthcare naturally exhibit multimodal data (e.g., images and captions, video and audio, multimodal medical readings) which are difficult to label [370, 497, 551, 703, 721]. As such, rather than the full distribution from p, we only have partial datasets:

- *Labeled unimodal* data $\mathcal{D}_1 = \{(x_1, y) : \mathcal{X}_1 \times \mathcal{Y}\}$, $\mathcal{D}_2 = \{(x_2, y) : \mathcal{X}_2 \times \mathcal{Y}\}$.
- *Unlabeled multimodal* data $\mathcal{D}_M = \{(x_1, x_2) : \mathcal{X}_1 \times \mathcal{X}_2\}$.

$\mathcal{D}_1$, $\mathcal{D}_2$ and $\mathcal{D}_M$ follow the *pairwise marginals* $p(x_1, y)$, $p(x_2, y)$ and $p(x_1, x_2)$. We define $\Delta_{p_{1,2}} = \{q \in \Delta : q(x_i, y) = p(x_i, y)\ \forall y \in \mathcal{Y}, x_i \in \mathcal{X}_i, i \in [2]\}$ as the set of joint distributions which agree with the labeled unimodal data $\mathcal{D}_1$ and $\mathcal{D}_2$, and $\Delta_{p_{1,2,12}} = \{r \in \Delta : r(x_1, x_2) = p(x_1, x_2), r(x_i, y) = p(x_i, y)\}$ as the set of joint distributions which agree with all $\mathcal{D}_1, \mathcal{D}_2$ and $\mathcal{D}_M$.

6.2.2 Multimodal interactions and information theory

The study of **multimodal interactions** aims to quantify the information shared between both modalities, in each modality alone, and how modalities can combine to form new information not present in either modality, eventually using these insights to design machine learning models to capture interactions from large-scale multimodal datasets [371]. Existing literature has primarily studied the interactions captured by trained models, such as using Shapley values [272] and Integrated gradients [375, 579, 619] to measure the importance a model assigns to each modality, or approximating trained models with additive or non-additive functions to determine what functions are best suited to capture interactions [180, 235, 562]. However, these measure interactions captured by a trained model - *our work is fundamentally different in that interactions are properties of data*. Quantifying the interactions in data, independent of trained models, allows us to characterize datasets, predict model performance, and perform model selection, prior to choosing and training a model altogether. Prior work in understanding data interactions to design multimodal models is often driven by intuition, such as using contrastive learning [486, 497, 608], correlation analysis [27], and agreement [147] for shared information (e.g., images and descriptive captions), or using tensors and multiplicative interactions [283, 712] for higher-order interactions (e.g., in expressions of sarcasm from speech and gestures).

To fill the gap in data quantification, **information theory** has emerged as a theoretical foundation since it naturally formalizes information and its sharing as statistical properties of data distributions. Information theory studies the information that one random variable (X_1) provides about another (X_2), as quantified by Shannon's mutual information (MI) and conditional MI:

$$I(X_1; X_2) = \int p(x_1, x_2) \log \frac{p(x_1, x_2)}{p(x_1)p(x_2)} d\boldsymbol{x}, \quad I(X_1; X_2|Y) = \int p(x_1, x_2, y) \log \frac{p(x_1, x_2|y)}{p(x_1|y)p(x_2|y)} d\boldsymbol{x}dy.$$

$I(X_1; X_2)$ measures the amount of information (in bits) obtained about X_1 by observing X_2, and by extension, $I(X_1; X_2|Y)$ is the expected value of MI given the value of a third (e.g., task Y).

To generalize information theory for multimodal interactions, Partial information decomposition (PID) [662] decomposes the total information that two modalities X_1, X_2 provide about a task Y into 4 quantities: $I_p(\{X_1, X_2\}; Y) = R + U_1 + U_2 + S$, where $I_p(\{X_1, X_2\}; Y)$ is the MI between the joint random variable (X_1, X_2) and Y. These 4 quantities are: redundancy R for the task-relevant information shared between X_1 and X_2, uniqueness U_1 and U_2 for the information present in only X_1 or X_2 respectively, and synergy S for the emergence of new information only when both X_1 and X_2 are present [59, 210]:

Definition 5. *(Multimodal interactions) Given X_1, X_2, and a target Y, we define their redundant (R), unique (U_1 and U_2), and synergistic (S) interactions as:*

$$R = \max_{q \in \Delta_{p_{1,2}}} I_q(X_1; X_2; Y), \quad U_1 = \min_{q \in \Delta_{p_{1,2}}} I_q(X_1; Y|X_2), \quad U_2 = \min_{q \in \Delta_{p_{1,2}}} I_q(X_2; Y|X_1), \tag{6.1}$$

$$S = I_p(\{X_1, X_2\}; Y) - \min_{q \in \Delta_{p_{1,2}}} I_q(\{X_1, X_2\}; Y), \tag{6.2}$$

where the notation $I_p(\cdot)$ and $I_q(\cdot)$ disambiguates mutual information (MI) under p and q respectively.

$I(X_1; X_2; Y) = I(X_1; X_2) - I(X_1; X_2|Y)$ is a multivariate extension of information theory [51, 412]. Most importantly, R, U_1, and U_2 can be computed exactly using convex programming over distributions $q \in \Delta_{p_{1,2}}$ with access only to the marginals $p(x_1, y)$ and $p(x_2, y)$ by solving a convex optimization problem with linear marginal-matching constraints $q^* = \arg\max_{q \in \Delta_{p_{1,2}}} H_q(Y|X_1, X_2)$ [59, 372]. This gives us an elegant interpretation that we need only labeled unimodal data in each feature from $\mathcal{D}_1$ and $\mathcal{D}_2$ to estimate redundant and unique interactions. Unfortunately, S is impossible to compute via equation (6.2) when we do not have access to the full joint distribution p, since the first term $I_p(\{X_1, X_2\}; Y)$ is unknown.

It is worth noting that other valid information-theoretic definitions of multimodal interactions also exist, but are known to suffer from issues regarding over- and under-estimation, and may even be negative; these are critical problems with the application of information theory for shared $I(X_1; X_2; Y)$ and unique information $I(X_1; Y|X_2)$, $I(X_2; Y|X_1)$ often quoted in the co-training [44, 65] and multi-view learning [563, 605, 608, 616] literature. We refer the reader to Griffith and Koch [210] for a full discussion. We choose the one in Definition 5 above since it fulfills several desirable properties, but our results can be extended to other definitions as well.

6.3 Estimating Semi-supervised Multimodal Interactions

Our goal is to estimate multimodal interactions R, U_1, U_2, and S assuming access to only semi-supervised multimodal data $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_M$. Our first insight is that while S cannot be computed exactly, R, U_1, and U_2 can be computed from equation 6.1 with access to only semi-supervised data. Therefore, studying the relationships between S and other multimodal interactions is key to its estimation. Using these relationships, we will then derive lower and upper bounds for synergy in the form $\underline{S} \leq S \leq \overline{S}$. Crucially, $\underline{S}$ and $\overline{S}$ depend *only* on $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_M$.

6.3.1 Understanding relationships between interactions

We start by identifying two important relationships, between S and R, and between S and U.

Synergy and redundancy Our first relationship stems from the case when two modalities contain shared information about the task. In studying these situations, a driving force for estimating S is the amount of shared information $I(X_1; X_2)$ between modalities, with the intuition that more shared information naturally leads to redundancy which gives less opportunity for new synergistic interactions. Mathematically, we formalize this by relating S to R,

$$S = R - I_p(X_1; X_2; Y) = R - I_p(X_1; X_2) + I_p(X_1; X_2|Y). \tag{6.3}$$

implying that synergy exists when there is high redundancy and low (or even negative) three-way MI $I_p(X_1; X_2; Y)$. By comparing the difference in X_1, X_2 dependence with and without the task (i.e., $I_p(X_1; X_2)$ vs $I_p(X_1; X_2|Y)$), 2 cases naturally emerge (see left side of Figure 6.1):

1. **S > R**: When both modalities do not share a lot of information as measured by low $I(X_1; X_2)$, but conditioning on Y *increases* their dependence: $I(X_1; X_2|Y) > I(X_1; X_2)$, then there is synergy between modalities when combining them for task Y. This setting is reminiscent of common cause structures. Examples of these distributions in the real world are multimodal question answering, where the image and question are less dependent (some questions like 'what is the color of the car' or 'how many people are there' can be asked for many images), but the answer (e.g., 'blue car') connects the two modalities, resulting in dependence given the label. As expected, $S = 4.92, R = 0.79$ for the VQA 2.0 dataset [206].
2. **R > S**: Both modalities share a lot of information but conditioning on Y *reduces* their dependence: $I(X_1; X_2) > I(X_1; X_2|Y)$, which results in more redundant than synergistic information. This setting is reminiscent of common effect structures. A real-world example is in detecting sentiment from multimodal videos, where text and video are highly dependent since they are emitted by the same speaker, but the sentiment label explains away some of the dependencies between both modalities. Indeed, for multimodal sentiment analysis from text, video, and audio of monologue videos on MOSEI [717], $R = 0.26$ and $S = 0.04$.

Synergy and uniqueness The second relationship arises when two modalities contain disagreeing information about the task, and synergy arises due to this disagreement in information. To illustrate this, suppose $y_1 = \arg\max_y p(y|x_1)$ is the most likely prediction from the first modality, $y_2 = \arg\max_y p(y|x_2)$ for the second modality, and $y = \arg\max_y p(y|x_1, x_2)$ is the true multimodal prediction. There are again 2 cases (see right side of Figure 6.1):

1. **U > S**: Multimodal prediction $y = \arg\max_y p(y|x_1, x_2)$ is the same as one of the unimodal predictions (e.g., $y = y_2$), in which case unique information in modality 2 leads to the outcome and there is no synergy. A real-world dataset is MIMIC involving mortality and disease prediction from tabular patient data and time-series medical sensors [286] which primarily shows unique information in the tabular modality. The disagreement on MIMIC is high at 0.13, but since disagreement is due to a lot of unique information, there is less synergy $S = 0.01$.
2. **S > U**: Multimodal prediction y is different from both y_1 and y_2, then both modalities interact synergistically to give rise to a final outcome different from both disagreeing unimodal

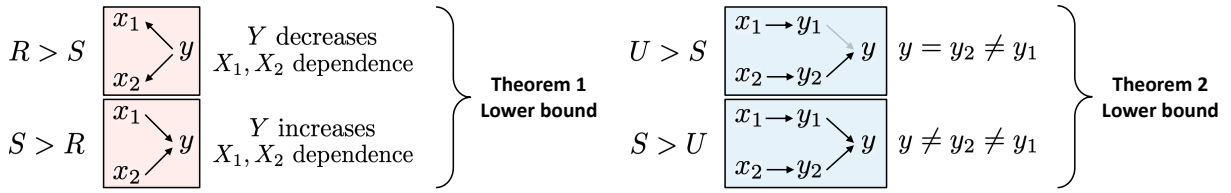


Figure 6.1: We study the relationships between (left) *synergy and redundancy* as a result of the task Y either increasing or decreasing the shared information between X_1 and X_2 (i.e., common cause structures as opposed to redundancy in common effect), as well as (right) *synergy and uniqueness* due to the disagreement between unimodal predictors resulting in a new prediction $y \neq y_1 \neq y_2$ (rather than uniqueness where $y = y_2 \neq y_1$).

predictions. This type of joint distribution is indicative of real-world expressions of sarcasm from language and speech - the presence of sarcasm is typically detected due to a contradiction between what is expressed in language and speech, as we observe from the experiments on MUSTARD [83] where $S = 0.44$ and disagreement = 0.12 are both large.

6.3.2 Lower and upper bounds on synergy

Given these relationships between synergy and other interactions, we now derive bounds on S. We present two lower bounds $\underline{S}_{\text{R}}$ and $\underline{S}_{\text{U}}$, which are based on redundancy and uniqueness, as well as an upper bound $\overline{S}$. We also describe the computational complexity for computing each bound.

Remark on high dimensional, continuous modalities. Our theoretical results are concerned with *finite* spaces for features and labels. However, this may be restrictive when working with real-world datasets (e.g., images, video, text) which are often continuous and/or high-dimensional. In such situations, we preprocess by performing discretization of each modality via clustering to estimate $p(x_1, y)$, $p(x_2, y)$, $p(x_1, x_2)$, each with a small, finite support. These are subsequently used for the computation of $\underline{S}_{\text{R}}$, $\underline{S}_{\text{U}}$ and $\overline{S}$. Discretization is a common way to approximate information theoretic quantities like mutual information [130, 372] and for learning representations over high-dimensional modalities [453].

Lower bound using redundancy Our first lower bound uses the relationship between synergy, redundancy, and dependence in equation 6.3. In semi-supervised settings, we can compute R exactly from $p(x_1, y), p(x_2, y)$, as well as the shared information $I(X_1; X_2)$ from $p(x_1, x_2)$. However, $I_p(X_1; X_2|Y)$ cannot be computed without access to the full distribution p. In Theorem 4, we obtain a lower bound on $I_p(X_1; X_2|Y)$, resulting in a lower bound $\underline{S}_{\text{R}}$ for synergy.

Theorem 4. *(Lower-bound on synergy via redundancy) We relate S to modality dependence*

$$\underline{S}_R = R - I_p(X_1; X_2) + \min_{r \in \Delta_{p_{1,2,12}}} I_r(X_1; X_2|Y) \leq S \tag{6.4}$$

We include a proof in the full paper [376]. This bound compares S to R via the difference of their dependence $I_p(X_1; X_2)$ and their dependence given the task $I_p(X_1; X_2|Y)$. Since the full distribution p is not available to compute $I_p(X_1; X_2|Y)$, we prove a lower bound using conditional MI computed with respect to a set of auxiliary distributions $r \in \Delta_{p_{1,2,12}}$ that are close to p, as

measured by matching both unimodal marginals $r(x_i, y) = p(x_i, y)$ and modality marginals $r(x_1, x_2) = p(x_1, x_2)$. If conditioning on the task increases the dependence and $I_r(X_1; X_2|Y)$ is large relative to $I_p(X_1; X_2)$ then we obtain a larger value of $\underline{S}_{\text{R}}$, otherwise if conditioning on the task decreases the dependence and $I_r(X_1; X_2|Y)$ is small relative to $I_p(X_1; X_2)$ then we obtain a smaller value of $\underline{S}_{\text{R}}$.

Computational complexity. R and $\min_{r \in \Delta_{p_{1,2,12}}} I_r(X_1; X_2|Y)$ are convex optimization problems solvable in polynomial time with off-the-shelf solvers. $I_p(X_1; X_2)$ can be computed directly.

Lower bound using uniqueness Our second bound formalizes the relationship between disagreement, uniqueness, and synergy. The key insight is that while labeled multimodal data is unavailable, the output of unimodal classifiers may be compared against each other. Consider unimodal classifiers $f_i : \mathcal{X}_i \to \mathcal{Y}$ and multimodal classifiers $f_M : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathcal{Y}$. Define *modality disagreement* as:

Definition 6. *(Modality disagreement) Given X_1, X_2, and a target Y, as well as unimodal classifiers f_1 and f_2, we define modality disagreement as $\alpha(f_1, f_2) = \mathbb{E}_{p(x_1,x_2)}[d(f_1, f_2)]$ where $d : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^{\geq 0}$ is a distance function in label space scoring the disagreement of f_1 and f_2's predictions.*

Connecting *modality disagreement* and synergy via Theorem 5 yields a lower bound $\underline{S}_{\text{U}}$:

Theorem 5. *(Lower-bound on synergy via uniqueness, informal) We can relate synergy S and uniqueness U to modality disagreement $\alpha(f_1, f_2)$ of optimal unimodal classifiers f_1, f_2 as follows:*

$$\underline{S}_U = \alpha(f_1, f_2) \cdot c - \max(U_1, U_2) \leq S \tag{6.5}$$

for some constant c depending on the label dimension $|\mathcal{Y}|$ and choice of label distance function d.

Theorem 5 implies that if there is substantial disagreement $\alpha(f_1, f_2)$ between unimodal classifiers, it must be due to the presence of unique or synergistic information. If uniqueness is small, then disagreement must be accounted for by synergy, thereby yielding a lower bound $\underline{S}_{\text{U}}$. Note that the optimality of unimodal classifiers is important: poorly trained unimodal classifiers could show high disagreement but would be uninformative about true interactions. We include the formal version of the theorem based on Bayes' optimality and a proof in the full paper [376].

Computational complexity. Lower bound $\underline{S}_{\text{U}}$ can also be computed efficiently by estimating $p(y|x_1)$ and $p(y|x_2)$ over modality clusters or training unimodal classifiers $f_\theta(y|x_1)$ and $f_\theta(y|x_2)$. U_1 and U_2 can be computed using a convex solver in polynomial time.

Hence, the relationships between S, R, and U yield two lower bounds $\underline{S}_{\text{R}}$ and $\underline{S}_{\text{U}}$. Note that these bounds *always* hold, so we could take $\underline{S} = \max\{\underline{S}_{\text{R}}, \underline{S}_{\text{U}}\}$.

Upper bound on synergy By definition, $S = I_p(\{X_1, X_2\}; Y) - R - U_1 - U_2$. However, $I_p(\{X_1, X_2\}; Y)$ cannot be computed exactly without the full distribution p. Using the same idea as lower bound 1, we upper bound synergy by *considering the worst-case maximum* $I_r(\{X_1, X_2\}; Y)$ computed over a set of auxiliary distributions $r \in \Delta_{p_{1,2,12}}$ that match both

unimodal marginals $r(x_i, y) = p(x_i, y)$ and modality marginals $r(x_1, x_2) = p(x_1, x_2)$:

$$\max_{r \in \Delta_{p_{1,2,12}}} I_r(\{X_1, X_2\}; Y) = \max_{r \in \Delta_{p_{1,2,12}}} \{H_r(X_1, X_2) + H_r(Y) - H_r(X_1, X_2, Y)\} \tag{6.6}$$

$$= H_p(X_1, X_2) + H_p(Y) - \min_{r \in \Delta_{p_{1,2,12}}} H_r(X_1, X_2, Y), \tag{6.7}$$

where the second line follows from the definition of $\Delta_{p_{1,2,12}}$. While the first two terms are easy to compute, the third may be difficult, as shown in the following theorem:

Theorem 6. *Solving* $r^* = \arg\min_{r \in \Delta_{p_{1,2,12}}} H_r(X_1, X_2, Y)$ *is NP-hard, even for a fixed* $|\mathcal{Y}| \geq 4$.

Theorem 6 suggests we cannot tractably find a joint distribution which tightly upper bounds synergy when the feature spaces are large. Fortunately, a relaxation of $r \in \Delta_{p_{1,2,12}}$ to $r \in \Delta_{p_{12,y}}$, where $r(x_1, x_2) = p(x_1, x_2)$ and $r(y) = p(y)$, recovers the classic *min-entropy coupling* problem over (X_1, X_2) and Y, which is still NP-hard but admits good approximations [118, 119, 123, 309]. Our final upper bound $\overline{S}$ is:

Theorem 7. *(Upper-bound on synergy)*

$$S \leq H_p(X_1, X_2) + H_p(Y) - \min_{r \in \Delta_{p_{12,y}}} H_r(X_1, X_2, Y) - R - U_1 - U_2 = \overline{S} \tag{6.8}$$

Proofs of Theorem 6, 7, and detailed approximation algorithms for min-entropy couplings are included in the full paper [376].

Computational complexity. The upper bound $\overline{S}$ can be computed efficiently since solving the variant of the min-entropy problem in Theorem 7 admits approximations that can be computed in time $O(k \log k)$ where $k = \max(|\mathcal{X}_1|, |\mathcal{X}_2|)$. All other entropy and R, U_1, U_2 terms are easy to compute (or have been computed via convex optimization from the lower bounds).

Practically, calculating all three bounds is extremely simple, with just a few lines of code. The computation takes < 1 minute and < 180 MB memory space on average for our large datasets ($1,000$-$20,000$ datapoints), more efficient than training even the smallest multimodal prediction model which takes at least 3x time and 15x memory. As a result, *these bounds scale to large and high-dimensional multimodal datasets found in the real world*, which we verify in the following experiments.

6.4 Experiments

We design comprehensive experiments to validate these estimated bounds and relationships between different multimodal interactions. Using these results, we describe applications in estimating optimal multimodal performance before training the model itself, which can be used to guide data collection and select appropriate multimodal models for various tasks.

6.4.1 Verifying interaction estimation in semi-supervised learning

Synthetic bitwise datasets Let $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{Y} = \{0, 1\}$. We generate joint distributions Δ by sampling $100,000$ vectors from the 8-dim probability simplex and assigning them to $p(x_1, x_2, y)$.

Large real-world multimodal datasets We use a collection of 10 real-world datasets from MultiBench [367] which add up to a size of more than $700,000$ datapoints.

1. MOSI: $2,199$ videos for sentiment analysis [710],
2. MOSEI: $23,000$ videos for sentiment and emotion analysis [717],
3. MUSTARD: 690 videos for sarcasm detection [83],
4. UR-FUNNY: a dataset of humor detection from $16,000$ TED talk videos [225],
5. MIMIC: $36,212$ examples predicting patient mortality and diseases from tabular patient data and medical sensors [286],
6. ENRICO: $1,460$ examples classifying mobile user interfaces and screenshots [340].
7. IRFL: $6,697$ images and figurative captions (e.g, 'the car is as fast as a cheetah' describing an image with a fast car in it) [700].
8. NYCaps: $1,820$ New York Yimes cartoon images and humorous captions describing these images [236].
9. VQA: $614,000$ questions and answers about natural images [29].
10. ScienceQA: $21,000$ questions and answers about science problems with scientific diagrams [392].

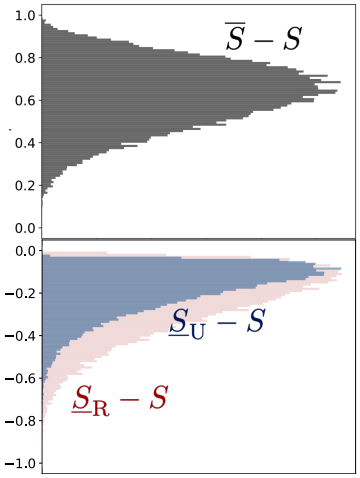


Figure 6.2: Our two lower bounds $\underline{S}_{\mathrm{R}}$ and $\underline{S}_{\mathrm{U}}$ track actual synergy S from below, and the upper bound $\overline{S}$ tracks S from above. We find that $\underline{S}_{\mathrm{R}}, \underline{S}_{\mathrm{U}}$ tend to approximate S better than $\overline{S}$.

These high-dimensional and continuous modalities require approximating disagreement and mutual information: we train unimodal classifiers $\hat{f}_\theta(y|x_1)$ and $\hat{f}_\theta(y|x_2)$ to estimate disagreement, and we cluster modality features to approximate continuous modalities by discrete distributions with finite support to compute the lower and upper bounds. We summarize the following regarding the validity of each bound:

1. Overall trends For the $100,000$ bitwise distributions, we compute S, the true value of synergy assuming oracle knowledge of the full multimodal distribution, and compute $\underline{S}_{\mathrm{R}} - S$, $\underline{S}_{\mathrm{U}} - S$, and $S - \overline{S}$ for each point. Plotting these points as a histogram in Figure 6.2, we find that the two lower bounds track synergy from below ($\underline{S}_{\mathrm{R}} - S$ and $\underline{S}_{\mathrm{U}} - S$ approaching 0 from below), and the upper bound tracks synergy from above ($S - \overline{S}$ approaching 0 from above). The two lower bounds are quite tight, as we see that for many points $\underline{S}_{\mathrm{R}} - S$ and $\underline{S}_{\mathrm{U}} - S$ are approaching close to 0, with an average gap of 0.18. $\underline{S}_{\mathrm{U}}$ seems to be tighter empirically than $\underline{S}_{\mathrm{R}}$: for half the points, $\underline{S}_{\mathrm{U}}$ is within 0.14 and $\underline{S}_{\mathrm{R}}$ is within 0.2 of S. For the upper bound, there is an average gap of 0.62. However, it performs especially well on high synergy data: when $S > 0.6$, the average gap is 0.24, with more than half of the points within 0.25 of S.

On real-world MultiBench datasets, we show the estimated bounds and actual S computed assuming knowledge of full p in Table 6.1. The lower and upper bounds track true S: as estimated $\underline{S}_{\mathrm{R}}$ and $\underline{S}_{\mathrm{U}}$ increases from MOSEI to UR-FUNNY to MOSI to MUSTARD, true S also increases. For datasets like MIMIC with disagreement but high uniqueness, $\underline{S}_{\mathrm{U}}$ can be negative, but we can rely on $\underline{S}_{\mathrm{R}}$ to give a tight estimate on low synergy. Unfortunately, our bounds do not track synergy well on ENRICO. We believe this is because ENRICO displays all interactions:

Table 6.1: We compute lower bounds $\underline{S}_{\mathrm{R}}$, $\underline{S}_{\mathrm{U}}$, and upper bound $\overline{S}$ in semi-supervised multimodal settings and compare them to S assuming knowledge of the full joint distribution p. The bounds always hold and track S well on MOSEI, UR-FUNNY, MOSI, and MUSTARD: true S increases as estimated $\underline{S}_{\mathrm{R}}$ and $\underline{S}_{\mathrm{U}}$ increases.

	MOSEI	UR-FUNNY	MOSI	MUSTARD	MIMIC	ENRICO	NYCAPS	IRFL	VQA	SCIENCEQA
$\overline{S}$	0.97	0.97	0.92	0.79	0.41	2.09	0.68	0.01	0.97	1.67
S	0.03	0.18	0.24	0.44	0.02	1.02	0.09	0	0.05	0.16
$\underline{S}_{\mathrm{R}}$	0	0	0.01	0.04	0	0.01	0	0	0	0.01
$\underline{S}_{\mathrm{U}}$	0.01	0.01	0.03	0.11	−0.12	−0.55	−0.03	−0.01	0	0

Table 6.2: Four representative examples: (a) disagreement XOR has high disagreement and high synergy, (b) agreement XOR has no disagreement and high synergy, (c) $y = x_1$ has high disagreement and uniqueness but no synergy, and (d) $y = x_1 = x_2$ has high agreement and redundancy but no synergy.

x_1	x_2	y	p
0	0	0	0
0	0	1	0.05
0	1	0	0.03
0	1	1	0.28
1	0	0	0.53
1	0	1	0.03
1	1	0	0.01
1	1	1	0.06

(a) Disagreement XOR

x_1	x_2	y	p
0	0	0	0.25
0	1	1	0.25
1	0	1	0.25
1	1	0	0.25

(b) Agreement XOR

x_1	x_2	y	p
0	0	0	0.25
0	1	0	0.25
1	0	1	0.25
1	1	1	0.25

(c) $y = x_1$

x_1	x_2	y	p
0	0	0	0.5
1	1	1	0.5

(d) $y = x_1 = x_2$

$R = 0.73, U_1 = 0.38, U_2 = 0.53, S = 0.34$, which makes it difficult to distinguish between R and S using $\underline{S}_{\mathrm{R}}$ or U and S using $\underline{S}_{\mathrm{U}}$ since no interaction dominates over others, and $\overline{S}$ is also quite loose. Given these general observations, we now carefully analyze the relationships between redundancy, uniqueness, and synergy.

2. Guidelines We provide a guideline to decide whether a lower or upper bound on synergy can be considered 'close enough'. It is close enough if the maximum interaction can be consistently estimated - often the exact value of synergy is not the most important (e.g, whether S is 0.5 or 0.6) but rather synergy relative to other interactions (e.g., if we estimate $S \in [0.2, 0.5]$, and exactly compute $R = U_1 = U_2 = 0.1$, then we know for sure that S is the most important interaction and can collect data or design models based on that). We find that our bounds accurately identify the same highest interaction on all 10 real-world datasets as the true synergy does. Furthermore, we observed that the estimated synergy correlates very well with true synergy: as high as 1.05 on ENRICO (true $S = 1.02$) and as low as 0.21 on MIMIC (true $S = 0.02$).

3. The relationship between S and R In Table 6.2b we show the classic AGREEMENT XOR distribution where X_1 and X_2 are independent, but $Y = 1$ sets $X_1 \neq X_2$ to increase their dependence. $I(X_1; X_2; Y)$ is negative, and $\underline{S}_{\mathrm{R}} = 1 \leq 1 = S$ is tight. On the other hand, Table 6.2d is an extreme example where the probability mass is distributed uniformly only when $y = x_1 = x_2$ and 0 elsewhere. As a result, X_1 is always equal to X_2 (perfect dependence), and yet Y perfectly

Table 6.3: Estimated lower, upper, and average bounds on optimal multimodal performance in comparison with the actual best unimodal model, the best simple fusion model, and the best complex fusion model. Our performance estimates closely predict actual model performance, *despite being computed only on semi-supervised data and never training the model itself.*

	MOSEI	UR-FUNNY	MOSI	MUSTARD	MIMIC	ENRICO
Estimated upper bound	1.07	1.21	1.29	1.63	1.27	0.88
Best complex multimodal	0.88	0.77	0.86	0.79	0.92	0.51
Best simple multimodal	0.85	0.76	0.84	0.74	0.92	0.49
Best unimodal	0.82	0.74	0.83	0.74	0.92	0.47
Estimated lower bound	0.52	0.58	0.62	0.78	0.76	0.48
Estimated average	0.80	0.90	0.96	1.21	1.02	0.68

explains away the dependence between X_1 and X_2 so $I(X_1; X_2|Y) = 0$: $\underline{S}_{\text{R}} = 0 \leq 0 = S$. A real-world example is multimodal sentiment analysis from text, video, and audio on MOSEI, $R = 0.26$ and $S = 0.03$, and as expected the lower bound is small $\underline{S}_{\text{R}} = 0 \leq 0.03 = S$ (Table 6.1).

4. The relationship between S and U In Table 6.2a we show an example called DISAGREEMENT XOR. There is maximum disagreement between $p(y|x_1)$ and $p(y|x_2)$: the likelihood for y is high when y is the opposite bit as x_1, but reversed for x_2. Given both x_1 and x_2: y takes a 'disagreement' XOR of the individual marginals, i.e. $p(y|x_1, x_2) = \arg\max_y p(y|x_1)$ XOR $\arg\max_y p(y|x_2)$, which indicates synergy (note that an exact XOR would imply perfect agreement and high synergy). The actual disagreement is 0.15, S is 0.16, and U is 0.02, indicating a very strong lower bound $\underline{S}_{\text{U}} = 0.14 \leq 0.16 = S$. A real-world equivalent dataset is MUSTARD, where the presence of sarcasm is often due to a contradiction between what is expressed in language and speech, so disagreement $\alpha = 0.12$ is the highest out of all the video datasets, giving a lower bound $\underline{S}_{\text{U}} = 0.11 \leq 0.44 = S$.

The lower bound is low when all disagreement is explained by uniqueness (e.g., $y = x_1$, Table 6.2c), which results in $\underline{S}_{\text{U}} = 0 \leq 0 = S$ (α and U cancel each other out). A real-world equivalent is MIMIC: from Table 6.1, disagreement is high $\alpha = 0.13$ due to unique information $U_1 = 0.25$, so the lower bound informs us about the lack of synergy $\underline{S}_{\text{U}} = -0.12 \leq 0.02 = S$. Finally, the lower bound is loose when there is synergy without disagreement, such as AGREEMENT XOR ($y = x_1$ XOR x_2, Table 6.2b) where the marginals $p(y|x_i)$ are both uniform, but there is full synergy: $\underline{S}_{\text{U}} = 0 \leq 1 = S$. Real-world datasets include UR-FUNNY where there is low disagreement in predicting humor $\alpha = 0.03$, and relatively high synergy $S = 0.18$, which results in a loose lower bound $\underline{S}_{\text{U}} = 0.01 \leq 0.18 = S$.

5. On upper bounds for synergy The upper bound for MUSTARD is close to real synergy, $\overline{S} = 0.79 \geq 0.44 = S$. On MIMIC, the upper bound is the lowest $\overline{S} = 0.41$, matching the lowest $S = 0.02$. Some of the other examples in Table 6.1 show weaker bounds. This could be because (i) there exists high synergy distributions that match $\mathcal{D}_i$ and $\mathcal{D}_M$, but these are rare in the real world, or (ii) our approximation used in Theorem 7 is loose. We leave these as directions for future work.

Additional results In the full paper [376], we also study the effect of imperfect unimodal predictors and disagreement measurements on our derived bounds, by perturbing the label by

various noise levels (from no noise to very noisy) and examining the changes in estimated upper and lower bounds. We found these bounds are quite robust to label noise, still giving close trends of S. We also include more discussions studying the relationships between various interactions, and how the relationship between disagreement and synergy can inspire new self-supervised learning methods.

6.4.2 Implications towards performance, data collection, model selection

Now that we have validated the accuracy of these bounds, we apply them to estimate multimodal performance in semi-supervised settings. This serves as a strong signal for deciding (1) whether to collect paired and labeled data from a second modality, and (2) what type of multimodal fusion method should be used. To estimate performance given $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_M$, we first compute our lower and upper bounds $\underline{S}$ and $\overline{S}$. Combined with the exact computation of R and U, we obtain the total information $I_p(\{X_1, X_2\}; Y)$, and combine a result from Feder and Merhav [172] with Fano's inequality [170] to yield tight bounds of performance as a function of total information.

Theorem 8. *Let $P_{acc}(f_M^*) = \mathbb{E}_p[\mathbf{1}[f_M^*(x_1, x_2) = y]]$ denote the accuracy of the Bayes' optimal multimodal model f_M^* (i.e., $P_{acc}(f_M^*) \geq P_{acc}(f_M')$ for all $f_M' \in \mathcal{F}_M$). We have that*

$$2^{I_p(\{X_1,X_2\};Y)-H(Y)} \leq P_{acc}(f_M^*) \leq \frac{I_p(\{X_1, X_2\}; Y) + 1}{\log |\mathcal{Y}|}, \tag{6.9}$$

and we can plug in $R + U_1, U_2 + \underline{S} \leq I_p(\{X_1, X_2\}; Y) \leq R + U_1, U_2 + \overline{S}$ to obtain lower $\underline{P}_{acc}(f_M^*)$ and upper $\overline{P}_{acc}(f_M^*)$ bounds on optimal multimodal performance.

We show the proof in the full paper [376]. Finally, we summarize estimated multimodal performance as the average $\hat{P}_M = (\underline{P}_{\text{acc}}(f_M^*) + \overline{P}_{\text{acc}}(f_M^*))/2$. A high $\hat{P}_M$ suggests the presence of important joint information from both modalities (not present in each) which could boost accuracy, so it is worthwhile to collect the full distribution p and explore multimodal fusion.

Setup For each MultiBench dataset, we implement a suite of unimodal and multimodel models spanning simple and complex fusion. Unimodal models are trained and evaluated separately on each modality. Simple fusion includes ensembling by taking an additive or majority vote between unimodal models [228]. Complex fusion is designed to learn higher-order interactions as exemplified by bilinear pooling [182], multiplicative interactions [283], tensor fusion [712], and cross-modal self-attention [613]. See our full paper [376] for models and training details. We include unimodal, simple and complex multimodal performance, as well as estimated lower and upper bounds on performance in Table 6.3.

RQ1: Estimating multimodal fusion performance *How well could my multimodal model perform?* We find that estimating interactions enables us to *closely predict multimodal model performance, before even training a model*. For example, on MOSEI, we estimate the performance to be 52% based on the lower bound and 107% based on the upper bound, for an average of 80% which is very close to true model performance ranging from 82% for the best unimodal model, and $85\% - 88\%$ for various multimodal model. Estimated performances for ENRICO, UR-FUNNY, and MOSI are 68%, 90%, 96%, which track true performances 51%, 77%, 86%.

RQ2: Data collection *Should I collect multimodal data?* We compare estimated performance $\hat{P}_M$ with the actual difference between unimodal and best multimodal performance in Figure 6.3 (left). Higher estimated $\hat{P}_M$ correlates with a larger gain from unimodal to multimodal (correlation $\rho = 0.21$ and rises to 0.53 if ignoring the outlier in MIMIC). MUSTARD and ENRICO show the most opportunity for multimodal modeling. Therefore, a rough guideline is that if the estimated multimodal performance based on semi-supervised data is higher, then collecting the full labeled multimodal data is worth it.

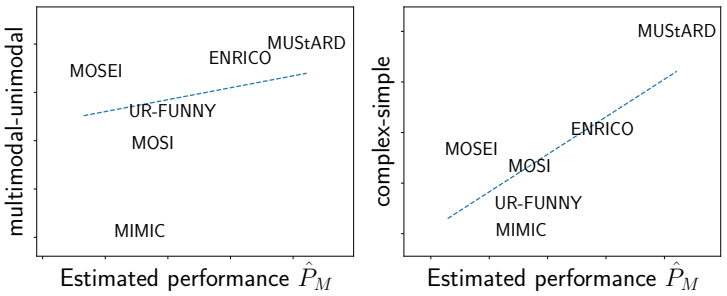


Figure 6.3: Datasets with higher estimated multimodal performance $\hat{P}_M$ tend to show improvements from unimodal to multimodal (left) and from simple to complex multimodal fusion (right).

RQ3: Model selection *What model should I choose for multimodal fusion?* We note strong relationships between estimated performance and the performance of different fusion methods. From Table 6.3, synergistic datasets like MUSTARD and ENRICO show best multimodal performance only slightly above our estimated lower bound, indicating that there is a lot of room for improvement in better fusion methods. Indeed, more complex fusion methods such as multimodal transformer designed to capture synergy is the best on MUSTARD which matches its high synergy (72% accuracy). For datasets with less synergy like MOSEI and MIMIC, the best multimodal performance is much higher than the estimated lower bound, indicating that existing fusion methods may already be quite optimal. Indeed, simpler fusion methods such as feature alignment, designed to capture redudnancy, are the best on MOSEI which matches its high redundancy (80% accuracy).

Figure 6.3 (right) shows a visual comparison, where plotting the performance gap between complex and simple fusion methods against estimated performance $\hat{P}_M$ shows a correlation coefficient of 0.77. We again observe positive trends between higher estimated performance and improvements with complex fusion, with large gains on MUSTARD and ENRICO. We expect new methods to further improve the state-of-the-art on these datasets due to their generally high interaction values and low multimodal performance relative to estimated lower bound $\underline{P}_{\text{acc}}(f_M^*)$. Therefore, a rough guideline is that if the estimated multimodal performance based on semi-supervised data is higher, then there is more potential for improvement by trying more complex multimodal fusion strategies.

6.5 Conclusion and Broader Impacts

We proposed estimators of multimodal interactions when observing only *labeled unimodal data* and some *unlabeled multimodal data*, a general semi-supervised setting that encompasses many real-world constraints involving partially observable modalities, limited labels, and privacy concerns. Our key results draw new connections between multimodal interactions, the disagreement

of unimodal classifiers, and min-entropy couplings, which yield new insights for estimating multimodal model performance, data analysis, and model selection. We are aware of some potential **limitations**:

1. These estimators only approximate real interactions due to cluster preprocessing or unimodal models, which naturally introduce optimization and generalization errors. We expect progress in density estimators, generative models, and unimodal classifiers to address these problems.
2. It is harder to quantify interactions for certain datasets, such as ENRICO which displays all interactions which makes it difficult to distinguish between R and S or U and S.
3. Finally, there exist challenges in quantifying interactions since the data generation process is never known for real-world datasets, so we have to resort to human judgment, other automatic measures, and downstream tasks such as estimating model performance and model selection.

