Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions

PAUL PU LIANG, Carnegie Mellon University, Pittsburgh, United States
AMIR ZADEH, Carnegie Mellon University, Pittsburgh, United States
LOUIS-PHILIPPE MORENCY, Carnegie Mellon University, Pittsburgh, United States

Multimodal machine learning is a vibrant multi-disciplinary research field that aims to design computer agents with intelligent capabilities such as understanding, reasoning, and learning through integrating multiple communicative modalities, including linguistic, acoustic, visual, tactile, and physiological messages. With the recent interest in video understanding, embodied autonomous agents, text-to-image generation, and multisensor fusion in application domains such as healthcare and robotics, multimodal machine learning has brought unique computational and theoretical challenges to the machine learning community given the heterogeneity of data sources and the interconnections often found between modalities. 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 application domains and theoretical frameworks from both historical and recent perspectives, this article is designed to provide an overview of the computational and theoretical foundations of multimodal machine learning. We start by defining three key principles of modality *heterogeneity*, *connections*, and *interactions* that have driven subsequent innovations, and propose a taxonomy of six core technical challenges: *representation*, *alignment*, *reasoning*, *generation*, *transference*, and *quantification* covering historical and recent trends. Recent technical achievements will be presented through the lens of this taxonomy, allowing researchers to understand the similarities and differences across new approaches. We end by motivating several open problems for future research as identified by our taxonomy.

CCS Concepts: • **Computing methodologies** → **Machine learning**; **Artificial intelligence**; *Computer vision*; *Natural language processing;*

Additional Key Words and Phrases: Multimodal machine learning, representation learning, data heterogeneity, feature interactions, language and vision, multimedia

ACM Reference Format:
Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2024. Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. *ACM Comput. Surv.* 56, 10, Article 264 (June 2024), 42 pages. https://doi.org/10.1145/3656580

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

Authors' Contact Information: Paul Pu Liang, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States; e-mail: pliang@cs.cmu.edu; Amir Zadeh, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States; e-mail: abagherz@cs.cmu.edu; Louis-Philippe Morency, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States; e-mail: morency@cs.cmu.edu.

ACM 0360-0300/2024/06-ART264
https://doi.org/10.1145/3656580

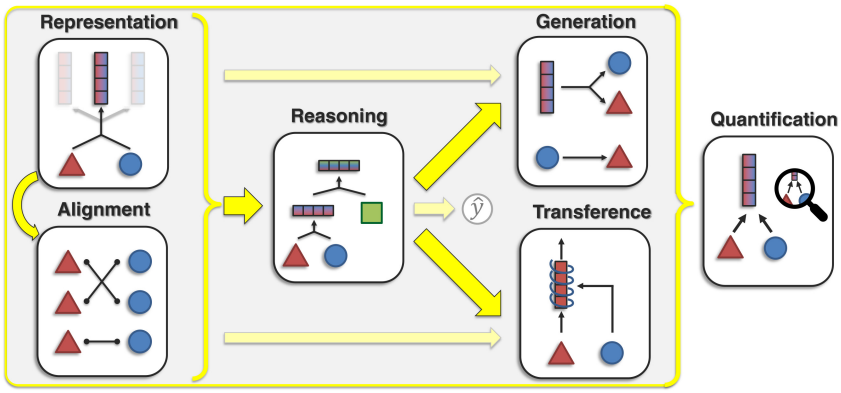

Fig. 1. Core research challenges in multimodal learning: Every multimodal problem typically requires tackling representation and alignment: (1) *Representation* studies how to summarize multimodal data to reflect the heterogeneity and interconnections between individual modality elements, before (2) *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.

data, similar to how humans perceive and interact with our world using multiple sensory modalities. With recent advances in embodied autonomous agents [43, 286], self-driving cars [374], image and video understanding [15, 314], image and video generation [273, 302], and multisensor fusion in application domain such as robotics [172, 221] and healthcare [150, 195], 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 [237], affective computing [265], robotics [160, 172], human–computer interaction [244, 296], and healthcare [47, 233].

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 article is designed to provide an overview of the methodological, computational, and theoretical foundations of multimodal machine learning. We begin by defining (in Section 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 principles, we propose a new taxonomy of six core challenges in multimodal learning: *representation*, *alignment*, *reasoning*, *generation*, *transference*, and *quantification* (see Figure 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 (Section 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) *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 (Section 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 (Section 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 (Section 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 (Section 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 (Section 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 article 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 article is designed to complement other surveys that belong broadly to the study of multiple modalities or views: multi-view learning [241, 315, 384] 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 [86, 99] 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 [345], language and **reinforcement learning** (**RL**) [211], multimedia analysis [24], and multimodal human-computer interaction [145] 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.

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 [263]. Some commonly studied tasks involve predicting sentiment [305, 399], emotions [400], humor [117], and sarcasm [52] 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 [4, 17]. Multimodal physiological signals recorded regularly from smartphones and wearable devices can also provide non-invasive health monitoring [74, 102, 191]. Public datasets include MIMIC [150] with patient tabular data, medical reports, and medical sensor readings, question answering on pathology [118] and radiology [169] images, and multi-omics data integration [334].

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 [172]. These multi-sensor robots have been successfully applied in haptic [249, 291] and surgical robots [3, 37]. More generally, language [211] and audio [75] 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 [97] as well as in the social world through virtual agents [257]. 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 [389] and WebAreana [412] 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 [24], such as classification [409], retrieval [274], and recommendation [411] of multimedia content, image and video question answering [9, 165, 174] and captioning [85, 355]), multimedia and entertainment content description [293] (including movies [23], memes [158, 295], and cartoons [122]), and more recently in automatic creation of text [408], images [278], videos [369], music [7], and more.

Human–computer interaction has sought to endow computers with multimodal capabilities to provide more natural, powerful, and compelling interactive user experiences [342]. These systems have leveraged speech, touch, vision, gestures, affective states [251] and affordable wearable and mobile sensors [145, 248, 342]. Public datasets have enabled the study of multimodal user interfaces [175, 357], speech and gesture interactions [89], and human sensing [57, 77, 289].

Science and environment: Deepening our knowledge of the natural sciences and physical environments can bring about impactful changes in scientific discovery, sustainability, and

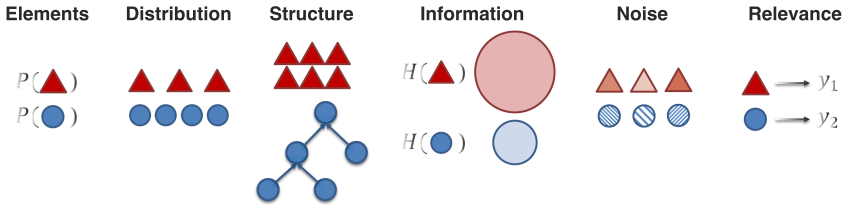

Fig. 2. The information present in different modalities will often show diverse qualities, structures, and representations. **Dimensions of heterogeneity** can be measured via differences in individual elements and their distribution, the structure of elements, as well as modality information, noise, and task relevance.

conservation. This requires processing modalities such as chemical molecules [310], protein structures [406], satellite images [66, 386], remote sensing [125, 178], wildlife movement [206], scientific diagrams and texts [209], and various physical sensors [231].

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 [171], and modeling student learning via asked questions, spoken feedback, and non-verbal gestures [55, 313, 378].

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* (connections + interactions) 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.

2.1 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 are different depending on 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 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 [32, 188]. 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 elements in modalities. Elements typically follow a unique distribution, with words in a linguistic corpus following Zipf's Law [419] as an example. Distribution heterogeneity then 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 [45]. For example, images exhibit spatial structure across individual object elements, language is hierarchically composed of individual 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 [294].

(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 [195]. 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 [103]. 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 [45]. In the multimodal case, modeling heterogeneity is useful when learning representations and capturing alignment [401], and is a key subchallenge in quantifying multimodal models [194].

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 [367]. 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 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 [331].

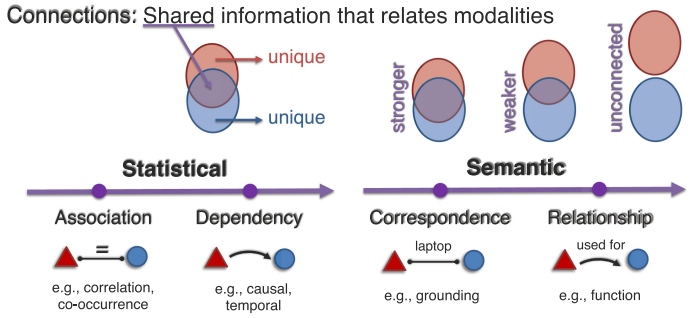

Fig. 3. Modality connections describe how modalities are related and share commonalities, such as correspondences between the same concept in language and images or dependencies across spatial and temporal dimensions. Connections can be studied through both statistical and semantic perspectives.

(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 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 [242, 343].
(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 [247]. Identifying correspondences is fundamental in many problems related to language grounding [54], translation and retrieval [264], and cross-modal alignment [320].
(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 [32, 223].

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 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 [189, 367].
(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 [148], as well as from a semantic perspective where two elements interact through a logical, causal, or temporal operation [344].

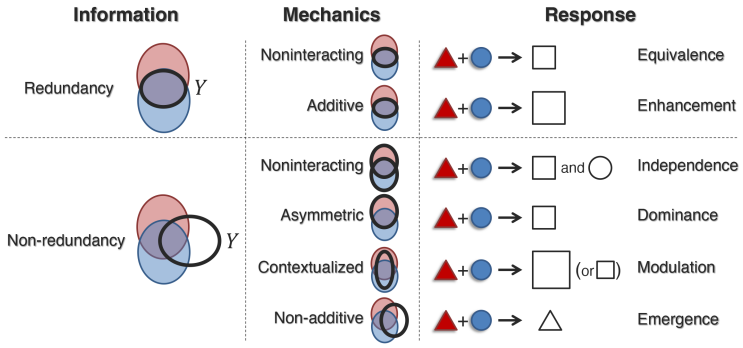

Fig. 4. Several dimensions of modality interactions: (1) Interaction information studies whether common redundant information or unique non-redundant information is involved in interactions; (2) interaction mechanics study the manner in which interaction occurs, and (3) interaction response studies how the inferred task changes in the presence of multiple modalities.