Future work should investigate more applications of multivariate information theory in designing self-supervised models, predicting multimodal performance, and other tasks involving feature interactions such as privacy-preserving and fair representation learning from high-dimensional data [161, 219]. Being able to provide guarantees for fairness and privacy-preserving learning, especially for semi-supervised pretraining datasets, can be particularly impactful.

Chapter 7

MultiBench: Large-scale Resources for Multisensory Learning

7.1 Introduction

Current multimodal research has led to impressive advances in benchmarking and modeling for specific domains such as language and vision [11, 360, 381, 502]. However, other domains, modalities, and tasks are relatively understudied. The future will lie in multisensory foundation models that are *grounded in the world*: being able to simultaneously process a large number of modalities beyond language, to vision, audio [11, 360, 381, 502], and leveraging advances in sensing technologies such as cellphones [366], wearable devices [218], autonomous vehicles [697], healthcare technologies [287], and robots [53, 304] that give a wealth of sensor data about the world.

MULTIBENCH: In order to accelerate research in building general-purpose multimodal foundation models, this chapter describes MULTIBENCH (Figure 7.1), a systematic and unified large-scale benchmark that brings us closer to the requirements of real-world multimodal applications. MULTIBENCH is designed to comprehensively evaluate generalization across domains and modalities. To that end, MULTIBENCH contains a diverse set of 28 datasets spanning 14 modalities and testing for more than 30 prediction tasks across 6 distinct research areas and 5 technical challenges of multimodal machine learning. These research areas include important tasks understudied from a multimodal learning perspective, such as healthcare, finance, and HCI. Building upon extensive data-collection efforts by domain experts, we worked with them to adapt datasets that reflect real-world relevance, present unique challenges to multimodal learning, and enable opportunities in algorithm design and evaluation.

Together, MULTIBENCH unifies efforts across separate research areas in multimodal learning to enable quick and accurate benchmarking across a wide range of datasets and metrics. To help the community accurately compare performance and ensure reproducibility, MULTIBENCH includes an end-to-end pipeline including data preprocessing, dataset splits, multimodal algorithms, evaluation metrics, and cross-validation protocols. This includes an implementation of 20 core multimodal approaches spanning innovations in fusion paradigms, optimization objectives, and training approaches in a standard public toolkit called MULTIZOO. We perform a systematic eval-

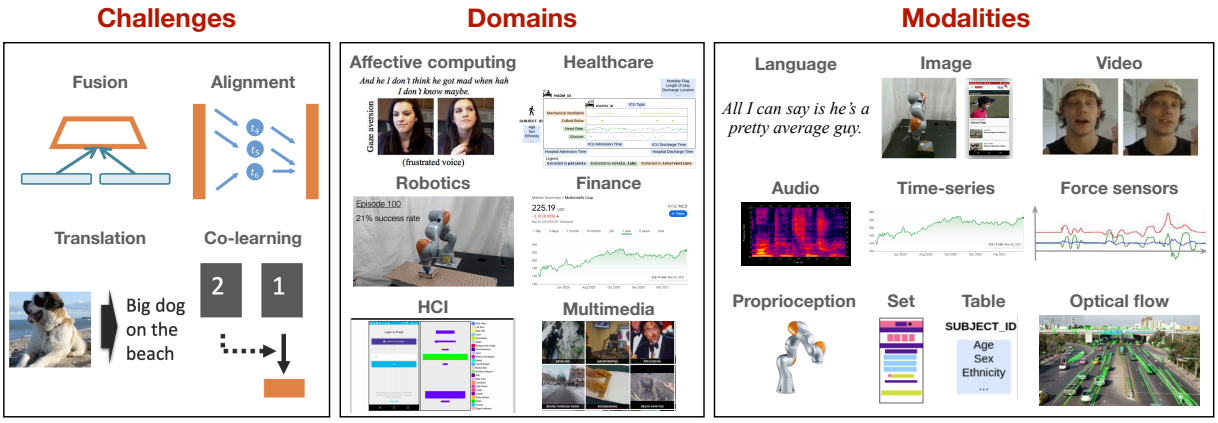


Figure 7.1: MULTIBENCH contains a diverse set of 28 datasets spanning 14 modalities and testing for more than 30 prediction tasks across 6 distinct research areas and 5 technical challenges of multimodal machine learning, thereby enabling standardized, reliable, and reproducible large-scale benchmarking of multimodal models. To reflect real-world requirements, MULTIBENCH is designed to holistically evaluate generalization performance across domains and modalities.

uation and show that directly applying these methods can improve the state-of-the-art performance on 9 out of the 15 datasets. Therefore, MULTIBENCH presents a step towards unifying disjoint efforts in multimodal research and paves a way towards a deeper understanding of multimodal models. Most importantly, our public zoo of multimodal benchmarks and models will ensure ease of use, accessibility, and reproducibility. Finally, we outline our plans to ensure the continual availability, maintenance, and expansion of MULTIBENCH, including using it as a theme for future workshops and competitions and to support the multimodal learning courses taught around the world.

7.2 MULTIBENCH: The Multiscale Multimodal Benchmark

Background: We define a modality as a single particular mode in which a signal is expressed or experienced. Multiple modalities then refer to a combination of multiple heterogeneous signals [46]. The first version of MULTIBENCH focuses on benchmarking algorithms for *multimodal fusion*, where the main challenge is to join information from two or more modalities to perform a prediction (e.g., classification, regression). Classic examples for multimodal fusion include audio-visual speech recognition where visual lip motion is fused with speech signals to predict spoken words [160]. Multimodal fusion can be contrasted with multimodal translation where the goal is to generate a new and different modality [640], grounding and question answering where one modality is used to query information in another (e.g., visual question answering [11]), and unsupervised or self-supervised multimodal representation learning [390, 571]. We plan future versions of MULTIBENCH to study these important topics in multimodal research.

Each of the following 15 datasets in MULTIBENCH contributes a unique perspective to the various technical challenges in multimodal learning involving learning and aligning complemen-

Table 7.1: MULTIBENCH provides a comprehensive suite of 28 multimodal datasets to benchmark current and proposed approaches in multimodal machine learning. It covers a diverse range of technical challenges, research areas, dataset sizes, input modalities (in the form of a: audio, e: embodied environment, f: force sensor, g: graph, i: image ℓ: language, o: optical flow, p: proprioception sensor, π: policy/action, q: question (for question-answering tasks), s: set, t: time-series, ta: tabular, v: video), and prediction tasks. We provide a standardized data loader for datasets in MULTIBENCH, along with a set of state-of-the-art multimodal models.

Challenge	Research Area	Size	Dataset	Modalities	# Samples	Prediction task
Fusion	Affect	S	MUSTARD [83]	$\{\ell, v, a\} \to y$	690	sarcasm
		M	CMU-MOSI [710]	$\{\ell, v, a\} \to y$	2,199	sentiment
		L	UR-FUNNY [225]	$\{\ell, v, a\} \to y$	16,514	humor
		L	CMU-MOSEI [717]	$\{\ell, v, a\} \to y$	22,777	sentiment, emotions
	Healthcare	L	MIMIC [287]	$\{t, ta\} \to y$	36,212	mortality, ICD-9 codes
		L	MIMIC-CXR [288]	$\{\ell, i\} \to y$	377,110	mortality, ICD-9 codes
	Robotics	M	MUJOCO PUSH [334]	$\{i, f, p\} \to y$	37,990	object pose
		L	VISION&TOUCH [335]	$\{i, f, p\} \to y$	147,000	contact, robot pose
	Finance	M	STOCKS-F&B	$\{t \times 18\} \to y$	5,218	stock price, volatility
		M	STOCKS-HEALTH	$\{t \times 63\} \to y$	5,218	stock price, volatility
		M	STOCKS-TECH	$\{t \times 100\} \to y$	5,218	stock price, volatility
	HCI	S	ENRICO [340]	$\{i, s\} \to y$	1,460	design interface
	Multimedia	M	HATEFUL MEMES [301]	$\{\ell, i\} \to y$	10,000	hate speech
		M	MM-IMDB [32]	$\{\ell, i\} \to y$	25,959	movie genre
		M	AV-MNIST [638]	$\{i, a\} \to y$	70,000	digit
		L	KINETICS400 [296]	$\{v, a, o\} \to y$	306,245	human action
Question Answering	Affect	M	SOCIAL IQ [715]	$\{v, a, \ell, q\} \to y$	7,500	QA
	Multimedia	L	CLEVR [289]	$\{i, q\} \to y$	853,554	QA
		L	VQA 2.0 [206]	$\{i, q\} \to y$	1,100,000	QA
Retrieval	Multimedia	S	CIFAR-ESC [369]	$i \leftrightarrow a$	2,080	image-audio retrieval
		M	CLOTHO [156]	$a \leftrightarrow \ell$	24,905	audio-caption retrieval
		M	YUMMLY-28K [421]	$i \leftrightarrow \ell$	27,638	image-caption retrieval
		L	FLICKR-30K [485]	$i \leftrightarrow \ell$	158,000	image-caption retrieval
RL	Simulation	L	RTFM [733]	$\{e, \ell \to \pi\}$	-	multimodal RL
Co-learning	Affect	M	CMU-MOSI → SST [716]	$\{\ell, v, a\} \to \ell$	11,855	video → text
		L	CMU-MOSEI → SST [716]	$\{\ell, v, a\} \to \ell$	11,855	video → text
	Multimedia	M	GLOVE → CIFAR10 [553]	$\{i, \ell\} \to i$	60,000	text → image
		L	Visual Genome [317, 409]	$\{i, g\} \to i$	100,000	knowledge graph → image

Figure 7.2: MULTIBENCH provides a standardized machine learning pipeline across data processing, data loading, multimodal models, evaluation metrics, and a public leaderboard to encourage future research in multimodal representation learning. MULTIBENCH aims to present a milestone in unifying disjoint efforts in multimodal machine learning research and paves the way towards a better understanding of the capabilities and limitations of multimodal models, all the while ensuring ease of use, accessibility, and reproducibility.

tary information, scalability to a large number of modalities, and robustness to realistic real-world imperfections.

MULTIBENCH provides a standardized machine learning pipeline that starts from data loading to running multimodal models, providing evaluation metrics, and a public leaderboard to encourage future research in multimodal representation learning (see Figure 7.2). Table 7.1 shows an overview of these datasets. We provide a brief overview of the research areas, modalities, and tasks for each of these datasets.

7.2.1 Research areas

Affective computing studies the perception of human affective states (emotions, sentiment, and personalities) from our display of multimodal signals spanning language (spoken words), visual (facial expressions, gestures), and acoustic (prosody, speech tone) [483]. It has impacts towards building emotionally intelligent computers, human behavior analysis, and AI-assisted education.

Healthcare: Modern medical decision-making often involves integrating complementary information and signals from several sources such as lab tests, imaging reports, and patient-doctor conversations. Multimodal models can help doctors make sense of high-dimensional data and assist them in the diagnosis process [21].

Robotics: Modern robot systems are equipped with multiple sensors to aid in their decision-making. Some systems also have a large number of heterogeneous sensors deployed in the real world with realistic noise and imperfections. These present scalability and robustness challenges for multimodal machine learning.

Finance: The field of machine learning for finance studies the use of algorithms to make better automatic trading decisions through historical data, news and document understanding, social media analytics, and other multimodal signals. This field presents challenges in time-series analysis on high-frequency multimodal signals, a dynamic and large number of possible modalities, as well as robustness and compute efficiency for real-world deployment.

Human Computer Interaction (HCI) studies the design of computer technology and interactive interfaces between humans and computers [151]. Many real-world human-centric problems involve multimodal inputs such as language, visual, and audio interfaces. Designing multimodal models that actively interact with humans further necessitates guarantees on their fairness and robustness in real-world scenarios.

Multimedia: A significant body of research in multimodal learning has been fueled by the large availability of multimedia data (language, image, video, and audio) on the internet. Multimedia research is exemplified by the research tasks of media description, multimodal question answering, and cross-modal retrieval.

Simulated environments: Finally, simulated interactive environments such as Atari games [52], Minecraft [216], and NetHack [323] present scalable opportunities for research in reinforcement learning while also enabling rich programming of multimodal environments involving text [394], audio [135], and video [88]. By way of their flexible design, these environments can often provide richer interactions between text and embodied environments, more difficult planning and exploration challenges, and procedurally generated tasks of increasing difficulty.

7.2.2 Fusion datasets

In multimodal fusion, the main challenge is to join information from two or more modalities to perform a prediction. Classic examples include audio-visual speech recognition where visual lip motion is fused with speech signals to predict spoken words [160]. Information coming from different modalities have varying predictive power by themselves and also when complemented by each other (i.e., higher-order interactions). In order to capture higher-order interactions, there is also a need to identify the relations between granular units from two or more different modalities (i.e., alignment). When dealing with temporal data, it also requires capturing possible long-range dependencies across time (i.e., temporal alignment). MULTIBENCH contains the following datasets for multimodal fusion spanning several research areas:

Affective computing: MULTIBENCH contains 4 datasets involving fusing *language*, *video*, and *audio* time-series data to predict sentiment (CMU-MOSI [710]), emotions (CMU-MOSEI [717]), humor (UR-FUNNY [225]), and sarcasm (MUSTARD [83]). Complementary information may occurs at different moments, requiring models to address the multimodal challenges of grounding and alignment.

Healthcare: MULTIBENCH includes the large-scale MIMIC dataset [287] which records ICU patient data including *time-series* data measured every hour and other demographic variables (e.g., age, gender, ethnicity in the form of *tabular numerical* data). These are used to predict the disease ICD-9 code and mortality rate. MIMIC poses unique challenges in integrating time-varying and static modalities, reinforcing the need of aligning multimodal information at correct granularities. Extending MIMIC, we also include the MIMIC-CXR [288] datasets of de-identified publicly available chest radiographs and free-text reports

Robotics: We include MUJOCO PUSH [334] and VISION&TOUCH [335] which record the manipulation of simulated and real robotic arms equipped with *visual* (RGB and depth), *force*, and *proprioception* sensors. In MUJOCO PUSH, the goal is to predict the pose of the object being pushed by the robot end-effector. In VISION&TOUCH, the goal is to predict action-conditional learning objectives that capture forward dynamics of contact prediction and robot end-effector pose. Robustness is important due to the risk of real-world sensor failures [336].

Finance: We gathered historical stock data from the internet to create our own dataset for financial time-series prediction across 3 groups of correlated stocks: STOCKS-F&B, STOCKS-HEALTH, and STOCKS-TECH. Within each group, the previous stock prices of a set of stocks are used as multimodal *time-series* inputs to predict the price and volatility of a related stock (e.g.,

using Apple, Google, and Microsoft data to predict future Microsoft prices). Multimodal stock prediction [520] presents scalability issues due to a large number of modalities ($18/63/100$ vs $2/3$ in most datasets), as well as robustness challenges arising from real-world data with an inherently low signal-to-noise ratio.

HCI: We use the ENRICO (Enhanced Rico) dataset [137, 340] of Android app screens (consisting of an *image* as well as a *set* of apps and their locations) categorized by their design motifs and collected for data-driven design applications such as design search, user interface (UI) layout generation, UI code generation, and user interaction modeling.

Multimedia: MULTIBENCH includes 4 popular large-scale multimedia datasets with varying sizes and levels of difficulty: (1) the hateful memes challenge [301] as a core challenge in multimedia to ensure safer learning from ubiquitous text and images from the internet, (2) AV-MNIST [638] is assembled from *images* of handwritten digits [332] and *audio* samples of spoken digits [341], (3) MM-IMDB [32] uses movie *titles*, *metadata*, and movie *posters* to perform multi-label classification of movie genres, and (4) KINETICS [296] contains *video*, *audio*, and *optical flow* of $306,245$ video clips annotated for 400 human actions.

7.2.3 Question answering datasets

Within the domain of language and vision, there has been growing interest in language-based question answering (i.e., "query" modality) of entities in the visual, video, or embodied domain (i.e., "queried" modality). Datasets such as Visual Question Answering [11], Social IQ [715], and Embodied Question Answering [131] have been proposed to benchmark the performance of multimodal models in these settings. A core challenge lies in aligning words asked in the question with entities in the queried modalities, which typically take the form of visual entities in images or videos (i.e., alignment). MULTIBENCH contains the following datasets for multimodal question answering spanning several research areas:

Affective computing: SOCIAL IQ [715] is an unconstrained benchmark specifically designed to train and evaluate socially intelligent AI through a rich source of open-ended questions and answers. It contains $1,250$ videos of natural social situations, $7,500$ questions and $52,500$ correct and incorrect answers

Multimedia: CLEVR [289] is a diagnostic dataset for studying the ability of VQA systems to perform visual reasoning. It contains $100,000$ rendered images and about $853,000$ unique automatically generated questions that test visual reasoning abilities such as counting, comparing, logical reasoning, and storing information in memory. VQA 2.0 [206] is a balanced version of the popular VQA [11] dataset by collecting complementary images such that every question is associated with not just a single image, but rather a pair of similar images that result in two different answers to the question. The reduces the occurrence of spurious correlations in the dataset and enables training of more robust models.

7.2.4 Retrieval datasets

Another area of great interest lies in cross-modal retrieval [369, 732], where the goal is to retrieve semantically similar data from a new modality using a modality as a query (e.g., given a phrase, retrieve the closest image describing that phrase). The core challenge is to perform alignment

of representations across both modalities. MULTIBENCH contains the following datasets for multimodal retrieval and grounding:

Multimedia: CIFAR-ESC [369] is an image-audio retrieval dataset constructed by combining CIFAR-100, CIFAR-10 [319], and ESC-50 [484] into 17 shared classes using concept ontologies from WordNet [420]. CLOTHO [156] is a dataset for audio captioning with 4981 audio samples of 15 to 30 seconds duration and 24,905 captions of 8 to 20 words length. YUMMLY-28K [421] contains parallel text descriptions and images of recipes with 27,638 recipes in total. Each recipe contains one recipe image, the ingredients, the cuisine and the course information. FLICKR-30K [485] contains 32,000 images collected from Flickr, together with 5 reference sentences provided by human annotators enabling the tasks of text-to-image reference resolution, localizing textual entity mentions in an image, and bidirectional image-caption retrieval.

7.2.5 Reinforcement learning environments

Learning from multiple modalities in an interactive setting is an area of interest towards building more intelligent embodied agents that can perceive the visual world, language instructions, auditory feedback, and other sensor modalities [394]. Recent work has also explored audio as a modality in an agent's multisensory interaction with the world [135]. These multimodal problems are fundamentally different from those that are concerned with prediction tasks. Alongside the core challenges in learning complementary information and aligning entities in language instructions to those in the visual environment, there also lies the core challenge of learning *actionable* representations that link to the set of actions that can be taken and their associated long-term rewards [394]. MULTIBENCH contains the following datasets for multimodal reinforcement learning in both real-world and simulated environments:

Simulated environments: We choose the RTFM [733] (Reading to Fight Monsters) simulated text and visual environment. RTFM requires an agent to jointly reason over a language goal, a document that specifies environment dynamics, and environment observations. It can also be procedurally generated for increasing difficult interactions between environment dynamics and natural language. RTFM is also part of the larger SILG benchmark [734] of 5 similar diverse grounded language learning environments under a common interface, so it enables generalization to these other environments as well.

7.2.6 Co-learning datasets

Co-learning aims to transfer knowledge between modalities and their representations. Exemplified by algorithms of fine-tuning, co-training, and contrastive learning, how can knowledge learned from an additional secondary modality (e.g., predicted labels or representation) help a computational model trained on a primary modality? This challenge is particularly relevant when the primary modality has limited resources such as lack of annotated data, noisy input, and unreliable labels.

Affective computing: In affective computing, we investigate transferring information from CMU-MOSI to SST, as well as the larger CMU-MOSEI to SST [716]. The former 2 are multimodal (language + vision + audio) datasets annotated for sentiment, while SST is a language-only sentiment analysis dataset.

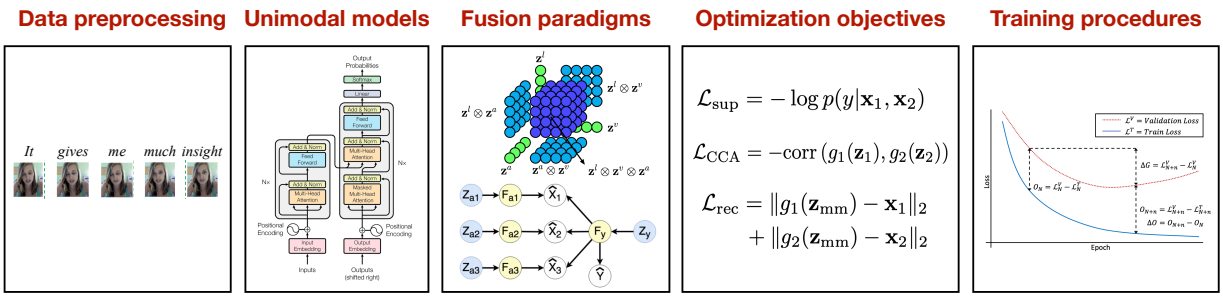


Figure 7.3: MULTIZOO provides a standardized implementation of a suite of multimodal methods in a modular fashion to enable accessibility for new researchers, compositionality of approaches, and reproducibility of results.

Multimedia: In multimedia, we transfer information from GLOVE word embeddings for CIFAR10 image classification [553]. We also transfer information from knowledge graphs to image classification by providing the Visual Genome dataset [317, 409].

7.3 Evaluation protocol

MULTIBENCH provides standardized evaluation using metrics designed for each dataset, ranging from MSE and MAE for regression to accuracy, micro & macro F1-score, and AUPRC for classification on each dataset. To assess for generalization, we compute the variance of a particular model's performance across all datasets in MULTIBENCH on which it is tested. We split these results on multiple datasets into *in-domain* datasets and *out-domain* datasets. *In-domain* datasets refer to model performance on datasets that it was initially proposed and tested on, while *out-domain* datasets refer to model performance on the remaining datasets. Comparing out-domain vs in-domain performance, as well as variance in performance across datasets as a whole, allow us to summarize the generalization statistics of each multimodal model.

7.4 MULTIZOO: A Zoo of Multimodal Algorithms

To complement MULTIBENCH, we release a comprehensive toolkit, MULTIZOO, as starter code for multimodal algorithms which implements 20 methods spanning different methodological innovations in (1) data preprocessing, (2) fusion paradigms, (3) optimization objectives, and (4) training procedures (see Figure 7.3). To introduce these algorithms, we use the simple setting with 2 modalities for notational convenience. We use $\mathbf{x}_1, \mathbf{x}_2$ for input modalities, $\mathbf{z}_1, \mathbf{z}_2$ for unimodal representations, $\mathbf{z}_{\text{mm}}$ for the multimodal representation, and $\hat{y}$ for the predicted label.

7.4.1 Data preprocessing

Temporal alignment [101] has been shown to help tackle the multimodal alignment problem for time-series data. This approach assumes a temporal granularity of the modalities (e.g., at the level

of words for text) and aligns information from the remaining modalities to the same granularity. We call this approach WORDALIGN [101] for temporal data where text is one of the modalities.

7.4.2 Fusion paradigms

Early and late fusion have been the de-facto first-approach when tackling new multimodal problems. Early fusion performs concatenation at the input data level before using a suitable prediction model (i.e., $\mathbf{z}_{\text{mm}} = [\mathbf{x}_1, \mathbf{x}_2]$) and late fusion applies suitable unimodal models to each modality to obtain their feature representations, concatenates these features, and defines a classifier to the label (i.e., $\mathbf{z}_{\text{mm}} = [\mathbf{z}_1, \mathbf{z}_2]$) [46]. MULTIZOO includes their implementations denoted as EF and LF respectively.

Tensors are specifically designed to tackle the multimodal complementarity challenge by explicitly capturing higher-order interactions across modalities [712]. Given unimodal representations $\mathbf{z}_1, \mathbf{z}_2$, a multimodal tensor representation is defined as $\mathbf{z}_{\text{mm}} = \begin{bmatrix} \mathbf{z}_1 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{z}_2 \\ 1 \end{bmatrix}$ where $\otimes$ denotes an outer product. However, computing tensor products is expensive since their dimension scales exponentially with the number of modalities. Several efficient variants have been proposed to approximate expensive full tensor products with cheaper variants while maintaining performance [245, 364, 388]. MULTIZOO includes Tensor Fusion (TF) [712] as well as approximate Low-rank Tensor Fusion (LRTF) [388]. As future work, we also plan to include more expressive higher-order tensor fusion methods [245].

Multiplicative Interactions (MI) further generalize tensor products to include learnable parameters that capture the interactions between streams of information [284]. In its most general form, MI defines a bilinear product $\mathbf{z}_{\text{mm}} = \mathbf{z}_1 \mathbb{W} \mathbf{z}_2 + \mathbf{z}_1^\intercal \mathbf{U} + \mathbf{V}\mathbf{z}_2 + \mathbf{b}$ where $\mathbb{W}, \mathbf{U}, \mathbf{Z}$, and $\mathbf{b}$ are trainable parameters. By appropriately constraining the rank and structure of these parameters, MI recovers HyperNetworks [217] (unconstrained parameters resulting in a matrix output), Feature-wise linear modulation (FiLM) [476, 733] (diagonal parameters resulting in vector output), and Sigmoid units [133] (scalar parameters resulting in scalar output). MULTIZOO includes all 3 as MI-MATRIX, MI-VECTOR, and MI-SCALAR respectively.

We also referred to the implementation of Feature-wise linear modulation (FiLM) [476] and added it as a module in MULTIBENCH, which we call FILM. While MI-VECTOR (i.e., diagonal parameters in a MI layer which results in a vector output) corresponds to the most basic implementation of FILM, the original FILM layer uses multiple non-linear layers instead of a single linear transformation in MI-VECTOR which has been shown to improve performance [476].

Multimodal gated units are prevalent in learning combinations of two representations that dynamically change for every input [88, 651, 680]. Its general form can be written as $\mathbf{z}_{\text{mm}} = \mathbf{z}_1 \odot h(\mathbf{z}_2)$, where h represents a function with sigmoid activation and $\odot$ denotes the element-wise product. The output $h(\mathbf{z}_2)$ is commonly referred to as "attention weights" learned from $\mathbf{z}_2$ used to attend on $\mathbf{z}_1$. We implement the Query-Key-Value mechanism as NL GATE as proposed in [651]. This attention mechanism is conceptually similar to the MI-VECTOR case above but recent work has explored more expressive forms of h such as using a Query-Key-Value mechanism [651] or several fully-connected layers [88] rather than a linear transformation in MI-VECTOR.

Multimodal transformers are useful in tackling the challenge of multimodal alignment and complementarity. Transformer models [631] have been shown to be useful for temporal

multimodal data by automatically aligning and capturing complementary features at different time-steps [613, 694]. We include the Multimodal Transformer (MULT) [613] which uses a Crossmodal Transformer block that uses $\mathbf{z}_1$ to attend to $\mathbf{z}_2$ (and vice-versa), before concatenating both representations to obtain $\mathbf{z}_{\text{mm}} = [\mathbf{z}_{1\to2}, \mathbf{z}_{2\to1}] = [\text{CM}(\mathbf{z}_1, \mathbf{z}_2), \text{CM}(\mathbf{z}_2, \mathbf{z}_1)]$.

To extend this to 3 modalities, the crossmodal transformer block is repeated across all 3 sets of modality pairs (i.e., $\mathbf{z}_{\text{mm}} = [\mathbf{z}_{1\to2}, \mathbf{z}_{2\to1}, \mathbf{z}_{1\to3}, \mathbf{z}_{3\to1}, \mathbf{z}_{2\to3}, \mathbf{z}_{3\to2}]$). While this is still computationally feasible for 3 modalities such as the language, video, and audio datasets that MULT was originally designed for, this quickly becomes intractable for problems involving more than 3 modalities. To adapt MULT for the financial prediction task involving more than 10 modalities, we cluster all modalities into 3 groups based on similarities in their data and perform early fusion on the data within each cluster before applying MULT only on the 3 clusters of modalities. While MULT is a strong model based on performance, it poses scalability issues that should be the subject of future work (i.e., since the number of cross-modal attention blocks grows quadratically with the number of modalities).

Architecture search: Finally, instead of hand-designing multimodal architectures, several approaches define a set of atomic neural operations (e.g., linear transformation, activation, attention, etc.) and use architecture search to automatically learn the best order of these operations for a given multimodal task [478, 682]. We focus on the more general approach, MFAS [478], designed for language and vision datasets. While this approach is categorized under innovations in model architecture (since it primarily targets better architectures for multimodal fusion), its code in the MULTIZOO toolkit is implemented under training structures, since architecture search requires an outer loop to learn model architectures over multiple inner supervised learning loops that train an individual model architecture. Therefore, we are unable to integrate MFAS directly with the basic supervised learning training loops like we do for the other fusion paradigms described above.

7.4.3 Optimization objectives

In addition to the standard supervised losses (e.g., cross entropy for classification, MSE/MAE for regression), several proposed methods have proposed new objective functions based on:

Prediction-level alignment: There has been extensive research in defining objective functions to tackle the challenge of multimodal alignment: capturing a representation space where semantically similar concepts from different modalities are close together. While primarily useful for cross-modal retrieval [369, 732], recent work has also shown its utility in learning representations for prediction [39, 126, 335, 604]. These alignment objectives have been applied at both prediction and feature levels. In the former, we implement Canonical Correlation Analysis (CCA) [27, 650], which computes $\mathcal{L}_{\text{CCA}} = \text{corr}\left(g_1(\mathbf{z}_1), g_2(\mathbf{z}_2)\right)$ where g_1, g_2 are auxiliary classifiers mapping each unimodal representation to the label. This method corresponds to prediction-level alignment since they aim to learn representations of each modality that agree on the label, as measured by the correlation of label predictions made by each modality across a batch of samples. We refer to the paper that most closely implements CCA-based alignment for multimodal data (specifically directly testing on the CMU-MOSI dataset) [578].

Feature-level alignment: In the latter, contrastive learning has emerged as a popular approach that brings similar concepts close in feature space and different concepts far away [126, 335, 604]. MULTIZOO includes REFNET [517] which includes a self-supervised contrastive loss

Algorithm 3 PyTorch code integrating MULTIBENCH datasets and MULTIZOO models.

```
from datasets.get_data import get_dataloader
from unimodals.common_models import ResNet, Transformer
from fusions.common_fusions import MultInteractions
from training_structures.gradient_blend import train, test

# loading Multimodal IMDB dataset
traindata, validdata, testdata = get_dataloader('multimodal_imdb')
out_channels = 3
# defining ResNet and Transformer unimodal encoders
encoders = [ResNet(in_channels=1, out_channels, layers=5),
            Transformer(in_channels=1, out_channels, layers=3)]
# defining a Multiplicative Interactions fusion layer
fusion = MultInteractions([out_channels*8, out_channels*32], out_channels*32, 'matrix')
classifier = MLP(out_channels*32, 100, labels=23)
# training using Gradient Blend algorithm
model = train(encoders, fusion, classifier, traindata, validdata,
        epochs=100, optimtype=torch.optim.SGD, lr=0.01, weight_decay=0.0001)
# testing
performance, complexity, robustness = test(model, testdata)
```

between unimodal representations $\mathbf{z}_1, \mathbf{z}_2$ and the multimodal representation $\mathbf{z}_{\text{mm}}$, i.e., $\mathcal{L}_{\text{contrast}} = 1 - \cos(\mathbf{z}_{\text{mm}}, g_1(\mathbf{z}_1)) + 1 - \cos(\mathbf{z}_{\text{mm}}, g_2(\mathbf{z}_2))$ where g_1, g_2 is an auxiliary layer mapping each modality's representation into the joint multimodal space. The intuition here is that the unimodal representations $\mathbf{z}_1, \mathbf{z}_2$ and the multimodal representation $\mathbf{z}_{\text{mm}}$ should be aligned in the multimodal feature space as measured by cosine similarity. While the original REFNET method does not use negative samples, closely related work in multi-view contrastive learning has extended this idea to use negative samples which is more closely in line with recent work in contrastive learning [604].

Reconstruction objectives: Methods based on generative-discriminative models (e.g., VAEs) include an objective to reconstruct the input (or some part of the input) [335, 614]. These have been shown to better preserve task-relevant information learned in the representation, especially in settings with sparse supervised signals such as robotics [335] and long videos [614]. We include the Multimodal Factorized Model (MFM) [614] which is a general approach that learns a representation $\mathbf{z}_{\text{mm}}$ that can reconstruct input data $\mathbf{x}_1, \mathbf{x}_2$ while also predicting the label. The multimodal representation is a concatenation of factorized representations $\mathbf{z}_1$, $\mathbf{z}_2$, ..., $\mathbf{z}_M$, and $\mathbf{z}_y$.

Since MFM optimizes a variational lower-bound to the log likelihood, the overall objective consists of 3 terms - generative, discriminative, and prior regularization:

$$\min_{f_i, f_{\text{mm}}, g_i, g_y} \mathbf{E}_{P_{\mathbf{x}_{1:M},\mathbf{y}}} \mathbf{E}_{f_1(\mathbf{z}_1|\mathbf{x}_1)} \cdots \mathbf{E}_{f_M(\mathbf{z}_M|\mathbf{x}_M)} \mathbf{E}_{f_{\text{mm}}(\mathbf{z}_y|\mathbf{x}_{1:M})} \left[\sum_{i=1}^{M} \|\mathbf{x}_i, g_i(\mathbf{z}_i, \mathbf{z}_\mathbf{y})\|_2 + \ell\left(\mathbf{y}, g_y(\mathbf{z}_\mathbf{y})\right) \right] + \lambda \text{MMD}(Q_\mathbf{z}, P_\mathbf{z}), \tag{7.1}$$

where f_i's are encoders from each modality to representations, f_{mm} is a multimodal encoder to the joint representation $\mathbf{z}_y$, g_i's are decoders from latent representations back into input data, and g_y is a classification head to the label. The final MMD term is a regularizer to bring the representations close to a unit Gaussian prior. The multimodal encoder f_{mm} in MFM can be instantiated with any multimodal model (e.g., learning $\mathbf{z}_y$ via tensors and adding a term to reconstruct input data). We use the public implementation in `https://github.com/pliang279/factorized`, which uses a temporal attention model as f_{mm} for multimodal time-series data. For the remaining experiments we replace f_{mm} with a simple late fusion but also run some experiments with multimodal methods that are state-of-the-art in each domain.

Improving robustness: These approaches modify the objective function to account for robustness to noisy [364] or missing [336, 398, 482] modalities. MULTIZOO includes MCTN [482]

which uses cycle-consistent translation to predict the noisy/missing modality from present ones. The key insight is that a joint representation between modalities $\mathbf{x}_1$ and $\mathbf{x}_2$ can be learned by using $\mathbf{x}_1$ to predict $\mathbf{x}_2$, in a vein similar to machine translation or image/text style transfer. MCTN defines a cyclic translation path $\mathbf{x}_1 \rightarrow \mathbf{z}_{\text{mm}} \rightarrow \hat{\mathbf{x}}_2 \rightarrow \mathbf{z}_{\text{mm}} \rightarrow \hat{\mathbf{x}}_1$ and adds additional reconstruction losses $\mathcal{L}_{\text{rec}} = \|\mathbf{x}_1 - \hat{\mathbf{x}}_1\|_2 + \|\mathbf{x}_2 - \hat{\mathbf{x}}_2\|_2$ on top of the supervised learning loss. The representations $\mathbf{z}_{\text{mm}}$ learned via translation are then used to predict the label. Surprisingly, the model needs to take in only $\mathbf{x}_1$ at test time and is therefore robust to all levels of noise or missingness in $\mathbf{x}_2$.

7.4.4 Training procedures

Improving generalization: Recent work has found that directly training a multimodal model with all modalities using supervised learning is sub-optimal since different modalities overfit and generalize at different rates. MULTIZOO includes an approach to solve this, called Gradient Blending (GRADBLEND), that computes generalization statistics for each modality to determine their weights during multimodal fusion [651]. We also include a similar work, Regularization by Maximizing Functional Entropies (RMFE), which uses functional entropy to balance the contribution of each modality to the classification result [191].

7.4.5 Putting everything together

In Algorithm 3, we show a sample code snippet in Python that loads a dataset from MULTIBENCH, defines the unimodal and multimodal architectures, optimization objective, and training procedures, before running the evaluation protocol. Our MULTIZOO toolkit is easy to use and trains entire multimodal models in less than 10 lines of code. By standardizing the implementation of each module and disentangling the individual effects of models, optimizations, and training, MULTIZOO ensures both accessibility and reproducibility of its algorithms.

7.5 Experiments and Discussion

Setup: Using MULTIBENCH, we load each of the datasets and test the multimodal approaches in MULTIZOO. We only vary the contributed method of interest and keep all other possibly confounding factors constant (i.e., using the exact same training loop when testing a new multimodal fusion paradigm), a practice unfortunately not consistent in previous work. Our code is available at `https://github.com/pliang279/MultiBench`. MULTIBENCH allows for careful analysis of multimodal models and we summarize the main take-away messages below.

7.5.1 Benefits of standardization

From Table 7.2, simply applying methods in a research different area achieves state-of-the-art performance on 9 out of the 15 fusion tasks. We find that this is especially true for domains and modalities that have been relatively less studied in multimodal research (i.e., healthcare, finance, HCI). Performance gains can be obtained using multimodal methods *outside* of that research area. Therefore, this motivates the benefits of standardizing and unifying areas of research in multimodal

Table 7.2: Standardizing methods and datasets enables quick application of methods from different research areas which achieves stronger performance on 9/15 datasets in MULTIBENCH, especially in healthcare, HCI, robotics, and finance. *In-domain* refers to the best performance across methods previously proposed on that dataset and *out-domain* shows best performance across remaining methods. ↑ indicates metrics where higher is better (Acc, AUPRC), ↓ indicates lower is better (MSE).

Dataset	MUSTARD ↑	CMU-MOSI ↑	UR-FUNNY ↑	CMU-MOSEI ↑	MIMIC ↑
Unimodal	68.6 ± 0.4	74.2 ± 0.5	58.3 ± 0.2	78.8 ± 1.5	76.7 ± 0.3
In-domain	66.3 ± 0.3	**83.0 ± 0.1**	62.9 ± 0.2	**82.1 ± 0.5**	77.9 ± 0.3
Out-domain	**71.8 ± 0.3**	75.5 ± 0.5	**66.7 ± 0.3**	78.1 ± 0.3	**78.2 ± 0.2**
Improvement	**4.7%**	-	**6.0%**	-	**0.4%**

Dataset	MUJOCO PUSH ↓	V&T EE ↓	STOCKS-F&B ↓	STOCKS-HEALTH ↓	STOCKS-TECH ↓
Unimodal	0.334 ± 0.034	0.202 ± 0.022	1.856 ± 0.093	0.541 ± 0.010	0.125 ± 0.004
In-domain	**0.290 ± 0.018**	0.258 ± 0.011	1.856 ± 0.093	0.541 ± 0.010	0.125 ± 0.004
Out-domain	0.402 ± 0.026	**0.185 ± 0.011**	**1.820 ± 0.138**	**0.526 ± 0.017**	**0.120 ± 0.008**
Improvement	-	**8.4%**	**1.9%**	**2.8%**	**4.0%**

Dataset	ENRICO ↑	MM-IMDB ↑	AV-MNIST ↑	KINETICS-S ↑	KINETICS-L ↑
Unimodal	47.0 ± 1.6	45.6 ± 4.5	65.1 ± 0.2	**56.5**	72.6
In-domain	47.0 ± 1.6	49.8 ± 1.7	**72.8 ± 0.2**	56.1	**74.7**
Out-domain	**51.0 ± 1.4**	**50.2 ± 0.9**	72.3 ± 0.2	23.7	71.7
Improvement	**8.5%**	**0.8%**	-	-	-

machine learning. We believe that the ever-expanding diversity of datasets in MULTIBENCH can greatly accelerate research in multimodal learning.

7.5.2 Generalization across domains and modalities

MULTIBENCH offers an opportunity to analyze algorithmic developments across a large suite of modalities, domains, and tasks. We illustrate these observations through 2 summary plots of the generalization performance of multimodal models. Firstly, in Figure 7.4, we plot the performance of each multimodal method across all datasets that it is tested on, using the color red to indicate performance on datasets that it was initially proposed and tested on (which we label as *in-domain*), and blue to indicate its performance on the remaining datasets (which we label as *out-domain*). Secondly, in Figure 7.5, we color-code the performance on each dataset depending on which research area the dataset belongs to (one of the 6 research areas covered in MULTIBENCH).

We summarize several observations regarding generalization across modalities and tasks:

1. Many multimodal methods still do not generalize across domains and datasets. For examples, MFAS [478] works well on domains it was designed for (AV-MNIST and MM-IMDB in the multimedia domain), but does not generalize to other domains such as healthcare (MIMIC). Similarly, the method designed for robustness, MCTN [482], does not generalize to datasets within the affective computing domain (UR-FUNNY and MUSTARD). Finally, GRADBLEND [651], an approach specifically designed to improve generalization in multimodal learning and tested on video and audio datasets (e.g., Kinetics), does not perform well on other datasets. Therefore, there still does not exist a one-size-fits-all model, especially on understudied modalities and tasks.

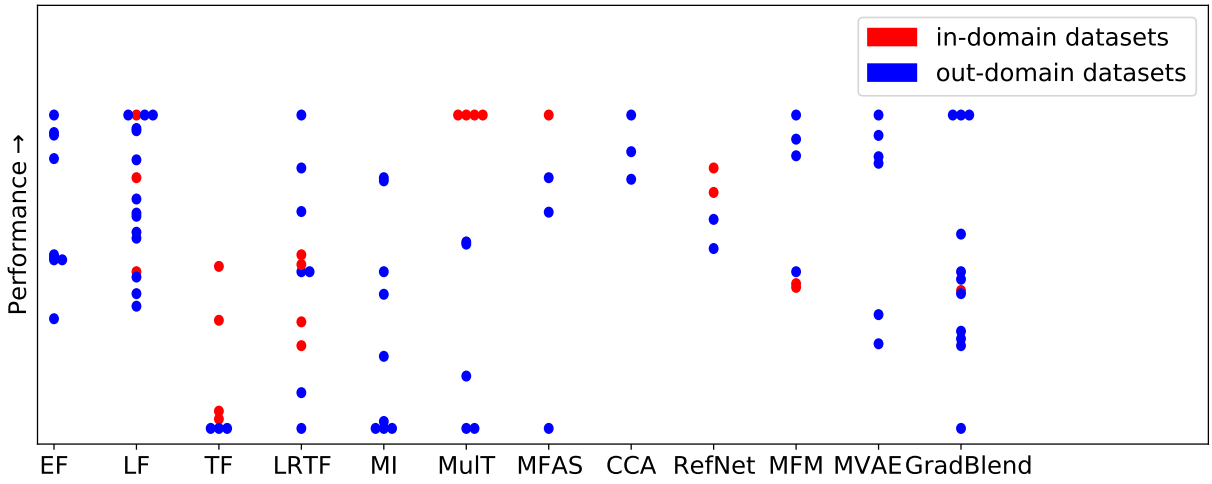


Figure 7.4: Relative performance of each model across in-domain (red dots) and out-domain datasets (blue dots). *In-domain* refers to the performance on datasets that the method was previously proposed for and *out-domain* shows performance on the remaining datasets. We find that many methods show strongest performance on in-domain datasets which drop when tested on different domains, modalities, and tasks. In general, we also observe high variance in the performance of multimodal methods across datasets in MULTIBENCH, which suggest open questions in building more generalizable models.

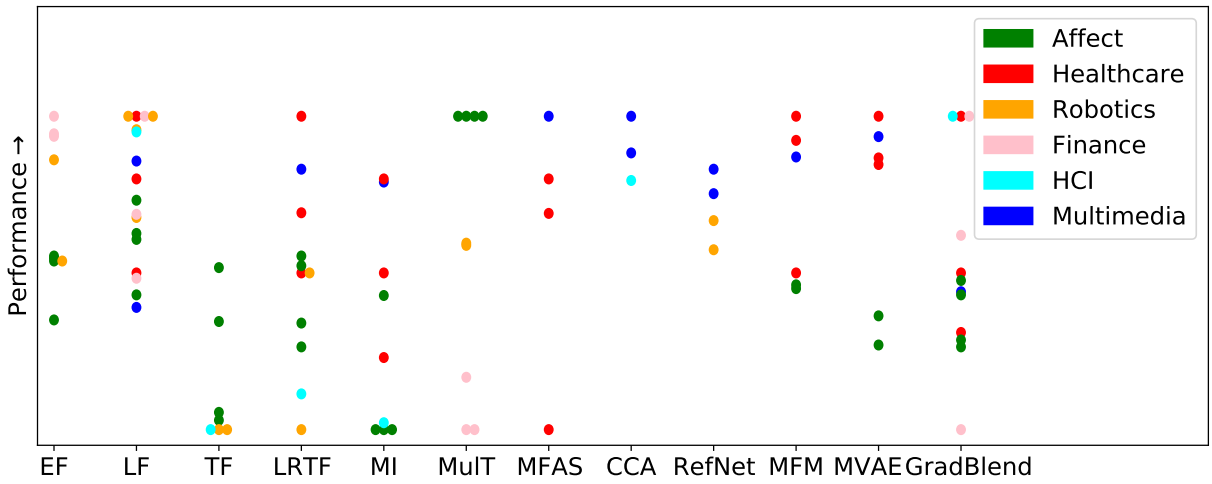


Figure 7.5: Relative performance of each model across different domains. We find that the performance of multimodal models varies significantly across datasets spanning different research areas and modalities. Similarly, the best-performing methods on each domain are also different. Therefore, there still does not exist a one-size-fits-all model, especially for understudied modalities and tasks.

2. From Figure 7.4, we find that many methods show their strongest performance on in-domain datasets, and their performance drops when tested on different domains, modalities, and tasks. Some interesting observations are that MULT performs extremely well on the affect recognition datasets it was designed for but struggles on other multimodal time-series in the finance and robotics domains. On the other hand, MFM shows an impressive performance in generalizing to new domains, although its in-domain performance has been exceeded by several other methods.
3. From Figure 7.4, we also observe high variance in the performance of multimodal methods across datasets in MULTIBENCH, which suggest open questions in building more generalizable models. We find that LF is quite stable and always achieves above-average performance.
4. There are methods that are surprisingly generalizable across datasets. These are typically general modality-agnostic methods such as LF. While simple, it is a strong method that balances simplicity, performance, and low complexity. However, it does not achieve the best performance on any dataset, which suggests that it is a good starting point but perhaps not the best eventual method.
5. From Figure 7.5, we find that performance also varies significantly across research areas.
6. Several methods such as MFAS and CCA are designed for only 2 modalities (usually image and text), and TF and MI do not scale efficiently beyond $2/3$ modalities. Therefore, we were unable to directly adapt these approaches to other datasets. We encourage the community to generalize these approaches across datasets and modalities on MULTIBENCH.

7.5.3 Tradeoffs between modalities

How far can we go with unimodal methods? Surprisingly far! From Table 7.2, we observe that decent performance can be obtained with the best performing modality. Further improvement via multimodal models may come at the expense of around $2-3\times$ the parameters.

7.6 Related Work

We review related work on standardizing datasets and methods in multimodal learning.

Comparisons with related benchmarks: To the best of our knowledge, MULTIBENCH is the first multimodal benchmark with such a large number of datasets, modalities, and tasks. Most previous multimodal benchmarks have focused on a single research area such as within affective computing [199], human multimodal language [360], language and vision-based question answering [174, 536], text classification with external multimodal information [212], and multimodal learning for education [227]. MULTIBENCH is specifically designed to go beyond the commonly studied language, vision, and audio modalities to encourage the research community to explore relatively understudied modalities (e.g., tabular data, time-series, sensors, graph and set data) and build general multimodal methods that can handle a diverse set of modalities.

Our work is also inspired by recent progress in better evaluation benchmarks for a suite of important tasks in ML such as language representation learning [642, 643], long-range sequence modeling [595], multilingual representation learning [251], graph representation learning [256],

and robustness to distribution shift [312]. These well-crafted benchmarks have accelerated progress in new algorithms, evaluation, and analysis in their respective research areas.

Standardizing multimodal learning: There have also been several attempts to build a single model that works well on a suite of multimodal tasks [348, 390, 571]. However, these are limited to the language and vision space, and multimodal training is highly tailored for text and images. Transformer architectures have emerged as a popular choice due to their suitability for both language and image data [108, 253] and a recent public toolkit was released for incorporating multimodal data on top of text-based Transformers for prediction tasks [212]. By going beyond Transformers and text data, MULTIBENCH opens the door to important research questions involving a much more diverse set of modalities and tasks while holistically evaluating performance, complexity, and robustness.

Analysis of multimodal representations: Recent work has carefully analyzed and challenged long-standing assumptions in multimodal learning. They have shown that certain models do not actually learn cross-modal interactions but rather rely on ensembles of unimodal statistics [235] and that certain datasets and models are biased to the most dominant modality [75, 206], sometimes ignoring others completely [10]. These observations are currently only conducted on specific datasets and models without testing their generalization to others, a shortcoming we hope to solve using MULTIBENCH which enables scalable analysis over modalities, tasks, and models.

7.7 Conclusion

Limitations: While MULTIBENCH can help to accelerate research in multimodal ML, we are aware of the following possible limitations:

1. *Tradeoffs between generality and specificity*: While it is desirable to build models that work across modalities and tasks, there is undoubtedly merit in building modality and task-specific models that can often utilize domain knowledge to improve performance and interpretability (e.g., see neuro-symbolic VQA [632], or syntax models for the language modality [120]). By easing access to data, models, and evaluation, we hope that MULTIBENCH will challenge researchers to design interpretable models leveraging domain knowledge for many multimodal tasks. It remains an open question to define "interpretability" for other modalities beyond image and text, a question we hope MULTIBENCH will drive research in.

2. *Scale of datasets, models, and metrics*: We plan for MULTIBENCH to be a continuously-growing community effort with regular maintenance and expansion. While MULTIBENCH currently does not include several important research areas outside of multimodal fusion (e.g., question answering [11, 223], retrieval [732], grounding [121], and reinforcement learning [394]), and is also limited by the models and metrics it supports, we have plans to expand MULTIBENCH towards a wider scale of datasets, models, and metrics.

Projected expansions of MULTIBENCH: In this subsection, we describe concrete ongoing and future work towards expanding MULTIBENCH:

1. *Other multimodal research problems*: We are genuinely committed to building a community around these resources and continue improving it over time. While we chose to focus on multimodal fusion by design for this first version to have a more coherent way to standardize and evaluate methods across datasets, we acknowledge the breadth of multimodal learning and

are looking forward to expanding it in other directions in collaboration with domain experts. We have already included 2 datasets in captioning (and more generally for non-language outputs, retrieval): (1) Yummly-28K of paired videos and text descriptions of food recipes [421], and (2) Clotho dataset for audio-captioning [156] as well as a language-guided RL environment Read to Fight Monsters (RTFM) [733] and are also working towards more datasets in QA, retrieval, and multimodal RL.

To help in scalable expansion, we plan for an open call to the community for suggestions and feedback about domains, datasets, and metrics. As a step in this direction, we have concrete plans to use MULTIBENCH as a theme for future workshops and competitions (building on top of the multimodal workshops we have been organizing at NAACL 2021, ACL 2020, and ACL 2019, and in multimodal learning courses (starting with the course taught annually at CMU). Since MULTIBENCH is public and will be regularly maintained, the existing benchmark, code, evaluation, and experimental protocols can greatly accelerate any dataset and modeling innovations added in the future. In our public GitHub, we have included a section on contributing through task proposals or additions of datasets and algorithms. The authors will regularly monitor new proposals through this channel.

2. *New evaluation metrics*: We also plan to include evaluation for distribution shift, uncertainty estimation, tests for fairness and social biases, as well as labels/metrics for interpretable multimodal learning. In the latter, we plan to include the EMAP score [235] as an interpretability metric assessing whether cross-modal interactions improve performance.

3. *Multimodal transfer learning and co-learning*: Can training data in one dataset help learning on other datasets? MULTIBENCH enables easy experimentation of such research questions: our initial experiments on transfer learning found that pre-training on larger datasets in the same domain can improve performance on smaller datasets when fine-tuned on a smaller dataset: performance on the smaller CMU-MOSI dataset improved from 75.2 to 75.8 using the same late fusion model with transfer learning from the larger UR-FUNNY and CMU-MOSEI datasets. Furthermore, recent work has shown that multimodal training can help improve unimodal performance as well [553, 675, 716]. While previous experiments were on a small scale and limited to a single domain, we plan to expand significantly on this phenomenon (multimodal co-learning) in future versions of MULTIBENCH.

4. *Multitask learning across modalities*: Multitask learning across multimodal tasks with a shared set of input modalities is a promising direction that can enable statistical strength sharing across datasets and efficiency in training a single model. Using MULTIBENCH, we also ran an extra experiment on multi-dataset multitask learning. We used the 4 datasets in the affective computing domain and trained a single model across all 4 of them with adjustable input embedding layers if the input features were different and separate classification heads for each dataset's task. We found promising initial results with performance on the largest CMU-MOSEI dataset improving from 79.2 to 80.9 for a late fusion model and from 82.1 to 82.9 using a multimodal transformer model, although performance on the smaller CMU-MOSI dataset decreased from 75.2 to 70.8. We believe that these potential future studies in multitask and transfer learning are strengths of MULTIBENCH since it shows the potential of interesting experiments and usage.

In conclusion, we present MULTIBENCH, a large-scale benchmark unifying previously disjoint efforts in multimodal research with a focus on ease of use, accessibility, and reproducibility, thereby paving the way towards a deeper understanding of multimodal models. Through its

unprecedented range of research areas, datasets, modalities, tasks, and evaluation metrics, MULTIBENCH highlights several future directions in building more generalizable, lightweight, and robust multimodal models.

Chapter 8

Neural Architectures for Multisensory Foundation Models

8.1 Introduction

To build general multisensory foundation models that work across the diverse modalities and tasks in MULTIBENCH, this chapter of the thesis presents two architectures that are broadly generalizable across diverse modalities. The large number of heterogeneous modalities creates challenges in building multisensory foundation models. For example, the healthcare domain typically collects tabular data and high-frequency sensors [287], and it remains an open question how to best combine large language models with tabular data and sensors [546]. To tackle the heterogeneity across many different modalities, we treat modalities in their most general form as sequences of elements, and study how to learn interactions between multiple elements across modalities. As motivated in the first part of the thesis, these local interactions between two elements can be *redundant*, *unique*, and *synergistic*: redundancy quantifies information shared between modalities, uniqueness quantifies the information present in only one of the modalities, and synergy quantifies the emergence of new information not previously present in either modality.

Treating modalities as sequences of elements now introduces a new challenge due to asynchrony in time: for example, the simultaneous co-occurrence between a smile and a positive word, or the delayed occurrence of laughter after the end of a sentence. Modeling these interactions lie at the heart of analyzing human communication, audio-video data, sensor fusion, and medical modalities. We now present two approaches to learn interactions from heterogeneous modality elements across sequences: the *cross-modal attention* [101, 359] and *multimodal transformer* [613] architectures.

The first architecture is called RECURRENT MULTISTAGE FUSION NETWORK, or RMFN for short. This method automatically decomposes the multimodal fusion problem into multiple recursive stages across the sequence. At each stage, a subset of multimodal signals is highlighted and fused with previous fusion representations (see Figure 9.1). This divide-and-conquer approach decreases the burden on each fusion stage, allowing each stage to be performed in a more specialized and effective way. This is in contrast with conventional fusion approaches which usually model interactions over multimodal sequences altogether in one iteration (e.g., early or late

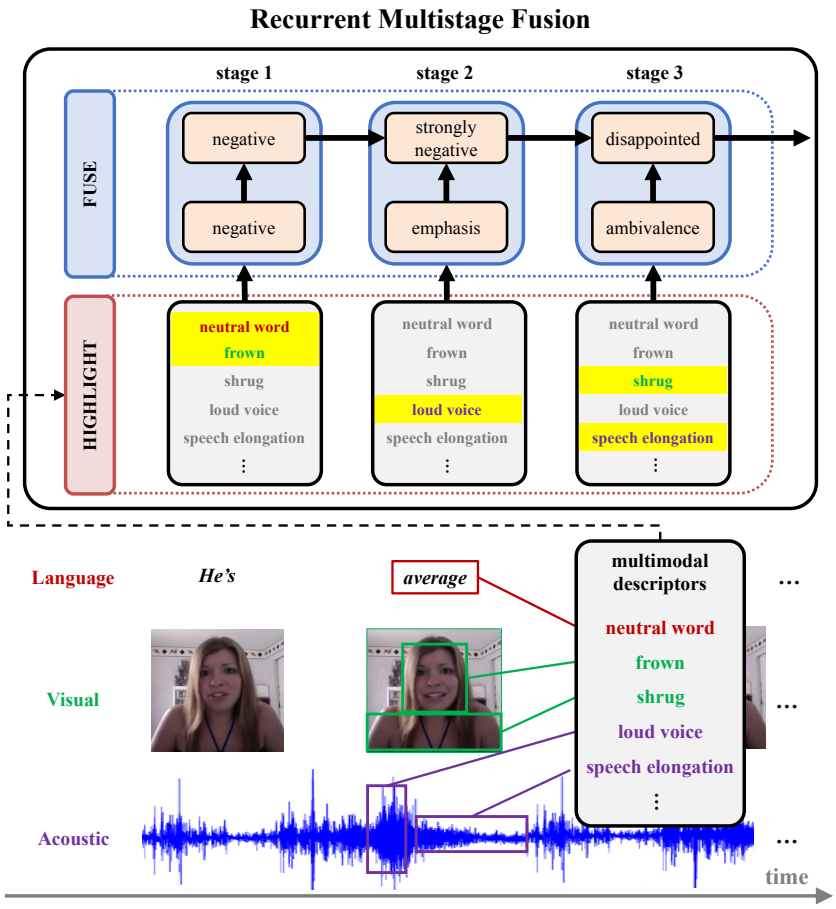


Figure 8.1: An illustrative example for Recurrent Multistage Fusion. At each recursive stage, a subset of multimodal signals is highlighted and then fused with previous fusion representations. The first fusion stage selects the neutral word and frowning behaviors which create an intermediate representation reflecting negative emotion when fused together. The second stage selects the loud voice behavior which is locally interpreted as emphasis before being fused with previous stages into a strongly negative representation. Finally, the third stage selects the shrugging and speech elongation behaviors that reflect ambivalence and when fused with previous stages is interpreted as a representation for the disappointed emotion.

fusion [45]). In RMFN, multimodal interactions are modeled by integrating our new multistage fusion process with a system of recurrent neural networks. Overall, RMFN recursively models all forms of redundant, unique, and synergistic multimodal interactions across the sequence and is differentiable end-to-end.

The second architecture we propose is the MULTIMODAL TRANSFORMER (MULT), an end-to-end model that extends the standard Transformer network [631] to learn representations directly from unaligned multimodal sequences. At the heart of MULT is the crossmodal attention module, which learns multimodal interactions between all elements in the first modality with all elements in the second modality. As a result, all multimodal interactions across the entire sequence are learned simultaneously, and can be parallelized efficiently over GPUs as compared to the first recurrent fusion approach. This makes MULT extremely scalable and effective, especially in settings where modality elements are asynchronous and where obtaining alignment information is difficult (e.g.,

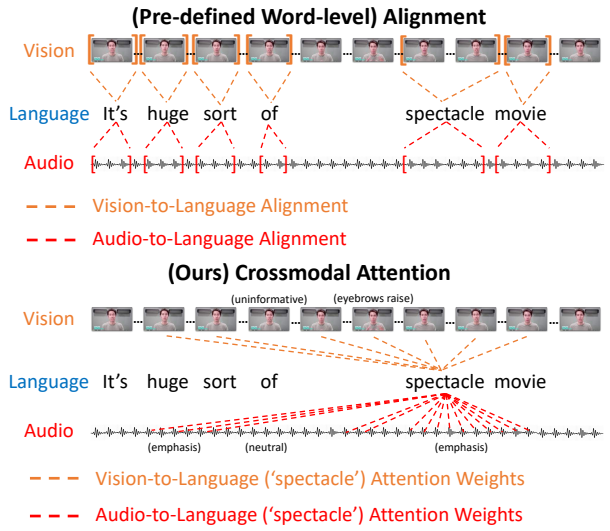


Figure 8.2: Example video clip from movie reviews. [Top]: Illustration of word-level alignment where video and audio features are averaged across the time interval of each spoken word. [Bottom] Illustration of crossmodal attention weights between text ("spectacle") and vision/audio.

by forced word-aligning before training [482, 717], see Figure 8.2 for a comparison).

We evaluate RMFN and MULT on three different tasks related to human multimodal language: sentiment analysis, emotion recognition, and speaker traits recognition across three public multimodal datasets. RMFN achieves state-of-the-art performance in all three tasks. Through a comprehensive set of ablation experiments and visualizations, we demonstrate the advantages of explicitly defining multiple recursive stages for multimodal fusion.

8.2 Related Work

Previous approaches in human multimodal language modeling can be categorized as follows:

Non-temporal Models: These models simplify the problem by using feature-summarizing temporal observations [490]. Each modality is represented by averaging temporal information through time, as shown for language-based sentiment analysis [105, 273] and multimodal sentiment analysis [2, 427, 449, 711]. Conventional supervised learning methods are utilized to discover intra-modal and cross-modal interactions without specific model design [488, 645]. These approaches have trouble modeling long sequences since the average statistics do not properly capture the temporal intra-modal and cross-modal dynamics [677].

Multimodal Temporal Graphical Models: The application of graphical models in sequence modeling has been an important research problem. Hidden Markov Models (HMMs) [50], Conditional Random Fields (CRFs) [324], and Hidden Conditional Random Fields (HCRFs) [494] were shown to work well on modeling sequential data from the language [264, 400, 422] and acoustic [707] modalities. These temporal graphical models have also been extended for modeling multimodal data. Several methods have been proposed including multi-view HCRFs where the potentials of the HCRF are designed to model data from multiple views [559], multi-layered CRFs

with latent variables to learn hidden spatio-temporal dynamics from multi-view data [559], and multi-view Hierarchical Sequence Summarization models that recursively build up hierarchical representations [560].

Multimodal Temporal Neural Networks: More recently, with the advent of deep learning, Recurrent Neural Networks [164, 279] have been used extensively for language and speech based sequence modeling [557, 742], sentiment analysis [78, 153, 202, 554], and emotion recognition [58, 220, 326]. Long-short Term Memory (LSTM) networks [242] have also been extended for multimodal settings [501] and by learning binary gating mechanisms to remove noisy modalities [101]. Recently, more advanced models were proposed to model both intra-modal and cross-modal interactions. These use Bayesian ranking algorithms [234] to model both person-independent and person-dependent features [361], generative-discriminative objectives to learn either joint [481] or factorized multimodal representations [614], external memory mechanisms to synchronize multimodal data [713], or low-rank tensors to approximate expensive tensor products [388]. All these methods assume that cross-modal interactions should be discovered all at once rather than across multiple stages, where each stage solves a simpler fusion problem. Our empirical evaluations show the advantages of the multistage fusion approach.

Transformer Network: The Transformer network [631] was first introduced for neural machine translation, where the encoder and decoder side each leverages a *self-attention* [382, 464, 631] transformer. After each layer of self-attention, the encoder and decoder are connected by an additional decoder sublayer where the decoder attends to each element of the source text for each element of the target text. In addition to translation, transformer networks have also been successfully applied to other tasks, including language modeling [41, 129], semantic role labeling [568], word sense disambiguation [590], learning sentence representations [144], and video activity recognition [652].

8.3 RECURRENT MULTISTAGE FUSION NETWORK

We first describe the RECURRENT MULTISTAGE FUSION NETWORK (RMFN for short) for multimodal language analysis (Figure 9.2). Given a set of modalities $\{l(anguage), v(isual), a(coustic)\}$, the signal from each modality $m \in \{l, v, a\}$ is represented as a temporal sequence $\mathbf{X}^m = \{\mathbf{x}_1^m, \mathbf{x}_2^m, \mathbf{x}_3^m, \cdots, \mathbf{x}_T^m\}$, where $\mathbf{x}_t^m$ is the input at time t. Each sequence $\mathbf{X}^m$ is modeled with an intra-modal recurrent neural network (see subsection 8.3.3 for details). At time t, each intra-modal recurrent network will output a unimodal representation $\mathbf{h}_t^m$. The Multistage Fusion Process uses a recursive approach to fuse all unimodal representations $\mathbf{h}_t^m$ into a cross-modal representation $\mathbf{z}_t$ which is then fed back into each intra-modal recurrent network.

8.3.1 Multistage fusion process

The Multistage Fusion Process (MFP) is a modular neural approach that performs multistage fusion to model cross-modal interactions. Multistage fusion is a divide-and-conquer approach which decreases the burden on each stage of multimodal fusion, allowing each stage to be performed in a more specialized and effective way. The MFP has three main modules: `HIGHLIGHT`, `FUSE` and `SUMMARIZE`.

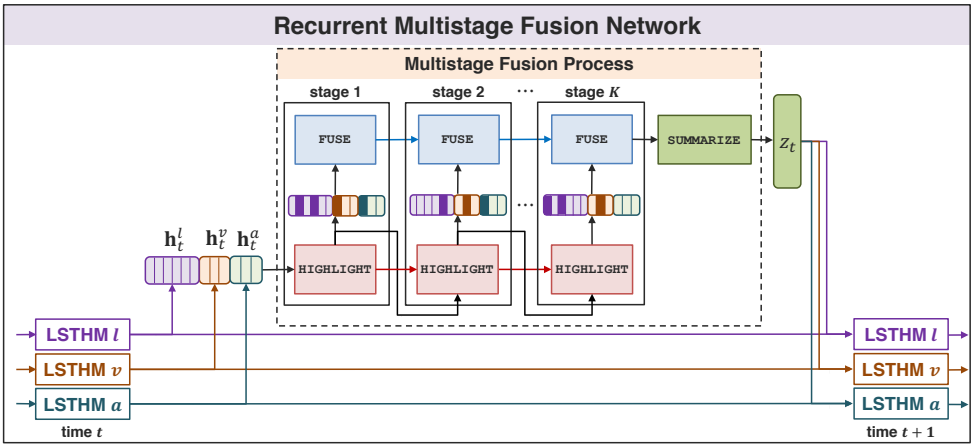


Figure 8.3: The RECURRENT MULTISTAGE FUSION NETWORK for multimodal language analysis. The Multistage Fusion Process has three modules: HIGHLIGHT, FUSE and SUMMARIZE. Multistage fusion begins with the concatenated intra-modal network outputs $\mathbf{h}_t^l, \mathbf{h}_t^v, \mathbf{h}_t^a$. At each stage, the HIGHLIGHT module identifies a subset of multimodal signals and the FUSE module performs local fusion before integration with previous fusion representations. The SUMMARIZE module translates the representation at the final stage into a cross-modal representation $\mathbf{z}_t$ to be fed back into the intra-modal recurrent networks.

Two modules are repeated at each stage: HIGHLIGHT and FUSE. The HIGHLIGHT module identifies a subset of multimodal signals from $[\mathbf{h}_t^l, \mathbf{h}_t^v, \mathbf{h}_t^a]$ that will be used for that stage of fusion. The FUSE module then performs two subtasks simultaneously: a local fusion of the highlighted features and integration with representations from previous stages. Both HIGHLIGHT and FUSE modules are realized using memory-based neural networks which enable coherence between stages and storage of previously modeled cross-modal interactions. As a final step, the SUMMARIZE module takes the multimodal representation of the final stage and translates it into a cross-modal representation $\mathbf{z}_t$.

Figure 9.1 shows an illustrative example for multistage fusion. The HIGHLIGHT module selects "neutral words" and "frowning" expression for the first stage. The local and integrated fusion at this stage creates a representation reflecting negative emotion. For stage 2, the HIGHLIGHT module identifies the acoustic feature "loud voice". The local fusion at this stage interprets it as an expression of emphasis and is fused with the previous fusion results to represent a strong negative emotion. Finally, the highlighted features of "shrug" and "speech elongation" are selected and are locally interpreted as "ambivalence". The integration with previous stages then gives a representation closer to "disappointed".

8.3.2 Module descriptions

In this section, we present the details of the three multistage fusion modules: HIGHLIGHT, FUSE and SUMMARIZE. Multistage fusion begins with the concatenation of intra-modal network outputs $\mathbf{h}_t = \bigoplus_{m \in M} \mathbf{h}_t^m$. We use superscript $^{[k]}$ to denote the indices of each stage $k = 1, \cdots, K$ during K total stages of multistage fusion. Let Θ denote the neural network parameters across all modules.

HIGHLIGHT: At each stage k, a subset of the multimodal signals represented in $\mathbf{h}_t$ will be

automatically highlighted for fusion. Formally, this module is defined by the process function f_H:

$$\mathbf{a}_t^{[k]} = f_H(\mathbf{h}_t \; ; \; \mathbf{a}_t^{[1:k-1]}, \Theta) \tag{8.1}$$

where at stage k, $\mathbf{a}_t^{[k]}$ is a set of attention weights which are inferred based on the previously assigned attention weights $\mathbf{a}_t^{[1:k-1]}$. As a result, the highlights at a specific stage k will be dependent on previous highlights. To fully encapsulate these dependencies, the attention assignment process is performed in a recurrent manner using a LSTM which we call the HIGHLIGHT LSTM. The initial HIGHLIGHT LSTM memory at stage 0, $\mathbf{c}_t^{\texttt{HIGHLIGHT}[0]}$, is initialized using a network $\mathcal{M}$ that maps $\mathbf{h}_t$ into LSTM memory space:

$$\mathbf{c}_t^{\texttt{HIGHLIGHT}[0]} = \mathcal{M}(\mathbf{h}_t \; ; \; \Theta) \tag{8.2}$$

This allows the memory mechanism of the HIGHLIGHT LSTM to dynamically adjust to the intra-modal representations $\mathbf{h}_t$. The output of the HIGHLIGHT LSTM $\mathbf{h}_t^{\texttt{HIGHLIGHT}[k]}$ is softmax activated to produce attention weights $\mathbf{a}_t^{[k]}$ at every stage k of the multistage fusion process:

$$\mathbf{a}_t^{[k]}{}_j = \frac{\exp\left(\mathbf{h}_t^{\texttt{HIGHLIGHT}[k]}{}_j\right)}{\sum_{d=1}^{|\mathbf{h}_t^{\texttt{HIGHLIGHT}[k]}|} \exp\left(\mathbf{h}_t^{\texttt{HIGHLIGHT}[k]}{}_d\right)} \tag{8.3}$$

and $\mathbf{a}_t^{[k]}$ is fed as input into the HIGHLIGHT LSTM at stage $k+1$. Therefore, the HIGHLIGHT LSTM functions as a decoder LSTM [115, 581] in order to capture the dependencies on previous attention assignments. Highlighting is performed by element-wise multiplying the attention weights $\mathbf{a}_t^{[k]}$ with the concatenated intra-modal representations $\mathbf{h}_t$:

$$\tilde{\mathbf{h}}_t^{[k]} = \mathbf{h}_t \odot \mathbf{a}_t^{[k]} \tag{8.4}$$

where $\odot$ denotes the Hadamard product and $\tilde{\mathbf{h}}_t^{[k]}$ are the attended multimodal signals that will be used for the fusion at stage k.

FUSE: The highlighted multimodal signals are simultaneously fused in a local fusion and then integrated with fusion representations from previous stages. Formally, this module is defined by the process function f_F:

$$\mathbf{s}_t^{[k]} = f_F(\tilde{\mathbf{h}}_t^{[k]} \; ; \; \mathbf{s}_t^{[1:k-1]}, \Theta) \tag{8.5}$$

where $\mathbf{s}_t^{[k]}$ denotes the integrated fusion representations at stage k. We employ a FUSE LSTM to simultaneously perform the local fusion and the integration with previous fusion representations. The FUSE LSTM input gate enables a local fusion while the FUSE LSTM forget and output gates enable integration with previous fusion results. The initial FUSE LSTM memory at stage 0, $\mathbf{c}_t^{\texttt{FUSE}[0]}$, is initialized using random orthogonal matrices [33, 330].

SUMMARIZE: After completing K recursive stages of HIGHLIGHT and FUSE, the SUMMARIZE operation generates a cross-modal representation using all final fusion representations $\mathbf{s}_t^{[1:K]}$. Formally, this operation is defined as:

$$\mathbf{z}_t = \mathcal{S}(\mathbf{s}_t^{[1:K]} \; ; \; \Theta) \tag{8.6}$$

where $\mathbf{z}_t$ is the final output of the multistage fusion process and represents all cross-modal interactions discovered at time t. The summarized cross-modal representation is then fed into the intra-modal recurrent networks as described in the subsection 8.3.3.

8.3.3 System of long short-term hybrid memories

To integrate the cross-modal representations $\mathbf{z}_t$ with the temporal intra-modal representations, we employ a system of Long Short-term Hybrid Memories (LSTHMs) [714]. The LSTHM extends the LSTM formulation to include the cross-modal representation $\mathbf{z}_t$ in a hybrid memory component:

$$\mathbf{i}^m_{t+1} = \sigma(\mathbf{W}^m_i \ \mathbf{x}^m_{t+1} + \mathbf{U}^m_i \ \mathbf{h}^m_t + \mathbf{V}^m_i \ \mathbf{z}_t + \mathbf{b}^m_i) \tag{8.7}$$

$$\mathbf{f}^m_{t+1} = \sigma(\mathbf{W}^m_f \ \mathbf{x}^m_{t+1} + \mathbf{U}^m_f \ \mathbf{h}^m_t + \mathbf{V}^m_f \ \mathbf{z}_t + \mathbf{b}^m_f) \tag{8.8}$$

$$\mathbf{o}^m_{t+1} = \sigma(\mathbf{W}^m_o \ \mathbf{x}^m_{t+1} + \mathbf{U}^m_o \ \mathbf{h}^m_t + \mathbf{V}^m_o \ \mathbf{z}_t + \mathbf{b}^m_o) \tag{8.9}$$

$$\bar{\mathbf{c}}^m_{t+1} = \mathbf{W}^m_{\bar{c}} \ \mathbf{x}^m_{t+1} + \mathbf{U}^m_{\bar{c}} \ \mathbf{h}^m_t + \mathbf{V}^m_{\bar{c}} \ \mathbf{z}_t + \mathbf{b}^m_{\bar{c}} \tag{8.10}$$

$$\mathbf{c}^m_{t+1} = \mathbf{f}^m_t \odot \mathbf{c}^m_t + \mathbf{i}^m_t \odot tanh(\bar{\mathbf{c}}^m_{t+1}) \tag{8.11}$$

$$\mathbf{h}^m_{t+1} = \mathbf{o}^m_{t+1} \odot tanh(\mathbf{c}^m_{t+1}) \tag{8.12}$$

where σ is the (hard-)sigmoid activation function, $tanh$ is the tangent hyperbolic activation function, $\odot$ denotes the Hadamard product. $\mathbf{i}, \mathbf{f}$ and $\mathbf{o}$ are the input, forget and output gates respectively. $\bar{\mathbf{c}}^m_{t+1}$ is the proposed update to the hybrid memory $\mathbf{c}^m_t$ at time $t+1$ and $\mathbf{h}^m_t$ is the time distributed output of each modality. The cross-modal representation $\mathbf{z}_t$ is modeled by the Multistage Fusion Process as discussed in subsection 8.3.2. The hybrid memory $\mathbf{c}^m_t$ contains both intra-modal interactions from individual modalities $\mathbf{x}^m_t$ as well as the cross-modal interactions captured in $\mathbf{z}_t$.

8.3.4 Optimization

The multimodal prediction task is performed using a final representation $\mathcal{E}$ which integrate (1) the last outputs from the LSTHMs and (2) the last cross-modal representation $\mathbf{z}_T$. Formally, $\mathcal{E}$ is defined as:

$$\mathcal{E} = (\bigoplus_{m \in M} \mathbf{h}^m_T) \bigoplus \mathbf{z}_T \tag{8.13}$$

where $\oplus$ denotes vector concatenation. $\mathcal{E}$ can then be used as a multimodal representation for supervised or unsupervised analysis of multimodal language. It summarizes all modeled intra-modal and cross-modal representations from the multimodal sequences. RMFN is differentiable end-to-end which allows the network parameters Θ to be learned using gradient descent approaches.

8.4 MULTIMODAL TRANSFORMER

We next describe our second proposed architecture, the MULTIMODAL TRANSFORMER (MULT) (Figure 9.6) for modeling unaligned multimodal language sequences. At the high level, MULT merges multimodal time-series via a feed-forward fusion process from multiple directional pairwise crossmodal transformers. Specifically, each crossmodal transformer (introduced in Section 8.4.2) serves to repeatedly reinforce a *target modality* with the low-level features from another *source modality* by learning the attention across the two modalities' features. A MULT

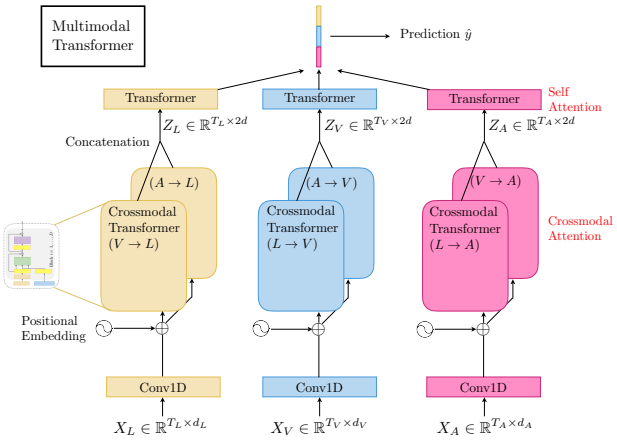


Figure 8.4: Overall architecture for MULT on modalities (L, V, A). The crossmodal transformers, which suggests latent crossmodal adaptations, are the core components of MULT for multimodal fusion.

architecture hence models all pairs of modalities with such crossmodal transformers, followed by sequence models (e.g., self-attention transformer) that predicts using the fused features.