(3) **Interaction response** studies how the inferred response changes in the presence of multiple 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 [254].

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 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.

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 with the goal of 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 5).

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 6). In fusion with abstract modalities, suitable unimodal encoders are

Table 1. This Table Summarizes Our Taxonomy of Six Core Challenges in Multimodal Machine Learning, Their Subchallenges, Categories of Corresponding Approaches, and Representative Examples

Challenge	Subchallenge	Approaches & key examples
Representation (Section 3)	Fusion (Section 3.1)	Abstract [148, 396] & raw [30, 272] fusion
	Coordination (Section 3.2)	Strong [95, 268] & partial [352, 407] coordination
	Fission (Section 3.3)	Modality-level [121, 337] & fine-grained [1, 58] fission
Alignment (Section 4)	Discrete connections (Section 4.1)	Local [71, 129] & global [182] alignment
	Continuous alignment (Section 4.2)	Warping [115, 132] & segmentation [314]
	Contextualization (Section 4.3)	Joint [180], cross-modal [120, 207] & graphical [385]
Reasoning (Section 5)	Structure modeling (Section 5.1)	Hierarchical [20], temporal [376], interactive [211] & discovery [260]
	Intermediate concepts (Section 5.2)	Attention [379], discrete symbols [18, 350] & language [140, 404]
	Inference paradigm (Section 5.3)	Logical [107, 318] & causal [6, 243, 390]
	External knowledge (Section 5.4)	Knowledge graphs [111, 417] & commonsense [253, 402]
Generation (Section 6)	Summarization (Section 6.1)	Extractive [62, 346] & abstractive [177, 250]
	Translation (Section 6.2)	Exemplar-based [153, 170] & generative [10, 146, 273]
	Creation (Section 6.3)	Conditional decoding [79, 245, 414]
Transference (Section 7)	Cross-modal transfer (Section 7.1)	Tuning [271, 341], multitask [194, 303] & transfer [208]
	Co-learning (Section 7.2)	Representation [149, 398] & generation [262, 321]
	Model Induction (Section 7.3)	Co-training [40, 87] & co-regularization [308, 387]
Quantification (Section 8)	Heterogenity (Section 8.1)	Importance [103, 252], bias [119, 258] & noise [214]
	Interconnections (Section 8.2)	Connections [5, 49, 327] & interactions [121, 193, 363]
	Learning (Section 8.3)	Generalization [194, 275], optimization [362, 371] & tradeoffs [195]

We believe that this taxonomy can help to catalog rapid progress in this field and better identify the open research questions.

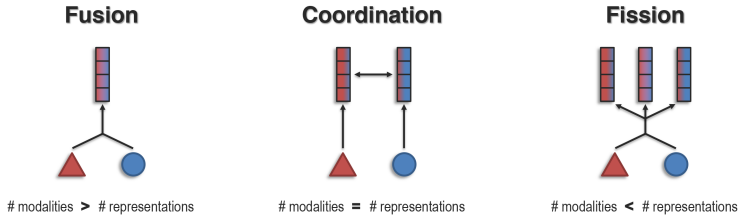

Fig. 5. Challenge 1 aims to learn **representations** that reflect cross-modal interactions between individual modality elements, through (1) *fusion*: integrating information to reduce the number of separate representations, (2) *coordination*: interchanging cross-modal information with the goal of 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.

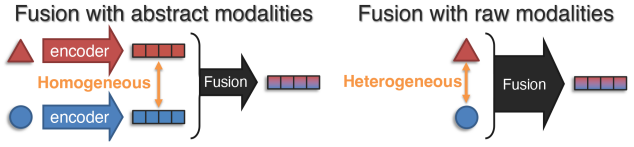

Fig. 6. We categorize **representation fusion** approaches into (1) *fusion with abstract modalities*, where unimodal encoders first capture a holistic representation of each element before fusion at relatively homogeneous representations, and (2) *fusion with raw modalities* which entails representation fusion at very early stages, perhaps directly involving heterogeneous raw modalities.

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 [94]. Otherwise, the additive representation $\mathbf{z}_{\text{mm}}$ can also undergo subsequent unimodal or multimodal processing [29]. **Multiplicative interaction**s (**MI**) 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 be interpreted as a *moderation* effect of $\mathbf{x}_1$ affecting the linear relationship between $\mathbf{x}_2$ and y [31]. 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 [396]. 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 [34, 96]. Tensor products of higher order represent polynomial interactions of higher order between elements [127]. 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 [127, 205]. Finally, *MI* generalize additive and multiplicative operators to include learnable parameters that capture second-order interactions [148]. In its most general form, MI defines a bilinear product $\mathbf{z}_{\text{mm}} = \mathbf{x}_1\mathbb{W}\mathbf{x}_2 + \mathbf{x}_1^\top\mathbf{U} + \mathbf{V}\mathbf{x}_2 + \mathbf{b}$ 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 [56, 362]. 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 [336], fully connected neural network layers [23, 56], or even hard gated units for sharper attention [65].

Fusion with raw modalities entails representation fusion at very early stages, perhaps even involving raw modalities themselves. These approaches typically bear resemblance to early fusion [29], 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. [30] demonstrated robustness benefits of fusion at early stages, while Gadzicki et al. [98] 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 [30, 98] and Transformers [194, 198]. However, do these complex non-additive fusion models actually learn non-additive interactions between modality elements? Not necessarily, according to Hessel and Lee [121]. We cover these fundamental analysis questions and more in the quantification challenge (Section 8).

3.2 Subchallenge 1b: Representation Coordination

Representation coordination aims to learn multimodal contextualized representations that are coordinated through their interconnections (Figure 7). In contrast to representation fusion, coordination keeps the same number of representations but improves multimodal contextualization. We

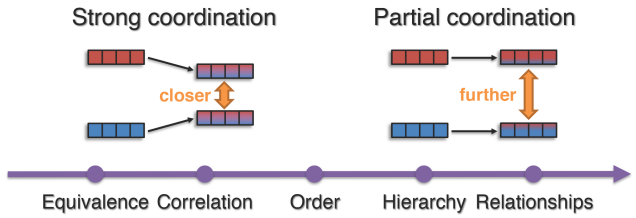

Fig. 7. There is a spectrum of **representation coordination** functions: *strong coordination* aims to enforce strong equivalence in all dimensions, whereas in *partial coordination* only certain dimensions may be coordinated to capture more general connections such as correlation, order, hierarchies, or relationships.

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) [95]. The coordination distance is typically cosine distance [225] or max-margin losses [131]. Recent work has explored large-scale representation coordination by scaling up contrastive learning of image and text pairs [268], and also found that contrastive learning provably captures redundant information across the two views [328, 332] (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 [88]. For example, Socher et al. [304] map image embeddings into word embedding spaces for zero-shot image classification. Similar ideas were used to learn coordinated representations between text, video, and audio [262], as well as between pretrained language models and image features [321].

Partial coordination: Instead of capturing strong equivalences, partial coordination captures 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 [326]. CCA models have been used extensively for cross-modal retrieval [274] audio-visual signal analysis [285], and emotion recognition [239]. To increase the expressiveness of CCA, several nonlinear extensions have been proposed including Kernel CCA [168], Deep CCA [21], and CCA Autoencoders [361].

Ordered and hierarchical spaces: Another example of representation coordination comes from order-embeddings of images and language [352], 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. [392] 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. [407] use structured representations of text and images to create multimodal concept taxonomies. Delaherche and Chetouani [76] learn coordinated representations capturing hierarchical relationships, while Alviar et al. [16] apply multiscale coordination of speech and music using partial correlation measures. Finally, Xu et al. [377] learn coordinated representations using a Cauchy loss to strengthen robustness to outliers.

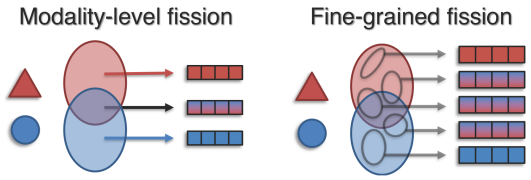

Fig. 8. Representation fission creates a larger set of decoupled representations that reflects knowledge about internal structure. (1) *Modality-level fission* factorizes into modality-specific information primarily in each modality, and multimodal information redundant in both modalities, while (2) *fine-grained fission* attempts to further break multimodal data down into individual subspaces.

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 8).

Modality-level fission aims to factorize into modality-specific information primarily in each modality and multimodal information redundant in both modalities [130, 190, 337]. *Disentangled representation learning* aims to learn mutually independent latent variables that each explain a particular variation of the data [36, 123], and has been useful for modality-level fission by enforcing independence constraints on modality-specific and multimodal latent variables [130, 337]. Tsai et al. [337] and Hsu and Glass [130] study factorized multimodal representations and demonstrate the importance of modality-specific and multimodal factors toward generation and prediction. Shi et al. [299] study modality-level fission in multimodal variational autoencoders using a mixture-of-experts layer, while Wu and Goodman [370] 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**) [121] 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 [227], 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 [353]. *Clustering* approaches that group data based on semantic similarity [216] have been integrated with multimodal networks for end-to-end representation fission and prediction. For example, Hu et al. [131] combine k-means clustering in representations with unsupervised audiovisual learning. Chen et al. [58] combine k-means clustering with self-supervised contrastive learning on videos. Subspace clustering [1, 157], manifold learning [186] approximate graph Laplacians [156], and dictionary learning [159] have also been integrated with multimodal models. *Matrix factorization* techniques have also seen several applications in multimodal fission for prediction [14] and cross-modal retrieval [48].

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

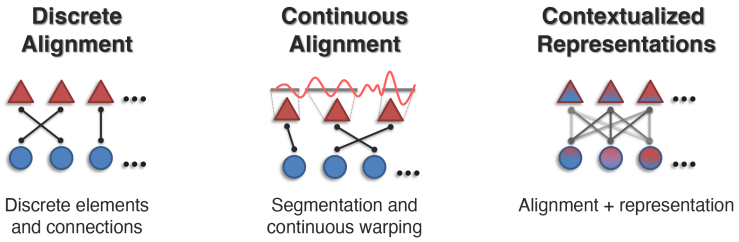

Fig. 9. Alignment aims to identify cross-modal connections and interactions between modality elements. Recent work has involved (1) *discrete alignment* to identify connections among discrete elements, (2) *continuous alignment* of continuous signals with ambiguous segmentation, and (3) *contextualized representation* learning to capture these cross-modal interactions between connected elements.