The core of our proposed model is crossmodal attention module, which we first introduce in Section 8.4.1. Then, in Section 8.4.2 and 8.4.3, we present in details the various ingredients of the MULT architecture (see Figure 9.6) and discuss the difference between crossmodal attention and classical multimodal alignment.

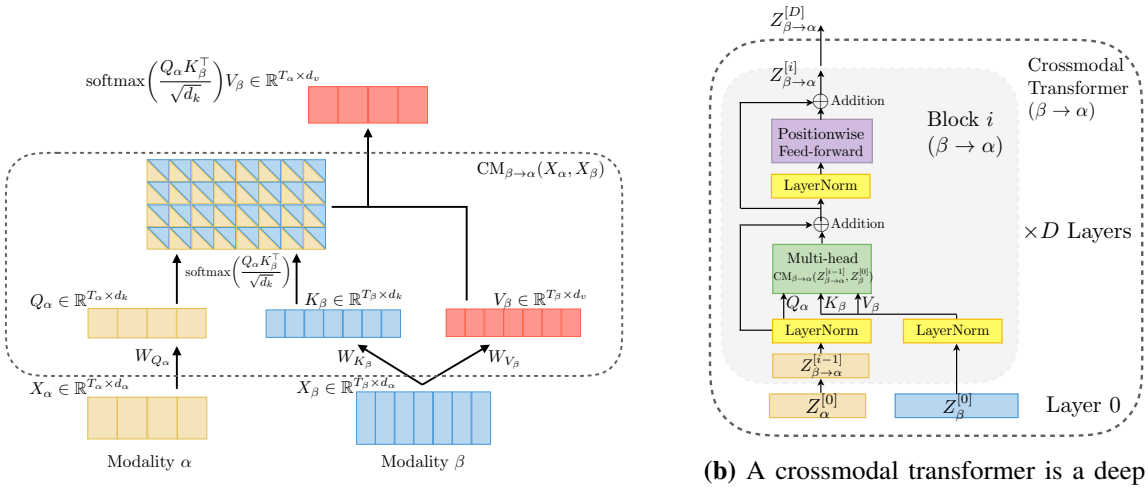
8.4.1 Crossmodal attention

We consider two modalities α and β, with two (potentially non-aligned) sequences from each of them denoted $X_\alpha \in \mathbb{R}^{T_\alpha \times d_\alpha}$ and $X_\beta \in \mathbb{R}^{T_\beta \times d_\beta}$, respectively. For the rest of the paper, $T_{(\cdot)}$ and $d_{(\cdot)}$ are used to represent sequence length and feature dimension, respectively. Inspired by the decoder transformer in NMT [631] that translates one language to another, we hypothesize a good way to fuse crossmodal information is providing a latent adaptation across modalities; i.e., β to α. Note that the modalities consider in our paper may span very different domains such as facial attributes and spoken words.

We define the Queries as $Q_\alpha = X_\alpha W_{Q_\alpha}$, Keys as $K_\beta = X_\beta W_{K_\beta}$, and Values as $V_\beta = X_\beta W_{V_\beta}$, where $W_{Q_\alpha} \in \mathbb{R}^{d_\alpha \times d_k}, W_{K_\beta} \in \mathbb{R}^{d_\beta \times d_k}$ and $W_{V_\beta} \in \mathbb{R}^{d_\beta \times d_v}$ are weights. The latent adaptation from β to α is presented as the crossmodal attention $Y_\alpha := \text{CM}_{\beta \to \alpha}(X_\alpha, X_B) \in \mathbb{R}^{T_\alpha \times d_v}$:

$$\begin{aligned} Y_\alpha &= \text{CM}_{\beta \to \alpha}(X_\alpha, X_\beta) \\ &= \text{softmax}\left(\frac{Q_\alpha K_\beta^\top}{\sqrt{d_k}}\right) V_\beta \\ &= \text{softmax}\left(\frac{X_\alpha W_{Q_\alpha} W_{K_\beta}^\top X_\beta^\top}{\sqrt{d_k}}\right) X_\beta W_{V_\beta}. \end{aligned} \tag{8.14}$$

Note that Y_α has the same length as Q_α (i.e., T_α), but is meanwhile represented in the feature space of V_β. Specifically, the scaled (by $\sqrt{d_k}$) softmax in Equation (8.14) computes a score matrix



(a) Crossmodal attention $\mathrm{CM}_{\beta\to\alpha}(X_\alpha, X_\beta)$ between sequences X_α, X_β from distinct modalities.

(b) A crossmodal transformer is a deep stacking of several crossmodal attention blocks.

Figure 8.5: Architectural elements of a crossmodal transformer between two time-series from modality α and β.

$\mathrm{softmax}(\cdot) \in \mathbb{R}^{T_\alpha \times T_\beta}$, whose (i, j)-th entry measures the attention given by the i-th time step of modality α to the j-th time step of modality β. Hence, the i-th time step of Y_α is a weighted summary of V_β, with the weight determined by i-th row in $\mathrm{softmax}(\cdot)$. We call Equation eq (8.14) a *single-head* crossmodal attention, which is illustrated in Figure 8.5a.

Following prior works on transformers [100, 129, 144, 631], we add a residual connection to the crossmodal attention computation. Then, another positionwise feed-forward sublayer is injected to complete a *crossmodal attention block* (see Figure 8.5b). Each crossmodal attention block adapts directly from the low-level feature sequence (i.e., $Z^{[0]}_\beta$ in Figure 8.5b) and does not rely on self-attention, which makes it different from the NMT encoder-decoder architecture [540, 631] (i.e., taking intermediate-level features). We argue that performing adaptation from low-level feature benefits our model to preserve the low-level information for each modality.

8.4.2 Overall architecture

Three major modalities are typically involved in multimodal language sequences: language (L), video (V), and audio (A) modalities. We denote with $X_{\{L,V,A\}} \in \mathbb{R}^{T_{\{L,V,A\}} \times d_{\{L,V,A\}}}$ the input feature sequences (and the dimensions thereof) from these 3 modalities. With these notations, in this subsection, we describe in greater details the components of Multimodal Transformer and how crossmodal attention modules are applied.

Temporal Convolutions. To ensure that each element of the input sequences has sufficient awareness of its neighborhood elements, we pass the input sequences through a 1D temporal convolutional layer:

$$\hat{X}_{\{L,V,A\}} = \mathrm{Conv1D}(X_{\{L,V,A\}}, k_{\{L,V,A\}}) \in \mathbb{R}^{T_{\{L,V,A\}} \times d} \tag{8.15}$$

where $k_{\{L,V,A\}}$ are the sizes of the convolutional kernels for modalities $\{L, V, A\}$, and d is a common dimension. The convolved sequences are expected to contain the local structure of the sequence, which is important since the sequences are collected at different sampling rates. Moreover, since the temporal convolutions project the features of different modalities to the same dimension d, the dot-products are admittable in the crossmodal attention module.

Positional Embedding. To enable the sequences to carry temporal information, following prior work [631], we augment positional embedding (PE) to $\hat{X}_{\{L,V,A\}}$:

$$Z^{[0]}_{\{L,V,A\}} = \hat{X}_{\{L,V,A\}} + \text{PE}(T_{\{L,V,A\}}, d) \tag{8.16}$$

where $\text{PE}(T_{\{L,V,A\}}, d) \in \mathbb{R}^{T_{\{L,V,A\}} \times d}$ computes the (fixed) embeddings for each position index, and $Z^{[0]}_{\{L,V,A\}}$ are the resulting low-level position-aware features for different modalities. We leave more details of the positional embedding to the full paper [613].

Crossmodal Transformers. Based on the crossmodal attention blocks, we design the crossmodal transformer that enables one modality for receiving information from another modality. In the following, we use the example for passing vision (V) information to language (L), which is denoted by "$V \to L$". We fix all the dimensions ($d_{\{\alpha,\beta,k,v\}}$) for each crossmodal attention block as d.

Each crossmodal transformer consists of D layers of crossmodal attention blocks (see Figure 8.5b). Formally, a crossmodal transformer computes feed-forwardly for $i = 1, \ldots, D$ layers:

$$\begin{aligned}
Z^{[0]}_{V\to L} &= Z^{[0]}_{L} \\
\hat{Z}^{[i]}_{V\to L} &= \text{CM}^{[i],\text{mul}}_{V\to L}(\text{LN}(Z^{[i-1]}_{V\to L}), \text{LN}(Z^{[0]}_{V})) + \text{LN}(Z^{[i-1]}_{V\to L}) \\
Z^{[i]}_{V\to L} &= f_{\theta^{[i]}_{V\to L}}(\text{LN}(\hat{Z}^{[i]}_{V\to L})) + \text{LN}(\hat{Z}^{[i]}_{V\to L})
\end{aligned} \tag{8.17}$$

where f_θ is a positionwise feed-forward sublayer parametrized by θ, and $\text{CM}^{[i],\text{mul}}_{V\to L}$ means a multi-head (see prior work [631] for more details) version of $\text{CM}_{V\to L}$ at layer i (note: d should be divisible by the number of heads). LN means layer normalization [38].

In this process, each modality keeps updating its sequence via low-level external information from the multi-head crossmodal attention module. At every level of the crossmodal attention block, the low-level signals from source modality are transformed to a different set of Key/Value pairs to interact with the target modality. Empirically, we find that the crossmodal transformer learns to correlate meaningful elements across modalities. The eventual MULT is based on modeling every pair of crossmodal interactions. Therefore, with 3 modalities (i.e., L, V, A) in consideration, we have 6 crossmodal transformers in total (see Figure 9.6).

Self-Attention Transformers and Prediction. As a final step, we concatenate the outputs from the crossmodal transformers that share the same target modality to yield $Z_{\{L,V,A\}} \in \mathbb{R}^{T_{\{L,V,A\}} \times 2d}$. For example, $Z_L = [Z^{[D]}_{V\to L}; Z^{[D]}_{A\to L}]$. Each of them is then passed through a sequence model to collect temporal information to make predictions. We choose the self-attention transformer [631]. Eventually, the last elements of the sequences models are extracted to pass through fully-connected layers to make predictions.

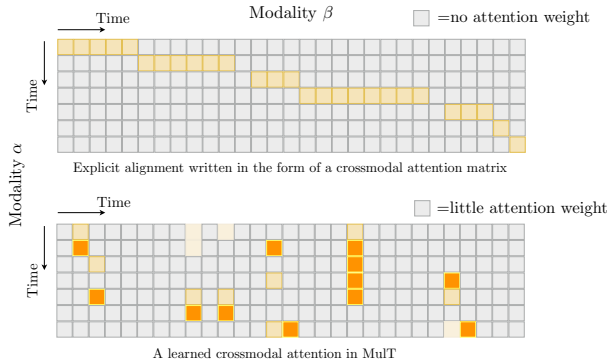


Figure 8.6: An example of visualizing alignment using attention matrix from modality β to α. Multimodal alignment is a special (monotonic) case for crossmodal attention.

8.4.3 Discussion about attention & alignment

When modeling unaligned multimodal language sequences, MULT relies on crossmodal attention blocks to merge signals across modalities. While the multimodal sequences were (manually) aligned to the same length in prior works before training [359, 482, 614, 655, 717], we note that MULT looks at the non-alignment issue through a completely different lens. Specifically, for MULT, the correlations between elements of multiple modalities are purely based on attention. In other words, MULT does not handle modality non-alignment by (simply) aligning them; instead, the crossmodal attention encourages the model to directly attend to elements in other modalities where strong signals or relevant information is present. As a result, MULT can capture long-range crossmodal contingencies in a way that conventional alignment could not easily reveal. Classical crossmodal alignment, on the other hand, can be expressed as a special (step diagonal) crossmodal attention matrix (i.e., monotonic attention [704]). We illustrate their differences in Figure 8.6.

8.5 Experimental Setup

To evaluate the performance and generalization of RMFN and MULT, three domains of human multimodal language were selected: multimodal sentiment analysis, emotion recognition, and speaker traits recognition. Our goal is to compare MULT with prior competitive approaches on both *word-aligned* (by word, which almost all prior works employ) and *unaligned* (which is more challenging, and which MULT is generically designed for) multimodal language sequences.

8.5.1 Datasets

All datasets consist of monologue videos. The speaker's intentions are conveyed through three modalities: language, visual and acoustic.

Multimodal Sentiment Analysis involves analyzing speaker sentiment based on video content. Multimodal sentiment analysis extends conventional language-based sentiment analysis to a multimodal setup where both verbal and non-verbal signals contribute to the expression of sentiment. We use **CMU-MOSI** [711] which consists of 2199 opinion segments from online videos each annotated with sentiment in the range [-3,3].

Multimodal Emotion Recognition involves identifying speaker emotions based on both verbal and nonverbal behaviors. We perform experiments on the **IEMOCAP** dataset [74] which consists of 7318 segments of recorded dyadic dialogues annotated for the presence of human emotions happiness, sadness, anger and neutral.

Multimodal Speaker Traits Recognition involves recognizing speaker traits based on multimodal communicative behaviors. **POM** [467] contains 903 movie review videos each annotated for 12 speaker traits: confident (con), passionate (pas), voice pleasant (voi), credible (cre), vivid (viv), expertise (exp), reserved (res), trusting (tru), relaxed (rel), thorough (tho), nervous (ner), persuasive (per) and humorous (hum).

Each task consists of a *word-aligned* (processed in the same way as in prior works) and an *unaligned* version. For both versions, the multimodal features are extracted from the textual (GloVe word embeddings [475]), visual (Facet [270]), and acoustic (COVAREP [136]) data modalities. A more detailed introduction to the features is included in the full paper [359].

For the word-aligned version, following [482, 614, 713], we first use P2FA [707] to obtain the aligned timesteps (segmented w.r.t. words) for audio and vision streams, and we then perform averaging on the audio and vision features within these time ranges. All sequences in the word-aligned case have length 50. The process remains the same across all the datasets. On the other hand, for the unaligned version, we keep the original audio and visual features as extracted, without any word-segmented alignment or manual subsampling. As a result, the lengths of each modality vary significantly, where audio and vision sequences may contain up to $> 1,000$ time steps. We elaborate on the three tasks below.

8.5.2 Multimodal features and alignment

GloVe word embeddings [475], Facet [270] and COVAREP [136] are extracted for the language, visual and acoustic modalities respectively [1]. Forced alignment is performed using P2FA [707] to obtain the exact utterance times of each word. We obtain the aligned video and audio features by computing the expectation of their modality feature values over each word utterance time interval [614].

8.5.3 Baseline models

We compare to the following models for multimodal machine learning: MFN [713] synchronizes multimodal sequences using a multi-view gated memory. It is the current state of the art on CMU-MOSI and POM. MARN [714] models intra-modal and cross-modal interactions using multiple attention coefficients and hybrid LSTM memory components. GME-LSTM(A) [101] learns binary gating mechanisms to remove noisy modalities that are contradictory or redundant for prediction. TFN [712] models unimodal, bimodal and trimodal interactions using tensor products. BC-LSTM [490] performs context-dependent sentiment analysis and emotion recognition, currently state of the art on IEMOCAP. EF-LSTM concatenates the multimodal inputs and uses that as input to a single LSTM [241]. We also implement the Stacked, (EF-SLSTM) [208] Bidirectional (EF-BLSTM) [528] and Stacked Bidirectional (EF-SBLSTM) LSTMs.

[1] Details on feature extraction are in supplementary.

Table 8.1: Results for multimodal sentiment analysis on CMU-MOSI with aligned and non-aligned multimodal sequences. ^{h} means higher is better and $^\ell$ means lower is better. EF stands for early fusion, and LF stands for late fusion.

Metric	Acc_7 ↑	Acc_2 ↑	F1 ↑	MAE ↓	Corr ↑
(Word Aligned) **CMU-MOSI Sentiment**					
EF-LSTM	33.7	75.3	75.2	1.023	0.608
LF-LSTM	35.3	76.8	76.7	1.015	0.625
RMFN [359]	38.3	78.4	78.0	0.922	0.681
MFM [614]	36.2	78.1	78.1	0.951	0.662
RAVEN [655]	33.2	78.0	76.6	0.915	**0.691**
MCTN [482]	35.6	79.3	79.1	0.909	0.676
MulT (ours)	**40.0**	**83.0**	**82.8**	**0.871**	**0.698**
(Unaligned) **CMU-MOSI Sentiment**					
CTC [209] + EF-LSTM	31.0	73.6	74.5	1.078	0.542
LF-LSTM	33.7	77.6	77.8	0.988	0.624
CTC + MCTN [482]	32.7	75.9	76.4	0.991	0.613
CTC + RAVEN [655]	31.7	72.7	73.1	1.076	0.544
MulT (ours)	**39.1**	**81.1**	**81.0**	**0.889**	**0.686**

Table 8.2: Results for multimodal sentiment analysis on (relatively large scale) CMU-MOSEI with aligned and non-aligned multimodal sequences.

Metric	Acc_7 ↑	Acc_2 ↑	F1 ↑	MAE ↓	Corr ↑
(Word Aligned) **CMU-MOSEI Sentiment**					
EF-LSTM	47.4	78.2	77.9	0.642	0.616
LF-LSTM	48.8	80.6	80.6	0.619	0.659
Graph-MFN [717]	45.0	76.9	77.0	0.71	0.54
RAVEN [655]	50.0	79.1	79.5	0.614	0.662
MCTN [482]	49.6	79.8	80.6	0.609	0.670
MulT (ours)	**51.8**	**82.5**	**82.3**	**0.580**	**0.703**
(Unaligned) **CMU-MOSEI Sentiment**					
CTC [209] + EF-LSTM	46.3	76.1	75.9	0.680	0.585
LF-LSTM	48.8	77.5	78.2	0.624	0.656
CTC + RAVEN [655]	45.5	75.4	75.7	0.664	0.599
CTC + MCTN [482]	48.2	79.3	79.7	0.631	0.645
MulT (ours)	**50.7**	**81.6**	**81.6**	**0.591**	**0.694**

8.5.4 Evaluation metrics

For classification, we report accuracy Ac where c denotes the number of classes and F1 score. For regression, we report Mean Absolute Error MAE and Pearson's correlation r. For MAE lower values indicate stronger performance. For all remaining metrics, higher values indicate stronger performance.

Table 8.3: Results for multimodal emotions analysis on IEMOCAP with aligned and non-aligned multimodal sequences.

Task	Happy		Sad		Angry		Neutral	
Metric	$Acc_2 \uparrow$	F1↑	$Acc_2 \uparrow$	F1 ↑	$Acc_2 \uparrow$	F1 ↑	$Acc_2 \uparrow$	F1 ↑
(Word Aligned) **IEMOCAP Emotions**								
EF-LSTM	86.0	84.2	80.2	80.5	85.2	84.5	67.8	67.1
LF-LSTM	85.1	86.3	78.9	81.7	84.7	83.0	67.1	67.6
RMFN [359]	87.5	85.8	83.8	82.9	85.1	84.6	69.5	69.1
MFM [614]	90.2	85.8	**88.4**	**86.1**	**87.5**	86.7	72.1	68.1
RAVEN [655]	87.3	85.8	83.4	83.1	**87.3**	86.7	69.7	69.3
MCTN [482]	84.9	83.1	80.5	79.6	79.7	80.4	62.3	57.0
MulT (ours)	**90.7**	**88.6**	86.7	**86.0**	**87.4**	**87.0**	**72.4**	**70.7**
(Unaligned) **IEMOCAP Emotions**								
CTC [209] + EF-LSTM	76.2	75.7	70.2	70.5	72.7	67.1	58.1	57.4
LF-LSTM	72.5	71.8	72.9	70.4	68.6	67.9	59.6	56.2
CTC + RAVEN [655]	77.0	76.8	67.6	65.6	65.0	64.1	**62.0**	**59.5**
CTC + MCTN [482]	80.5	77.5	72.0	71.7	64.9	65.6	49.4	49.3
MulT (ours)	**84.8**	**81.9**	**77.7**	**74.1**	**73.9**	**70.2**	**62.5**	**59.7**

Table 8.4: Results for personality trait recognition on POM. Best results are highlighted in bold and Δ_{SOTA} shows improvement over previous SOTA. Symbols denote baseline model which achieves the reported performance: MFN: $\star$, MARN: §, BC-LSTM: $\bullet$, TFN: †, MV-LSTM: #, EF-LSTM: $\flat$, RF: $\heartsuit$, SVM: $\times$. The MFP outperforms the current SOTA across all evaluation metrics except the Δ_{SOTA} entries highlighted in gray. Improvements are highlighted in green.

Dataset	**POM Speaker Personality Traits**											
Task	Con	Pas	Voi	Cre	Viv	Exp	Res	Rel	Tho	Ner	Per	Hum
Metric	$Acc_7 \uparrow$	$Acc_7 \uparrow$	$Acc_7 \uparrow$	$Acc_7 \uparrow$	$Acc_7 \uparrow$	$Acc_7 \uparrow$	$Acc_5 \uparrow$	$Acc_5 \uparrow$	$Acc_5 \uparrow$	$Acc_5 \uparrow$	$Acc_7 \uparrow$	$Acc_6 \uparrow$
EF-LSTM	25.1	30.5	34.0	36.9	29.6	32.5	31.0	48.3	42.4	44.8	25.6	39.4
MV-LSTM	25.6	28.6	28.1	25.6	32.5	29.6	33.0	50.7	37.9	42.4	26.1	38.9
BC-LSTM	26.6	26.6	31.0	27.6	36.5	30.5	33.0	47.3	45.8	36.0	27.1	36.5
TFN	24.1	31.0	31.5	24.6	25.6	27.6	30.5	35.5	33.0	42.4	27.6	33.0
MFN	34.5	35.5	**37.4**	34.5	36.9	36.0	38.4	53.2	47.3	47.8	34.0	**47.3**
RMFN (ours)	**37.4**	**38.4**	**37.4**	**37.4**	**38.9**	**38.9**	**39.4**	**53.7**	**48.3**	**48.3**	**35.0**	46.8

8.6 Results and Discussion

8.6.1 Overall performance on multimodal language

Word-Aligned Experiments. We first evaluate RMFN and MULT on the *word-aligned sequences*— the "home turf" of prior approaches modeling human multimodal language [482, 541, 614, 655]. The upper part of the Table 8.1, 8.2, 8.3, and 8.4 show the results of our proposed approaches and previous baselines on the word-aligned task. With similar model sizes (around 200K parameters), MULT outperforms the other competitive approaches on different metrics on all tasks, with the exception of the "sad" class results on IEMOCAP. We also observe that RMFN does not improve results on IEMOCAP neutral emotion and the model outperforming RMFN is a memory-based fusion baseline [713]. We believe that this is because neutral expressions are quite idiosyncratic. Some people may always look angry given their facial configuration (e.g., natural eyebrow raises of actor Jack Nicholson). In these situations, it becomes useful to compare the

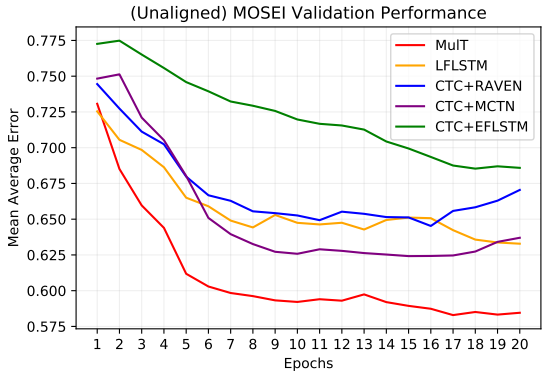


Figure 8.7: Validation set convergence of MULT when compared to other baselines on the unaligned CMU-MOSEI task.

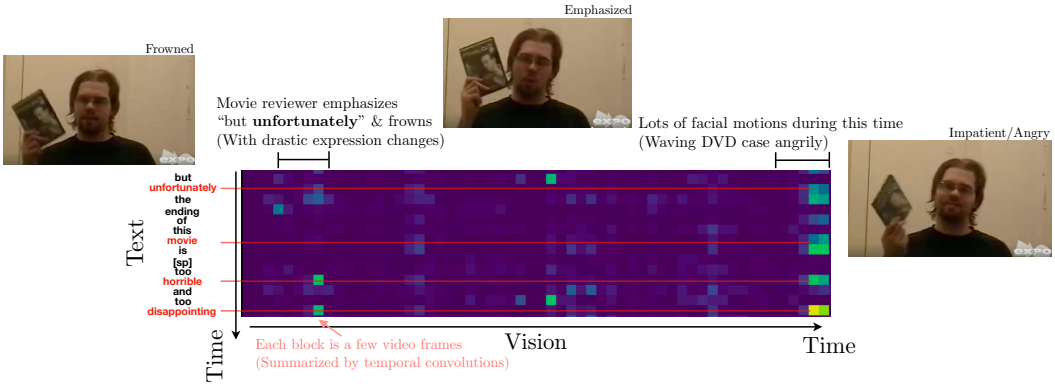


Figure 8.8: Visualization of sample crossmodal attention weights from layer 3 of $[V \rightarrow L]$ crossmodal transformer on CMU-MOSEI. We found that the crossmodal attention has learned to correlate certain meaningful words (e.g., "movie", "disappointing") with segments of stronger visual signals (typically stronger facial motions or expression change), despite the lack of alignment between original L/V sequences. Note that due to temporal convolution, each textual/visual feature contains the representation of nearby elements.

current image with a memorized or aggregated representation of the speaker's face. Our proposed multistage fusion approach can easily be extended to memory-based fusion methods.

Unaligned Experiments. Next, we evaluate MULT on the same set of datasets in the unaligned setting. Note that MULT can be directly applied to unaligned multimodal stream, while the baseline models (except for LF-LSTM) require the need of additional alignment module (e.g., CTC module).

The results are shown in the bottom part of Table 8.1, 8.2, and 8.3. On the three benchmark datasets, MULT improves upon the prior methods (some with CTC) by 10%-15% on most attributes. Empirically, we find that MULT converges faster to better results at training when compared to other competitive approaches (see Figure 8.7). In addition, while we note that in general there is a performance drop on all models when we shift from the word-aligned to unaligned multimodal time-series, the impact MULT takes is much smaller than the other approaches. We hypothesize such performance drop occurs because the asynchronous (and much

Table 8.5: Effect of varying the number of stages on CMU-MOSI sentiment analysis performance. Multistage fusion improves performance as compared to single stage fusion.

Dataset	CMU-MOSI				
Task	Sentiment				
Metric	A2 ↑	F1 ↑	A7 ↑	MAE ↓	Corr ↑
RMFN-R1	75.5	75.5	35.1	0.997	0.653
RMFN-R2	76.4	76.4	34.5	0.967	0.642
RMFN-R3	**78.4**	**78.0**	**38.3**	**0.922**	**0.681**
RMFN-R4	76.0	76.0	36.0	0.999	0.640
RMFN-R5	75.5	75.5	30.9	1.009	0.617
RMFN-R6	70.4	70.5	30.8	1.109	0.560
RMFN	**78.4**	**78.0**	**38.3**	**0.922**	**0.681**

Table 8.6: Comparison studies of RMFN on CMU-MOSI. Modeling cross-modal interactions using multistage fusion and attention weights are crucial in multimodal language analysis.

Dataset	CMU-MOSI				
Task	Sentiment				
Metric	A2 ↑	F1 ↑	A7 ↑	MAE ↓	Corr ↑
MARN	77.1	77.0	34.7	0.968	0.625
RMFN (no MFP)	76.5	76.5	30.8	0.998	0.582
RMFN (no `HIGHLIGHT`)	77.9	77.9	35.9	0.952	0.666
RMFN	**78.4**	**78.0**	**38.3**	**0.922**	**0.681**

longer) data streams introduce more difficulty in recognizing important features and computing the appropriate attention.

8.6.2 Deeper analysis of RMFN

Ablation studies: To achieve a deeper understanding of the multistage fusion process, we study five research questions. (Q1): whether modeling cross-modal interactions across multiple stages is beneficial. (Q2): the effect of the number of stages K during multistage fusion on performance. (Q3): the comparison between multistage and independent modeling of cross-modal interactions. (Q4): whether modeling cross-modal interactions are helpful. (Q5): whether attention weights from the `HIGHLIGHT` module are required for modeling cross-modal interactions.

Q1: To study the effectiveness of the multistage fusion process, we test the baseline RMFN-R1 which performs fusion in only one stage instead of across multiple stages. This model makes the strong assumption that all cross-modal interactions can be modeled during only one stage. From Table 8.5, RMFN-R1 underperforms as compared to RMFN which performs multistage fusion.

Q2: We test baselines RMFN-RK which perform K stages of fusion. From Table 8.5, we observe that increasing the number of stages K increases the model's capability to model cross-modal interactions up to a certain point ($K = 3$) in our experiments. Further increases led to decreases in performance and we hypothesize this is due to overfitting on the dataset.

Q3: To compare multistage against independent modeling of cross-modal interactions, we pay close attention to the performance comparison with respect to MARN which models multiple cross-modal interactions all at once (see Table 8.6). RMFN shows improved performance, indicating that multistage fusion is both effective and efficient for human multimodal language modeling.

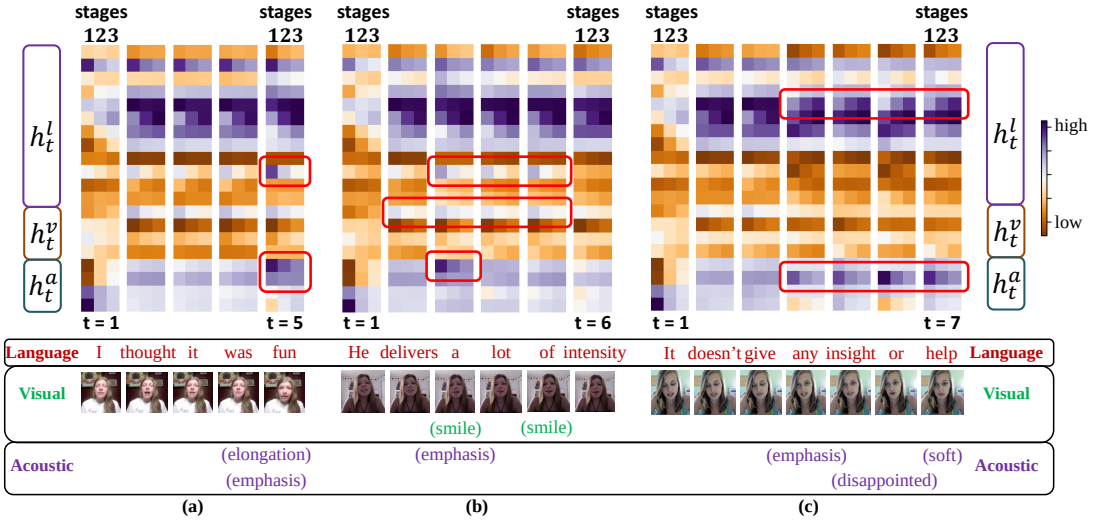


Figure 8.9: Visualization of learned attention weights across stages 1,2 and 3 of the multistage fusion process and across time of the multimodal sequence. We observe that the attention weights are diverse and evolve across stages and time. In these three examples, the red boxes emphasize specific moments of interest. (a) Synchronized interactions: the positive word "fun" and the acoustic behaviors of emphasis and elongation ($t = 5$) are synchronized in both attention weights for language and acoustic features. (b) Asynchronous trimodal interactions: the asynchronous presence of a smile ($t = 2 : 5$) and emphasis ($t = 3$) help to disambiguate the language modality. (c) Bimodal interactions: the interactions between the language and acoustic modalities are highlighted by alternating stages of fusion ($t = 4 : 7$).

Q4: RMFN (no MFP) represents a system of LSTHMs without the integration of $\mathbf{z}_t$ from the MFP to model cross-modal interactions. From Table 8.6, RMFN (no MFP) is outperformed by RMFN, confirming that modeling cross-modal interactions is crucial in analyzing human multimodal language.

Q5: RMFN (no `HIGHLIGHT`) removes the `HIGHLIGHT` module from MFP during multistage fusion. From Table 8.6, RMFN (no `HIGHLIGHT`) underperforms, indicating that highlighting multimodal representations using attention weights are important for modeling cross-modal interactions.

Visualizations of learned fusion patterns: Using an attention assignment mechanism during the `HIGHLIGHT` process gives more interpretability to the model since it allows us to visualize the attended multimodal signals at each stage and time step (see Figure 8.9). Using RMFN trained on the CMU-MOSI dataset, we plot the attention weights across the multistage fusion process for three videos in CMU-MOSI. Based on these visualizations we first draw the following general observations on multistage fusion:

Across stages: Attention weights change their behaviors across the multiple stages of fusion. Some features are highlighted by earlier stages while other features are used in later stages. This supports our hypothesis that RMFN learns to specialize in different stages of the fusion process.

Across time: Attention weights vary over time and adapt to the multimodal inputs. We observe that the attention weights are similar if the input contains no new information. As soon as new multimodal information comes in, the highlighting mechanism in RMFN adapts to these new inputs.

Priors: Based on the distribution of attention weights, we observe that the language and acoustic modalities seem the most commonly highlighted. This represents a prior over the expression of sentiment in human multimodal language and is closely related to the strong connections between language and speech in human communication [322].

Inactivity: Some attention coefficients are not active (always orange) throughout time. We hypothesize that these corresponding dimensions carry only intra-modal dynamics and are not involved in the formation of cross-modal interactions.

In addition to the general observations above, Figure 8.9 shows three examples where multi-stage fusion learns cross-modal representations across three different scenarios.

Synchronized Interactions: In Figure 8.9(a), the language features are highlighted corresponding to the utterance of the word "fun" that is highly indicative of sentiment (t = 5). This sudden change is also accompanied by a synchronized highlighting of the acoustic features. We also notice that the highlighting of the acoustic features lasts longer across the 3 stages since it may take multiple stages to interpret all the new acoustic behaviors (elongated tone of voice and phonological emphasis).

Asynchronous Trimodal Interactions: In Figure 8.9(b), the language modality displays ambiguous sentiment: "delivers a lot of intensity" can be inferred as both positive or negative. We observe that the circled attention units in the visual and acoustic features correspond to the asynchronous presence of a smile (t = $2:5$) and phonological emphasis (t = 3) respectively. These nonverbal behaviors resolve ambiguity in language and result in an overall display of positive sentiment. We further note the coupling of attention weights that highlight the language, visual and acoustic features across stages (t = $3:5$), further emphasizing the coordination of all three modalities during multistage fusion despite their asynchronous occurrences.

Bimodal Interactions: In Figure 8.9(c), the language modality is better interpreted in the context of acoustic behaviors. The disappointed tone and soft voice provide the nonverbal information useful for sentiment inference. This example highlights the bimodal interactions (t = $4:7$) in alternating stages: the acoustic features are highlighted more in earlier stages while the language features are highlighted increasingly in later stages.

8.6.3 Deeper analysis of MULT

Ablation studies: To further study the influence of the individual components in MULT, we perform comprehensive ablation analysis using the unaligned version of CMU-MOSEI. The results are shown in Table 8.7.

First, we consider the performance for only using unimodal transformers (i.e., language, audio or vision only). We find that the language transformer outperforms the other two by a large margin. For example, for the Acc_2^h metric, the model improves from 65.6 to 77.4 when comparing audio only to language only unimodal transformer. This fact aligns with the observations in prior work [482], where the authors found that a good language network could already achieve good performance at inference time.

Second, we consider 1) a late-fusion transformer that feature-wise concatenates the last elements of three self-attention transformers; and 2) an early-fusion self-attention transformer that takes in a temporal concatenation of three asynchronous sequences $[\hat{X}_L, \hat{X}_V, \hat{X}_A] \in \mathbb{R}^{(T_L+T_V+T_A)\times d_q}$

Table 8.7: An ablation study on the benefit of MulT's crossmodal transformers using CMU-MOSEI).

Description	(Unaligned) CMU-MOSEI Sentiment				
	Acc_7^h	Acc_2^h	F1^h	MAE^ℓ	Corr^h
Unimodal Transformers					
Language only	46.5	77.4	78.2	0.653	0.631
Audio only	41.4	65.6	68.8	0.764	0.310
Vision only	43.5	66.4	69.3	0.759	0.343
Late Fusion by using Multiple Unimodal Transformers					
LF-Transformer	47.9	78.6	78.5	0.636	0.658
Temporally Concatenated Early Fusion Transformer					
EF-Transformer	47.8	78.9	78.8	0.648	0.647
Multimodal Transfomers					
Only $[V, A \rightarrow L]$ (ours)	**50.5**	80.1	80.4	0.605	0.670
Only $[L, A \rightarrow V]$ (ours)	48.2	79.7	80.2	0.611	0.651
Only $[L, V \rightarrow A]$ (ours)	47.5	79.2	79.7	0.620	0.648
MulT mixing intermediate-level features (ours)	50.3	80.5	80.6	0.602	0.674
MulT (ours)	**50.7**	**81.6**	**81.6**	**0.591**	**0.691**

(see Section 8.4.2). Empirically, we find that both EF- and LF-Transformer (which fuse multimodal signals) outperform unimodal transformers.

Finally, we study the importance of individual crossmodal transformers according to the target modalities (i.e., using $[V, A \rightarrow L]$, $[L, A \rightarrow V]$, or $[L, V \rightarrow A]$ network). As shown in Table 8.7, we find crossmodal attention modules consistently improve over the late- and early-fusion transformer models in most metrics on unaligned CMU-MOSEI. In particular, among the three crossmodal transformers, the one where language(L) is the target modality works best. We also additionally study the effect of adapting intermediate-level instead of the low-level features from source modality in crossmodal attention blocks (similar to the NMT encoder-decoder architecture but without self-attention; see Section 8.4.1). While MULT leveraging intermediate-level features still outperform models in other ablative settings, we empirically find adapting from low-level features works best. The ablations suggest that crossmodal attention concretely benefits MULT with better representation learning.

Qualitative analysis of learned cross-modal attention: To understand how crossmodal attention works while modeling unaligned multimodal data, we empirically inspect what kind of signals MULT picks up by visualizing the attention activations. Figure 8.8 shows an example of a section of the crossmodal attention matrix on layer 3 of the $V \rightarrow L$ network of MULT (the original matrix has dimension $T_L \times T_V$; the figure shows the attention corresponding to approximately a 6-sec short window of that matrix). We find that crossmodal attention has learned to attend to meaningful signals across the two modalities. For example, stronger attention is given to the intersection of words that tend to suggest emotions (e.g., "movie", "disappointing") and drastic facial expression changes in the video (start and end of the above vision sequence). This observation advocates one of the aforementioned advantage of MULT over conventional alignment (see Section 8.4.3): crossmodal attention enables MULT to directly capture potentially long-range signals, including those off-diagonals on the attention matrix.

8.7 Conclusion

This chapter proposed the RECURRENT MULTISTAGE FUSION NETWORK (RMFN) and Multimodal Transformer (MULT) architectures or analyzing human multimodal language. RMFN which recursively decomposes the multimodal fusion problem into multiple stages, each focused on learning interactions from a subset of attended multimodal signals. MULT uses the crossmodal attention module to learn multimodal interactions between all elements in the first modality with all elements in the second modality. As a result, all multimodal interactions across the entire sequence are learned simultaneously, and can be parallelized efficiently over GPUs.

Both methods show strong results on multiple datasets with multimodal temporal data (e.g., human communication), displaying capabilities to capture long-range multimodal interactions, handling unaligned multimodal data, and learning redundant, unique, and synergistic interactions.

Chapter 9

Training High-modality Foundation Models

9.1 Introduction

Finally, using MULTIBENCH, we scale multimodal transformers to the high-modality setting where there are a large number of modalities partially observed for different tasks [370]. While there have been impressive advances in modeling language, vision, and audio [11, 502], advances in sensing technologies have resulted in many real-world platforms such as cellphones, smart devices, self-driving cars, healthcare technologies, and robots now integrating a much larger number of sensors such as time-series, proprioception, sets, tables, and high-frequency sensors [53, 179, 335, 340, 366, 697]. This new setting of *high-modality learning* involves learning representations over many diverse modality inputs. As more modalities are introduced, adding new model parameters for every new modality or task [283, 390, 613] becomes prohibitively expensive and not scalable [371]. A critical technical challenge for efficient high-modality learning, therefore, is *heterogeneity quantification*: how can we measure which modalities encode *similar information* and *similar interactions* in order to permit parameter sharing with previous modalities (see Figure 9.1)? For example, how can one determine whether the same modality encoder can be shared when processing language and speech, or that the same fusion network can be shared when fusing human speech and gestures as well as robot visual and force sensors?

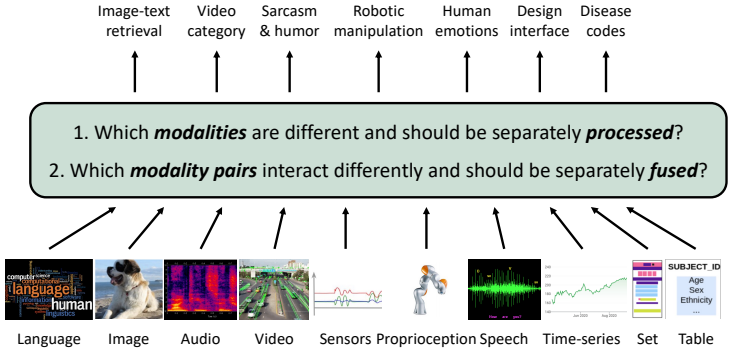


Figure 9.1: Heterogeneity quantification: Efficiently learning from many modalities requires measuring (1) *modality heterogeneity*: which modalities are different and should be separately processed, and (2) *interaction heterogeneity*: which modality pairs interact differently and should be separately fused. HIGHMMT uses these measurements to dynamically group parameters balancing performance and efficiency.

In this paper, we propose a principled approach for heterogeneity quantification via modality information transfer, an approach that measures the amount of transferable information from one modality to another. Our first proposed metric, (1) *modality heterogeneity* studies how similar

2 modalities $\{X_1, X_2\}$ are by measuring how much usable information can be transferred from X_1 to X_2, and our second metric, (2) *interaction heterogeneity* studies how similarly 2 modality pairs $\{X_1, X_2\}, \{X_3, X_4\}$ interact by measuring how much usable interaction information can be transferred from $\{X_1, X_2\}$ to $\{X_3, X_4\}$. We show the importance of these 2 proposed metrics in high-modality scenarios as a way to automatically prioritize the fusion of modalities that contain unique information or unique interactions, and otherwise sharing parameters across similar modalities displaying similar information or interactions.

Operationalizing these ideas on a suite of 10 modalities, 15 prediction tasks, and 5 research areas, we show how to train a single model, HIGHMMT, that (1) improves the tradeoff between performance and efficiency over task-specific state-of-the-art models [283, 367], and general multimodal models with full parameter sharing [15, 253, 276, 505], (2) enables cross-modal transfer by pretraining on source tasks before transferring to new target modalities and tasks, and (3) is especially beneficial for low-resource scenarios (less training data and partially-observable modalities). Beyond these empirical results, we believe that our insights on quantifying heterogeneity and information sharing in multimodal models are independently useful for future work.

9.2 HIGH-MODALITY MULTIMODAL TRANSFORMER

In this section, we describe our overall approach for high-modality representation learning (see Figure 9.2). In §9.2.1, we formalize modality and interaction heterogeneity to understand whether modalities should be processed similarly or differently. Using these insights, §9.2.2 describes our proposed HIGHMMT model with dynamic parameter sharing based on heterogeneity measurements.

9.2.1 Measuring heterogeneity via modality information transfer

We begin our motivation by formalizing two important sources of heterogeneity in multimodal tasks. Firstly, *modality heterogeneity* occurs because the information present in different modalities often shows diverse qualities, structures, and representations. Secondly, *interaction heterogeneity* occurs because different modalities interact differently to give rise to new information when used for task inference. Formalizing and measuring these two sources of heterogeneity results in actionable insights for building multimodal models: measuring modality heterogeneity enables us to answer: should I use the same unimodal model to encode X_1 and X_2? Measuring interaction heterogeneity enables us to answer: should I use the same fusion model to fuse $\{X_1, X_2\}$ and $\{X_3, X_4\}$? We will formalize heterogeneity via *modality transfer*, an approach that measures the amount of transferable information from one modality to another.

Estimating modality heterogeneity via unimodal information transfer. We propose to measure heterogeneity between modalities X_1 and X_2 via unimodal transfer. Given a task Y defined over X_1 and X_2, how well does an unimodal model trained on the task $(X_1; Y)$ transfer to $(X_2; Y)$? We choose model transfer as our focus of heterogeneity since it is captured at the level of features extracted via representation learning, rather than at the data-level. Even though the input data may be very different (e.g., images from different cameras or paraphrased sentences), effective feature extractors may be able to learn similar representations from them. Furthermore,

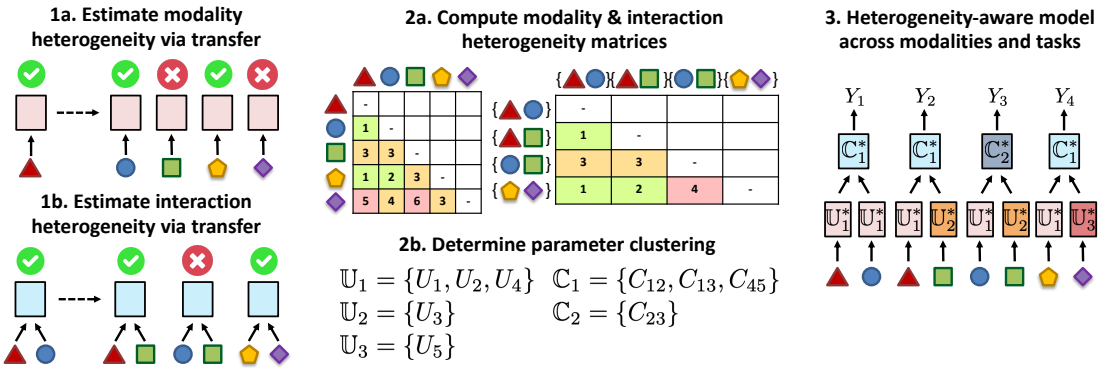


Figure 9.2: HIGHMMT workflow: (1) We estimate modality and interaction heterogeneity via modality transfer to determine which modalities should be processed and fused differently. (2) Using the inferred heterogeneity, we determine the optimal grouping of parameters balancing both total performance and parameter efficiency, which (3) informs our design of a heterogeneity-aware model with dynamic parameter sharing across many modalities and tasks. HIGHMMT enables statistical strength sharing, efficiency, and generalization to new modalities and tasks.

it directly models task-relevance: the degree of heterogeneity depends on the end task, which enables using these heterogeneity measures subsequently for end-task optimization.

We formalize unimodal transfer as the difference in performance between unimodal models trained on X_1 before transfer to X_2, versus those trained directly on X_2. Specifically, we represent an unimodal model using modality X_2 with parameters θ as $\hat{y} = f(y|x_2;\theta)$. For a suitably chosen loss function $\ell(\hat{y}, y)$, define the loss of a model as $\mathbb{E}_{p(x_2,y)}\ell(f(y|x_2;\theta), y)$ which measures the expected error over the joint distribution $p(x_2, y)$. To measure transfer, we train 2 models to obtain an approximation of task performance: the first randomly initialized and trained on the target task giving loss $\mathcal{L}_2^*$,

$$\mathcal{L}_2^* = \min_\theta \mathbb{E}_{p(x_2,y)}\ell(f(y|x_2;\theta), y), \tag{9.1}$$

and the second using initialization from model parameters θ_1 trained on the source task $(X_1; Y)$ before fine-training on the target task giving loss $\mathcal{L}_{1\to2}^*$.

$$\theta_1 = \arg\min_\theta \mathbb{E}_{p(x_1,y)}\ell(f(y|x_1;\theta), y), \tag{9.2}$$

$$\mathcal{L}_{1\to2}^* = \min_\theta \mathbb{E}_{p(x_2,y)}\ell(f(y|x_2;\theta \leftarrow \theta_1), y), \tag{9.3}$$

where $\theta \leftarrow \theta_1$ denotes parameter initialization with θ_1. Intuitively, $\mathcal{L}_2^*$ measures the (baseline) task-relevant information in X_2, while $\mathcal{L}_{1\to2}^*$ measures the task-relevant information transferable from X_1 to X_2. The differences between these 2 losses,

$$T(X_1 \to X_2; Y) = \mathcal{L}_{1\to2}^* - \mathcal{L}_2^*, \tag{9.4}$$

therefore measures the difficulty of transferring a model trained on the source task $(X_1; Y)$ to a target task $(X_2; Y)$. Note that computing $T(X_1 \to X_2; Y)$ only requires the training or fine-tuning of 2 models across the source and target modalities, which is efficient. In practice, the expectations

over $p(x_1, y)$ and $p(x_2, y)$ are approximated using empirical samples from the training set (for model fine-tuning) and validation dataset (for final evaluation of performance).

What are some properties of $T(X_1 \to X_2; Y)$? For very different modalities X_1 and X_2, we typically expect a source task (X_1, Y) to contain less usable information for a target task $(X_2; Y)$, which would imply that $\mathcal{L}^*_{1\to 2} \geq \mathcal{L}^*_2$ and therefore $T(X_1 \to X_2; Y) \geq 0$ (i.e., positive distance). This is consistent with work demonstrating negative transfer across different modalities [367, 369, 623, 657]. Under these scenarios, the larger the positive magnitude of $T(X_1 \to X_2; Y)$, the more different modalities X_1 and X_2 are in the context of task Y (more difficult to transfer). However, there can also be cases of zero or even positive transfer (i.e., $T(X_1 \to X_2; Y) \leq 0$), even in the surprising case of very different modalities [391]. These cases reinforce the benefits of feature-based approaches to measure heterogeneity: while the raw modalities themselves seem very different, they can still be processed by similar models resulting in positive transfer, and should be assigned a difference of 0. Our final heterogeneity measure $d(X_1; X_2)$ aggregates the non-negative value (to account for positive transfer) of transfer difficulty statistics across both transfer directions $X_1 \to X_2$ and $X_2 \to X_1$:

$$d(X_1; X_2) = T(X_1 \to X_2; Y)_{\geq 0} + T(X_2 \to X_1; Y)_{\geq 0}. \tag{9.5}$$

where $x_{\geq 0} = \max(x, 0)$. Under certain assumptions on the modalities and tasks, our modality heterogeneity measure $d(X_1; X_2)$ is a metric: it satisfies *non-negativity*: $d(X_1; X_2) \geq 0$, with $d(X_1; X_2) = 0$ when $X_1 = X_2$, and *symmetry*: $d(X_1; X_2) = d(X_2; X_1)$, *positivity*, $X_1 \neq X_2$ implies that $d(X_1; X_2) > 0$, and a relaxed version of the *triangle inequality*: $d(X_1; X_3) \leq d(X_1; X_2) + d(X_2; X_3)$. However, in the most general case, there may be settings where positivity and the triangle inequality are not satisfied since the exact dynamics of transfer learning is still not well understood for general deep networks: positive transfer can happen (which would imply cases of $X_1 \neq X_2$ but $d(X_1; X_2) = 0$), and in practice, the relaxed triangle inequality is satisfied 96% of the time from a real heterogeneity matrix in Figure 9.5.

Estimating interaction heterogeneity via crossmodal information transfer. We are also interested in interaction heterogeneity: specifically, how differently should I fuse modalities $\{X_1, X_2\}$ versus $\{X_3, X_4\}$? We therefore extend to crossmodal transfer by comparing the difference in performance between a multimodal model pretrained on $(X_1, X_2; Y)$ before transfer to $(X_3, X_4; Y)$, versus those trained directly on the target task $(X_3, X_4; Y)$. In other words, we measure the difference in loss between

$$\theta_{12} = \arg\min_\theta \mathbb{E}_{p(x_1,x_2,y)} \ell(f(y|x_1, x_2; \theta), y), \tag{9.6}$$

$$\mathcal{L}^*_{12\to 34} = \min_\theta \mathbb{E}_{p(x_3,x_4,y)} \ell(f(y|x_3, x_4; \theta \leftarrow \theta_{12}), y), \tag{9.7}$$

and direct training

$$\mathcal{L}^*_{34} = \min_\theta \mathbb{E}_{p(x_3,x_4,y)} \ell(f(y|x_3, x_4; \theta), y), \tag{9.8}$$

to obtain $T(X_1, X_2 \to X_3, X_4; Y) = \mathcal{L}^*_{12\to 34} - \mathcal{L}^*_{34}$. The distance $d(X_1, X_2; X_3, X_4)$ after aggregation over tasks and transfer directions estimates the interaction heterogeneity between $\{X_1, X_2\}$ and $\{X_3, X_4\}$.

Modality and interaction heterogeneity matrix. Finally, we construct a modality heterogeneity matrix $M_U(i,j) = d(X_i; X_j)$ and an interaction heterogeneity matrix (technically 4D-tensor) $M_C(i,j,k,\ell) = d(X_i, X_j; X_k, X_\ell)$. Determining parameter groupings to balance both total performance and parameter efficiency can be solved via agglomerative hierarchical clustering where modalities are nodes and heterogeneity measurements are edges. The number of clusters k is treated as a hyperparameter dependent on the parameter budget (see clustering examples in §9.3.1). Clustering on the modality heterogeneity matrix M_U results in a grouping of modalities based on similarity (e.g., $\mathcal{U}_1 = \{X_1, X_2, X_4\}, \mathcal{U}_2 = \{X_3\}, \mathcal{U}_3 = \{X_5\}$), and likewise for the crossmodal matrix M_C (e.g., $\mathcal{C}_1 = \{\{X_1, X_2\}, \{X_1, X_3\}, \{X_4, X_5\}\}, \mathcal{C}_2 = \{\{X_2, X_3\}, \mathcal{C}_3 = \{\{X_4, X_6\}, \{X_5, X_6\}\}$, and so on.

Computational complexity. In a high-modality setting, suppose we are given a suite of modalities and tasks of the form $\{(X_1, X_2, Y_1), (X_1, X_3, X_4, Y_2), ...\}$ and so on, where there are a total of M unique modality and task pairs $\{(X_1, Y_1), (X_2, Y_1), (X_1, Y_2), (X_3, Y_2), (X_4, Y_2), ...\}$. In practice, the number of unique (pairwise) interaction and task pairs $\{(X_1, X_2, Y_1), (X_1, X_3, Y_2), ...\}$ is also $O(M)$, since the maximum number of modalities jointly observed for a task is never above a constant (at most 4 in all real-world datasets, and often 2 or 3). As an example in Figure 9.5, our experiments involve $M = 10$ modality and task pairs (across 4 tasks defined on $2, 2, 3$ and 3 modalities respectively), and $8 = \binom{2}{2} + \binom{2}{2} + \binom{3}{2} + \binom{3}{2}$ interaction and task pairs.

The modality heterogeneity matrix for M unique modality and task pairs has $M(M-1)/2$ unique entries after removing the upper triangular portion due to symmetry and diagonal entries since $d(X_i, X_i) = 0$. Computing these $M(M-1)/2$ entries exactly requires one to first train M unimodal models (to estimate the M $\mathcal{L}^*_m$ terms) before fine-tuning $M(M-1)$ transfer models (to estimate the $M(M-1)$ $\mathcal{L}^*_{m\to n}$ terms), for a total of M^2 pre-trained and fine-tuned models. The interaction heterogeneity matrix also requires $O(M^2)$ models for exact computation. However, we find that a key approximation can be made in practice: the heterogeneity matrices are highly structured due to distances approximately satisfying the triangle inequality, which implies that we do not need to compute all entries and instead rely on low-rank reconstruction from partial entries in practice. In our experiments, even using a low-rank approximation of $r = 3$ is sufficient to approximate the entire matrix. This suggests that we do not need to exhaustively measure unimodal and interaction transfer between all modality pairs to enjoy the benefits of our proposed approach. Instead, running a random sample of $O(M)$ pairs of heterogeneity values, and imputing the rest of the heterogeneity matrix, is sufficient in practice. Please see an example heterogeneity quantification for real-world datasets in §9.3.1.

9.2.2 Capturing heterogeneity and homogeneity in HIGHMMT

Using these insights, we now describe our architecture for a general model HIGHMMT suitable for high-modality representation across many modalities and tasks (see Figure 9.3). Training the HIGHMMT model consists of 2 main steps (see Figure 9.4): (1) *homogeneous pre-training* of a fully shared model across all modalities, before (2) *heterogeneity-aware fine-tuning* to respect modality and interaction heterogeneity.

Homogeneous pre-training. We first design a homogeneous multimodal model fully shared across all modalities and tasks with the following key components (see Figure 9.3)

1. Standardized input sequence: We first standardize modalities as a sequence of embeddings,

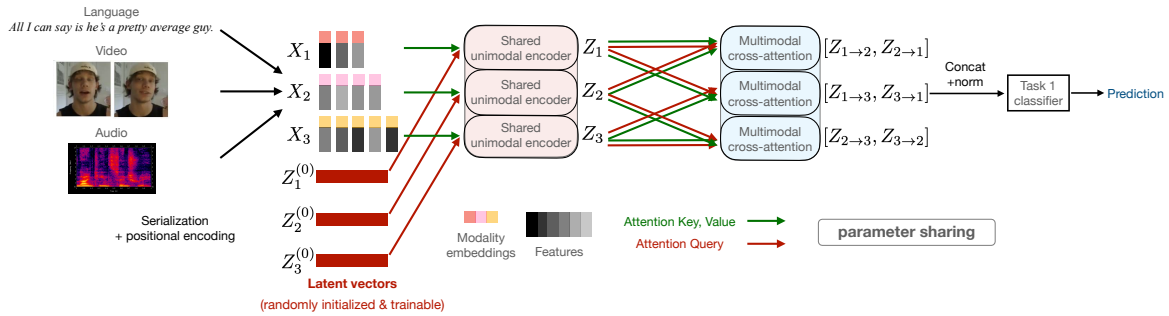


Figure 9.3: HIGHMMT architecture: Given arbitrary modalities, (1) the inputs are standardized into a sequence and padded, (2) modality embeddings and positional encodings are added to the input sequence, (3) a single shared unimodal Perceiver encoder is applied to all modalities to learn modality-agnostic representations, (4) each pair of unimodal representations is fed through a shared multimodal cross-attention layer to learn multimodal representations, and finally (5) all outputs are concatenated, batch-normalized, and fed into task-specific classification heads.

as is already done for sequential data such as text, audio, and time series, and recently adapted for image patches too [154]. For tables, sets, and graphs we treat each element in the table/set/graph as an element in the sequence. The end result is a standardized input data X_m of dimension $t_m \times d_m$, where t_m is a modality and task-specific input sequence length, and d_m is a modality and task-specific input dimension.

2. Modality-specific embedding and positional encoding. For each distinct modality $m \in M$ (which may appear across multiple tasks), we define a one-hot modality embedding $\mathbf{e}_m \in \mathbb{R}^{|M|}$, where $|M|$ is the total number of distinct modalities, to identify common modalities across different tasks for information sharing. We also introduce Fourier feature positional encodings $\mathbf{p}_m \in \mathbb{R}^{t_m \times d_{pm}}$, where d_{pm} is the positional encoding dimension, to capture positional information across each modality. For multimodal tasks where a common dimension is shared across time (e.g., videos/time series), we apply a common positional encoding to capture the common time dimension (i.e., the first image frame occurs at the same time as the first word and first audio segment). Finally, the processed modality m is given by concatenating $X_m = X_m \oplus \mathbf{e}_m \oplus \mathbf{p}_m \oplus \mathbf{0}_m$ (i.e., the input sequence, modality embedding, positional encodings and zero-padding) into a standard dimension $t_m \times d_{all}$. $d_{all} = \max_{m \in M}(d_m + |M| + d_{pm})$ where d_m is the channel size of modality m, d_{pm} is the positional encoding size of modality m, and $|M|$ is the modality encoding size (i.e., the total number of involved modalities).

3. Shared unimodal networks. Now that we have standardized all modalities into a common format, we design a general unimodal encoder with parameters $\mathbb{U}$ via a Transformer-based Perceiver block [276]. Our model recursively trains a latent array Z_m of shape $d_{LN} \times d_{LS}$ (where d_{LN} is the sequence length/number of latent vectors and d_{LS} is the latent dimension) that is random initialized as $Z_m^{(0)}$. For each layer L starting with a previously-computed representation $Z_m^{(L-1)}$, we first perform cross-attention from the processed input (X_m of shape $t_m \times d_{all}$) to $Z_m^{(L-1)}$ obtaining an intermediate representation $\tilde{Z}_m^{(L)}$, before self-attention and feed-forward layers on

$\tilde{Z}_m^{(L)}$ resulting in a new representation $Z_m^{(L)}$ for input to the next layer:

$$\tilde{Z}_m^{(L)} = \text{CROSS ATTENTION}(Z_m^{(L-1)}, X_m) = \text{softmax}\left(\frac{Q_c K_c^\top}{\sqrt{d_{LS}}}\right) V_c = \text{softmax}\left(\frac{Z_m^{(L-1)} W_{Q_c} W_{V_c}^\top X_m^\top}{\sqrt{d_{LS}}}\right) X_m W_{V_c}, \tag{9.9}$$

$$Z_m^{(L)} = \text{SELF ATTENTION}(\tilde{Z}_m^{(L)}) = \text{softmax}\left(\frac{Q_s K_s^\top}{\sqrt{d_{LS}}}\right) V_s = \text{softmax}\left(\frac{\tilde{Z}_m^{(L)} W_{Q_s} W_{V_s}^\top \tilde{Z}_m^{(L)\top}}{\sqrt{d_{LS}}}\right) \tilde{Z}_m^{(L)} W_{V_s}, \tag{9.10}$$

with trainable cross-attention parameters $W_{Q_c} \in \mathbb{R}^{d_{LS} \times d_{LS}}, W_{K_c} \in \mathbb{R}^{d_{all} \times d_{LS}}, W_{V_c} \in \mathbb{R}^{d_{all} \times d_{LS}}$ and self-attention parameters $W_{Q_s} \in \mathbb{R}^{d_{LS} \times d_{LS}}, W_{K_s} \in \mathbb{R}^{d_{LS} \times d_{LS}}, W_{V_s} \in \mathbb{R}^{d_{LS} \times d_{LS}}$. Repeating cross- and self-attention between the latent vector and the input modality summarizes the relationships between modality elements into the latent vector, resulting in a final unimodal representation $Z_m \in \mathbb{R}^{d_{LN} \times d_{LS}}$. Summarizing all information into a common $d_{LN} \times d_{LS}$ latent array regardless of the input shape $t_m \times d_{all}$ results in total runtime only linear with respect to the size of t_m and d_{all} which scales to high-modality scenarios.

4. Shared crossmodal networks. To learn multimodal representations, we use a shared Crossmodal Transformer block with parameters $\mathbb{C}$ [390, 613]. Given 2 unimodal representations Z_1 and Z_2 of common shape $d_{LN} \times d_{LS}$ learned from unimodal Perceiver encoders, a Crossmodal Transformer (CT) block uses crossmodal self-attention by setting the input layer query $Q = Z_1$ and keys and values $K, V = Z_2$ to learn attention from Z_2 to Z_1, and a separate block to capture the attention in the opposite direction.

$$Z_{2\to1} = \text{CROSS ATTENTION}(Z_1, Z_2) = \text{softmax}\left(\frac{Q_1 K_2^\top}{\sqrt{d_k}}\right) V_2 = \text{softmax}\left(\frac{Z_1 W_{Q_1} W_{V_2}^\top Z_2^\top}{\sqrt{d_k}}\right) Z_2 W_{V_2}, \tag{9.11}$$

and vice-versa for $Z_{1\to2}$, with parameters $W_{Q_1}, W_{Q_2} \in \mathbb{R}^{d_{LS} \times d_k}, W_{K_1}, W_{K_2} \in \mathbb{R}^{d_{LS} \times d_k}, W_{V_1}, W_{V_2} \in \mathbb{R}^{d_{LS} \times d_k}$. This step enables one modality's elements to discover bidirectional interactions with another, resulting in a final multimodal representation $Z_{\text{mm}} = [Z_{1\to2}, Z_{2\to1}]$ of shape $d_{LS} \times 2d_k$. For each layer, we first perform cross-attention followed by self-attention and feed-forward functions. For tasks with more than 2 modalities, a Crossmodal Transformer block is applied for each pair of modalities before concatenating all representations.

5. Task-specific classifier and multitask pre-training. Finally, on top of Z_{mm}, we use a separate linear classification layer per task. To enable information sharing across modalities and tasks, homogeneous pre-training is performed across a diverse set of datasets in a multitask manner by optimizing a weighted sum of losses over tasks. The result is a single set of shared unimodal parameters $\mathbb{U}^*$ that encodes all modalities, and a single set of shared crossmodal parameters $\mathbb{C}^*$ that captures all pairwise interactions between modality pairs, along with all modality-specific embeddings $\mathbb{E}^*$ and task-specific classifiers $\mathbb{T}^*$.

Heterogeneity-aware fine-tuning. Finally, we account for heterogeneity by grouping unimodal parameters based on modalities that we know to be similar from §9.2.1 (e.g., setting $\mathbb{U}_1 = \{U_1, U_2\}, \mathbb{U}_2 = \{U_3\}, \mathbb{U}_3 = \{U_4, U_5, U_6\}$), and likewise for the crossmodal parameters (e.g.,

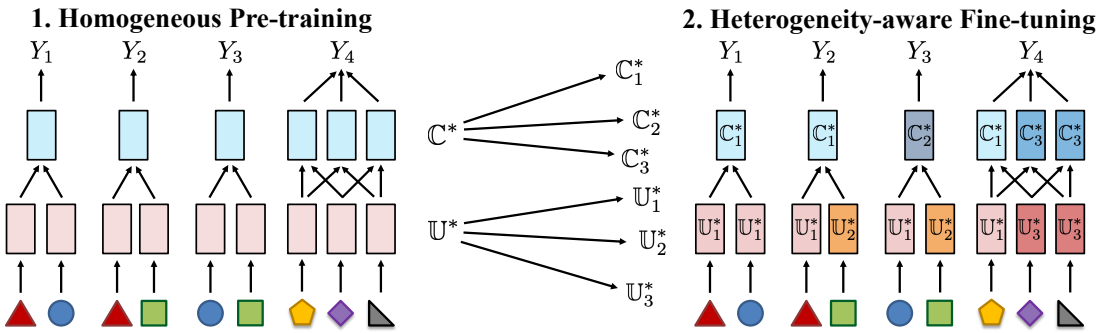


Figure 9.4: HIGHMMT training involves 2 steps: (1) *homogeneous pre-training* of a fully shared model across all modalities, before (2) *heterogeneity-aware fine-tuning* of modality and interaction parameters in different groups to respect modality and interaction heterogeneity respectively.

Table 9.1: We investigate a multitask setup to evaluate the performance of HIGHMMT across different modality inputs and prediction objectives. The total size of datasets involved in our experiments exceeds $370,000$ and covers diverse modalities, tasks, and research areas.

Datasets	Modalities	Size	Prediction task	Research Area
ENRICO	{image, set}	$1,460$	design interface	HCI
UR-FUNNY	{text, video, audio}	$16,514$	humor	Affective Computing
MOSEI	{text, video, audio}	$22,777$	sentiment, emotions	Affective Computing
MIMIC	{time-series, table}	$36,212$	mortality, ICD-9 codes	Healthcare
PUSH	{image, force, proprioception, control}	$37,990$	object pose	Robotics
AV-MNIST	{image, audio}	$70,000$	digit	Multimedia
V&T	{image, force, proprioception, depth}	$147,000$	contact, robot pose	Robotics

$\mathbb{C}_1 = \{C_{12}, C_{13}, C_{14}\}, \mathbb{C}_2 = \{C_{23}, C_{15}\}, \mathbb{C}_3 = \{C_{24}, ...\}$). From Figure 9.4, these parameter groups are first initialized with the homogeneous model $\mathbb{U}^*$ and $\mathbb{C}^*$ before separate fine-tuning, which results in final parameters $\mathbb{U}^* \rightarrow \{\mathbb{U}_1^*, \mathbb{U}_2^*, ...\}$ and $\mathbb{C}^* \rightarrow \{\mathbb{C}_1^*, \mathbb{C}_2^*, ...\}$. The modality embeddings $\mathbb{E}^*$ and task classifiers $\mathbb{T}^*$ are jointly fine-tuned as well. Fine-tuning is also performed in a multitask manner by optimizing a weighted sum of supervised losses across all modalities and tasks.

9.3 Experiments

Setup: In this section, we design experiments to analyze the multitask, transfer, and generalization capabilities of HIGHMMT. We use a large collection of multimodal datasets provided in MultiBench [367] spanning 10 modalities, 15 prediction tasks, and 5 research areas. We trained 3 multitask models across combinations of these datasets (see Table 9.1 for details). Overall, the total size of datasets involved in our experiments exceeds $370,000$ and covers diverse modalities such as images, video, audio, text, time-series, robotics sensors, sets, and tables, prediction tasks spanning the image-caption matching, robot pose, object pose, robot contact, design interfaces, digits, humor, sentiment, emotions, mortality rate, and ICD-9 codes from the research areas of affective computing, healthcare, multimedia, robotics, and HCI.

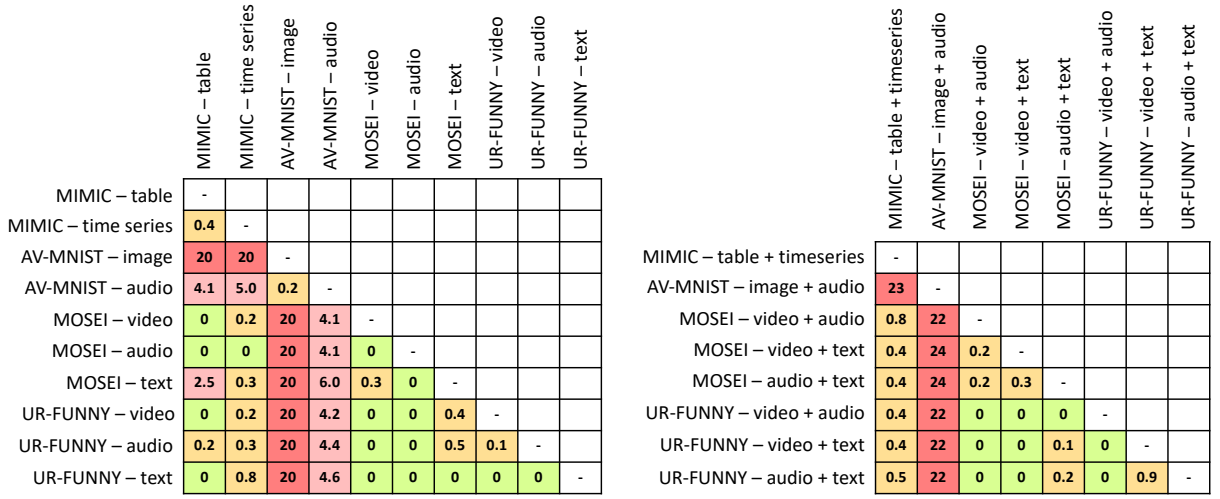


Figure 9.5: Modality and interaction heterogeneity matrices color coded by distances, with green showing smaller distances and dark red larger distances. We find clear task outliers (AV-MNIST has high difficulty transferring to others), and that there is generally more interaction heterogeneity than unimodal heterogeneity. Otherwise, the same modality and modality pairs across different tasks are generally similar to each other.

9.3.1 Heterogeneity measurements and parameter groups

We begin with a study of the heterogeneity matrices in Figure 9.5 and the resulting parameter groups.

Modality heterogeneity: We first notice that the modalities from AV-MNIST only transfer well to each other and has high difficulty transferring to other modalities from the other datasets. The same modality across different tasks is generally similar to each other (e.g., text between UR-FUNNY and MOSEI, audio between UR-FUNNY and MOSEI). The text modality in UR-FUNNY seems to be close to most other modalities, and likewise for the tabular modality in MIMIC. It is also worth noting that the video and audio modalities are not the most informative in MOSEI, and predictions are dominated by language [712], which may explain their general homogeneity with respect to other modalities.

Interaction heterogeneity: There is generally more interaction heterogeneity than unimodal, implying that the interactions between modality pairs tend to be more unique. Again, we notice the general poor transfer from the modality pair (image+audio) in AV-MNIST to other pairs, and the general strong transfer from (audio+text) in UR-FUNNY to the rest, which shows a relationship between modality and interaction heterogeneity. We also find that the same modality pairs (video+text) and (video+audio) shows crossmodal similarity across both datasets they appear in: MOSEI and UR-FUNNY. Finally, while the triplet of crossmodal pairs in MOSEI are quite different from each other, those in UR-FUNNY are more similar.

Using these measurements, we show the final groups of parameters obtained after clustering the matrices for different values of k. As an example, for $|\mathbb{U}| = 3, |\mathbb{C}| = 3, k = 6$, the groups are $\mathcal{U}_1 = \{$AV-MNIST image, AV-MNIST audio$\}$, $\mathcal{U}_2 = \{$MIMIC table, MOSEI video, MOSEI audio$\}$, $\mathcal{U}_3 = \{$MIMIC timeseries, MOSEI text, UR-FUNNY text, UR-FUNNY video, UR-

FUNNY audio}, and $\mathcal{C}_1$ = {AV-MNIST image+audio}, $\mathcal{C}_2$ = {MOSEI video+audio}, and $\mathcal{C}_3$ = {MIMIC table+timeseries, MOSEI video+text, MOSEI audio+text, UR-FUNNY video+text, UR-FUNNY video+audio, UR-FUNNY audio+text}.

Finally, we observe the low-rank nature of the heterogeneity matrices due to symmetry and approximate triangle inequality, such that even using a low-rank approximation of $r = 3$ is sufficient to approximate the entire matrix. This suggests that we do not need to exhaustively measure unimodal and interaction transfer between all modality pairs to enjoy the benefits of our proposed approach.

9.3.2 Qualitative results

We now present our results on the multitask, transfer, and generalization capabilities of HIGHMMT using performance and efficiency metrics. Henceforth, we will refer to the following models:

(1) **HIGHMMT share none** refers to individual copies of HIGHMMT models, one for each task.

(2) **HIGHMMT share all** refers to one single HIGHMMT model fully shared across all modalities and tasks.

(3) **HIGHMMT** refers to the full heterogeneity-aware HIGHMMT model across all modalities and tasks with learned parameter groupings based on heterogeneity measurements.

Multitask performance and efficiency. In Figure 9.6, we summarize the overall tradeoff between performance and efficiency using existing task-specific models and variants of HIGHMMT. The blue dots represent all possible combinations of task-specific models across multiple datasets (summarized in MultiBench [367], $> 10^5$ total combinations) with their overall performance (scaled to a $0-1$ range before averaging across datasets) and overall efficiency (inverted total number of parameters). The red dots represent the state-of-the-art Pareto front: points that are not strictly dominated in both performance and efficiency. In light green, separate single-task HIGHMMT models (share none) already improve parameter efficiency as compared to standard Multimodal Transformers [390, 613]. In dark green is HIGHMMT (share all) trained in a homogeneous multitask manner (i.e., with full parameter sharing across unimodal and multimodal layers within and across tasks), which further pushes forward the Pareto front by improving both performance and efficiency. Finally, in orange, HIGHMMT with heterogeneity-aware fine-tuning achieves significantly better tradeoffs between performance and efficiency, with efficiency and consistently high performance across multiple modalities and tasks.

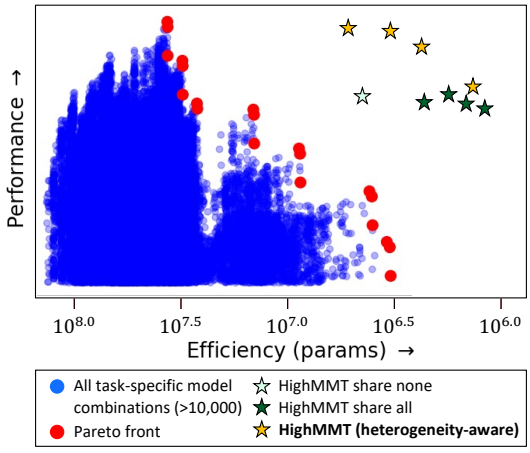


Figure 9.6: Overall tradeoff. HIGHMMT pushes forward the Pareto front of performance and efficiency as compared to all possible ($> 10^5$) combinations of task-specific models across multiple datasets [367]. The x-axis denotes (inverted) total parameters and y-axis denotes performance scaled to a $0-1$ range before averaging across datasets.

Table 9.2: Tuning the number of parameter groups results in controlled tradeoffs between parameters and performance.

Clusters	Performance ↑	Params (M) ↓
2 (share all)	68.4 ± 0.4	1.07
4	68.8 ± 0.5	1.24
6	70.1 ± 0.2	2.47
7	71.0 ± 0.1	3.11
9	71.2 ± 0.2	4.23

Table 9.3: Cross-modal few-shot transfer to new modalities and tasks. We train multitask HIGHMMT on $1/2/3$ datasets and find that it generalizes few-shot to new modalities and tasks on the 4th dataset, with improved performance over single-task training on the 4th dataset. Cross-modal transfer improves with more pretraining tasks and works best on the smallest target tasks (UR-FUNNY).