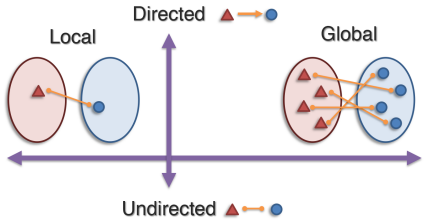

Fig. 10. Discrete alignment identifies connections between discrete elements, spanning (1) *local alignment* to discover connections given matching pairs, and (2) *global alignment* where alignment must be performed globally to learn both the connections and matchings between modality elements.

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 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 9).

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 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 [164], visual referring expression recognition [69, 70], and cross-modal retrieval [95, 264]). 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 [29]. Several objective functions for learning aligned spaces from varying quantities of paired [50, 136] and unpaired [110] data have been proposed. Many of the ideas that enforce strong [95, 196] or partial [21, 352, 407] representation coordination (Section 3.2) are also applicable for local alignment. Several examples include aligning books with their corresponding movies/scripts [416], matching

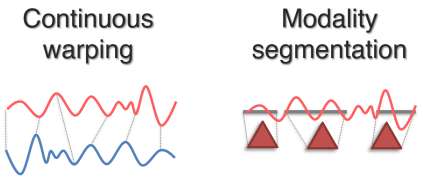

Fig. 11. Continuous alignment tackles the difficulty of aligning continuous signals where element segmentation is not readily available. We cover related work in (1) *continuous warping* of representation spaces and (2) *modality segmentation* of continuous signals into discrete elements at an appropriate granularity.

referring expressions to visual objects [220], and finding similarities between image regions and their descriptions [134]. 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 [136], sound [71, 304], and multimedia [416] 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 [354] (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 [182, 266]. To alleviate computational issues, several recent advances have integrated them with neural networks [64], approximated OT with entropy regularization [365], and formulated convex relaxations for efficient learning [110].

4.2 Subchallenge 2b: Continuous Alignment

So far, we have assumed 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 segmentation is not readily provided, such as in continuous signals (e.g, financial or medical time-series), spatio-temporal data (e.g., satellite or weather images), or data without clear semantic boundaries (e.g., MRI images). In these settings, it is important to perform continuous alignment based on warping or segmentation (see Figure 11):

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 [101, 329, 330]. *Adversarial training* is a popular approach to warp one representation space into another. Initially used in domain adaptation [33], adversarial training learns a domain-invariant representation across domains where a domain classifier is unable to identify which domain a feature came from [12]. These ideas have been extended to align multimodal spaces [129, 132, 234]. Hsu et al. [129] use adversarial training to align images and medical reports, Hu et al. [132] design an adversarial network for cross-modal retrieval, and Munro and Damen [234] design both self-supervised alignment and adversarial alignment objectives for multimodal action recognition. ***Dynamic time warping (DTW)*** [167] 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 [22, 323], such as combining DTW with CCA or other coordination functions [335].

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 [395]. Malmaud et al. [218] explore multimodal alignment using a

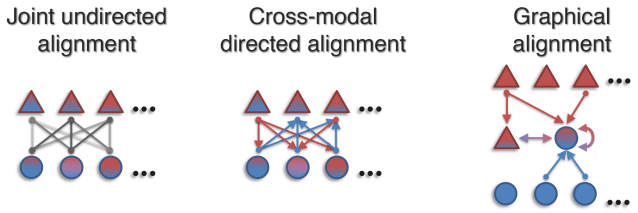

Fig. 12. Contextualized representation learning aims to model modality connections to learn better representations. Recent directions include (1) *joint undirected alignment* that captures undirected symmetric connections, (2) *cross-modal directed alignment* that models asymmetric connections in a directed manner, and (3) *graphical alignment* that generalizes the sequential pattern into arbitrary graph structures.

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 [216]. Clustering-based discretization has recently emerged as an important preprocessing step for generalizing language-based pretraining (with clear word/byte pair segmentation boundaries and 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 [314] perform masked pretraining on raw video and audio data. Similarly, approaches such as DALL.E [273], VQ-VAE [347], and CMCM [202] also utilize discretized intermediate layers obtained via vector quantization and showed benefits in modality alignment.

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 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 [215]. 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 [56], tensor products [205, 396], and MI [148]. More recently, transformer models [349] have emerged as powerful encoders for sequential data by automatically aligning and capturing complementary features at different timesteps. 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) [180, 314]. 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 [376, 405]. 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

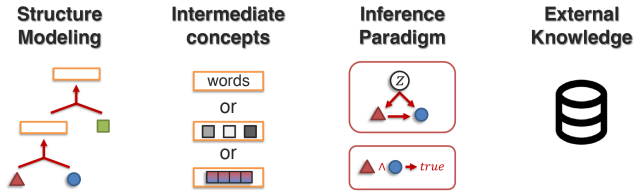

Fig. 13. Reasoning aims to combine knowledge, usually through multiple inferential steps, exploiting the problem structure. Reasoning involves (1) *structure modeling*: defining or learning the relationships over which reasoning occurs, (2) the *intermediate concepts* used in reasoning, (3) *inference* 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 [207, 320, 336]. These methods are useful for sequential data by automatically aligning and capturing complementary features at different timesteps [336].

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 [99, 194, 246, 268, 275]. These models typically build on top of pretrained language models [269], pretrained visual encoders [83] combined with an alignment layer. Alignment can be done via end-to-end training with multimodal transformers [380] (e.g., Flamingo [15], OpenFlamingo [27], Kosmos [259]), or keeping the language and vision parts frozen and only training a post-hoc alignment layer (e.g., MiniGPT-4 [413], BLIP-2 [179], InstructBLIP [73], LLaMA-Adapter V2 [100]). 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 [84, 180, 413]. Pretraining objectives typically consist of unimodal language modeling [269, 270], image-to-text or text-to-image alignment [120, 413], and multimodal instruction tuning [73, 203, 210]. We refer the reader to recent survey papers discussing these large vision-language models in more detail [86, 99].

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 [351] can be used to recursively learn element representations contextualized with the elements in locally connected neighborhoods [287, 351], such as in MTAG [385] and F2F-CL [366] for multimodal and multi-speaker videos.

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 toward multimodal reasoning into four subchallenges of structure modeling, intermediate concepts, inference paradigm, and external knowledge (Figure 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 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.

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

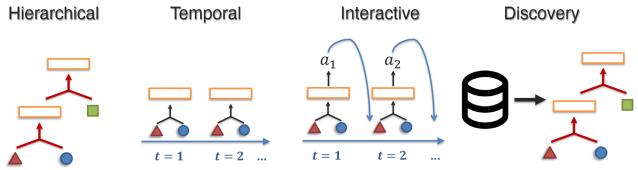

Fig. 14. Structure modeling aims to define the relationship over which composition occurs, which can be (1) *hierarchical* (i.e., more abstract concepts are defined as a function of less abstract ones), (2) *temporal* (i.e., organized across time), (3) *interactive* (i.e., where the state changes depending on each step's decision), and (4) *discovered* when the latent structure is unknown and instead directly inferred from data and optimization.

used data structures include trees [126], graphs [394], or neural modules [20]. We cover recent work in modeling latent *hierarchical*, *temporal*, and *interactive* structure, as well as *structure discovery* when the latent structure is unknown (Figure 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 [298], such as using language syntactic structure to guide visual modules that discover specific information in images [20, 69]. Graph-based reasoning approaches have been applied for visual commonsense reasoning [200], visual question answering [284], machine translation [391], recommendation systems [322], web image search [358], and social media analysis [288].

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. [272] explore various memory representations including multimodal fusion, coordination, and factorization. Insights from key-value memory [376] and attention-based memory [397] 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 [316]. To tackle the challenges of interactive reasoning, the growing research field of multimodal RL has emerged from the intersection of language understanding, embodiment in the visual world, deep RL, and robotics. We refer the reader to the extensive survey paper by Luketina et al. [211] and the position paper by Bisk et al. [39] for a full review of this field. Luketina et al. [211] separate the literature into multimodal-conditional RL (in which multimodal interaction is necessitated by the problem formulation itself, such as instruction following [56, 364]) and language-assisted RL (in which multimodal data is optionally used to facilitate learning, such as reading instruction manuals [238]).

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 [260, 381]. 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 [141]. Finally, Hu et al. [133] extend the predefined reasoning structure obtained through language parsing in Andreas et al. [20] by instead using policy gradients to automatically optimize a compositional structure over a discrete set of neural modules.

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 toward 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. [20] 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. [379] explore both soft and hard attention mechanisms for reasoning in image captioning generation. Related work has also used attention maps through dual attention architectures [235] or stacked latent attention architectures [91] for multimodal reasoning. These are typically applied for problems involving complex reasoning steps such as CLEVR [151] or VQA [410].

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 [20, 219, 350] or referring expression recognition [69]. A core challenge in this approach lies in maintaining the differentiability of discrete symbols, which has been tackled via logic-based differentiable reasoning [18, 292].

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. [404] 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 [111, 417]. Hudson and Manning [140] designed a neural state machine to simulate the execution of a question being asked about an image, while using discrete words as intermediate concepts.

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 Section 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 [18, 292]. Many of these approaches use differentiable fuzzy logic which provides a probabilistic interpretation of logical predicates, functions, and constants to ensure differentiability. These logical operators have been applied for visual question

answering [107] and visual reasoning [18]. Among the greatest benefits of logical reasoning lies in its ability to perform interpretable and compositional multi-step reasoning [142]. Logical frameworks have also been useful for visual-textual entailment [318] and geometric numerical reasoning [60], fields where logical inductive biases are crucial toward strong performance.

Causal inference extends the associational level of reasoning to interventional and counterfactual levels [255], which requires extensive knowledge of the world to imagine counterfactual effects. For example, Yi et al. [390] 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 [6] and Counterfactual VQA [243] to measure the robustness of VQA models under controlled interventions to the question as a step toward 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 [360].