# Source tasks	Target task			
	UR-FUNNY	MOSEI	MIMIC	AV-MNIST
0 (no transfer)	63.1 ± 0.5	79.0 ± 0.5	67.7 ± 0.6	70.3 ± 0.4
1	63.5 ± 0.5	79.2 ± 0.3	67.9 ± 0.5	70.5 ± 0.4
2	64.0 ± 0.7	79.3 ± 0.5	68.0 ± 0.8	70.5 ± 0.4
3	**64.7 ± 0.4**	**79.6 ± 0.6**	**68.4 ± 0.6**	70.6 ± 0.4

The suite of HIGHMMT models is obtained by tuning k, the total number of unimodal and crossmodal parameter groups (i.e., the number of clusters when clustering heterogeneity matrices). k can be seen as a hyper-parameter depending on the computational budget, with smaller k implying more parameter sharing on lower budgets and vice-versa. In Table 9.2, we show the effect of k on average performance and total parameters. We test k in the range $\{2, 4, 6, 7, 9\}$, with $|\mathbb{U}| = 1, |\mathbb{C}| = 1$, $|\mathbb{U}| = 3, |\mathbb{C}| = 1$, $|\mathbb{U}| = 3, |\mathbb{C}| = 3$, $|\mathbb{U}| = 3, |\mathbb{C}| = 4$, and $|\mathbb{U}| = 4, |\mathbb{C}| = 5$ respectively where $|\mathbb{U}|, |\mathbb{C}|$ denote the number of unimodal and crossmodal parameter groups. We see a controllable tradeoff: starting with a fully shared model and increasing the number of parameter groups, we also see steadily improving performance approaching task-specific state-of-the-art models. Overall, optimizing for performance results in a model as strong as current state-of-the-art models while using $8\times$ fewer total parameters. Optimizing for efficiency results in a model that reaches within 96% of current state-of-the-art performance but using $30\times$ fewer total parameters (mean and deviation over 10 runs).

Positive transfer to new modalities and tasks. HIGHMMT also offers opportunities to study whether we can *transfer* knowledge between completely different modalities and tasks. Starting with the collection of 4 datasets in the order MOSEI, AV-MNIST, MIMIC, and UR-FUNNY ranked by largest dataset size (total of datapoints and memory storage per datapoint), we pre-train a fully-shared HIGHMMT model on $1/2/3$ of the 4 tasks before fine-tuning on the fourth task only (e.g., train on MOSEI and transfer to UR-FUNNY, on MOSEI+AV-MNIST then transfer to UR-FUNNY, and on MOSEI+AV-MNIST+MIMIC then transfer to UR-FUNNY, and likewise for transfer to the other 3 datasets.

From Table 9.3, we found that on all four combinations of multitask pretraining and fine-tuning, weights learned from other multimodal tasks generalize well to new modalities and tasks, improving performance over single target-task training (mean and standard deviation over 10 runs). When we increase the number of pretraining datasets, we observe a consistent

Table 9.4: HIGHMMT achieves strong performance on overall performance and efficiency (mean and deviation over 10 runs), sometimes even beating (shown in **bold**) the task-specific state-of-the-art, especially on the relatively understudied modalities (time-series, robotics sensors, and sets) from the robotics (PUSH, V&T) HCI (ENRICO), and healthcare (MIMIC) research areas, while using **10× fewer parameters** due to parameter sharing and multitask learning. SOTA captures the max performance and parameters of more than 20 task-specific multimodal models: [1] GRADBLEND [651], [2] LF-LSTM [148], [3] LF [184], [4] MULT [613], [5] MFAS [478], [6] MFM [686], and [7] LRTF [723].

Model	ENRICO ↑	PUSH ↓	V&T ↑	UR-FUNNY ↑	MOSEI ↑	MIMIC ↑	AV-MNIST ↑
SOTA	51.0 ± 1.4[1]	0.290 ± 0.1[2]	93.6 ± 0.1[3]	66.7 ± 0.3[4]	**82.1 ± 0.5**[4]	68.9 ± 0.5[6,7]	**72.8 ± 0.2**[5]
HIGHMMT	**52.7 ± 0.6**	**0.277 ± 0.1**	**96.3 ± 0.2**	66.2 ± 0.4	80.2 ± 0.2	68.2 ± 0.3	71.1 ± 0.2

Model	Params (M) ↓
SOTA	32.3
HIGHMMT	**3.01**

improvement in fine-tuned target task performance. There is an inverse correlation between target task size and performance improvement: the smallest dataset, UR-FUNNY, benefited the most (+2.4%) from transfer learning from 0 to 3 multitask datasets. This implies that our multimodal pretraining-fine-tuning paradigm is useful for low-resource target modalities and tasks.

Finally, we compare transfer learning performance across different levels of partial observability. While one would expect the transfer to MIMIC to be the hardest due to its modality set {time-series, table} being completely disjoint from the remaining 3 datasets, we still observe a +0.8% gain as compared to single-task training. Therefore, HIGHMMT can generalize to new modalities and tasks. Unsurprisingly, for datasets with more overlap (e.g., UR-FUNNY with complete overlap in {text, video, audio} with respect to pretraining), we find larger improvements using transfer learning over single-task models (+2.4%).

Comparison with task-specific state-of-the-art. In Table 9.4, we compare multitask performance and efficiency with task-specific state-of-the-art models. We achieve performance within the range of published models (and usually close to the individual task-specific state-of-the-art) in MultiBench, which tallies more than 20 recent multimodal models in each task's literature [367]. In fact, HIGHMMT even sets new state-of-the-art results on several datasets, especially on the relatively understudied modalities (time-series, force and proprioception sensors, and sets) from the robotics (PUSH, V&T) and HCI (ENRICO) research areas. On top of strong performance, the main benefit lies in using fewer total parameters as compared to separate task-specific models - more than 10× reduction. Since this reduction grows with the number of tasks, our approach is scalable to high-modality scenarios.

Partial-observability. Observe HIGHMMT performance on partially-observable modality subsets (i.e., target task involving modalities not present in the other tasks): from Table 9.4, we find that the model performs well on the MIMIC dataset despite its modality set {time-series, table} being completely disjoint from the remaining 3 datasets - we obtain similar performance across both multitask and single-task models (68.2 ± 0.3% vs 68.9 ± 0.5%). We find that HIGHMMT multitask also works on ENRICO dataset in HCI (52.7±0.6% multitask vs 51.0±1.4% single-task) despite it having completely disjoint modality inputs.

Multitask fusion and retrieval. We perform multitask training over multimodal fusion in

Table 9.5: We conduct in-depth **ablation studies** and find strong evidence for (1) having separate unimodal and interaction layers, (2) determining parameter sharing via feature transfer, and (3) homogeneous pretraining before heterogeneity-aware fine-tuning into parameter groups (mean and standard deviation over 10 runs).

	Model	UR-FUNNY ↑	MOSEI ↑	MIMIC ↑	AV-MNIST ↑	Ave ↑
Full model	HIGHMMT	$\mathbf{66.2 \pm 0.4}$	$\mathbf{80.2 \pm 0.2}$	$\mathbf{68.2 \pm 0.3}$	$\mathbf{71.1 \pm 0.2}$	$\mathbf{71.4 \pm 0.3}$
Architecture ablations	- w/o embeddings	63.0 ± 1.2	79.0 ± 0.7	67.1 ± 1.2	70.3 ± 0.7	69.8 ± 0.3
	- w/o unimodal	57.9 ± 0.3	61.9 ± 2.1	63.0 ± 0.9	59.5 ± 1.4	60.6 ± 0.7
	- w/o crossmodal [505]	63.8 ± 1.0	79.5 ± 0.5	$\mathbf{67.9 \pm 0.4}$	70.4 ± 0.5	70.4 ± 0.5
Param sharing ablations	- share none [367]	63.7 ± 0.7	79.4 ± 0.4	67.7 ± 0.7	70.4 ± 0.1	70.2 ± 0.3
	- share unimodal [505]	62.5 ± 1.3	79.0 ± 1.1	63.4 ± 1.4	70.1 ± 0.7	68.8 ± 0.8
	- share crossmodal [15]	63.0 ± 1.1	79.5 ± 0.3	64.3 ± 0.3	70.1 ± 0.9	69.2 ± 0.3
	- share all [551]	63.1 ± 0.7	79.2 ± 0.3	63.7 ± 1.6	68.6 ± 0.6	68.7 ± 0.5
	- random difference	62.9 ± 0.9	79.5 ± 0.6	67.6 ± 0.3	70.4 ± 0.2	70.1 ± 0.3
	- feature difference [575]	64.0 ± 1.0	79.4 ± 0.3	$\mathbf{67.9 \pm 0.3}$	70.1 ± 0.4	70.4 ± 0.2
Training ablations	- w/o homogeneous pretraining	61.2 ± 0.1	78.5 ± 0.1	64.8 ± 0.1	$\mathbf{71.1 \pm 0.2}$	69.9 ± 0.1

AV-MNIST and retrieval in CIFAR-ESC. While fusion emphasizes information integration, retrieval focuses on aligning corresponding elements expressed through different views of the data [371]. Even across these vastly different prediction tasks, we find that multitask training (60.5% retrieval accuracy) improves upon single-task training (58.8%). Not only have the unimodal networks simultaneously processed different modalities, but the crossmodal network has captured correspondences useful for both fusion and retrieval.

9.3.3 Ablation studies

In this subsection, we carefully ablate the model architectures, parameter sharing, and training decisions.

Architectural ablations. We first analyze each architectural component of HIGHMMT: (1) *w/o embeddings* removes the only modality-specific component in the model - the modality embeddings. We set embeddings for all modalities to be the same to test whether a modality-specific component is necessary to capture heterogeneity across input data sources, (2) *w/o unimodal* removes the unimodal encoder and directly applies the cross-attention layer, and *w/o crossmodal* replaces the crossmodal layer with a concatenation of unimodal features and a linear classification layer. The latter resembles the most direct multimodal extension of existing work in shared unimodal encoders like Perceiver [276], MultiModel [291], ViT-BERT [353] or PolyViT [378]. From Table 9.5, removing any of the 3 components in HIGHMMT results in worse performance. The unimodal encoder is particularly important.

Param sharing ablations. We further ablate with respect to possible parameter sharing settings in HIGHMMT: (1) *share none* uses separate unimodal and multimodal layers reminiscent of typical single-task multimodal transformers [232, 390, 613], (2-3) *share unimodal (crossmodal)* only shares the unimodal (crossmodal) layer during multitask training, (4) *share all* shares all parameters without accounting for possible heterogeneity [505], (5) *random difference* determines k parameter groups randomly rather than via heterogeneity measurements, (6) *feature difference* uses feature-level divergences on jointly trained unimodal encoders (i.e., $\|U(X_1) - U(X_2)\|_2^2$) rather than transfer performance to measure heterogeneity as is commonly done in transfer

learning and domain adaptation [132, 575]. From Table 9.5, our proposed heterogeneity-aware parameter grouping results in the best overall performance as compared to fully shared, fully separate, or parameter grouping informed by other heterogeneity measures such as random or feature distance.

Training ablations. Finally, we explore *w/o homogeneous pretraining*: directly learning a model with parameter groups as selected by our approach as opposed to performing homogeneous pre-training before fine-tuning them into parameter groups. From Table 9.5, we find that this ablation underperforms - training parameter groups from scratch overfits to smaller datasets which hurts overall performance.

9.3.4 Understanding homogeneity and heterogeneity in HIGHMMT

We now take a deeper empirical analysis to better understand HIGHMMT, through parameter overlap and interference experiments.

Parameter overlap. Starting with a trained multitask HIGHMMT, we use a gradient-based method [221] to determine how much each parameter is involved in a specific task. For each task T and parameter $\theta \in \Theta$ in multitask model M_Θ, we compute the involvement $I_T(\theta) = \mathbb{E}_{(x,y)\epsilon T}|\nabla_\theta M_\Theta(y|x)|$ where $M_\Theta(y|x)$ is the predicted probability of correct target y by M_Θ given x as input. In other words, this measures the absolute gradient with respect to θ when predicting y given x in task T. A higher absolute gradient implies "activated" neurons and vice-versa for gradients closer to 0. This enables us to compute the extent a parameter θ is involved for each task. The *number of tasks* a given parameter θ is involved in can then be approximated by thresholding and summing up $n(\theta) = \sum_T (\mathbb{1}\{I_T(\theta) > \epsilon \max(I_1(\theta), I_2(\theta), I_3(\theta), I_4(\theta)\})$ which returns an integer from 1 to 4. We chose a threshold ϵ such that parameters are classified as active about half the time on average, which occurs at $\epsilon = 0.2$.

Since we are interested in the level of parameter overlap in the shared unimodal encoder and multimodal layer, we set θ as these 2 modules and report results in Table 9.6. There is evidence of significant parameter overlap across unimodal encoders: more than 92% of neurons are involved in at least 3 of the 4 tasks. On the other hand, there is not nearly as much parameter overlap in the multimodal layer: only 10% of neurons are involved in 3 or 4 tasks. Hence, it seems like the unimodal encoders learn task-agnostic representations, but the subsequent multimodal layers (closer to task-specific classifiers) capture more task-specific information. This also reinforces our observation in §9.3.1 that there is generally more interaction heterogeneity than modality heterogeneity, which suggests using fewer unimodal parameter groups and more crossmodal parameter groups.

Table 9.6: We find evidence of significant **parameter overlap** across unimodal encoders: $> 92\%$ of neurons are involved in at least 3 of the 4 tasks, while the multimodal layers are more task-specific: only 10% of neurons are involved in 3 or 4 tasks.

Component	Number of involved tasks			
	1	2	3	4
Unimodal layers	2.8%	5.1%	**61.1%**	**31.1%**
Crossmodal layers	**48.8%**	**39.7%**	9.9%	1.6%

Parameter interference. Another empirical proof for parameter sharing in multitask models is the phenomenon of *parameter interference*: to what extent do parameters interfere with each other across tasks? We perform an experiment to investigate parameter interference: we pick one task and flip the labels in its training set, train the multitask model on the modified training set,

Table 9.7: Parameter interference: we observe different performance drops on each task (columns) after training on one task with flipped labels (rows). Training the shared unimodal encoders causes the most harm, which implies that unimodal encoders contain more shared neurons sensitive to task changes. Red for drops greater than 20%, yellow for drops between 10 and 20%, and green for drops below 10%.

(a) Training entire model

Flipped task	UR-FUNNY	MOSEI	MIMIC	AV-MNIST
UR-FUNNY	−24.6	−8.83	−10.6	−57.7
MOSEI	−4.07	−59.7	−20.3	−53.2
MIMIC	−3.59	−5.83	−33.1	−37.5
AV-MNIST	−3.50	−1.23	−4.87	−68.9

(b) Only training unimodal encoder

Flipped task	UR-FUNNY	MOSEI	MIMIC	AV-MNIST
UR-FUNNY	−23.8	−10.1	−12.8	−58.4
MOSEI	−5.77	−57.6	−21.1	−52.7
MIMIC	−3.03	−3.54	−35.0	−56.3
AV-MNIST	−2.94	−7.82	−53.6	−69.3

(c) Only training multimodal layer

Flipped task	UR-FUNNY	MOSEI	MIMIC	AV-MNIST
UR-FUNNY	−25.2	−8.34	−2.67	−8.16
MOSEI	0.47	−59.6	−19.8	−8.19
MIMIC	0.19	−0.76	−35.2	−4.87
AV-MNIST	−1.61	−1.48	−2.23	−69.1

and see how the incorrectly labeled task affects performance on other tasks. This experiment provides evidence of information sharing: if the multitask model does not share information (i.e., the model learns independent subspaces for each task), then one would not observe negative interference from one noisy dataset. We study negative interference under 3 configurations of training (a) the whole model; (b) only the unimodal encoder, and (c) only the multimodal layer on the flipped training set.

From Table 9.7, certain tasks are more affected by negative interference (e.g., AV-MNIST), while some tasks are not influenced as much (e.g., UR-FUNNY). Again, this reflects our heterogeneity measurements in §9.3.1, where AV-MNIST displays high heterogeneity. Furthermore, performance drops due to training the unimodal encoders are the most significant, which corroborates with our parameter overlap and heterogeneity analysis that unimodal encoders contain more entangled parameters which are more sensitive to task changes. On the other hand, multimodal layers contain more disentangled parameters, which results in higher heterogeneity measurements and needs more separate parameter groups.

9.4 Related Work

Multimodal Transformers have emerged as strong models for representation learning. Building upon the Transformer [631], multimodal extensions use either full self-attention over modalities concatenated across the sequence dimension [108, 348, 571, 576] or a cross-modal attention layer [390, 588, 613], and are useful for sequential data by automatically aligning and capturing complementary features at different time-steps [337, 613, 694]. Self-supervised multimodal pretraining has emerged as an effective way to train these architectures, with the aim of learning

representations from large-scale unlabeled multimodal data before transferring to downstream tasks via fine-tuning [348, 390, 571]. These pretraining objectives typically consist of unimodal masked prediction, crossmodal masked prediction, and multimodal alignment prediction [232].

Unified encoder for unimodal learning. Several works such as Perceiver [275, 276], MultiModel [291], ViT-BERT [353], and PolyViT [378] have explored the possibility of using the same architecture for different inputs on unimodal tasks (i.e., language, image, video, or audio-only). The Transformer architecture has emerged as a popular choice due to its suitability for serialized inputs such as text [144], images [154], video [576], and time-series data [379], a phenomenon further observed by Lu et al. [391] where a single Transformer pretrained on text transfers to sequence modeling and image classification. While these serve as building blocks in our model, our focus is on a general-purpose multimodal model for multitask and transfer learning across different subsets of modalities rather than unimodal tasks.

Multimodal multitask and transfer learning. There have also been several attempts to build a single model that works well on a suite of multimodal tasks [113, 348, 390, 505, 571]. For example, UniT [253], VLBERT [571], ViLBERT [390], and VL-T5 [113] are all unifying models for vision-and-language tasks. VATT [15] jointly trains a shared model on video, audio, and text data to perform audio-only, video-only, and image-text retrieval tasks. FLAVA [551] found that pretraining a shared model with unpaired images, unpaired text, and image-text pairs results in strong performance on image-only, text-only, and image-text multimodal tasks, while Reed et al. [505] scales up a single Transformer model for image, text, and decision-making tasks. However, all of these train a single model for all tasks, without investigating how heterogeneity can necessitate partial parameter sharing. On the transfer side, while more research has focused on transfer within the same modality with external information [158, 553, 675, 716], Liang et al. [369] is the only work that studies transfer to completely new modalities. However, they require paired data collection and modality-specific modeling. Our work goes beyond the commonly studied language, vision, and audio modalities to relatively understudied ones (e.g., tabular data, time-series, sensors, graphs, and set data). Furthermore, we show the possibility of generalizing to new modality subsets. Finally, our work also complements studies of transfer learning in a single modality [566, 673, 718], where insights from task heterogeneity have informed multitask approaches, as well as multisensor fusion in various domains such as healthcare [429] and robotics [584, 591].

9.5 Conclusion

We propose an information transfer approach for estimating modality and interaction heterogeneity, a key component towards automatically determining which modalities should be processed and fused jointly for efficient representation learning in high-modality scenarios. Our resulting model, HIGHMMT dynamically determines the optimal parameter groupings balancing total performance and parameter efficiency, simultaneously achieves strong results on modalities (text, image, video, audio, time-series, sensors, tables, and sets) and tasks from different research areas, and transfers to new modalities and tasks during fine-tuning. We release our code and benchmarks which we hope will present a unified platform for subsequent analysis.

Chapter 10

Conclusion

In this thesis, we advanced the foundations of multimodal machine learning by highlighting its key principles and core challenges. In the bulk of the thesis, we outlined our progress towards understanding the foundations of multimodal interactions and new modeling methods for generalizable representation learning across many input modalities and tasks. This concluding chapter provides a summary of the main contributions, discusses potential limitations, and outlines future research directions in multimodal artificial intelligence.

10.1 Summary of Thesis Contributions

Multimodal artificial intelligence is one of the most exciting subareas of artificial intelligence research today, and has the potential to make major impacts in autonomous agents with digital, physical, and social capabilities. This thesis aims to pave a foundation for multimodal artificial intelligence so that future students and researchers are able to better understand the breadth and depth of multimodal research today, are equipped with the scientific fundamentals required to perform cutting-edge research in this field, and are up-to-date with practical methods for machine learning from real-world multimodal datasets.

To summarize the contributions of this thesis, we began (in Section 2) by outlining the theoretical and computational foundations of multimodal machine learning by synthesizing a broad range of theoretical frameworks and application domains from both historical and recent perspectives. This foundation involves three key principles of modality *heterogeneity*, *connections*, and *interactions* often present in multimodal problems which brings unique challenges to machine learning, which we outline through a taxonomy of six core challenges: *representation*, *alignment*, *reasoning*, *generation*, *transference*, and *quantification*. This taxonomy enables researchers to navigate the breadth of recent technical achievements and enables us to identify key open problems for future research.

In this first major part of this thesis, we build a foundation for multimodal interactions: the basic principle of how modalities combine to give rise to new information for a task. Section 3 presented an information-theoretic framework formalizing how *modalities interact* with each other to give rise to new information for a task, which can be decomposed into redundancy, uniqueness, and synergy [372]. Using this theoretical framework, we proposed two practical estimators to

quantify the interactions in real-world datasets. Quantifying the types of interactions a multimodal task requires enables researchers to understand their data and choose the right model to learn interactions in a principled way. Using this foundation of multimodal interactions, we design new self-supervised approaches to learn these interactions [374] (Section 4), visualization tools for practitioners to analyze whether their model has succeeded in learning [375] (Section 5), and new guidelines for practitioners to decide which modality to collect for maximum increase in performance [376] (Section 6).

In the second major part of this thesis, we design practical multimodal foundation models that generalize over many modalities and tasks, which presents a step toward grounding large language models to real-world sensory modalities such as videos, physical sensors, and medical data. Section 7 introduced MULTIBENCH, a unified large-scale benchmark across a wide range of modalities, tasks, and research areas enabling research towards multimodal foundation models [367]. Section 8 presented the *cross-modal attention* [101, 359] and *multimodal transformers* [613] architectures that are suitable for learning the interactions across many elements in modality sequences such as text, videos, time-series, and sensors. Finally, Section 9 showed how we can scale these architectures on MULTIBENCH to create general-purpose multimodal multitask models across a variety of tasks, including collaborating with practitioners to apply these models for real-world impact on affective computing, mental health, and cancer prognosis.

Together, our contributions deliver fundamental methodological and practical insights in multimodal learning, presenting approaches that are principled and explainable to practitioners while also capturing the benefits of scale across many modalities and tasks. Some of the work done during the PhD but not included in this thesis also paves a way towards improving the robustness, safety, and efficiency of multimodal models for real-world deployment.

10.2 Limitations and Future Directions

Finally, we conclude this thesis by identifying the following future research challenges in multimodal artificial intelligence:

Representation: Learning multimodal representations is the cornerstone of multimodal machine learning. There has been substantial progress towards increasingly expressive and performant multimodal representations. However, there remain key challenges in their theoretical understanding and generalization beyond image and text.

Theoretical and empirical frameworks: How can we formally define the three core principles of heterogeneity, connections, and interactions? Can we quantify their presence in multimodal datasets and models, and understand whether current multimodal representation learning methods are suitable for learning different interactions? Answering these fundamental questions will lead to a better understanding of the capabilities and limitations of current multimodal representations, and inspire the development of new methods in a principled manner.

Beyond additive and multiplicative cross-modal interactions: While recent work has been successful at modeling multiplicative interactions of increasing order, how can we capture causal, logical, and temporal connections and interactions? What is the right type of data and domain knowledge necessary to model these relationships? Modeling these interactions in a principled manner could lead to systems that are more robust, compositional, and explainable than those

based fully on neural networks.

Tabular, sensors, and time-series: Existing work has shown success in learning image, text, and audio-visual representations. However, tabular and time-series data are prevalent in many real-world applications such as healthcare and autonomous vehicles. How can we learn multimodal interactions between the best encoders for tabular and sensor data, which may not be based on deep learning (e.g., decision trees, time-series analysis), and neural network representations that are state-of-the-art for the text and image modalities?

Brain and multimodal perception. There are many core insights regarding multimodal processing to be gained from human cognition, including the brain's multimodal properties [314] and mental imagery [435]. How does the human brain represent different modalities, how is multisensory integration performed, and how can these insights inform multimodal learning? In the other direction, what are opportunities in processing high-resolution brain signals such as fMRI and MEG/EEG, and how can multimodal learning help in the future analysis of data collected in neuroscience?

Alignment: There remain important challenges in aligning modality elements when these elements are extremely fine-grained in nature and exhibit long-range patterns across time.

Memory and long-term interactions. Many current multimodal benchmarks only have a short temporal dimension, which has limited the demand for models that can accurately process long-range sequences and learn long-range interactions. Capturing long-term interactions presents challenges since it is difficult to semantically relate information when they occur very far apart in time or space and raises complexity issues. How can we design models (perhaps with memory mechanisms) to ensure that these long-term cross-modal interactions are captured?

Reasoning: Today's multimodal systems, especially those based on deep learning or large language models, are still not capable of robust and complex reasoning. We outline two challenges in compositional and interactive reasoning.

Multimodal compositionality. How can we understand the reasoning process of trained models, especially regarding how they combine information from modality elements? This challenge of compositional generalization is difficult since many compositions of elements are typically not present during training, and the possible number of compositions increases exponentially with the number of elements [601]. How can we best test for compositionality, and what reasoning approaches can enable compositional generalization?

Multimodal embodiment and interaction. Most of today's multimodal systems are trained to make predictions without the capability to take actions in the world. The next generation of these systems will be those that can plan actions, imagine the effect these actions will have on the world, and choose the right sequence of actions over a long period of time to solve complex tasks. We have begun to build these interactive multimodal agents for the virtual world, such as processing multimedia web data to help humans with web tasks like online shopping, travel bookings, and content management. Building multisensory robotic systems that can actions in the real world, while respecting safety and robustness, is another long-term future direction.

Generation: The incredible advances of generative AI have inspired many future directions in generating multimedia content.

Multimodal creation. Synchronized creation of realistic video, text, and audio remains a challenge. These systems can be applied for entertainment, such as generating music videos, virtual avatar characters, virtual humans, and more. It is also likely that better multimodal

generative models of the world can serve as world models to train planning and sequential decision making agents.

Real-world ethical concerns. However, the recent success in generation has brought ethical concerns regarding their use. For example, large-scale pretrained language models can generate text denigrating to particular social groups [542], toxic speech [193], and sensitive pretraining data [81]. Future work should study how these risks are potentially amplified or reduced when the dataset is multimodal, and whether there are ethical issues specific to multimodal generation.

Transference: Advances in foundation models have also enabled increasingly general-purpose models that can transfer information and knowledge across a wide range of modalities and tasks. This opens up new directions in high-modality learning.

High-modality learning aims to learn representations from an especially large number of heterogeneous data sources, which is a common feature of many real-world multimodal systems such as self-driving cars and IoT [263]. More modalities introduce more dimensions of heterogeneity, incur complexity challenges in unimodal and multimodal processing, and require dealing with non-parallel data (i.e., not all modalities are present at the same time).

Quantification: Finally, we highlight several important lines of future work in quantifying and understanding key design decisions in the multimodal learning process.

Modality utility, tradeoffs, and selection. How can we formalize why modalities can be useful or potentially harmful for a task? There are also challenges in quantifying *modality and social biases* and *robustness* to imperfect, noisy, and out-of-distribution modalities. Future work should come up with formal guidelines to compare these tradeoffs and select the optimal set of modalities balancing performance with these other potential concerns, which can help practitioners decide the right modalities to work with.

Explainability and interpretability. Before models can be safely used by real-world stakeholders in domains such as medicine, autonomous systems, and user interfaces, we need to understand how to interpret their inner workings. How can we evaluate whether these phenomena are accurately interpreted? These challenges are exacerbated for relatively understudied modalities beyond language and vision, where the modalities themselves are not easy to visualize. Finally, how can we tailor these explanations, possibly in a *human-in-the-loop* manner, to inform real-world decision-making?

In conclusion, we believe that this thesis can lay the theoretical and practical foundations for multimodal machine learning and inspire future work towards these open problems.

Bibliography

[1] Mahdi Abavisani and Vishal M Patel. Deep multimodal subspace clustering networks. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1601–1614, 2018.

[2] Harika Abburi, Rajendra Prasath, Manish Shrivastava, and Suryakanth V Gangashetty. Multimodal sentiment analysis using deep neural networks. In *International Conference on Mining Intelligence and Knowledge Exploration*, pages 58–65. Springer, 2016.

[3] Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306, 2021.

[4] Ahmad Abiri, Jake Pensa, Anna Tao, Ji Ma, Yen-Yi Juo, Syed J Askari, et al. Multi-modal haptic feedback for grip force reduction in robotic surgery. *Scientific reports*, 9(1):1–10, 2019.

[5] Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. Multimodal biomedical ai. *Nature Medicine*, 28(9):1773–1784, 2022.

[6] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *CoRR*, abs/1810.03292, 2018. URL `http://arxiv.org/abs/1810.03292`.

[7] Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. In *CVPR*, pages 21406–21415, 2022.

[8] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *CVPR*, pages 9690–9698, 2020.

[9] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

[10] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. In *EMNLP*, pages 1955–1960, 2016.

[11] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. VQA: Visual question answering. *International Journal of Computer Vision*, 2017.

[12] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *3DV*, pages 719–728. IEEE, 2019.

[13] Chaitanya Ahuja, Dong Won Lee, Yukiko I Nakano, and Louis-Philippe Morency. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *ECCV*, pages 248–265. Springer, 2020.

[14] Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014.

[15] Hassan Akbari, Liangzhe Yuan, Rui Qian, et al. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *arXiv preprint arXiv:2104.11178*, 2021.

[16] Mehmet Aktukmak, Yasin Yilmaz, and Ismail Uysal. A probabilistic framework to incorporate mixed-data type features: Matrix factorization with multimodal side information. *Neurocomputing*, 367:164–175, 2019.

[17] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile

networks. *Advances in Neural Information Processing Systems*, 33:25–37, 2020.

[18] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

[19] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL https://www.aclweb.org/anthology/W19-1909.

[20] Camila Alviar, Rick Dale, Akeiylah Dewitt, and Christopher Kello. Multimodal coordination of sound and movement in music and speech. *Discourse Processes*, 57(8):682–702, 2020.

[21] Paras Malik Amisha, Monika Pathania, and Vyas Kumar Rathaur. Overview of artificial intelligence in medicine. *Journal of family medicine and primary care*, 8(7):2328, 2019.

[22] Saeed Amizadeh, Hamid Palangi, Alex Polozov, Yichen Huang, and Kazuhito Koishida. Neuro-symbolic visual reasoning: Disentangling visual from reasoning. In *ICML*, pages 279–290. PMLR, 2020.

[23] Brandon Amos. Tutorial on amortized optimization for learning to optimize over continuous domains. *arXiv preprint arXiv:2202.00665*, 2022.

[24] Ankesh Anand, Eugene Belilovsky, Kyle Kastner, Hugo Larochelle, and Aaron Courville. Blindfold baselines for embodied qa. *arXiv preprint arXiv:1811.05013*, 2018.

[25] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016.

[26] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *CVPR*, pages 39–48, 2016.

[27] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *ICML*, 2013.

[28] Xavier Anguera, Jordi Luque, and Ciro Gracia. Audio-to-text alignment for speech recognition with very limited resources. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[29] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, pages 2425–2433, 2015.

[30] MOSEK ApS. *MOSEK Optimizer API for Python 10.0.34*, 2022. URL https://docs.mosek.com/latest/pythonapi/index.html.

[31] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pages 609–617, 2017.

[32] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

[33] Martín Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. *CoRR*, abs/1511.06464, 2015.

[34] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.

[35] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.

[36] Benjamin Auffarth, Maite López, and Jesús Cerquides. Comparison of redundancy and relevance measures for feature selection in tissue classification of ct images. In *Industrial conference on data mining*, pages 248–262. Springer, 2010.

[37] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

[38] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[39] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pages 15535–15545, 2019.

[40] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11: 1803–1831, 2010.

[41] Alexei Baevski and Michael Auli. Adaptive input representations for neural language modeling. In *International Conference on Learning Representations (ICLR)*, 2019.

[42] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460, 2020.

[43] Ricardo Baeza-Yates. Data and algorithmic bias in the web. In *Proceedings of the 8th ACM Conference on Web Science*, pages 1–1, 2016.

[44] Maria-Florina Balcan, Avrim Blum, and Ke Yang. Co-training and expansion: Towards bridging theory and practice. *NeurIPS*, 17, 2004.

[45] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *arXiv preprint arXiv:1705.09406*, 2017.

[46] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE TPAMI*, 41(2):423–443, 2018.

[47] George Barnum, Sabera J Talukder, and Yisong Yue. On the benefits of early fusion in multimodal representation learning. In *NeurIPS 2020 Workshop SVRHM*, 2020.

[48] Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 1986.

[49] Roland Barthes. *Image-music-text*. Macmillan, 1977.

[50] Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563, 1966.

[51] Anthony J Bell. The co-information lattice. In *Proceedings of the fifth international workshop on independent component analysis and blind signal separation: ICA*, volume 2003, 2003.

[52] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

[53] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. Social robots for education: A review. *Science robotics*, 3(21), 2018.

[54] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *NeurIPS*, 19, 2006.

[55] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *ICCV*, pages 2612–2620, 2017.

[56] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *FaaCT*, pages 610–623, 2021.

[57] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *TPAMI*, 35(8), August 2013.

[58] Dario Bertero, Farhad Bin Siddique, Chien-Sheng Wu, Yan Wan, Ricky Chan, and Pascale Fung. Real-time speech emotion and sentiment recognition for interactive dialogue systems, 01 2016.

[59] Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique informa-

tion. *Entropy*, 16(4):2161–2183, 2014.

[60] Brian T Bethea, Allison M Okamura, Masaya Kitagawa, Torin P Fitton, Stephen M Cattaneo, Vincent L Gott, William A Baumgartner, and David D Yuh. Application of haptic feedback to robotic surgery. *Journal of Laparoendoscopic & Advanced Surgical Techniques*, 14(3):191–195, 2004.

[61] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

[62] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. Experience grounds language. In *EMNLP*, pages 8718–8735, 2020.

[63] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *AAAI*, volume 34, pages 7432–7439, 2020.

[64] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in nlp. In *ACL*, pages 5454–5476, 2020.

[65] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, pages 92–100, 1998.

[66] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *ACM SIGMOD*, pages 1247–1250, 2008.

[67] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NeurIPS*, pages 4349–4357, 2016.

[68] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloé Kiddon, Jakub Konecný, Stefano Mazzocchi, H. Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, and Jason Roselander. Towards federated learning at scale: System design. *CoRR*, abs/1902.01046, 2019. URL `http://arxiv.org/abs/1902.01046`.

[69] Simon Brodeur, Ethan Perez, Ankesh Anand, Florian Golemo, Luca Celotti, et al. Home: a household multimodal environment. In *NIPS 2017's Visually-Grounded Interaction and Language Workshop*, 2017.

[70] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[71] Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

[72] Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language berts. *Transactions of the Association for Computational Linguistics*, 9:978–994, 2021.

[73] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[74] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, 42(4):335–359, dec 2008. doi: 10.1007/s10579-008-9076-6.

[75] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. *NeurIPS*, 32:841–852, 2019.

[76] Qiong Cai, Hao Wang, Zhenmin Li, and Xiao Liu. A survey on multimodal data-driven smart healthcare systems: approaches and applications. *IEEE Access*, 7:133583–133599, 2019.

[77] Juan C Caicedo and Fabio A González. Online matrix factorization for multimodal image retrieval. In *Iberoamerican Congress on Pattern Recognition*, pages 340–347. Springer, 2012.

[78] E. Cambria. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107, Mar 2016. ISSN 1541-1672. doi: 10.1109/MIS.2016.31.

[79] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *European Conference on Computer Vision*, pages 565–580. Springer, 2020.

[80] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Qiang Yang. Transitive hashing network for heterogeneous multimedia retrieval. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 81–87, 2017.

[81] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, et al. Extracting training data from large language models. In *USENIX Security*, pages 2633–2650, 2021.

[82] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[83] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *ACL*, 2019.

[84] CDC. *Suicide Facts at a Glance 2015*, 2015 (accessed September 6, 2020). URL `https://www.cdc.gov/violencePrevention/pdf/suicide-datasheet-a.pdf`.

[85] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. Do explanations make vqa models more predictable to a human? In *EMNLP*, 2018.

[86] Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. Grounding 'grounding'in nlp. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4283–4305, 2021.

[87] Wilson Chango, Juan A Lara, Rebeca Cerezo, and Cristobal Romero. A review on data fusion in multimodal learning analytics and educational data mining. *Wiley Interdisciplinary Reviews*, 2022.

[88] Devendra Singh Chaplot, Kanthashree Mysore Sathyendra, Rama Kumar Pasumarthi, Dheeraj Rajagopal, and Ruslan Salakhutdinov. Gated-attention architectures for task-oriented language grounding. In *AAAI*, 2018.

[89] Anargyros Chatzitofis, Leonidas Saroglou, Prodromos Boutis, Petros Drakoulis, Nikolaos Zioulis, et al. Human4d: A human-centric multimodal dataset for motions and immersive media. *IEEE Access*, 8:176241–176262, 2020.

[90] Liwei Che, Jiaqi Wang, Yao Zhou, and Fenglong Ma. Multimodal federated learning: A survey. *Sensors*, 23 (15):6986, 2023.

[91] Gal Chechik, Amir Globerson, M Anderson, E Young, Israel Nelken, and Naftali Tishby. Group redundancy measures reveal redundancy reduction in the auditory pathway. *Advances in neural information processing systems*, 14, 2001.

[92] Brian Chen, Andrew Rouditchenko, Kevin Duarte, Hilde Kuehne, Samuel Thomas, Angie Boggust, et al. Multimodal clustering networks for self-supervised learning from unlabeled videos. In *ICCV*, pages 8012–8021, 2021.

[93] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *European Conference on Computer Vision*, pages 17–36. Springer, 2020.

[94] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In *ACL-IJCNLP Findings*, 2021.

[95] Jingqiang Chen and Hai Zhuge. Abstractive text-image summarization using multi-modal attentional hierarchical rnn. In *EMNLP*, 2018.

[96] Jingqiang Chen and Hai Zhuge. Extractive text-image summarization using multi-modal rnn. In *2018 14th International Conference on Semantics, Knowledge and Grids (SKG)*, pages 245–248. IEEE, 2018.

[97] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. *arXiv preprint arXiv:2102.10407*, 2021.

[98] Lele Chen, Guofeng Cui, Ziyi Kou, Haitian Zheng, and Chenliang Xu. What comprises a good talking-head video generation?: A survey and benchmark. *arXiv preprint arXiv:2005.03201*, 2020.

[99] Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. Graph optimal transport for cross-domain alignment. In *ICML*, pages 1542–1553. PMLR, 2020.

[100] Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. The best of both worlds: Combining recent advances in neural machine translation. In *ACL*, 2018.

[101] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 163–171, 2017.

[102] Richard J Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8):865–878, 2022.

[103] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[104] Valerie Chen, Jeffrey Li, Joon Sik Kim, Gregory Plumb, and Ameet Talwalkar. Interpretable machine learning: Moving from mythos to diagnostics. *Queue*, 19(6):28–56, 2022.

[105] Xilun Chen, Ben Athiwaratkun, Yu Sun, Kilian Q. Weinberger, and Claire Cardie. Adversarial deep averaging networks for cross-lingual sentiment classification. *CoRR*, abs/1606.01614, 2016.

[106] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[107] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[108] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[109] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

[110] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pages 1779–1788. PMLR, 2020.

[111] Yanhua Cheng, Xin Zhao, Rui Cai, Zhiwei Li, Kaiqi Huang, Yong Rui, et al. Semi-supervised multimodal deep learning for rgb-d object recognition. In *IJCAI*, pages 3345–3351, 2016.

[112] Jianfeng Chi, William Shand, Yaodong Yu, Kai-Wei Chang, Han Zhao, and Yuan Tian. Conditional supervised contrastive learning for fair text classification. *arXiv preprint arXiv:2205.11485*, 2022.

[113] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *ICML*, 2021.

[114] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053*, 2022.

[115] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078, 2014.

[116] C Mario Christoudias, Raquel Urtasun, and Trevor Darrell. Multi-view learning in the presence of view disagreement. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages

88–96, 2008.

[117] Jason Chuang, Daniel Ramage, Christopher Manning, and Jeffrey Heer. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 443–452, 2012.

[118] Ferdinando Cicalese and Ugo Vaccaro. Supermodularity and subadditivity properties of the entropy on the majorization lattice. *IEEE Transactions on Information Theory*, 48(4):933–938, 2002.

[119] Ferdinando Cicalese, Luisa Gargano, and Ugo Vaccaro. How to find a joint probability distribution of minimum entropy (almost) given the marginals. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2173–2177. IEEE, 2017.

[120] Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. Using syntax to ground referring expressions in natural images. In *AAAI*, volume 32, 2018.

[121] Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. Visual referring expression recognition: What do systems actually learn? In *NAACL*, pages 781–787, 2018.

[122] Volkan Cirik, Taylor Berg-Kirkpatrick, and L-P Morency. Refer360: A referring expression recognition dataset in 360 images. In *ACL*, 2020.

[123] Spencer Compton, Dmitriy Katz, Benjamin Qi, Kristjan Greenewald, and Murat Kocaoglu. Minimum-entropy coupling approximation guarantees beyond the majorization barrier. In *International Conference on Artificial Intelligence and Statistics*, pages 10445–10469. PMLR, 2023.

[124] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *arXiv preprint arXiv:2306.05284*, 2023.

[125] Thomas M Cover and Joy A Thomas. Information theory and statistics. *Elements of information theory*, 1(1): 279–335, 1991.

[126] Wanyun Cui, Guangyu Zheng, and Wei Wang. Unsupervised natural language inference via decoupled multimodal contrastive learning, 2020.

[127] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[128] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, et al. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, June 2023. arXiv:2305.06500 [cs].

[129] Zihang Dai, Zhilin Yang, Yiming Yang, William W Cohen, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Language modeling with longer-term dependency. 2018.

[130] Georges A Darbellay and Igor Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321, 1999.

[131] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *CVPR*, pages 1–10, 2018.

[132] Hal Daumé III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, 2007.

[133] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *ICML*, pages 933–941. PMLR, 2017.

[134] Debraj De, Pratool Bharti, Sajal K Das, and Sriram Chellappan. Multimodal wearable sensing for fine-grained activity recognition in healthcare. *IEEE Internet Computing*, 19(5):26–35, 2015.

[135] Victoria Dean, Shubham Tulsiani, and Abhinav Gupta. See, hear, explore: Curiosity via audio-visual association. *NeurIPS*, 33:14961–14972, 2020.

[136] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep—a collaborative voice analysis repository for speech technologies. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 960–964. IEEE, 2014.

[137] Biplab Deka, Zifeng Huang, Chad Franzen, Joshua Hibschman, Daniel Afergan, Yang Li, Jeffrey Nichols, and Ranjitha Kumar. Rico: A mobile app dataset for building data-driven design applications. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, pages 845–854, 2017.

[138] Emilie Delaherche and Mohamed Chetouani. Multimodal coordination: exploring relevant features and measures. In *Proceedings of the 2nd international workshop on Social signal processing*, pages 47–52, 2010.

[139] Joseph DelPreto, Chao Liu, Yiyue Luo, et al. Actionsense: A multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment. *NeurIPS*, 2022.

[140] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[141] Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *EMNLP*, pages 7580–7605, 2021.

[142] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *ICML*, pages 1174–1183. PMLR, 2018.

[143] Laurence Devillers, Laurence Vidrascu, and Lori Lamel. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422, 2005.

[144] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, 2019.

[145] Abhinav Dhall, Akshay Asthana, Roland Goecke, and Tom Gedeon. Emotion recognition using phog and lpq features. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 878–883. IEEE, 2011.

[146] Daisy Yi Ding and Robert Tibshirani. Cooperative learning for multi-view analysis. *arXiv preprint arXiv:2112.12337*, 2021.

[147] Daisy Yi Ding, Shuangning Li, Balasubramanian Narasimhan, and Robert Tibshirani. Cooperative learning for multiview analysis. *Proceedings of the National Academy of Sciences*, 119(38):e2202113119, 2022.

[148] Ning Ding, Sheng-wei Tian, and Long Yu. A multimodal fusion method for sarcasm detection based on late fusion. *Multimedia Tools and Applications*, 81(6):8597–8616, 2022.

[149] Wenhao Ding, Baiming Chen, Bo Li, Kim Ji Eun, and Ding Zhao. Multimodal safety-critical scenarios generation for decision-making algorithms evaluation. *IEEE Robotics and Automation Letters*, 6(2):1551–1558, 2021.

[150] Zhengming Ding, Shao Ming, and Yun Fu. Latent low-rank transfer subspace learning for missing modality recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

[151] Alan Dix, Janet Finlay, Gregory D Abowd, and Russell Beale. Human-computer interaction. *Harlow ua*, 2000.

[152] Alexander Domahidi, Eric Chu, and Stephen Boyd. Ecos: An socp solver for embedded systems. In *2013 European Control Conference (ECC)*, pages 3071–3076. IEEE, 2013.

[153] Cícero Nogueira dos Santos and Maira Gatti. Deep convolutional neural networks for sentiment analysis of short texts. In *COLING*, pages 69–78. ACL, 2014.

[154] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[155] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[156] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE, 2020.

[157] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. *arXiv*

preprint arXiv:2202.10936, 2022.

[158] Jared Dunnmon, Alexander Ratner, Nishith Khandwala, Khaled Saab, Matthew Markert, Hersh Sagreiya, Roger E. Goldman, Christopher Lee-Messer, Matthew P. Lungren, Daniel L. Rubin, and Christopher Ré. Cross-modal data programming enables rapid medical machine learning. *CoRR*, abs/1903.11101, 2019. URL http://arxiv.org/abs/1903.11101.

[159] Jared A Dunnmon, Alexander J Ratner, Khaled Saab, Nishith Khandwala, Matthew Markert, Hersh Sagreiya, Roger Goldman, et al. Cross-modal data programming enables rapid medical machine learning. *Patterns*, 2020.

[160] Stéphane Dupont and Juergen Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE transactions on multimedia*, 2(3):141–151, 2000.

[161] Sanghamitra Dutta, Praveen Venkatesh, Piotr Mardziel, Anupam Datta, and Pulkit Grover. An information-theoretic quantification of discrimination with exempt features. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3825–3833, 2020.

[162] Chris Dyer. Notes on noise contrastive estimation and negative sampling. *arXiv preprint arXiv:1410.8251*, 2014.

[163] R. Elliott, Z. Agnew, and J. F. W. Deakin. Medial orbitofrontal cortex codes relative rather than absolute value of financial rewards in humans. *European Journal of Neuroscience*, 27(9):2213–2218. doi: 10.1111/j.1460-9568.2008.06202.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-9568.2008.06202.x.

[164] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[165] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

[166] Sergio Escalera, Jordi Gonzàlez, Xavier Baró, Miguel Reyes, Oscar Lopes, Isabelle Guyon, Vassilis Athitsos, and Hugo Escalante. Multi-modal gesture recognition challenge 2013: Dataset and results. In *ICMI*, pages 445–452, 2013.

[167] Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, et al. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 2013.

[168] Benjamin Eysenbach, Tianjun Zhang, Sergey Levine, and Russ R Salakhutdinov. Contrastive learning as goal-conditioned reinforcement learning. *Advances in Neural Information Processing Systems*, 35:35603–35620, 2022.

[169] Haoqi Fan and Jiatong Zhou. Stacked latent attention for multimodal reasoning. In *CVPR*, pages 1072–1080, 2018.

[170] Robert M Fano. *Transmission of information: a statistical theory of communications*. Mit Press, 1968.

[171] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*, pages 15–29. Springer, 2010.

[172] Meir Feder and Neri Merhav. Relations between entropy and error probability. *IEEE Transactions on Information theory*, 40(1):259–266, 1994.

[173] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. *arXiv preprint arXiv:2002.07017*, 2020.

[174] Francis Ferraro, Nasrin Mostafazadeh, Ting-Hao Huang, Lucy Vanderwende, Jacob Devlin, Michel Galley, and Margaret Mitchell. A survey of current datasets for vision and language research. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 207–213, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1021. URL https://www.aclweb.org/anthology/D15-1021.

[175] Ross Flom and Lorraine E Bahrick. The development of infant discrimination of affect in multimodal and unimodal stimulation: The role of intersensory redundancy. *Developmental psychology*, 43(1):238, 2007.

[176] Andreas Foltyn and Jessica Deuschel. Towards reliable multimodal stress detection under distribution shift. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*, pages 329–333, 2021.

[177] Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9847–9857, 2021.

[178] Joseph C Franklin, Jessica D Ribeiro, Kathryn R Fox, Kate H Bentley, Evan M Kleiman, Xieyining Huang, Katherine M Musacchio, Adam C Jaroszewski, Bernard P Chang, and Matthew K Nock. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychological bulletin*, 143(2):187, 2017.

[179] Christos A Frantzidis, Charalampos Bratsas, Manousos A Klados, Evdokimos Konstantinidis, Chrysa D Lithari, Ana B Vivas, Christos L Papadelis, Eleni Kaldoudi, Costas Pappas, and Panagiotis D Bamidis. On the classification of emotional biosignals evoked while viewing affective pictures: an integrated data-mining-based approach for healthcare applications. *IEEE Transactions on Information Technology in Biomedicine*, 14(2): 309–318, 2010.

[180] Jerome H Friedman and Bogdan E Popescu. Predictive learning via rule ensembles. *The annals of applied statistics*, 2(3):916–954, 2008.

[181] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.

[182] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, pages 457–468. ACL, 2016.

[183] Hiroki Furuta, Ofir Nachum, Kuang-Huei Lee, Yutaka Matsuo, Shixiang Shane Gu, and Izzeddin Gur. Multimodal web navigation with instruction-finetuned foundation models. *arXiv preprint arXiv:2305.11854*, 2023.

[184] Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetzsche. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, pages 1–6. IEEE, 2020.

[185] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 2022.

[186] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

[187] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *CVPR*, pages 324–333, 2019.

[188] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

[189] Enrique Garcia-Ceja, Michael Riegler, Tine Nordgreen, Petter Jakobsen, et al. Mental health monitoring with multimodal sensing and machine learning: A survey. *Pervasive and Mobile Computing*, 2018.

[190] Wendell R Garner. Uncertainty and structure as psychological concepts. 1962.

[191] Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. *Advances in Neural Information Processing Systems*, 33, 2020.

[192] Itai Gat, Idan Schwartz, and Alex Schwing. Perceptual score: What data modalities does your model perceive? *NeurIPS*, 34, 2021.

[193] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *EMNLP Findings*, pages 3356–3369, 2020.

[194] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020.

[195] Mor Geva, Yoav Goldberg, and Jonathan Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *EMNLP-IJCNLP*, pages 1161–1166, 2019.

[196] AmirEmad Ghassami and Negar Kiyavash. Interaction information for causal inference: The case of directed triangle. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1326–1330. IEEE, 2017.

[197] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3681–3688, 2019.

[198] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[199] Dimitris Gkoumas, Qiuchi Li, Christina Lioma, Yijun Yu, and Dawei Song. What makes the difference? an empirical comparison of fusion strategies for multimodal language analysis. *Information Fusion*, 66:184–197.

[200] Catherine R Glenn and Matthew K Nock. Improving the short-term prediction of suicidal behavior. *American journal of preventive medicine*, 47(3):S176–S180, 2014.

[201] Amir Globerson and Tommi Jaakkola. Approximate inference using conditional entropy decompositions. In *Artificial Intelligence and Statistics*, pages 131–138. PMLR, 2007.

[202] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 513–520, USA, 2011. Omnipress. ISBN 978-1-4503-0619-5. URL `http://dl.acm.org/citation.cfm?id=3104482.3104547`.

[203] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Vqa-lol: Visual question answering under the lens of logic. In *European conference on computer vision*, pages 379–396. Springer, 2020.

[204] Scott A Golder, Dennis M Wilkinson, and Bernardo A Huberman. Rhythms of social interaction: Messaging within a massive online network. In *Communities and technologies 2007*, pages 41–66. Springer, 2007.

[205] Yash Goyal, Akrit Mohapatra, Devi Parikh, and Dhruv Batra. Towards transparent ai systems: Interpreting visual question answering models. *arXiv preprint arXiv:1608.08974*, 2016.

[206] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6904–6913, 2017.

[207] Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890. PMLR, 2019.

[208] A. Graves, A. r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649, May 2013. doi: 10.1109/ICASSP.2013.6638947.

[209] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *ICML*, 2006.

[210] Virgil Griffith and Christof Koch. Quantifying synergistic mutual information. In *Guided self-organization: inception*, pages 159–190. Springer, 2014.

[211] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284, 2020.

[212] Ken Gu. Multimodal toolkit. `https://github.com/georgian-io/Multimodal-Toolkit`,

2020.

[213] Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language. *arXiv preprint arXiv:2112.08614*, 2021.

[214] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Multimodal semi-supervised learning for image classification. In *2010 IEEE Computer society conference on computer vision and pattern recognition*, pages 902–909. IEEE, 2010.

[215] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pages 3608–3617, 2018.

[216] William H Guss, Brandon Houghton, Nicholay Topin, Phillip Wang, Cayden Codel, Manuela Veloso, and Ruslan Salakhutdinov. Minerl: a large-scale dataset of minecraft demonstrations. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2442–2448, 2019.

[217] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.

[218] Pascal Hamisu, Gregor Heinrich, Christoph Jung, Volker Hahn, Carlos Duarte, Pat Langdon, and Pradipta Biswas. Accessible ui design and multimodal interaction through hybrid tv platforms: towards a virtual-user centered design framework. In *International Conference on Universal Access in Human-Computer Interaction*, pages 32–41. Springer, 2011.

[219] Faisal Hamman and Sanghamitra Dutta. Demystifying local and global fairness trade-offs in federated learning using information theory. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities*, 2023.

[220] Kun Han, Dong Yu, and Ivan Tashev. Speech emotion recognition using deep neural network and extreme learning machine. September 2014.

[221] Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov. Explaining black box predictions and unveiling data artifacts through influence functions. In *ACL*, pages 5553–5563, 2020.

[222] Jeffrey T Hancock and Jeremy N Bailenson. The social impact of deepfakes, 2021.

[223] Darryl Hannan, Akshay Jain, and Mohit Bansal. Manymodalqa: Modality disambiguation and qa over diverse inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7879–7886, 2020.

[224] Sanjay Haresh, Sateesh Kumar, Huseyin Coskun, Shahram N Syed, Andrey Konin, Zeeshan Zia, and Quoc-Huy Tran. Learning by aligning videos in time. In *CVPR*, pages 5548–5558, 2021.

[225] Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, et al. Ur-funny: A multimodal language dataset for understanding humor. In *EMNLP-IJCNLP*, pages 2046–2056, 2019.

[226] Md Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. Humor knowledge enriched transformer for understanding multimodal humor. In *AAAI*, 2021.

[227] Javaria Hassan, Jovin Leong, and Bertrand Schneider. *Multimodal Data Collection Made Easy: The EZ-MMLA Toolkit: A Data Collection Website That Provides Educators and Researchers with Easy Access to Multimodal Data Streams.*, page 579–585. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450389358. URL https://doi.org/10.1145/3448139.3448201.

[228] Trevor Hastie and Robert Tibshirani. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.

[229] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[230] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

[231] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer*

Vision (ECCV), pages 771–787, 2018.

[232] Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. Decoupling the role of data, attention, and losses in multimodal transformers. *arXiv preprint arXiv:2102.00529*, 2021.

[233] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning AI with shared human values. In *ICLR*, 2021.

[234] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill™: A bayesian skill rating system. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 569–576. MIT Press, 2007. URL http://papers.nips.cc/paper/3079-trueskilltm-a-bayesian-skill-rating-system.pdf.

[235] Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 861–877, 2020.

[236] Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor" understanding" benchmarks from the new yorker caption contest. *arXiv preprint arXiv:2209.06293*, 2022.

[237] Shohei Hidaka and Chen Yu. Analyzing multimodal time series as dynamical systems. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, ICMI-MLMI '10, pages 53:1–53:8, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0414-6. doi: 10.1145/1891903.1891968. URL http://doi.acm.org/10.1145/1891903.1891968.

[238] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

[239] Ryota Hinami, Junwei Liang, Shin'ichi Satoh, and Alexander Hauptmann. Multimodal co-training for selecting good examples from webly labeled video. *arXiv preprint arXiv:1804.06057*, 2018.

[240] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2018.

[241] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[242] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[243] Danfeng Hong, Lianru Gao, Naoto Yokoya, Jing Yao, Jocelyn Chanussot, Qian Du, and Bing Zhang. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5):4340–4354, 2020.

[244] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *IEEE TPAMI*, 2019.

[245] Ming Hou, Jiajia Tang, Jianhai Zhang, Wanzeng Kong, and Qibin Zhao. Deep multimodal multilinear fusion with high-order polynomial pooling. *NeurIPS*, 32:12136–12145, 2019.

[246] Tsung-Yu Hsieh, Yiwei Sun, Suhang Wang, and Vasant Honavar. Adaptive structural co-regularization for unsupervised multi-view feature selection. In *2019 IEEE International Conference on Big Knowledge (ICBK)*. IEEE, 2019.

[247] Tzu-Ming Harry Hsu, Wei-Hung Weng, Willie Boag, Matthew McDermott, and Peter Szolovits. Unsupervised multimodal representation learning across medical images and reports. *arXiv preprint arXiv:1811.08615*, 2018.

[248] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution

for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

[249] Wei-Ning Hsu and James Glass. Disentangling by partitioning: A representation learning framework for multimodal sensory data. *arXiv preprint arXiv:1805.11264*, 2018.

[250] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *CVPR*, pages 9248–9257, 2019.

[251] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR, 2020.

[252] Peng Hu, Dezhong Peng, Xu Wang, and Yong Xiang. Multimodal adversarial network for cross-modal retrieval. *Knowledge-Based Systems*, 180:38–50, 2019.

[253] Ronghang Hu and Amanpreet Singh. Transformer is all you need: Multimodal multitask learning with a unified transformer. *arXiv preprint arXiv:2102.10772*, 2021.

[254] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *CVPR*, pages 4555–4564, 2016.

[255] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *ICCV*, pages 804–813, 2017.

[256] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *NeurIPS*, 2020.

[257] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *CoRR*, abs/1904.05342, 2019.

[258] Po-Yao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, and Alexander Hauptmann. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. *arXiv preprint arXiv:2103.08849*, 2021.

[259] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *arXiv preprint arXiv:2201.07207*, 2022.

[260] Xin Huang, Yuxin Peng, and Mingkuan Yuan. Cross-modal common representation learning by hybrid transfer network. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1893–1900, 2017.

[261] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *NeurIPS*, 34, 2021.

[262] Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). *arXiv preprint arXiv:2203.12221*, 2022.

[263] Zhenhua Huang, Xin Xu, Juan Ni, Honghao Zhu, and Cheng Wang. Multimodal representation learning for recommendation in internet of things. *IEEE Internet of Things Journal*, 6(6):10675–10685, 2019.

[264] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *CoRR*, abs/1508.01991, 2015.

[265] Drew Hudson and Christopher D Manning. Learning by abstraction: The neural state machine. *NeurIPS*, 2019.

[266] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*, 2018.

[267] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709, 2019.

[268] Thelma Hunt. The measurement of social intelligence. *Journal of Applied Psychology*, 12(3):317, 1928.

[269] HyeongJoo Hwang, Geon-Hyeong Kim, Seunghoon Hong, and Kee-Eung Kim. Variational interaction

information maximization for cross-domain disentanglement. *Advances in Neural Information Processing Systems*, 33:22479–22491, 2020.

[270] iMotions. Facial expression analysis, 2017. URL goo.gl/1rh1JN.

[271] Masha Itkina, B. Ivanovic, Ransalu Senanayake, Mykel J. Kochenderfer, and Marco Pavone. Evidential sparsification of multimodal latent spaces in conditional variational autoencoders. *ArXiv*, abs/2010.09164, 2020.

[272] Jan Ittner, Lukasz Bolikowski, Konstantin Hemker, and Ricardo Kennedy. Feature synergy, redundancy, and independence in global model explanations using shap vector decomposition. *arXiv preprint arXiv:2107.12436*, 2021.

[273] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *Association for Computational Linguistics*, 2015.

[274] Allan Jabri, Armand Joulin, and Laurens van der Maaten. Revisiting visual question answering baselines. In *ECCV*, pages 727–739. Springer, 2016.

[275] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

[276] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention. *arXiv preprint arXiv:2103.03206*, 2021.

[277] Alejandro Jaimes and Nicu Sebe. Multimodal human–computer interaction: A survey. *Computer vision and image understanding*, 108(1-2):116–134, 2007.

[278] Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. Mural: multimodal, multitask retrieval across languages. *arXiv preprint arXiv:2109.05125*, 2021.

[279] L. C. Jain and L. R. Medsker. *Recurrent Neural Networks: Design and Applications*. CRC Press, Inc., Boca Raton, FL, USA, 1st edition, 1999. ISBN 0849371813.

[280] Aleks Jakulin and Ivan Bratko. Quantifying and visualizing attribute interactions: An approach based on entropy. 2003.

[281] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. You said that?: Synthesising talking faces from audio. *International Journal of Computer Vision*, 127(11):1767–1779, 2019.

[282] Anubhav Jangra, Adam Jatowt, Mohammad Hasanuzzaman, and Sriparna Saha. Text-image-video summary generation using joint integer linear programming. In *European Conference on Information Retrieval*. Springer, 2020.

[283] Siddhant M. Jayakumar, Wojciech M. Czarnecki, Jacob Menick, Jonathan Schwarz, Jack Rae, Simon Osindero, Yee Whye Teh, Tim Harley, and Razvan Pascanu. Multiplicative interactions and where to find them. In *International Conference on Learning Representations*, 2020.

[284] Siddhant M. Jayakumar, Wojciech M. Czarnecki, Jacob Menick, Jonathan Schwarz, Jack Rae, Simon Osindero, Yee Whye Teh, Tim Harley, and Razvan Pascanu. Multiplicative interactions and where to find them. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rylnK6VtDH.

[285] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, et al. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021.

[286] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[287] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, et al. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[288] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-

ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8, 2019.

[289] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pages 2901–2910, 2017.

[290] Jonathan Kahana and Yedid Hoshen. A contrastive objective for learning disentangled representations. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 579–595. Springer, 2022.

[291] Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. One model to learn them all. *arXiv preprint arXiv:1706.05137*, 2017.

[292] Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. Mdetr–modulated detection for end-to-end multi-modal understanding. *arXiv preprint arXiv:2104.12763*, 2021.

[293] Atsushi Kanehira, Kentaro Takemoto, Sho Inayoshi, and Tatsuya Harada. Multimodal explanations by predicting counterfactuality in videos. In *CVPR*, pages 8594–8602, 2019.

[294] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *NeurIPS*, 27, 2014.

[295] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020.

[296] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[297] Vasil Khalidov, Florence Forbes, and Radu Horaud. Conjugate mixture models for clustering multimodal data. *Neural Computation*, 2011.

[298] Aparajita Khan and Pradipta Maji. Approximate graph laplacians for multimodal data clustering. *IEEE TPAMI*, 43(3):798–813, 2019.

[299] Aparajita Khan and Pradipta Maji. Multi-manifold optimization for multi-view subspace clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 33(8):3895–3907, 2021.

[300] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[301] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020.

[302] John F Kihlstrom and Nancy Cantor. Social intelligence. *Handbook of intelligence*, 2:359–379, 2000.

[303] Byoungjip Kim, Sungik Choi, Dasol Hwang, Moontae Lee, and Honglak Lee. Transferring pre-trained multimodal representations with cross-modal similarity matching. *Advances in Neural Information Processing Systems*, 35:30826–30839, 2022.

[304] Elizabeth S Kim, Lauren D Berkovits, Emily P Bernier, Dan Leyzberg, Frederick Shic, Rhea Paul, and Brian Scassellati. Social robots as embedded reinforcers of social behavior in children with autism. *Journal of autism and developmental disorders*, 43(5):1038–1049, 2013.

[305] Kyung Hwan Kim, Seok Won Bang, and Sang Ryong Kim. Development of person-independent emotion recognition system based on multiple physiological signals. In *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society Engineering in Medicine and Biology*, volume 1, pages 50–51. IEEE, 2002.

[306] Minjae Kim, David K Han, and Hanseok Ko. Joint patch clustering-based dictionary learning for multimodal image fusion. *Information Fusion*, 27:198–214, 2016.

[307] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[308] Elsa A Kirchner, Stephen H Fairclough, and Frank Kirchner. Embedded multimodal interfaces in robotics: applications, future trends, and societal implications. In *The Handbook of Multimodal-Multisensor Interfaces: Language Processing, Software, Commercialization, and Emerging Directions-Volume 3*, pages 523–576. 2019.

[309] Murat Kocaoglu, Alexandros Dimakis, Sriram Vishwanath, and Babak Hassibi. Entropic causal inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[310] Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Text-to-image generation grounded by fine-grained user attention. In *WACV*, 2021.

[311] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *ICML*, pages 5338–5348. PMLR, 2020.

[312] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, et al. Wilds: A benchmark of in-the-wild distribution shifts. *arXiv preprint arXiv:2012.07421*, 2020.

[313] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15):5802–5805, 2013.

[314] Stephen M Kosslyn, Giorgio Ganis, and William L Thompson. Multimodal images in the brain. *The neurophysiological foundations of mental and motor imagery*, pages 3–16, 2010.

[315] Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Visual coreference resolution in visual dialog using neural module networks. In *ECCV*, pages 153–169, 2018.

[316] Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2011.

[317] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017.

[318] Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*, 2022.

[319] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[320] Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti, and Siva Reddy. Image retrieval from contextual descriptions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3426–3440, 2022.

[321] Joseph B Kruskal. An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM review*, 25(2):201–237, 1983.

[322] Patricia K. Kuhl. A new view of language acquisition. *Proceedings of the National Academy of Sciences*, 97 (22):11850–11857, 2000. ISSN 0027-8424. doi: 10.1073/pnas.97.22.11850. URL `http://www.pnas.org/content/97/22/11850`.

[323] Heinrich Küttler, Nantas Nardelli, Alexander Miller, Roberta Raileanu, Marco Selvatici, Edward Grefenstette, and Tim Rocktäschel. The nethack learning environment. *Advances in Neural Information Processing Systems*, 33:7671–7684, 2020.

[324] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1. URL `http://dl.acm.org/citation.cfm?id=645530.655813`.

[325] Pei Ling Lai and Colin Fyfe. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 2000.

[326] Egor Lakomkin, Cornelius Weber, Sven Magg, and Stefan Wermter. Reusing neural speech representations for auditory emotion recognition. *CoRR*, abs/1803.11508, 2018.

[327] Karthik Lakshmanan, Patrick T. Sadtler, Elizabeth C. Tyler-Kabara, Aaron P. Batista, and Byron M. Yu. Extracting low-dimensional latent structure from time series in the presence of delays. *Neural Computation*, 27:1825–1856, 2015.

[328] Matthew Michael Large, Daniel Thomas Chung, Michael Davidson, Mark Weiser, and Christopher James Ryan. In-patient suicide: selection of people at risk, failure of protection and the possibility of causation. *BJPsych Open*, 3(3):102–105, 2017.

[329] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

[330] Quoc V. Le, Navdeep Jaitly, and Geoffrey E. Hinton. A simple way to initialize recurrent networks of rectified linear units. *CoRR*, abs/1504.00941, 2015.

[331] Rémi Lebret, Pedro Pinheiro, and Ronan Collobert. Phrase-based image captioning. In *ICML*, pages 2085–2094. PMLR, 2015.

[332] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[333] Dong Won Lee, Chaitanya Ahuja, Paul Pu Liang, Sanika Natu, and Louis-Philippe Morency. Lecture presentations multimodal dataset: Towards understanding multimodality in educational videos. In *ICCV*, pages 20087–20098, October 2023.

[334] Michelle A Lee, Brent Yi, Roberto Martín-Martín, Silvio Savarese, and Jeannette Bohg. Multimodal sensor fusion with differentiable filters. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10444–10451. IEEE.

[335] Michelle A Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, et al. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *ICRA*, pages 8943–8950. IEEE, 2019.

[336] Michelle A Lee, Matthew Tan, Yuke Zhu, and Jeannette Bohg. Detect, reject, correct: Crossmodal compensation of corrupted sensors. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[337] Sangho Lee, Youngjae Yu, Gunhee Kim, Thomas Breuel, Jan Kautz, and Yale Song. Parameter efficient multimodal transformers for video representation learning. In *ICLR*, 2020.

[338] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *CVPR*, pages 14943–14952, 2023.

[339] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, pages 1369–1379, 2018.

[340] Luis A Leiva, Asutosh Hota, and Antti Oulasvirta. Enrico: A dataset for topic modeling of mobile ui designs. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–4, 2020.

[341] R Gary Leonard and George Doddington. Tidigits speech corpus. *Texas Instruments, Inc*, 1993.

[342] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, pages 3045–3059, 2021.

[343] Willem J.M Levelt and Stephanie Kelter. Surface form and memory in question answering. *Cognitive Psychology*, 14(1):78 – 106, 1982. ISSN 0010-0285. doi: https://doi.org/10.1016/0010-0285(82)90005-6. URL `http://www.sciencedirect.com/science/article/pii/0010028582900056`.

[344] David D Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pages 148–156. Elsevier, 1994.

[345] Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. Multi-modal summarization for

asynchronous collection of text, image, audio and video. In *EMNLP*, pages 1092–1102, 2017.

[346] Jiaxin Li, Danfeng Hong, Lianru Gao, Jing Yao, Ke Zheng, Bing Zhang, and Jocelyn Chanussot. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation*, 112:102926, 2022.

[347] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[348] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[349] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does bert with vision look at? In *ACL*, pages 5265–5275, 2020.

[350] Manling Li, Lingyu Zhang, Heng Ji, and Richard J Radke. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *ACL*, pages 2190–2196, 2019.

[351] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting text and images with event structures. In *CVPR*, pages 16420–16429, 2022.

[352] Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. Vmsmo: Learning to generate multimodal summary for video-based news articles. *arXiv preprint arXiv:2010.05406*, 2020.

[353] Qing Li, Boqing Gong, Yin Cui, Dan Kondratyuk, Xianzhi Du, et al. Towards a unified foundation model: Jointly pre-training transformers on unpaired images and text. *arXiv preprint arXiv:2112.07074*, 2021.

[354] Shu Li, Wei Wang, Wen-Tao Li, and Pan Chen. Multi-view representation learning with manifold smoothness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8447–8454, 2021.

[355] Shuang Li, Xavier Puig, Yilun Du, Clinton Wang, Ekin Akyurek, Antonio Torralba, Jacob Andreas, and Igor Mordatch. Pre-trained language models for interactive decision-making. *arXiv preprint arXiv:2202.01771*, 2022.

[356] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2020.

[357] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL-IJCNLP*, 2021.

[358] Paul Pu Liang. Brainish: Formalizing a multimodal language for intelligence and consciousness. *arXiv preprint arXiv:2205.00001*, 2022.

[359] Paul Pu Liang, Ziyin Liu, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Multimodal language analysis with recurrent multistage fusion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 150–161, 2018.

[360] Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Computational modeling of human multimodal language: The mosei dataset and interpretable dynamic fusion. 2018.

[361] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Multimodal local-global ranking fusion for emotion recognition. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI 2018, 2018.

[362] Paul Pu Liang, Yao Chong Lim, Yao-Hung Hubert Tsai, Ruslan Salakhutdinov, and Louis-Philippe Morency. Strong and simple baselines for multimodal utterance embeddings. In *NAACL-HLT*, 2019.

[363] Paul Pu Liang, Terrance Liu, Liu Ziyin, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. *NeurIPS Workshop on Federated Learning*, 2019.

[364] Paul Pu Liang, Zhun Liu, Yao-Hung Hubert Tsai, Qibin Zhao, Ruslan Salakhutdinov, and Louis-Philippe Morency. Learning representations from imperfect time series data via tensor rank regularization. In *ACL*, 2019.

[365] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations. In *ACL*, pages 5502–5515, 2020.

[366] Paul Pu Liang, Terrance Liu, Anna Cai, Michal Muszynski, Ryo Ishii, Nicholas Allen, Randy Auerbach, et al. Learning language and multimodal privacy-preserving markers of mood from mobile data. In *ACL/IJCNLP (1)*, 2021.

[367] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Yufan Chen, et al. Multibench: Multiscale benchmarks for multimodal representation learning. In *NeurIPS Datasets and Benchmarks Track*, 2021.

[368] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In *ICML*, 2021.

[369] Paul Pu Liang, Peter Wu, Liu Ziyin, Louis-Philippe Morency, and Ruslan Salakhutdinov. Cross-modal generalization: Learning in low resource modalities via meta-alignment. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2680–2689, 2021.

[370] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Yudong Liu, Jeffrey Tsaw, Shengtong Mo, Dani Yogatama, Louis-Philippe Morency, and Ruslan Salakhutdinov. Highmmt: Towards modality and task generalization for high-modality representation learning. *Transactions on Machine Learning Research*, 2022.

[371] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*, 2022.

[372] Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency. Quantifying & modeling feature interactions: An information decomposition framework. In *Advances in Neural Information Processing Systems*, 2023.

[373] Paul Pu Liang, Yun Cheng, Ruslan Salakhutdinov, and Louis-Philippe Morency. Multimodal fusion interactions: A study of human and automatic quantification. In *ICMI*, 2023.

[374] Paul Pu Liang, Zihao Deng, Martin Ma, James Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. In *Advances in Neural Information Processing Systems*, 2023.

[375] Paul Pu Liang, Yiwei Lyu, Gunjan Chhablani, Nihal Jain, Zihao Deng, Xingbo Wang, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multiviz: Towards visualizing and understanding multimodal models. In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=i2_TvOFmEml`.

[376] Paul Pu Liang, Chun Kai Ling, Yun Cheng, Alex Obolenskiy, Yudong Liu, Rohan Pandey, Alex Wilf, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal learning without multimodal data: Guarantees and applications. In *International Conference on Learning Representations*, 2024.

[377] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *NeurIPS*, 35:17612–17625, 2022.

[378] Valerii Likhosherstov, Mostafa Dehghani, Anurag Arnab, Krzysztof Marcin Choromanski, Mario Lucic, Yi Tay, and Adrian Weller. Polyvit: Co-training vision transformers on images, videos and audio, 2022.

[379] Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 2021.

[380] Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In *EMNLP-IJCNLP*, pages 2829–2839, 2019.

[381] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[382] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

[383] Jana Lipkova, Richard J Chen, Bowen Chen, Ming Y Lu, Matteo Barbieri, Daniel Shao, et al. Artificial intelligence for multimodal data integration in oncology. *Cancer cell*, 2022.

[384] Alex Liu, SouYoung Jin, Cheng-I Lai, Andrew Rouditchenko, Aude Oliva, and James Glass. Cross-modal discrete representation learning. In *ACL*, pages 3013–3035, 2022.

[385] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[386] Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. Mitigating political bias in language models through reinforced calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14857–14866, 2021.

[387] Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S Rosenblum. Mmkg: multi-modal knowledge graphs. In *European Semantic Web Conference*, pages 459–474. Springer, 2019.

[388] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256, 2018.

[389] Carlos Eduardo Rodrigues Lopes and Linnyer Beatrys Ruiz. On the development of a multi-tier, multimodal wireless sensor network for wild life monitoring. In *2008 1st IFIP Wireless Days*, pages 1–5. IEEE, 2008.

[390] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

[391] Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. Pretrained transformers as universal computation engines. *arXiv preprint arXiv:2103.05247*, 2021.

[392] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[393] Yadong Lu, Chunyuan Li, Haotian Liu, Jianwei Yang, Jianfeng Gao, and Yelong Shen. An empirical study of scaling instruct-tuned large multimodal models. *arXiv preprint arXiv:2309.09958*, 2023.

[394] Jelena Luketina, Nantas Nardelli, Gregory Farquhar, Jakob N Foerster, Jacob Andreas, Edward Grefenstette, Shimon Whiteson, and Tim Rocktäschel. A survey of reinforcement learning informed by natural language. In *IJCAI*, 2019.

[395] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[396] Yiwei Lyu, Paul Pu Liang, Zihao Deng, Ruslan Salakhutdinov, and Louis-Philippe Morency. Dime: Fine-grained interpretations of multimodal models via disentangled local explanations. *arXiv preprint arXiv:2203.02013*, 2022.

[397] Martin Q Ma, Yao-Hung Hubert Tsai, Paul Pu Liang, Han Zhao, Kun Zhang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Conditional contrastive learning for improving fairness in self-supervised learning. *arXiv preprint arXiv:2106.02866*, 2021.

[398] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. *arXiv preprint arXiv:2103.05677*, 2021.

[399] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18177–18186, 2022.

[400] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf, 2016. URL `http://arxiv.org/abs/1603.01354`. ACL 2016.

[401] Emiliano Macaluso and Jon Driver. Multisensory spatial interactions: a window onto functional integration in

the human brain. *Trends in neurosciences*, 28(5):264–271, 2005.

[402] T Soni Madhulatha. An overview on clustering methods. *arXiv preprint arXiv:1205.1117*, 2012.

[403] Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *CVPR*, pages 6884–6893, 2017.

[404] Jonathan Malmaud, Jonathan Huang, Vivek Rathod, Nicholas Johnston, Andrew Rabinovich, and Kevin Murphy. What's cookin'? interpreting cooking videos using text, speech and vision. In *NAACL-HLT*, pages 143–152, 2015.

[405] Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1062. URL https://www.aclweb.org/anthology/N19-1062.

[406] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *ICLR*, 2018.

[407] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016.

[408] Matthew Marge, Carol Espy-Wilson, Nigel G Ward, Abeer Alwan, Yoav Artzi, Mohit Bansal, et al. Spoken language interaction with robots: Recommendations for future research. *Computer Speech & Language*, 2022.