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 [41], DBpedia [25], YAGO [311], WordNet [229]) to semantic networks containing multimodal concepts as nodes and multimodal relationships as edges [415]. Multimodal knowledge graphs are important because they enable the grounding of structured information in the visual and physical world. For example, Liu et al. [204] construct 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 and text [165]. These multimodal knowledge bases have been shown to benefit visual question answering [372, 417], knowledge base completion [261], and image captioning [226]. Gui et al. [111] integrates knowledge into vision-and-language transformers for automatic reasoning over both knowledge sources. Another promising approach is multimodal knowledge expansion [267, 373, 383] using knowledge distillation to expand knowledge from unimodal data to multimodal settings. We refer the reader to a comprehensive survey by Zhu et al. [415] 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 [402] and VisualCOMET [253], while other works have also introduced datasets with video and text inputs to test for temporal reasoning (e.g., MovieQA [324], MovieFIB [217], TVQA [174]). Benchmarks for multimodal commonsense typically require leveraging external knowledge from knowledge bases [306] or pretraining paradigms on large-scale datasets [207, 403].

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 15). These three categories are distinguished based on the information change from input to output modalities, following categorizations in text generation [78]. We will cover recent advances as well as the evaluation of generated content.

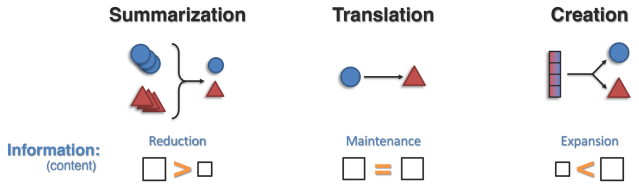

Fig. 15. How can we learn a generative process to produce raw modalities that reflect cross-modal interactions, structure, and coherence? **Generation** involves (1) *summarizing* multimodal data to highlight the most salient parts, (2) *translating* from one modality to another while being consistent with modality connections, and (3) *creating* multiple modalities simultaneously while maintaining coherence.

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 [61], video [181], and audio [90, 147, 177]. 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 [62, 147, 177]. 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 [282]. Li et al. [177] collected a dataset of news videos and articles paired with manually annotated summaries as a benchmark toward multimodal summarization. Finally, UzZaman et al. [346] 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 [61, 183]. Although most approaches only focus on generating a textual summary from multimodal data [250], several directions have also explored generating summarized images to supplement the generated textual summary [61, 181].

6.2 Subchallenge 4b: Translation

Translation aims to map one modality to another while respecting semantic connections and information content [355]. For example, generating a descriptive caption of an image can help improve the accessibility of visual content for blind people [113]. 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 [92], and *generative* models which can translate into arbitrary instances interpolating beyond the data but face challenges in quality, diversity, and evaluation [161, 273, 341]. Despite these challenges, recent progress in large-scale generative models has yielded impressive results in text-to-image [273, 278], text-to-video [302], audio-to-image [146], text-to-speech [276], speech-to-gesture [10], speaker-to-listener [240], language to pose [11], and speech and music generation [7, 72, 245].

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

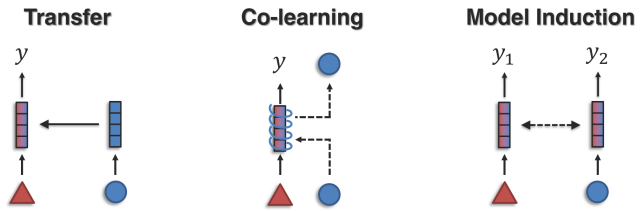

Fig. 16. Transference studies the transfer of knowledge between modalities, usually to help a noisy or limited primary modality, via (1) *cross-modal transfer* from models trained with abundant data in the secondary modality, (2) *multimodal co-learning* to share information across modalities by sharing representations, and (3) *model induction* that keeps individual unimodal models separate but induces behavior in separate models.

for a long time [269], and recent work has explored high-resolution speech and sound generation using neural networks [245]. Photorealistic image generation has also recently become possible due to advances in large-scale generative modeling [154]. Furthermore, there have been a number of attempts at generating abstract scenes [319], computer graphics [228], and talking heads [414]. While there has been some progress toward video generation [302], 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 [35], hate speech [2, 104], deepfakes [114], and lip-syncing videos [317] 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 [105]. Several automatic proxy metrics have been proposed [19, 63] by none are universally robust across many generation tasks.

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 16).

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 [187] and prompt tuning [176], recent work has also studied the tuning of pretrained language models to condition on visual and other modalities. For example, Tsimpoukelli et al. [341] quickly condition a pretrained, frozen language model on images for image captioning. Related work has also adapted prefix tuning for image captioning [59], multimodal fusion [116], and summarization [393]. 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. [418] include

additional self-attention layers between language model layers and external modalities. Rahman et al. [271] 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 [144], MultiModel [152], ViT-BERT [184], and PolyViT [198] 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) [81], images (sequence of patches) [83], video (sequence of images) [314], and other time-series data (sequence of timesteps) [199]. 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 [194], VATT [13], FLAVA [303], and Gato [275].

Transfer learning: While more research has focused on transfer within the same modality with external information [304, 375, 398], Liang et al. [196] studies transfer to new modalities using small amounts of paired but unlabeled data. Lu et al. [208] found that Transformers pretrained on language transfer to other sequential modalities as well. Liang et al. [194] build 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 [135], interactive decision-making [185], and robotics [44].

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 [95]. Marino et al. [222] use knowledge graphs for image classification via a graph-based joint representation. Jia et al. [149] improve image classifiers with contrastive learning between images and noisy captions. Finally, Zadeh et al. [398] 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 include language modeling by mapping contextualized text embeddings into images [321], image classification by projecting image embeddings into word embeddings [304], and language sentiment analysis by translating language into video and audio [262].

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 are used to pseudo-label new unlabeled examples in the other modality, thereby enlarging the training set of the other modality [40]. The second is co-regularization [301, 308], 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.

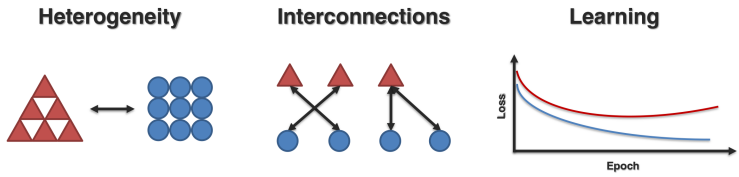

Fig. 17. Quantification: what are the empirical and theoretical studies we can design to better understand (1) the dimensions of *heterogeneity*, (2) the presence and type of *interconnections*, and (3) the *learning* and optimization challenges?

Multimodal co-training extends co-training by jointly learning classifiers for multiple modalities [124]. Guillaumin et al. [112] 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. [67] perform semi-supervised multimodal learning using a diversity-preserving co-training algorithm. Finally, Dunnmon et al. [87] apply 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 employ 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 [301]. Sridharan and Kakade [308] provide guarantees for these approaches using an information-theoretic framework. More recently, similar co-regularization approaches have also been applied for multimodal feature selection [128], semi-supervised multimodal learning [387], and video summarization [232].

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 three 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 17).

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 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 [252, 338, 400] or (2) post-hoc explanations of black-box models [53, 109]. In the former, methods such as Concept Bottleneck Models [162] and fitting sparse linear layers [368] or decision trees [356] on top of deep feature representations have emerged as promising choices. In the latter, gradient-based visualizations [109, 290, 300]) and feature attributions (e.g., modality contribution [103], LIME [277], and Shapley values [227]) have been used to highlight regions of modality importance.

Modality biases are unintended correlations between input and outputs [38, 42]. Modality biases can lead to unexpectedly poor performance in the real world [283], or even more dangerously, potential for harm toward underrepresented groups [119, 258]. For example, Goyal et al. [108] found *unimodal biases* in the language modality of VQA tasks, resulting in mistakes due

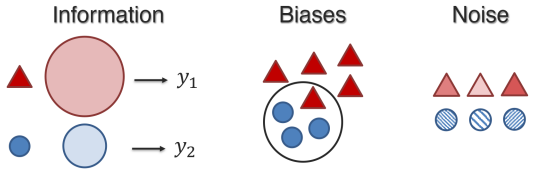

Fig. 18. The subchallenge of **heterogeneity** quantification aims to understand the dimensions of heterogeneity commonly encountered in multimodal research, such as (1) different quantities and usages of *modality information*, (2) the presence of *modality biases*, and (3) quantifying and mitigating *modality noise*.

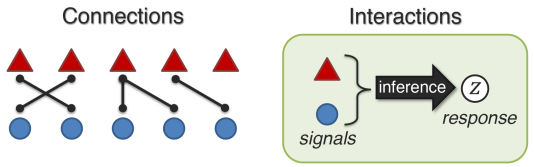

Fig. 19. Quantifying **modality interconnections** studies (1) *connections*: can we discover what modality elements are related to each other and why, and (2) *interactions*: can we understand how modality elements interact during inference?

to ignoring visual information [8]. Subsequent work has developed carefully curated diagnostic benchmarks to mitigate data collection biases, like VQA 2.0 [108], GQA [142], and NLVR2 [312]. Recent work has also found compounding *social biases* in multimodal systems [68, 279, 309] stemming from gender bias in both language and visual modalities [46, 297], which may cause danger when deployed [258].

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 timesteps. Liang et al. [195] collect a comprehensive set of targeted noisy distributions unique to each modality. In addition to natural noise topologies [173, 213], related work has also explored adversarial attacks [82] and distribution shifts [93] in multimodal systems. Finally, there have been recent efforts on incomplete multimodal learning [194, 214, 359, 388] to account for noisy or missing modalities, such as modality imputation using probabilistic models [214], autoencoders [333], translation models [262], low-rank approximations [192], or knowledge distillation [359], or training general models with a wide range of modalities so they can still operate on partial subsets [194, 275]. However, they may run the risk of possible error compounding and require knowing which modalities are imperfect beforehand.

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 modality elements interact during inference (Figure 19).

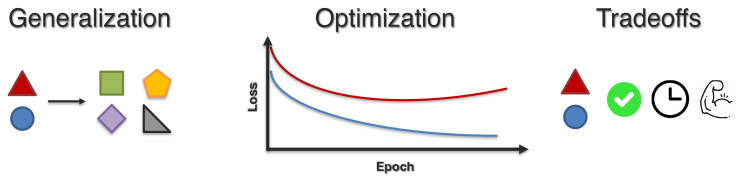

Fig. 20. Studying the multimodal **learning process** involves understanding (1) *generalization* across modalities and tasks, (2) *optimization* for balanced and efficient training, and (3) *tradeoffs* between performance, robustness, and complexity in the real-world deployment of multimodal models.

Connections: Recent work has explored the quantification of modality connections through visualization tools on joint representation spaces [143] or attention maps [5]. Perturbation-based analysis perturbs the input and observes changes in the output to understand internal connections [193, 243]. Finally, specifically curated diagnostic datasets are also useful in understanding semantic connections: Winoground [327] probes vision and language models for visio-linguistic compositionality, and PaintSkills [68] measures the connections necessary for visual reasoning.

Interactions: One common categorization of interactions involves redundancy, uniqueness, and synergy [367]. 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 [28, 40] and contrastive learning estimators [332, 339]. Other approaches have studied these measures in isolation, such as redundancy via distance between prediction logits using either feature [224], statistical distribution tests on input features [26], or via human annotations [281]. From the semantic view, recent work in Causal VQA [6] and Counterfactual VQA [243] 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 [307, 340, 382]. Parallel research such as EMAP [121], DIME [212], M2Lens [363], and MultiViz [193] 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 [106], difficulty in evaluation [166], and challenges in extending visualization methods to applications such as biomedical data integration, imaging, intelligent systems, and user interfaces.

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 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 much larger number of sensors beyond the prototypical text, video, and audio modalities [139]. Recent work has studied generalization across paired modality inputs [196, 268] and in unpaired scenarios where each task is defined over only a small subset of all modalities [194, 208, 275].

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 [362]. Subsequent work has studied why joint training of multimodal networks may be difficult and proposed methods to improve the optimization process via weighting approaches [371], adaptive learning [137, 138], or contrastive learning [197].

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 [195] as well as increased susceptibility to noise and imperfection in the additional modality [262]. 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 [188, 325, 339], while information-theoretic arguments have also provided useful insights [40, 308].

9 CONCLUSION

This article defined three core principles of modality heterogeneity, connections, and interactions central to multimodal machine learning research, before proposing a taxonomy of six core technical challenges: representation, alignment, reasoning, generation, transference, and quantification covering historical and recent directions. Despite the immense opportunities afforded by recent progress, there remain many unsolved challenges:

9.1 Future Directions

Representation: *Theoretical and empirical frameworks.* How can we formally define the three core principles of heterogeneity, connections, and interactions? What mathematical or empirical frameworks will enable us to taxonomize the dimensions of heterogeneity and interconnections, and subsequently quantify their presence in multimodal datasets and models? Answering these fundamental questions will lead to a better understanding of the capabilities and limitations of current multimodal representations. *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 interactions? *Brain and multimodal perception.* There are many core insights regarding multimodal processing to be gained from human cognition, including the brain's multimodal properties [163] and mental imagery [236]. 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: *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: *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 [327]. How can we best test for compositionality, and what reasoning approaches can enable compositional generalization?

Generation: *Creation and real-world ethical concerns.* Synchronized creation of realistic video, text, and audio remains a challenge. Furthermore, 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 [297], toxic speech [104], and sensitive pretraining data [51]. 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: *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 [139]. 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: *Modality utility, tradeoffs, and selection.* How can we formalize why modalities can be useful or potentially harmful for a task? Can we come up with formal guidelines to compare these tradeoffs and select modalities? *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? There are also challenges in quantifying *modality and social biases* and *robustness* to imperfect, noisy, and out-of-distribution modalities.

In conclusion, we believe that our taxonomy will help to catalog future research papers and better understand the remaining unresolved problems in multimodal machine learning.

REFERENCES

[1] Mahdi Abavisani and Vishal M. Patel. 2018. Deep multimodal subspace clustering networks. *IEEE Journal of Selected Topics in Signal Processing* 12, 6 (2018), 1601–1614.

[2] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *AIES*. 298–306.

[3] Ahmad Abiri, Jake Pensa, Anna Tao, Ji Ma, Yen-Yi Juo, and Syed J. Askari. 2019. Multi-modal haptic feedback for grip force reduction in robotic surgery. *Scientific Reports* 9, 1 (2019), 1–10.

[4] Julián N. Acosta, Guido J. Falcone, Pranav Rajpurkar, and Eric J. Topol. 2022. Multimodal biomedical AI. *Nature Medicine* 28, 9 (2022), 1773–1784.

[5] Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. 2022. VL-InterpreT: An interactive visualization tool for interpreting vision-language transformers. In *CVPR*. 21406–21415.

[6] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2020. Towards causal VQA: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *CVPR*. 9690–9698.

[7] Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, and Antoine Caillon. 2023. MusicLM: Generating music from text. arXiv:2301.11325. Retrieved from https://arxiv.org/abs/2301.11325

[8] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *EMNLP*. 1955–1960.

[9] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2017. VQA: Visual question answering: www.visualqa.org. *International Journal of Computer Vision* 123, 1 (2017), 4–31.

[10] Chaitanya Ahuja, Dong Won Lee, Yukiko I. Nakano, and Louis-Philippe Morency. 2020. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *ECCV*. Springer, 248–265.

[11] Chaitanya Ahuja and Louis-Philippe Morency. 2019. Language2pose: Natural language grounded pose forecasting. In *3DV*. IEEE, 719–728.

[12] Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. 2014. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446* (2014).
[13] Hassan Akbari, Liangzhe Yuan, and Rui Qian. 2021. VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. *arXiv preprint arXiv:2104.11178* (2021).
[14] Mehmet Aktukmak, Yasin Yilmaz, and Ismail Uysal. 2019. A probabilistic framework to incorporate mixed-data type features: Matrix factorization with multimodal side information. *Neurocomputing* 367 (2019), 164–175.
[15] Jean-Baptiste Alayrac, Ję Donahue, Pauline Luc, Antoine Miech, Iain Barr, and Yana Hasson. 2022. Flamingo: A visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198* (2022).
[16] Camila Alviar, Rick Dale, Akeiylah Dewitt, and Christopher Kello. 2020. Multimodal coordination of sound and movement in music and speech. *Discourse Processes* 57, 8 (2020), 682–702.
[17] Paras Malik Amisha, Monika Pathania, and Vyas Kumar Rathaur. 2019. Overview of artificial intelligence in medicine. *Journal of Family Medicine and Primary Care* 8, 7 (2019), 2328.
[18] Saeed Amizadeh, Hamid Palangi, Alex Polozov, Yichen Huang, and Kazuhito Koishida. 2020. Neuro-symbolic visual reasoning: Disentangling visual from reasoning. In *ICML*. PMLR, 279–290.
[19] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*. Springer, 382–398.
[20] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *CVPR*. 39–48.
[21] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *ICML*.
[22] Xavier Anguera, Jordi Luque, and Ciro Gracia. 2014. Audio-to-text alignment for speech recognition with very limited resources. In *INTERSPEECH*.
[23] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A. González. 2017. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992* (2017).
[24] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. 2010. Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems* 16, 6 (2010), 345–379.
[25] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a web of open data. In *The Semantic Web*. Springer, 722–735.
[26] Benjamin Auffarth, Maite López, and Jesús Cerquides. 2010. Comparison of redundancy and relevance measures for feature selection in tissue classification of CT images. In *ICDM*. Springer, 248–262.
[27] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, and Yusuf Hanafy. 2023. OpenFlamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390* (2023).
[28] Maria-Florina Balcan, Avrim Blum, and Ke Yang. 2004. Co-training and expansion: Towards bridging theory and practice. In *NeurIPS*.
[29] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE TPAMI* 41, 2 (2018), 423–443.
[30] George Barnum, Sabera J. Talukder, and Yisong Yue. 2020. On the benefits of early fusion in multimodal representation learning. In *NeurIPS 2020 Workshop SVRHM*.
[31] Reuben M. Baron and David A. Kenny. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* 51, 6 (1986), 1173.
[32] Roland Barthes. 1977. *Image-Music-Text*. Macmillan.
[33] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2006. Analysis of representations for domain adaptation. In *NeurIPS*.
[34] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. MUTAN: Multimodal tucker fusion for visual question answering. In *ICCV*. 2612–2620.
[35] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FaaCT*. 610–623.
[36] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 8 (2013), 1798–1828.
[37] Brian T. Bethea, Allison M. Okamura, Masaya Kitagawa, Torin P. Fitton, Stephen M. Cattaneo, Vincent L. Gott, William A. Baumgartner, and David D. Yuh. 2004. Application of haptic feedback to robotic surgery. *Journal of Laparoendoscopic & Advanced Surgical Techniques* 14, 3 (2004), 191–195.
[38] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: Misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963* (2021).
[39] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, and Aleksandr Nisnevich. 2020. Experience grounds language. In *EMNLP*. 8718–8735.
[40] Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *COLT*. 92–100.