[409] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. The more you know: Using knowledge graphs for image classification. In *CVPR*, pages 20–28. IEEE, 2017.

[410] Emily E Marsh and Marilyn Domas White. A taxonomy of relationships between images and text. *Journal of documentation*, 2003.

[411] Alessio Mazzetto, Dylan Sam, Andrew Park, Eli Upfal, and Stephen Bach. Semi-supervised aggregation of dependent weak supervision sources with performance guarantees. In *AISTATS*, 2021.

[412] William McGill. Multivariate information transmission. *Transactions of the IRE Professional Group on Information Theory*, 4(4):93–111, 1954.

[413] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, 2016.

[414] Dalila Mekhaldi. Multimodal document alignment: towards a fully-indexed multimedia archive. In *Proceedings of the Multimedia Information Retrieval Workshop, SIGIR, Amsterdam, the Netherlands*, 2007.

[415] Luke Melas-Kyriazi, Alexander M Rush, and George Han. Training for diversity in image paragraph captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 757–761, 2018.

[416] Yu Meng, Chenyan Xiong, Payal Bajaj, Paul Bennett, Jiawei Han, Xia Song, et al. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *Advances in Neural Information Processing Systems*, 34:23102–23114, 2021.

[417] Luke Merrick and Ankur Taly. The explanation game: Explaining machine learning models using shapley values. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 17–38. Springer, 2020.

[418] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*. Springer, 2020.

[419] George Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information, 1956. URL http://cogprints.org/730/.

[420] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[421] Weiqing Min, Shuqiang Jiang, Jitao Sang, Huayang Wang, Xinda Liu, and Luis Herranz. Being a supercook: Joint food attributes and multimodal content modeling for recipe retrieval and exploration. *IEEE Transactions on Multimedia*, 19(5):1100–1113, 2016.

[422] Shotaro Misawa, Motoki Taniguchi, Yasuhide Miura, and Tomoko Ohkuma. Character-based bidirectional lstm-crf with words and characters for japanese named entity recognition. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 97–102. Association for Computational Linguistics, 2017. URL http://aclweb.org/anthology/W17-4114.

[423] Dalibor Mitrović, Matthias Zeppelzauer, and Christian Breiteneder. Features for content-based audio retrieval. In *Advances in computers*, volume 78, pages 71–150. Elsevier, 2010.

[424] Shentong Mo, Paul Pu Liang, Russ Salakhutdinov, and Louis-Philippe Morency. Multiiot: Towards large-scale multisensory learning for the internet of things. *arXiv preprint arXiv:2311.06217*, 2023.

[425] Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A Gutman, Jill S Barnholtz-Sloan, José E Velázquez Vega, Daniel J Brat, and Lee AD Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970–E2979, 2018.

[426] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. *CoRR*, abs/1902.00146, 2019. URL http://arxiv.org/abs/1902.00146.

[427] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 169–176, 2011.

[428] Olivier Morere, Hanlin Goh, Antoine Veillard, Vijay Chandrasekhar, and Jie Lin. Co-regularized deep representations for video summarization. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 3165–3169. IEEE, 2015.

[429] Ghulam Muhammad, Fatima Alshehri, Fakhri Karray, Abdulmotaleb El Saddik, et al. A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. *Information Fusion*, 76:355–375, 2021.

[430] Sudipto Mukherjee, Himanshu Asnani, and Sreeram Kannan. Ccmi: Classifier based conditional mutual information estimation. In *Uncertainty in artificial intelligence*, pages 1083–1093. PMLR, 2020.

[431] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *CVPR*, pages 122–132, 2020.

[432] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.

[433] Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. Just-in-time adaptive interventions (jitais) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, 52(6):446–462, 2018.

[434] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *CVPR*, pages 299–307, 2017.

[435] Bence Nanay. Multimodal mental imagery. *Cortex*, 105:125–134, 2018.

[436] Milind Naphade, John R Smith, Jelena Tesic, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis. Large-scale concept ontology for multimedia. *IEEE multimedia*, 13(3):86–91, 2006.

[437] Karthik Narasimhan, Regina Barzilay, and Tommi Jaakkola. Grounding language for transfer in deep reinforcement learning. *Journal of Artificial Intelligence Research*, 63:849–874, 2018.

[438] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*, 2022.

[439] Shahla Nemati, Reza Rohani, Mohammad Ehsan Basiri, Moloud Abdar, Neil Y Yen, and Vladimir Makarenkov. A hybrid latent space data fusion method for multimodal emotion recognition. *IEEE Access*, 7:172948–172964, 2019.

[440] Cancer Genome Atlas Research Network. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *New England Journal of Medicine*, 372(26):2481–2498, 2015.

[441] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *CVPR*, pages 20395–20405, 2022.

[442] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011.

[443] Nam D Nguyen and Daifeng Wang. Multiview learning for understanding functional multiomics. *PLoS computational biology*, 16(4):e1007677, 2020.

[444] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[445] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2015.

[446] Adrian Nilsson, Simon Smith, Gregor Ulm, Emil Gustavsson, and Mats Jirstrand. A performance evaluation of federated learning algorithms. In *DIDL*, 2018.

[447] Sheila Nirenberg, Steve M Carcieri, Adam L Jacobs, and Peter E Latham. Retinal ganglion cells act largely as independent encoders. *Nature*, 411(6838):698–701, 2001.

[448] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *CVPR*, pages 12700–12710, 2021.

[449] Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI 2016, pages 284–288, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4556-9. doi: 10.1145/2993148.2993176. URL `http://doi.acm.org/10.1145/2993148.2993176`.

[450] Zeljko Obrenovic and Dusan Starcevic. Modeling multimodal human-computer interaction. *Computer*, 2004.

[451] Brendan O'Donoghue, Eric Chu, Neal Parikh, and Stephen Boyd. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068, June 2016. URL `http://stanford.edu/~boyd/papers/scs.html`.

[452] Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, et al. Parallel wavenet: Fast high-fidelity speech synthesis. In *ICML*, pages 3918–3926. PMLR, 2018.

[453] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[454] OpenAI. Gpt-4 technical report, 2023.

[455] Jahna Otterbacher, Alessandro Checco, Gianluca Demartini, and Paul Clough. Investigating user perception of gender bias in image search: the role of sexism. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 933–936, 2018.

[456] Christian Otto, Matthias Springstein, Avishek Anand, and Ralph Ewerth. Characterization and classification of semantic image-text relations. *International Journal of Multimedia Information Retrieval*, 9:31–45, 2020.

[457] Sharon Oviatt. Ten myths of multimodal interaction. *Communications of the ACM*, 42(11):74–81, 1999.

[458] Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aaron Van den Oord, Sergey Levine, and Pierre Sermanet. Wasserstein dependency measure for representation learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[459] Dinesh K Pai. Multisensory interaction: Real and virtual. In *Robotics Research. The Eleventh International*

Symposium, pages 489–498. Springer, 2005.

[460] Ari Pakman, Amin Nejatbakhsh, Dar Gilboa, Abdullah Makkeh, Luca Mazzucato, Michael Wibral, and Elad Schneidman. Estimating the unique information of continuous variables. *Advances in Neural Information Processing Systems*, 34:20295–20307, 2021.

[461] Shruti Palaskar, Jindrich Libovickỳ, Spandana Gella, and Florian Metze. Multimodal abstractive summarization for how2 videos. *arXiv preprint arXiv:1906.07901*, 2019.

[462] Maja Pantic and Leon JM Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, 2003.

[463] Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. Seeing past words: Testing the cross-modal capabilities of pretrained v&l models on counting tasks. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 32–44, 2021.

[464] Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*, 2016.

[465] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *CVPR*, pages 8779–8788, 2018.

[466] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image. In *European Conference on Computer Vision*, pages 508–524. Springer, 2020.

[467] Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pages 50–57, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2885-2. doi: 10.1145/2663204.2663260. URL `http://doi.acm.org/10.1145/2663204.2663260`.

[468] Sarah Partan and Peter Marler. Communication goes multimodal. *Science*, 283(5406):1272–1273, 1999.

[469] Sarah R Partan and Peter Marler. Issues in the classification of multimodal communication signals. *The American Naturalist*, 166(2):231–245, 2005.

[470] Judea Pearl. *Causality*. Cambridge university press, 2009.

[471] Catherine Pelachaud. Modelling multimodal expression of emotion in a virtual agent. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3539–3548, 2009.

[472] Catherine Pelachaud, Carlos Busso, and Dirk Heylen. Multimodal behavior modeling for socially interactive agents. In *The Handbook on Socially Interactive Agents*, pages 259–310. 2021.

[473] Alejandro Peña, Ignacio Serna, Aythami Morales, and Julian Fierrez. Faircvtest demo: Understanding bias in multimodal learning with a testbed in fair automatic recruitment. In *ICMI*, pages 760–761, 2020.

[474] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[475] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[476] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

[477] Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. Utterance-level multimodal sentiment analysis. In *ACL (1)*, pages 973–982, 2013.

[478] Juan-Manuel Pérez-Rúa, Valentin Vielzeuf, Stéphane Pateux, Moez Baccouche, and Frédéric Jurie. Mfas: Multimodal fusion architecture search. In *CVPR*, pages 6966–6975, 2019.

[479] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke

Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202.

[480] Pouya Pezeshkpour, Liyan Chen, and Sameer Singh. Embedding multimodal relational data for knowledge base completion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3208–3218, 2018.

[481] Hai Pham, Thomas Manzini, Paul Pu Liang, and Barnabas Poczos. Seq2seq2sentiment: Multimodal sequence to sequence models for sentiment analysis. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 53–63, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W18-3308.

[482] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *AAAI*, volume 33, pages 6892–6899, 2019.

[483] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[484] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015.

[485] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015.

[486] Petra Poklukar, Miguel Vasco, Hang Yin, Francisco S Melo, Ana Paiva, and Danica Kragic. Geometric multimodal contrastive representation learning. In *International Conference on Machine Learning*, pages 17782–17800. PMLR, 2022.

[487] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180. PMLR, 2019.

[488] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. Convolutional mkl based multimodal emotion recognition and sentiment analysis. 2016.

[489] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 2017.

[490] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *Association for Computational Linguistics*, 2017.

[491] Shraman Pramanick, Aniket Roy, and Vishal M Patel. Multimodal learning using optimal transport for sarcasm and humor detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022.

[492] Stefania Prendes-Alvarez and Charles B Nemeroff. Personalized medicine: Prediction of disease vulnerability in mood disorders. *Neuroscience letters*, 669:10–13, 2018.

[493] Alexandra M Proca, Fernando E Rosas, Andrea I Luppi, Daniel Bor, Matthew Crosby, and Pedro AM Mediano. Synergistic information supports modality integration and flexible learning in neural networks solving multiple tasks. *arXiv preprint arXiv:2210.02996*, 2022.

[494] Ariadna Quattoni, Sybor Wang, Louis-Philippe Morency, Michael Collins, and Trevor Darrell. Hidden conditional random fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(10):1848–1852, October 2007. ISSN 0162-8828. doi: 10.1109/TPAMI.2007.1124. URL http://dx.doi.org/10.1109/TPAMI.2007.1124.

[495] Gorjan Radevski, Dusan Grujicic, Matthew Blaschko, Marie-Francine Moens, and Tinne Tuytelaars. Multimodal distillation for egocentric action recognition. In *ICCV*, pages 5213–5224, 2023.

[496] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are

unsupervised multitask learners. 2019.

[497] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021.

[498] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020.

[499] Anil Rahate, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Information Fusion*, 81:203–239, 2022.

[500] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *ACL*, pages 2359–2369, 2020.

[501] Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrušaitis, and Roland Goecke. Extending long short-term memory for multi-view structured learning. In *European Conference on Computer Vision*, 2016.

[502] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831. PMLR, 2021.

[503] Benjamin J Raphael, Ralph H Hruban, Andrew J Aguirre, Richard A Moffitt, Jen Jen Yeh, Chip Stewart, A Gordon Robertson, Andrew D Cherniack, Manaswi Gupta, Gad Getz, et al. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer cell*, 32(2):185–203, 2017.

[504] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM Multimedia*, pages 251–260, 2010.

[505] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Giménez, Yury Sulsky, et al. One model to learn them all. *Deepmind Technical Report*, 2022.

[506] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[507] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *NeurIPS*, 32, 2019.

[508] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[509] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *KDD*, pages 1135–1144, 2016.

[510] John A Rice. *Mathematical statistics and data analysis*. Cengage Learning, 2006.

[511] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[512] Candace Ross, Boris Katz, and Andrei Barbu. Measuring social biases in grounded vision and language embeddings. *arXiv preprint arXiv:2002.08911*, 2020.

[513] Yong Rui, Thomas S Huang, and Shih-Fu Chang. Image retrieval: Current techniques, promising directions, and open issues. *Journal of visual communication and image representation*, 10(1):39–62, 1999.

[514] Natalie Ruiz, Ronnie Taib, and Fang Chen. Examining the redundancy of multimodal input. In *Proceedings of the 18th Australia conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments*, pages 389–392, 2006.

[515] Shagan Sah, Sourabh Kulhare, Allison Gray, Subhashini Venugopalan, Emily Prud'Hommeaux, and Raymond Ptucha. Semantic text summarization of long videos. In *WACV*, pages 989–997. IEEE, 2017.

[516] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *AAAI*, volume 34, pages 8732–8740, 2020.

[517] Sethuraman Sankaran, David Yang, and Ser-Nam Lim. Multimodal fusion refiner networks. *arXiv preprint arXiv:2104.03435*, 2021.

[518] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. In *ACL*, pages 5477–5490, 2020.

[519] Raeid Saqur and Karthik Narasimhan. Multimodal graph networks for compositional generalization in visual question answering. *NeurIPS*, 33:3070–3081, 2020.

[520] Marcelo Sardelich and Suresh Manandhar. Multimodal deep learning for short-term stock volatility prediction. *arXiv preprint arXiv:1812.10479*, 2018.

[521] Mehmet Emre Sargin, Yücel Yemez, Engin Erzin, and A. Murat Tekalp. Audiovisual synchronization and fusion using canonical correlation analysis. *IEEE Transactions on Multimedia*, 9(7):1396–1403, 2007.

[522] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *ICCV*, pages 9339–9347, 2019.

[523] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

[524] Bob R Schadenberg, Dennis Reidsma, Vanessa Evers, Daniel P Davison, Jamy J Li, Dirk KJ Heylen, Carlos Neves, Paulo Alvito, Jie Shen, Maja Pantić, et al. Predictable robots for autistic children—variance in robot behaviour, idiosyncrasies in autistic children's characteristics, and child–robot engagement. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 28(5):1–42, 2021.

[525] Manos Schinas, Symeon Papadopoulos, Georgios Petkos, Yiannis Kompatsiaris, and Pericles A Mitkas. Multimodal graph-based event detection and summarization in social media streams. In *ACM Multimedia*, 2015.

[526] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *ICMI*, pages 400–408, 2018.

[527] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

[528] M. Schuster and K.K. Paliwal. Bidirectional recurrent neural networks. *Trans. Sig. Proc.*, 45(11):2673–2681, November 1997. ISSN 1053-587X. doi: 10.1109/78.650093. URL `http://dx.doi.org/10.1109/78.650093`.

[529] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[530] Lucia Seminara, Paolo Gastaldo, Simon J Watt, Kenneth F Valyear, Fernando Zuher, and Fulvio Mastrogiovanni. Active haptic perception in robots: a review. *Frontiers in neurorobotics*, 13:53, 2019.

[531] Luciano Serafini and Artur d'Avila Garcez. Logic tensor networks: Deep learning and logical reasoning from data and knowledge. *arXiv preprint arXiv:1606.04422*, 2016.

[532] Rajiv Shah and Roger Zimmermann. *Multimodal analysis of user-generated multimedia content*. Springer, 2017.

[533] Dougal Shakespeare, Lorenzo Porcaro, Emilia Gómez, and Carlos Castillo. Exploring artist gender bias in music recommendation. *arXiv preprint arXiv:2009.01715*, 2020.

[534] Bin Shan, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil 2.0: Multi-view contrastive learning for image-text pre-training, 2022.

[535] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 1948.

[536] Naeha Sharif, Uzair Nadeem, Syed Afaq Ali Shah, Mohammed Bennamoun, and Wei Liu. Vision to language: Methods, metrics and datasets. In *Machine Learning Paradigms*, pages 9–62. Springer, 2020.

[537] Chhavi Sharma, William Paka, Scott, Deepesh Bhageria, Amitava Das, Soujanya Poria, et al. Task Report:

Memotion Analysis 1.0 @SemEval 2020: The Visuo-Lingual Metaphor! In *SemEval*, Sep 2020.

[538] Rajeev Sharma, Vladimir I Pavlović, and Thomas S Huang. Toward multimodal human–computer interface. In *Advances in image processing and understanding: A Festschrift for Thomas S Huang*, pages 349–365. World Scientific, 2002.

[539] Sagar Sharma, Keke Chen, and Amit Sheth. Toward practical privacy-preserving analytics for iot and cloud-based healthcare systems. *IEEE Internet Computing*, 22(2):42–51, 2018.

[540] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. 2018.

[541] Imran Sheikh, Sri Harsha Dumpala, Rupayan Chakraborty, and Sunil Kumar Kopparapu. Sentiment analysis using imperfect views from spoken language and acoustic modalities. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 35–39, 2018.

[542] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. In *EMNLP-IJCNLP*, pages 3398–3403, 2019.

[543] Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, and S Yu Philip. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering*, 29(1):17–37, 2016.

[544] Yuge Shi, Brooks Paige, and Philip Torr. Variational mixture-of-experts autoencoders for multimodal deep generative models. *NeurIPS*, 2019.

[545] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749, 2020.

[546] Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

[547] Ravid Shwartz-Ziv and Yann LeCun. To compress or not to compress–self-supervised learning and information theory: A review. *arXiv preprint arXiv:2304.09355*, 2023.

[548] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[549] Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of ICML workshop on learning with multiple views*, volume 2005, pages 74–79. Citeseer, 2005.

[550] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[551] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

[552] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet Talwalkar. Federated multi-task learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4427–4437, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[553] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. *NeurIPS*, 2013.

[554] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Stroudsburg, PA, October 2013. Association for Computational Linguistics.

[555] Kihyuk Sohn, Wenling Shang, and Honglak Lee. Improved multimodal deep learning with variation of information. In *Advances in Neural Information Processing Systems*, pages 2141–2149, 2014.

[556] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. A

survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14, 2017.

[557] Hagen Soltau, Hank Liao, and Hasim Sak. Neural speech recognizer: Acoustic-to-word LSTM model for large vocabulary speech recognition. *CoRR*, abs/1610.09975, 2016. URL http://arxiv.org/abs/1610.09975.

[558] Dandan Song, Siyi Ma, Zhanchen Sun, Sicheng Yang, and Lejian Liao. Kvl-bert: Knowledge enhanced visual-and-linguistic bert for visual commonsense reasoning. *Knowledge-Based Systems*, 230:107408, 2021.

[559] Yale Song, Louis-Philippe Morency, and Randall Davis. Multi-view latent variable discriminative models for action recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2120–2127. IEEE, 2012.

[560] Yale Song, Louis-Philippe Morency, and Randall Davis. Action recognition by hierarchical sequence summarization. In *CVPR*, pages 3562–3569, 2013.

[561] Alessandro Sordoni, Nouha Dziri, Hannes Schulz, Geoff Gordon, Philip Bachman, and Remi Tachet Des Combes. Decomposed mutual information estimation for contrastive representation learning. In *International Conference on Machine Learning*, pages 9859–9869. PMLR, 2021.

[562] Daria Sorokina, Rich Caruana, Mirek Riedewald, and Daniel Fink. Detecting statistical interactions with additive groves of trees. In *Proceedings of the 25th international conference on Machine learning*, pages 1000–1007, 2008.

[563] Karthik Sridharan and Sham M Kakade. An information theoretic framework for multi-view learning. 2008.

[564] Tejas Srinivasan and Yonatan Bisk. Worst of both worlds: Biases compound in pre-trained vision-and-language models. *arXiv preprint arXiv:2104.08666*, 2021.

[565] Nitish Srivastava and Russ R Salakhutdinov. Multimodal learning with deep boltzmann machines. *NeurIPS*, 25, 2012.

[566] Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International Conference on Machine Learning*, pages 9120–9132. PMLR, 2020.

[567] Neil Stewart, Gordon D. A. Brown, and Nick Chater. Absolute identification by relative judgment. *Psychological review*, 112 4:881–911, 2005.

[568] Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/D18-1548.

[569] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, 2019.

[570] Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, et al. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv preprint arXiv:2209.05481*, 2022.

[571] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2020.

[572] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *WWW*, 2007.

[573] Alane Suhr and Yoav Artzi. Nlvr2 visual bias analysis. *arXiv preprint arXiv:1909.10411*, 2019.

[574] Ömer Sümer, Patricia Goldberg, Sidney D'Mello, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. Multimodal engagement analysis from facial videos in the classroom. *IEEE Trans. on Affective Computing*, 2021.

[575] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[576] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, pages 7464–7473, 2019.

[577] Shiliang Sun. A survey of multi-view machine learning. *Neural computing and applications*, 23:2031–2038, 2013.

[578] Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8992–8999, 2020.

[579] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[580] Simon Šuster, Stéphan Tulkens, and Walter Daelemans. A short review of ethical challenges in clinical natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. ACL, 2017.

[581] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215, 2014.

[582] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[583] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

[584] Masahiro Suzuki and Yutaka Matsuo. A survey of multimodal deep generative models. *Advanced Robotics*, 36(5-6):261–278, 2022.

[585] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*, 2016.

[586] Riko Suzuki, Hitomi Yanaka, Masashi Yoshikawa, Koji Mineshima, and Daisuke Bekki. Multimodal logical inference system for visual-textual entailment. *arXiv preprint arXiv:1906.03952*, 2019.

[587] Fuwen Tan, Song Feng, and Vicente Ordonez. Text2scene: Generating compositional scenes from textual descriptions. In *CVPR*, pages 6710–6719, 2019.

[588] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*, pages 5100–5111, 2019.

[589] Hao Tan and Mohit Bansal. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *EMNLP*, pages 2066–2080, 2020.

[590] Gongbo Tang, Mathias Müller, Annette Rios, and Rico Sennrich. Why self-attention? a targeted evaluation of neural machine translation architectures. *arXiv preprint arXiv:1808.08946*, 2018.

[591] Tadahiro Taniguchi, Shingo Murata, Masahiro Suzuki, Dimitri Ognibene, Pablo Lanillos, Emre Ugur, Lorenzo Jamone, Tomoaki Nakamura, Alejandra Ciria, Bruno Lara, et al. World models and predictive coding for cognitive and developmental robotics: Frontiers and challenges. *arXiv preprint arXiv:2301.05832*, 2023.

[592] Zhulin Tao, Yinwei Wei, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat-Seng Chua. Mgat: Multimodal graph attention network for recommendation. *Information Processing & Management*, 57(5):102277, 2020.

[593] Makarand Tapaswi, Martin Bauml, and Rainer Stiefelhagen. Book2movie: Aligning video scenes with book chapters. In *CVPR*, pages 1827–1835, 2015.

[594] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, pages 4631–4640, 2016.

[595] Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena : A benchmark for efficient transformers. In *ICLR*, 2021.

[596] Han Te Sun. Multiple mutual informations and multiple interactions in frequency data. *Inf. Control*, 46:26–45, 1980.

[597] Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, et al. The language interpretability tool: Extensible, interactive visualizations and analysis for nlp models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 107–118, 2020.

[598] Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond J Mooney. Learning multi-modal grounded linguistic semantics by playing" i spy". In *IJCAI*, pages 3477–3483, 2016.

[599] Bruce Thompson. Canonical correlation analysis. 2000.

[600] Edward L Thorndike. Intelligence and its uses. *Harper's magazine*, 1920.

[601] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, pages 5238–5248, 2022.

[602] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 247–263, 2018.

[603] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *ECCV 2020*, pages 436–454. Springer, 2020.

[604] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020.

[605] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *NeurIPS*, 33:6827–6839, 2020.

[606] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.

[607] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015.

[608] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *ALT*, 2021.

[609] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. Missing modalities imputation via cascaded residual autoencoder. In *CVPR*, 2017.

[610] Nhat C Tran and Jean X Gao. Openomics: A bioinformatics api to integrate multi-omics datasets and interface with public databases. *Journal of Open Source Software*, 6(61):3249, 2021.

[611] George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. Deep canonical time warping for simultaneous alignment and representation learning of sequences. *IEEE TPAMI*, 40(5):1128–1138, 2017.

[612] Yao-Hung Hubert Tsai, Tianqin Li, Weixin Liu, Peiyuan Liao, Ruslan Salakhutdinov, and Louis-Philippe Morency. Learning weakly-supervised contrastive representations. In *International Conference on Learning Representations*.

[613] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, pages 6558–6569, 2019.

[614] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. In *ICLR*, 2019.

[615] Yao-Hung Hubert Tsai, Martin Ma, Muqiao Yang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Multimodal routing: Improving local and global interpretability of multimodal language analysis. In *EMNLP*, pages 1823–1833, 2020.

[616] Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning

from a multi-view perspective. In *ICLR*, 2020.

[617] Yao-Hung Hubert Tsai, Han Zhao, Makoto Yamada, Louis-Philippe Morency, and Russ R Salakhutdinov. Neural methods for point-wise dependency estimation. *Advances in Neural Information Processing Systems*, 33:62–72, 2020.

[618] Yao-Hung Hubert Tsai, Tianqin Li, Martin Q Ma, Han Zhao, Kun Zhang, Louis-Philippe Morency, and Ruslan Salakhutdinov. Conditional contrastive learning with kernel. *arXiv preprint arXiv:2202.05458*, 2022.

[619] Michael Tsang, Dehua Cheng, and Yan Liu. Detecting statistical interactions from neural network weights. In *ICLR*, 2018.

[620] Michael Tsang, Dehua Cheng, Hanpeng Liu, Xue Feng, Eric Zhou, and Yan Liu. Feature interaction interpretability: A case for explaining ad-recommendation systems via neural interaction detection. In *ICLR*, 2019.

[621] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2019.

[622] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *NeurIPS*, 34, 2021.

[623] Endel Tulving and Michael J Watkins. On negative transfer: Effects of testing one list on the recall of another. *Journal of Verbal Learning and Verbal Behavior*, 13(2):181–193, 1974.

[624] Matthew Turk. Multimodal interaction: A review. *Pattern recognition letters*, 36:189–195, 2014.

[625] Peter D Turney and Michael L Littman. Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60:251–278, 2005.

[626] Len Unsworth and Chris Cléirigh. Multimodality and reading: The construction of meaning through image-text interaction. Routledge, 2014.

[627] Shagun Uppal, Sarthak Bhagat, Devamanyu Hazarika, Navonil Majumder, et al. Multimodal research in vision and language: A review of current and emerging trends. *Information Fusion*, 77:149–171, 2022.

[628] Naushad UzZaman, Jeffrey P Bigham, and James F Allen. Multimodal summarization of complex sentences. In *Proceedings of the 16th international conference on Intelligent user interfaces*, pages 43–52. ACM, 2011.

[629] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 30, 2017.

[630] Emile van Krieken, Erman Acar, and Frank van Harmelen. Analyzing differentiable fuzzy logic operators. *Artificial Intelligence*, 2022.

[631] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[632] Ramakrishna Vedantam, Karan Desai, Stefan Lee, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Probabilistic neural symbolic models for interpretable visual question answering. In *ICML*, pages 6428–6437. PMLR, 2019.

[633] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

[634] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015.

[635] Jorge R Vergara and Pablo A Estévez. A review of feature selection methods based on mutual information. *Neural computing and applications*, 24:175–186, 2014.

[636] Philip E Vernon. Some characteristics of the good judge of personality. *The Journal of Social Psychology*, 4 (1):42–57, 1933.

[637] René Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011.

[638] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multilayer approach for multimodal fusion, 2018.

[639] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[640] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE TPAMI*, 2016.

[641] Alvin Wan, Lisa Dunlap, Daniel Ho, Jihan Yin, Scott Lee, Suzanne Petryk, Sarah Adel Bargal, and Joseph E Gonzalez. Nbdt: Neural-backed decision tree. In *ICLR*, 2020.

[642] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[643] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, 32, 2019.

[644] Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. Screen2words: Automatic mobile ui summarization with multimodal learning. In *ACM UIST*, pages 498–510, 2021.

[645] Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P Xing. Select-additive learning: Improving cross-individual generalization in multimodal sentiment analysis. *arXiv preprint arXiv:1609.05244*, 2016.

[646] Haoqing Wang, Xun Guo, Zhi-Hong Deng, and Yan Lu. Rethinking minimal sufficient representation in contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16041–16050, 2022.

[647] Meng Wang, Hao Li, Dacheng Tao, Ke Lu, and Xindong Wu. Multimodal graph-based reranking for web image search. *IEEE transactions on image processing*, 21(11):4649–4661, 2012.

[648] Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou. Multimodal learning with incomplete modalities by knowledge distillation. In *KDD*, pages 1828–1838, 2020.

[649] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense representation learning via causal inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[650] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *ICML*, pages 1083–1092. PMLR, 2015.

[651] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, pages 12695–12705, 2020.

[652] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[653] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, et al. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *CVPR*, pages 6629–6638, 2019.

[654] Xingbo Wang, Jianben He, Zhihua Jin, et al. M2lens: Visualizing and explaining multimodal models for sentiment analysis. *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[655] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. *AAAI*, 2019.

[656] Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. Multi-passage bert: A globally normalized bert model for open-domain question answering. *arXiv preprint arXiv:1908.08167*, 2019.

[657] Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11293–11302, 2019.

[658] Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of research and*

development, 4(1):66–82, 1960.

[659] Xiaofan Wei, Huibin Li, Jian Sun, and Liming Chen. Unsupervised domain adaptation with regularized optimal transport for multimodal 2d+ 3d facial expression recognition. In *FG 2018*, pages 31–37. IEEE, 2018.

[660] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

[661] Alex Wilf, Qianli M Ma, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Face-to-face contrastive learning for social intelligence question-answering. *arXiv preprint arXiv:2208.01036*, 2022.

[662] Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

[663] Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28 (3):46–53, 2013.

[664] Patricia Wollstadt, Joseph Lizier, Raul Vicente, Conor Finn, Mario Martinez-Zarzuela, Pedro Mediano, Leonardo Novelli, and Michael Wibral. Idtxl: The information dynamics toolkit xl: a python package for the efficient analysis of multivariate information dynamics in networks. *Journal of Open Source Software*, 4(34): 1081, 2019.

[665] Patricia Wollstadt, Sebastian Schmitt, and Michael Wibral. A rigorous information-theoretic definition of redundancy and relevancy in feature selection based on (partial) information decomposition. *arXiv preprint arXiv:2105.04187*, 2021.

[666] Eric Wong, Shibani Santurkar, and Aleksander Madry. Leveraging sparse linear layers for debuggable deep networks. In *ICML*, 2021.

[667] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, et al. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, pages 7623–7633, 2023.

[668] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *NeurIPS*, 31, 2018.

[669] Mike Wu, Chengxu Zhuang, Milan Mosse, Daniel Yamins, and Noah Goodman. On mutual information in contrastive learning for visual representations. *arXiv preprint arXiv:2005.13149*, 2020.

[670] Nan Wu, Stanisław Jastrzębski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. *arXiv preprint arXiv:2202.05306*, 2022.

[671] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *CVPR*, pages 4622–4630, 2016.

[672] Xindi Wu, Zhiwei Deng, and Olga Russakovsky. Multimodal dataset distillation for image-text retrieval. *arXiv preprint arXiv:2308.07545*, 2023.

[673] Xuetong Wu, Jonathan H Manton, Uwe Aickelin, and Jingge Zhu. An information-theoretic analysis for transfer learning: Error bounds and applications. *arXiv preprint arXiv:2207.05377*, 2022.

[674] Yi Xiao, Felipe Codevilla, Akhil Gurram, Onay Urfalioglu, and Antonio M López. Multimodal end-to-end autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[675] Chen Xing, Negar Rostamzadeh, Boris Oreshkin, and Pedro O O. Pinheiro. Adaptive cross-modal few-shot learning. In *NeurIPS*. 2019.

[676] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *ICML*, 2016.

[677] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

[678] Chang Xu, Dacheng Tao, and Chao Xu. Multi-view intact space learning. *IEEE TPAMI*, 2015.

[679] Fangli Xu, Lingfei Wu, KP Thai, Carol Hsu, Wei Wang, and Richard Tong. Mutla: A large-scale dataset for multimodal teaching and learning analytics. *arXiv preprint arXiv:1910.06078*, 2019.

[680] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015.

[681] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[682] Zhen Xu, David R So, and Andrew M Dai. Mufasa: Multimodal fusion architecture search for electronic health records. *arXiv preprint arXiv:2102.02340*, 2021.

[683] Zihui Xue, Sucheng Ren, Zhengqi Gao, and Hang Zhao. Multimodal knowledge expansion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 854–863, 2021.

[684] Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. *arXiv preprint arXiv:2206.06487*, 2022.

[685] Xiaoqiang Yan, Shizhe Hu, Yiqiao Mao, Yangdong Ye, and Hui Yu. Deep multi-view learning methods: A review. *Neurocomputing*, 448:106–129, 2021.

[686] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1642–1651, 2022.

[687] Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, et al. Mtag: Modal-temporal attention graph for unaligned human multimodal language sequences. In *NAACL-HLT*, 2021.

[688] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space, 2022.

[689] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning, 2022.

[690] Xiao Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, and C. Lee Giles. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[691] Yang Yang, De-Chuan Zhan, Xiang-Rong Sheng, and Yuan Jiang. Semi-supervised multi-modal learning with incomplete modalities. In *IJCAI*, pages 2998–3004, 2018.

[692] Yang Yang, Ke-Tao Wang, De-Chuan Zhan, Hui Xiong, and Yuan Jiang. Comprehensive semi-supervised multi-modal learning. In *IJCAI*, 2019.

[693] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 270–279, 2010.

[694] Shaowei Yao and Xiaojun Wan. Multimodal transformer for multimodal machine translation. In *ACL*, pages 4346–4350, 2020.

[695] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35: 20744–20757, 2022.

[696] Zesheng Ye and Lina Yao. Contrastive conditional neural processes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9687–9696, 2022.

[697] De Jong Yeong, Gustavo Velasco-Hernandez, John Barry, Joseph Walsh, et al. Sensor and sensor fusion technology in autonomous vehicles: A review. *Sensors*, 21(6):2140, 2021.

[698] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

[699] Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. A novel

graph-based multi-modal fusion encoder for neural machine translation. In *ACL*, pages 3025–3035, 2020.

[700] Ron Yosef, Yonatan Bitton, and Dafna Shahaf. Irfl: Image recognition of figurative language. *arXiv preprint arXiv:2303.15445*, 2023.

[701] Jason Yosinski, Jeff Clune, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. In *In ICML Workshop on Deep Learning*. Citeseer, 2015.

[702] M. H. Peter Young, Alice Lai, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–68, 2014.

[703] Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, 2004.

[704] Lei Yu, Jan Buys, and Phil Blunsom. Online segment to segment neural transduction. *arXiv preprint arXiv:1609.08194*, 2016.

[705] Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. Vision guided generative pre-trained language models for multimodal abstractive summarization. In *EMNLP*, pages 3995–4007, 2021.

[706] Weijiang Yu, Jingwen Zhou, Weihao Yu, Xiaodan Liang, and Nong Xiao. Heterogeneous graph learning for visual commonsense reasoning. In *NeurIPS*, 2019.

[707] Jiahong Yuan and Mark Liberman. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America*, 123(5):3878, 2008.

[708] Jiahong Yuan, Mark Liberman, et al. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America*, 2008.

[709] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6995–7004, 2021.

[710] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

[711] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016.

[712] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, 2017.

[713] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *AAAI*, volume 32, 2018.

[714] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[715] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *CVPR*, pages 8807–8817, 2019.

[716] Amir Zadeh, Paul Pu Liang, and Louis-Philippe Morency. Foundations of multimodal co-learning. *Information Fusion*, 64:188–193, 2020.

[717] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *ACL*, 2018.

[718] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, pages 3712–3722, 2018.

[719] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[720] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *NeurIPS*, 34, 2021.

[721] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022.

[722] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.

[723] Haijin Zeng, Jize Xue, Hiêp Q Luong, and Wilfried Philips. Multimodal core tensor factorization and its applications to low-rank tensor completion. *IEEE Transactions on Multimedia*, 2022.

[724] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. In *AAAI*, 2017.

[725] Da Zhang and Mansur Kabuka. Multimodal deep representation learning for protein interaction identification and protein family classification. *BMC bioinformatics*, 20(16):1–14, 2019.

[726] Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3):1–37, 2023.

[727] Hao Zhang, Zhiting Hu, Yuntian Deng, Mrinmaya Sachan, Zhicheng Yan, and Eric Xing. Learning concept taxonomies from multi-modal data. In *ACL*, pages 1791–1801, 2016.

[728] Tong Zhang and C-C Jay Kuo. Audio content analysis for online audiovisual data segmentation and classification. *IEEE Transactions on speech and audio processing*, 9(4):441–457, 2001.

[729] Weifeng Zhang, Jing Yu, Hua Hu, Haiyang Hu, and Zengchang Qin. Multimodal feature fusion by relational reasoning and attention for visual question answering. *Information Fusion*, 55:116–126, 2020.

[730] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, 2020.

[731] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

[732] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10394–10403, 2019.

[733] Victor Zhong, Tim Rocktäschel, and Edward Grefenstette. Rtfm: Generalising to new environment dynamics via reading. In *ICLR*, 2020.

[734] Victor Zhong, Austin W Hanjie, Karthik Narasimhan, Luke Zettlemoyer, Austin W Hanjie, Victor Zhong, Karthik Narasimhan, Machel Reid, Victor Zhong, Victor Zhong, et al. Silg: The multi-environment symbolic interactive language grounding benchmark. *NeurIPS*, 2021.

[735] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023.

[736] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[737] Hao Zhu, Huaibo Huang, Yi Li, Aihua Zheng, and Ran He. Arbitrary talking face generation via attentional audio-visual coherence learning. In *IJCAI*, pages 2362–2368, 2021.

[738] Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, et al. Multi-modal knowledge graph construction and application: A survey. *arXiv preprint arXiv:2202.05786*, 2022.

[739] Yuke Zhu, Ce Zhang, Christopher Ré, and Li Fei-Fei. Building a large-scale multimodal knowledge base system for answering visual queries. *arXiv preprint arXiv:1507.05670*, 2015.

[740] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja

Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*, 2015.

[741] Zachary M Ziegler, Luke Melas-Kyriazi, Sebastian Gehrmann, and Alexander M Rush. Encoder-agnostic adaptation for conditional language generation. *arXiv preprint arXiv:1908.06938*, 2019.

[742] Julian Georg Zilly, Rupesh Kumar Srivastava, Jan Koutník, and Jürgen Schmidhuber. Recurrent Highway Networks. *arXiv preprint arXiv:1607.03474*, 2016.

[743] George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. 2016.