[41] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *ACM SIGMOD*. 1247–1250.
[42] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *NeurIPS*. 4349–4357.
[43] Simon Brodeur, Ethan Perez, Ankesh Anand, Florian Golemo, and Luca Celotti. 2017. HoME: A household multimodal environment. In *NIPS 2017's Visually-Grounded Interaction and Language Workshop*.
[44] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, and Chelsea Finn. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818* (2023).
[45] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. 2021. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478* (2021).
[46] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAccT*. PMLR, 77–91.
[47] Qiong Cai, Hao Wang, Zhenmin Li, and Xiao Liu. 2019. A survey on multimodal data-driven smart healthcare systems: Approaches and applications. *IEEE Access* 7 (2019), 133583–133599.
[48] Juan C. Caicedo and Fabio A. González. 2012. Online matrix factorization for multimodal image retrieval. In *Iberoamerican Congress on Pattern Recognition*. Springer, 340–347.
[49] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *European Conference on Computer Vision*. Springer, 565–580.
[50] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Qiang Yang. 2017. Transitive hashing network for heterogeneous multimedia retrieval. In *AAAI*. 81–87.
[51] Nicholas Carlini, Florian Tramer, EricWallace, Matthew Jagielski, and Ariel Herbert-Voss. 2021. Extracting training data from large language models. In *USENIX Security*. 2633–2650.
[52] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (An _Obviously_ Perfect Paper). In *ACL*.
[53] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. 2018. Do explanations make VQA models more predictable to a human? In *EMNLP*.
[54] Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W. Black. 2021. Grounding 'Grounding' in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 4283–4305.
[55] Wilson Chango, Juan A. Lara, Rebeca Cerezo, and Cristobal Romero. 2022. A review on data fusion in multimodal learning analytics and educational data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12, 4 (2022), e1458.
[56] Devendra Singh Chaplot, Kanthashree Mysore Sathyendra, Rama Kumar Pasumarthi, Dheeraj Rajagopal, and Ruslan Salakhutdinov. 2018. Gated-attention architectures for task-oriented language grounding. In *AAAI*.
[57] Anargyros Chatzitofis, Leonidas Saroglou, Prodromos Boutis, Petros Drakoulis, and Nikolaos Zioulis. 2020. Human4d: A human-centric multimodal dataset for motions and immersive media. *IEEE Access* 8 (2020), 176241–176262.
[58] Brian Chen, Andrew Rouditchenko, Kevin Duarte, Hilde Kuehne, Samuel Thomas, and Angie Boggust. 2021. Multimodal clustering networks for self-supervised learning from unlabeled videos. In *ICCV*. 8012–8021.
[59] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. 2021. VisualGPT: Data-efficient adaptation of pretrained language models for image captioning. *arXiv preprint arXiv:2102.10407* (2021).
[60] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. 2021. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In *ACL-IJCNLP Findings*.
[61] Jingqiang Chen and Hai Zhuge. 2018. Abstractive text-image summarization using multi-modal attentional hierarchical RNN. In *EMNLP*.
[62] Jingqiang Chen and Hai Zhuge. 2018. Extractive text-image summarization using multi-modal RNN. In *SKG*. IEEE, 245–248.
[63] Lele Chen, Guofeng Cui, Ziyi Kou, Haitian Zheng, and Chenliang Xu. 2020. What comprises a good talking-head video generation?: A survey and benchmark. *arXiv preprint arXiv:2005.03201* (2020).
[64] Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. 2020. Graph optimal transport for cross-domain alignment. In *ICML*. PMLR, 1542–1553.
[65] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *ICMI*. 163–171.
[66] Gong Cheng, Junwei Han, and Xiaoqiang Lu. 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE* 105, 10 (2017), 1865–1883.
[67] Yanhua Cheng, Xin Zhao, Rui Cai, Zhiwei Li, Kaiqi Huang, and Yong Rui. 2016. Semi-supervised multimodal deep learning for RGB-D object recognition. In *IJCAI*. 3345–3351.

[68] Jaemin Cho, Abhay Zala, and Mohit Bansal. 2022. DALL-Eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053* (2022).
[69] Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. 2018. Using syntax to ground referring expressions in natural images. In *AAAI*, Vol. 32.
[70] Volkan Cirik, Taylor Berg-Kirkpatrick, and L.-P. Morency. 2020. Refer360: A referring expression recognition dataset in 360 images. In *ACL*.
[71] Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. 2018. Visual referring expression recognition: What do systems actually learn? In *NAACL*. 781–787.
[72] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation. *arXiv preprint arXiv:2306.05284* (2023).
[73] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, and Weisheng Wang. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500. Retrieved from https://arxiv.org/abs/2305.06500
[74] Debraj De, Pratool Bharti, Sajal K. Das, and Sriram Chellappan. 2015. Multimodal wearable sensing for fine-grained activity recognition in healthcare. *IEEE Internet Computing* 19, 5 (2015), 26–35.
[75] Victoria Dean, Shubham Tulsiani, and Abhinav Gupta. 2020. See, hear, explore: Curiosity via audio-visual association. In *NeurIPS*. 14961–14972.
[76] Emilie Delaherche and Mohamed Chetouani. 2010. Multimodal coordination: Exploring relevant features and measures. In *SSPW*. 47–52.
[77] Joseph DelPreto, Chao Liu, and Yiyue Luo. 2022. ActionSense: A multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment. In *NeurIPS*.
[78] Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *EMNLP*. 7580–7605.
[79] Emily Denton and Rob Fergus. 2018. Stochastic video generation with a learned prior. In *ICML*. PMLR, 1174–1183.
[80] Laurence Devillers, Laurence Vidrascu, and Lori Lamel. 2005. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks* 18, 4 (2005), 407–422.
[81] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.
[82] Wenhao Ding, Baiming Chen, Bo Li, Kim Ji Eun, and Ding Zhao. 2021. Multimodal safety-critical scenarios generation for decision-making algorithms evaluation. *IEEE Robotics and Automation Letters* 6, 2 (2021), 1551–1558.
[83] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, and Xiaohua Zhai. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.
[84] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, and Ayzaan Wahid. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378* (2023).
[85] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP*. IEEE, 736–740.
[86] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. 2022. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936* (2022).
[87] Jared A. Dunnmon, Alexander J. Ratner, Khaled Saab, Nishith Khandwala, Matthew Markert, Hersh Sagreiya, Roger Goldman, Christopher Lee-Messer, Matthew P. Lungren, and Daniel L. Rubin. 2020. Cross-modal data programming enables rapid medical machine learning. *Patterns* 1, 2 (2020).
[88] Chris Dyer. 2014. Notes on noise contrastive estimation and negative sampling. *arXiv preprint arXiv:1410.8251* (2014).
[89] Sergio Escalera, Jordi Gonzàlez, Xavier Baró, Miguel Reyes, Oscar Lopes, Isabelle Guyon, Vassilis Athitsos, and Hugo Escalante. 2013. Multi-modal gesture recognition challenge 2013: Dataset and results. In *ICMI*. 445–452.
[90] Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. 2013. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia* 15, 7 (2013), 1553–1568.
[91] Haoqi Fan and Jiatong Zhou. 2018. Stacked latent attention for multimodal reasoning. In *CVPR*. 1072–1080.
[92] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *ECCV*. Springer, 15–29.
[93] Andreas Foltyn and Jessica Deuschel. 2021. Towards reliable multimodal stress detection under distribution shift. In *ICMI*. 329–333.
[94] Jerome H. Friedman and Bogdan E. Popescu. 2008. Predictive learning via rule ensembles. *The Annals of Applied Statistics* 2, 3 (2008), 916–954.
[95] Andrea Frome, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *NeurIPS*. 2121–2129.

[96] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*. ACL, 457–468.
[97] Hiroki Furuta, Ofir Nachum, Kuang-Huei Lee, Yutaka Matsuo, Shixiang Shane Gu, and Izzeddin Gur. 2023. Multimodal web navigation with instruction-finetuned foundation models. *arXiv preprint arXiv:2305.11854* (2023).
[98] Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetzsche. 2020. Early vs late fusion in multimodal convolutional neural networks. In *FUSION*. IEEE, 1–6.
[99] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. 2022. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and TrendsR in Computer Graphics and Vision* 14, 3–4 (2022), 163–352.
[100] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, and Wei Zhang. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010* (2023).
[101] Ruohan Gao and Kristen Grauman. 2019. 2.5 d visual sound. In *CVPR*. 324–333.
[102] Enrique Garcia-Ceja, Michael Riegler, Tine Nordgreen, Petter Jakobsen, Ketil J. Oedegaard, and Jim Tørresen. 2018. Mental health monitoring with multimodal sensing and machine learning: A survey. *Pervasive and Mobile Computing* 51 (2018), 1–26.
[103] Itai Gat, Idan Schwartz, and Alex Schwing. 2021. Perceptual score: What data modalities does your model perceive? In *NeurIPS*.
[104] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *EMNLP Findings*. 3356–3369.
[105] Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. In *EMNLP-IJCNLP*. 1161–1166.
[106] Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of neural networks is fragile. In *AAAI*. Vol. 33, 3681–3688.
[107] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. VQA-LOL: Visual question answering under the lens of logic. In *ECCV*. Springer, 379–396.
[108] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*. 6904–6913.
[109] Yash Goyal, Akrit Mohapatra, Devi Parikh, and Dhruv Batra. 2016. Towards transparent AI systems: Interpreting visual question answering models. *arXiv preprint arXiv:1608.08974* (2016).
[110] Edouard Grave, Armand Joulin, and Quentin Berthet. 2019. Unsupervised alignment of embeddings with Wasserstein Procrustes. In *AISTATS*. PMLR, 1880–1890.
[111] Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2021. KAT: A knowledge augmented transformer for vision-and-language. *arXiv preprint arXiv:2112.08614* (2021).
[112] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. 2010. Multimodal semi-supervised learning for image classification. In *CVPR*. IEEE, 902–909.
[113] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*. 3608–3617.
[114] Jeffrey T. Hancock and Jeremy N. Bailenson. 2021. The Social Impact of Deepfakes. 149–152 pages.
[115] Sanjay Haresh, Sateesh Kumar, Huseyin Coskun, Shahram N. Syed, Andrey Konin, Zeeshan Zia, and Quoc-Huy Tran. 2021. Learning by aligning videos in time. In *CVPR*. 5548–5558.
[116] Md Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. 2021. Humor knowledge enriched transformer for understanding multimodal humor. In *AAAI*.
[117] Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, et al. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. In *EMNLP-IJCNLP*. 2046–2056.
[118] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. PathVQA: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286* (2020).
[119] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *ECCV*. 771–787.
[120] Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. 2021. Decoupling the role of data, attention, and losses in multimodal transformers. *arXiv preprint arXiv:2102.00529* (2021).
[121] Jack Hessel and Lillian Lee. 2020. Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think!. In *EMNLP*.
[122] Jack Hessel, Ana Marasović, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2022. Do androids laugh at electric sheep? Humor "Understanding" benchmarks from the New Yorker Caption Contest. *arXiv preprint arXiv:2209.06293* (2022).
[123] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2016. beta-VAE: Learning basic visual concepts with a constrained variational framework. (2016).

[124] Ryota Hinami, Junwei Liang, Shin'ichi Satoh, and Alexander Hauptmann. 2018. Multimodal co-training for selecting good examples from webly labeled video. *arXiv preprint arXiv:1804.06057* (2018).
[125] Danfeng Hong, Lianru Gao, Naoto Yokoya, Jing Yao, Jocelyn Chanussot, Qian Du, and Bing Zhang. 2020. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. *IEEE Transactions on Geoscience and Remote Sensing* 59, 5 (2020), 4340–4354.
[126] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. 2019. Learning to compose and reason with language tree structures for visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 2 (2019), 684–696.
[127] Ming Hou, Jiajia Tang, Jianhai Zhang, Wanzeng Kong, and Qibin Zhao. 2019. Deep multimodal multilinear fusion with high-order polynomial pooling. In *NeurIPS*. 12136–12145.
[128] Tsung-Yu Hsieh, Yiwei Sun, Suhang Wang, and Vasant Honavar. 2019. Adaptive structural co-regularization for unsupervised multi-view feature selection. In *ICBK'19*. IEEE.
[129] Tzu-Ming Harry Hsu, Wei-Hung Weng, Willie Boag, Matthew McDermott, and Peter Szolovits. 2018. Unsupervised multimodal representation learning across medical images and reports. *arXiv preprint arXiv:1811.08615* (2018).
[130] Wei-Ning Hsu and James Glass. 2018. Disentangling by partitioning: A representation learning framework for multimodal sensory data. *arXiv preprint arXiv:1805.11264* (2018).
[131] Di Hu, Feiping Nie, and Xuelong Li. 2019. Deep multimodal clustering for unsupervised audiovisual learning. In *CVPR*. 9248–9257.
[132] Peng Hu, Dezhong Peng, Xu Wang, and Yong Xiang. 2019. Multimodal adversarial network for cross-modal retrieval. *Knowledge-Based Systems* 180 (2019), 38–50.
[133] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. 2017. Learning to reason: End-to-end module networks for visual question answering. In *ICCV*. 804–813.
[134] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. In *CVPR*. 4555–4564.
[135] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: extracting actionable knowledge for embodied agents. *arXiv preprint arXiv:2201.07207* (2022).
[136] Xin Huang, Yuxin Peng, and Mingkuan Yuan. 2017. Cross-modal common representation learning by hybrid transfer network. In *IJCAI*. 1893–1900.
[137] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. 2021. What makes multimodal learning better than single (provably). In *NeurIPS*.
[138] Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. 2022. Modality competition: what makes joint training of multi-modal network fail in deep learning?(Provably). *arXiv preprint arXiv:2203.12221* (2022).
[139] Zhenhua Huang, Xin Xu, Juan Ni, Honghao Zhu, and Cheng Wang. 2019. Multimodal representation learning for recommendation in Internet of Things. *IEEE Internet of Things Journal* 6, 6 (2019), 10675–10685.
[140] Drew Hudson and Christopher D. Manning. 2019. Learning by abstraction: The neural state machine. In *NeurIPS*.
[141] Drew A. Hudson and Christopher D. Manning. 2018. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067* (2018).
[142] Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*. 6700–6709.
[143] Masha Itkina, B. Ivanovic, Ransalu Senanayake, Mykel J. Kochenderfer, and Marco Pavone. 2020. Evidential sparsification of multimodal latent spaces in conditional variational autoencoders. arXiv:2010.09164. Retrieved from https://arxiv.org/abs/2010.09164
[144] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. *arXiv preprint arXiv:2103.03206* (2021).
[145] Alejandro Jaimes and Nicu Sebe. 2007. Multimodal human–computer interaction: A survey. *Computer Vision and Image Understanding* 108, 1–2 (2007), 116–134.
[146] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. 2019. You said that?: Synthesising talking faces from audio. *International Journal of Computer Vision* 127, 11 (2019), 1767–1779.
[147] Anubhav Jangra, Adam Jatowt, Mohammad Hasanuzzaman, and Sriparna Saha. 2020. Text-image-video summary generation using joint integer linear programming. In *ECIR*. Springer.
[148] Siddhant M. Jayakumar, Wojciech M. Czarnecki, Jacob Menick, Jonathan Schwarz, Jack Rae, Simon Osindero, Yee Whye Teh, Tim Harley, and Razvan Pascanu. 2020. Multiplicative interactions and where to find them. In *ICLR*.
[149] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, and Hieu Pham. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*. PMLR, 4904–4916.
[150] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, H. Lehman Li-Wei, Mengling Feng, and Mohammad Ghassemi. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* 3, 1 (2016), 1–9.

[151] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*. 2901–2910.
[152] Lukasz Kaiser, Aidan N. Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. 2017. One model to learn them all. arXiv:1706.05137. Retrieved from https://arxiv.org/abs/1706.05137
[153] Andrej Karpathy, Armand Joulin, and Li F. Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *NeurIPS*.
[154] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of StyleGAN. In *CVPR*. 8110–8119.
[155] Vasil Khalidov, Florence Forbes, and Radu Horaud. 2011. Conjugate mixture models for clustering multimodal data. *Neural Computation* (2011).
[156] Aparajita Khan and Pradipta Maji. 2019. Approximate graph Laplacians for multimodal data clustering. *IEEE TPAMI* 43, 3 (2019), 798–813.
[157] Aparajita Khan and Pradipta Maji. 2021. Multi-manifold optimization for multi-view subspace clustering. *IEEE Transactions on Neural Networks and Learning Systems* 33, 8 (2021), 3895–3907.
[158] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *NeurIPS*. 2611–2624.
[159] Minjae Kim, David K. Han, and Hanseok Ko. 2016. Joint patch clustering-based dictionary learning for multimodal image fusion. *Information Fusion* 27 (2016), 198–214.
[160] Elsa A. Kirchner, Stephen H. Fairclough, and Frank Kirchner. 2019. Embedded multimodal interfaces in robotics: applications, future trends, and societal implications. *Association for Computing Machinery and Morgan & Claypool*, 523–576. https://doi.org/10.1145/3233795.3233810
[161] Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021. Text-to-image generation grounded by fine-grained user attention. In *WACV*.
[162] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *ICML*. PMLR, 5338–5348.
[163] Stephen M. Kosslyn, Giorgio Ganis, and William L. Thompson. 2010. Multimodal images in the brain. *The Neurophysiological Foundations of Mental and Motor Imagery* (2010), 3–16.
[164] Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual coreference resolution in visual dialog using neural module networks. In *ECCV*. 153–169.
[165] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, and Joshua Kravitz. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.
[166] Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. 2022. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602* (2022).
[167] Joseph B. Kruskal. 1983. An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM Review* 25, 2 (1983), 201–237.
[168] Pei Ling Lai and Colin Fyfe. 2000. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems* 10, 05 (2000), 365–377.
[169] Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data* 5, 1 (2018), 1–10.
[170] Rémi Lebret, Pedro Pinheiro, and Ronan Collobert. 2015. Phrase-based image captioning. In *ICML*. PMLR, 2085–2094.
[171] Dong Won Lee, Chaitanya Ahuja, Paul Pu Liang, Sanika Natu, and Louis-Philippe Morency. 2023. Lecture presentations multimodal dataset: Towards understanding multimodality in educational videos. In *ICCV*. 20087–20098.
[172] Michelle A. Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, et al. 2019. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *ICRA*. IEEE, 8943–8950.
[173] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. 2023. Multimodal prompting with missing modalities for visual recognition. In *CVPR*. 14943–14952.
[174] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *EMNLP*. 1369–1379.
[175] Luis A. Leiva, Asutosh Hota, and Antti Oulasvirta. 2020. Enrico: A dataset for topic modeling of mobile UI designs. In *MobileHCI*. 1–4.
[176] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *EMNLP*. 3045–3059.
[177] Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *EMNLP*. 1092–1102.

[178] Jiaxin Li, Danfeng Hong, Lianru Gao, Jing Yao, Ke Zheng, Bing Zhang, and Jocelyn Chanussot. 2022. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation* 112 (2022), 102926.
[179] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).
[180] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).
[181] Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2020. VMSMO: Learning to generate multimodal summary for video-based news articles. *arXiv preprint arXiv:2010.05406* (2020).
[182] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. 2022. Clip-event: Connecting text and images with event structures. In *CVPR*. 16420–16429.
[183] Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019. Keep meeting summaries on topic: Abstractive multimodal meeting summarization. In *ACL*. 2190–2196.
[184] Qing Li, Boqing Gong, Yin Cui, Dan Kondratyuk, and Xianzhi Du. 2021. Towards a unified foundation model: Jointly pre-training transformers on unpaired images and text. *arXiv preprint arXiv:2112.07074* (2021).
[185] Shuang Li, Xavier Puig, Yilun Du, Clinton Wang, Ekin Akyurek, Antonio Torralba, Jacob Andreas, and Igor Mordatch. 2022. Pre-trained language models for interactive decision-making. *arXiv preprint arXiv:2202.01771* (2022).
[186] Shu Li, Wei Wang, Wen-Tao Li, and Pan Chen. 2021. Multi-view representation learning with manifold smoothness. In *AAAI*. Vol. 35, 8447–8454.
[187] Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL-IJCNLP*.
[188] Paul Pu Liang. 2022. Brainish: Formalizing a multimodal language for intelligence and consciousness. *arXiv preprint arXiv:2205.00001* (2022).
[189] Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard J. Chen, and Zihao Deng. 2023. Quantifying & modeling multimodal interactions: An information decomposition framework. In *NeurIPS*.
[190] Paul Pu Liang, Zihao Deng, Martin Ma, James Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2023. Factorized contrastive learning: Going beyond multi-view redundancy. In *NeurIPS*.
[191] Paul Pu Liang, Terrance Liu, Anna Cai, Michal Muszynski, Ryo Ishii, Nicholas Allen, and Randy Auerbach. 2021. Learning language and multimodal privacy-preserving markers of mood from mobile data. In *ACL/IJCNLP (1)*.
[192] Paul Pu Liang, Zhun Liu, Yao-Hung Hubert Tsai, Qibin Zhao, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2019. Learning representations from imperfect time series data via tensor rank regularization. In *ACL*.
[193] Paul Pu Liang, Yiwei Lyu, Gunjan Chhablani, Nihal Jain, and Zihao Deng. 2023. MultiViz: Towards visualizing and understanding multimodal models. In *ICLR*.
[194] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Jeffrey Tsaw, Yudong Liu, Shentong Mo, Dani Yogatama, Louis-Philippe Morency, and Russ Salakhutdinov. 2023. High-modality multimodal transformer: Quantifying modality & interaction heterogeneity for high-modality representation learning. *Transactions on Machine Learning Research* (2023).
[195] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, and Leslie Yufan Chen. 2021. MultiBench: Multiscale benchmarks for multimodal representation learning. In *NeurIPS Datasets and Benchmarks Track*.
[196] Paul Pu Liang, Peter Wu, Liu Ziyin, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Cross-modal generalization: Learning in low resource modalities via meta-alignment. In *ACM Multimedia*. 2680–2689.
[197] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y. Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*. 17612–17625.
[198] Valerii Likhosherstov, Mostafa Dehghani, Anurag Arnab, Krzysztof Marcin Choromanski, Mario Lucic, Yi Tay, and Adrian Weller. 2022. PolyViT: Co-training Vision Transformers on Images, Videos and Audio.
[199] Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. 2021. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting* 37, 4 (2021), 1748–1764.
[200] Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. In *EMNLP-IJCNLP*. 2829–2839.
[201] Jana Lipkova, Richard J. Chen, Bowen Chen, Ming Y. Lu, Matteo Barbieri, and Daniel Shao. 2022. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell* (2022).
[202] Alex Liu, SouYoung Jin, Cheng-I. Lai, Andrew Rouditchenko, Aude Oliva, and James Glass. 2022. Cross-modal discrete representation learning. In *ACL*. 3013–3035.
[203] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485* (2023).
[204] Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S. Rosenblum. 2019. MMKG: Multi-modal knowledge graphs. In *ESWC*. Springer, 459–474.
[205] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *ACL*. 2247–2256.

[206] Carlos Eduardo Rodrigues Lopes and Linnyer Beatrys Ruiz. 2008. On the development of a multi-tier, multimodal wireless sensor network for wild life monitoring. In *2008 1st IFIP Wireless Days*. IEEE, 1–5.
[207] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*. 13–23.
[208] Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. 2021. Pretrained transformers as universal computation engines. *arXiv preprint arXiv:2103.05247* (2021).
[209] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*.
[210] Yadong Lu, Chunyuan Li, Haotian Liu, Jianwei Yang, Jianfeng Gao, and Yelong Shen. 2023. An empirical study of scaling instruct-tuned large multimodal models. *arXiv preprint arXiv:2309.09958* (2023).
[211] Jelena Luketina, Nantas Nardelli, Gregory Farquhar, Jakob N. Foerster, Jacob Andreas, Edward Grefenstette, Shimon Whiteson, and Tim Rocktäschel. 2019. A survey of reinforcement learning informed by natural language. In *IJCAI*.
[212] Yiwei Lyu, Paul Pu Liang, Zihao Deng, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. DIME: Fine-grained interpretations of multimodal models via disentangled local explanations. *arXiv preprint arXiv:2203.02013* (2022).
[213] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. Are multimodal transformers robust to missing modality? In *CVPR*. 18177–18186.
[214] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. 2021. Smil: Multimodal learning with severely missing modality. *arXiv preprint arXiv:2103.05677* (2021).
[215] Emiliano Macaluso and Jon Driver. 2005. Multisensory spatial interactions: A window onto functional integration in the human brain. *Trends in Neurosciences* 28, 5 (2005), 264–271.
[216] T. Soni Madhulatha. 2012. An overview on clustering methods. *arXiv preprint arXiv:1205.1117* (2012).
[217] Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. 2017. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *CVPR*. 6884–6893.
[218] Jonathan Malmaud, Jonathan Huang, Vivek Rathod, Nicholas Johnston, Andrew Rabinovich, and Kevin Murphy. 2015. What's Cookin'? Interpreting cooking videos using text, speech and vision. In *NAACL-HLT*. 143–152.
[219] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2018. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *ICLR*.
[220] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *CVPR*. 11–20.
[221] Matthew Marge, Carol Espy-Wilson, Nigel G. Ward, Abeer Alwan, Yoav Artzi, Mohit Bansal, Gil Blankenship, Joyce Chai, Hal Daumé III, Debadeepta Dey, et al. 2022. Spoken language interaction with robots: Recommendations for future research. *Computer Speech & Language* 71 (2022), 101255.
[222] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. 2017. The more you know: Using knowledge graphs for image classification. In *CVPR*. IEEE, 20–28.
[223] Emily E. Marsh and Marilyn Domas White. 2003. A taxonomy of relationships between images and text. *Journal of Documentation* 59, 6 (2003), 647–672.
[224] Alessio Mazzetto, Dylan Sam, Andrew Park, Eli Upfal, and Stephen Bach. 2021. Semi-supervised aggregation of dependent weak supervision sources with performance guarantees. In *AISTATS*.
[225] Dalila Mekhaldi. 2007. Multimodal document alignment: Towards a fully-indexed multimedia archive. In *SIGIR*.
[226] Luke Melas-Kyriazi, Alexander M. Rush, and George Han. 2018. Training for diversity in image paragraph captioning. In *EMNLP*. 757–761.
[227] Luke Merrick and Ankur Taly. 2020. The explanation game: Explaining machine learning models using Shapley values. In *CD-MAKE*. Springer, 17–38.
[228] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*. Springer.
[229] George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM* 38, 11 (1995), 39–41.
[230] Dalibor Mitrović, Matthias Zeppelzauer, and Christian Breiteneder. 2010. Features for content-based audio retrieval. In *Advances in Computers*. Vol. 78. Elsevier, 71–150.
[231] Shentong Mo, Paul Pu Liang, Russ Salakhutdinov, and Louis-Philippe Morency. 2023. MultiIoT: Towards large-scale multisensory learning for the Internet of Things. arXiv:2311.06217. Retrieved from https://arxiv.org/abs/2311.06217
[232] Olivier Morere, Hanlin Goh, Antoine Veillard, Vijay Chandrasekhar, and Jie Lin. 2015. Co-regularized deep representations for video summarization. In *ICIP*. IEEE, 3165–3169.
[233] Ghulam Muhammad, Fatima Alshehri, Fakhri Karray, and Abdulmotaleb El Saddik. 2021. A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. *Information Fusion* 76, 1 (2021), 355–375.
[234] Jonathan Munro and Dima Damen. 2020. Multi-modal domain adaptation for fine-grained action recognition. In *CVPR*. 122–132.

[235] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *CVPR*. 299–307.
[236] Bence Nanay. 2018. Multimodal mental imagery. *Cortex* 105 (2018), 125–134.
[237] Milind Naphade, John R. Smith, Jelena Tesic, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis. 2006. Large-scale concept ontology for multimedia. *IEEE Multimedia* 13, 3 (2006), 86–91.
[238] Karthik Narasimhan, Regina Barzilay, and Tommi Jaakkola. 2018. Grounding language for transfer in deep reinforcement learning. *Journal of Artificial Intelligence Research* 63 (2018), 849–874.
[239] Shahla Nemati, Reza Rohani, Mohammad Ehsan Basiri, Moloud Abdar, Neil Y. Yen, and Vladimir Makarenkov. 2019. A hybrid latent space data fusion method for multimodal emotion recognition. *IEEE Access* 7 (2019), 172948–172964.
[240] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. 2022. Learning to listen: Modeling non-deterministic dyadic facial motion. In *CVPR*. 20395–20405.
[241] Nam D. Nguyen and Daifeng Wang. 2020. Multiview learning for understanding functional multiomics. *PLoS Computational Biology* 16, 4 (2020), e1007677.
[242] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2015. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE* 104, 1 (2015), 11–33.
[243] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual VQA: A cause-effect look at language bias. In *CVPR*. 12700–12710.
[244] Zeljko Obrenovic and Dusan Starcevic. 2004. Modeling multimodal human-computer interaction. *Computer* 37, 9 (2004), 65–72.
[245] Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, and Oriol Vinyals. 2018. Parallel wavenet: Fast high-fidelity speech synthesis. In *ICML*. PMLR, 3918–3926.
[246] R. OpenAI. 2023. GPT-4 technical report. arXiv:2303–08774. Retrieved from https://arxiv.org/abs/2303--08774
[247] Christian Otto, Matthias Springstein, Avishek Anand, and Ralph Ewerth. 2020. Characterization and classification of semantic image-text relations. *International Journal of Multimedia Information Retrieval* 9 (2020), 31–45.
[248] Sharon Oviatt. 1999. Ten myths of multimodal interaction. *Communications of the ACM* 42, 11 (1999), 74–81.
[249] Dinesh K. Pai. 2005. Multisensory interaction: Real and Virtual. In *Robotics Research. The Eleventh International Symposium*, Paolo Dario and Raja Chatila (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 489–498.
[250] Shruti Palaskar, Jindrich Libovickỳ, Spandana Gella, and Florian Metze. 2019. Multimodal abstractive summarization for how2 videos. *arXiv preprint arXiv:1906.07901* (2019).
[251] Maja Pantic and Leon J. M. Rothkrantz. 2003. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE* 91, 9 (2003), 1370–1390.
[252] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *CVPR*. 8779–8788.
[253] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *ECCV*. Springer, 508–524.
[254] Sarah Partan and Peter Marler. 1999. Communication goes multimodal. *Science* 283, 5406 (1999), 1272–1273.
[255] Judea Pearl. 2009. *Causality*. Cambridge University Press.
[256] Catherine Pelachaud. 2009. Modelling multimodal expression of emotion in a virtual agent. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 1535 (2009), 3539–3548.
[257] Catherine Pelachaud, Carlos Busso, and Dirk Heylen. 2021. Multimodal behavior modeling for socially interactive agents (1ed.). Association for Computing Machinery, New York, NY, USA, 259–310. https://doi.org/10.1145/3477322.3477331
[258] Alejandro Peña, Ignacio Serna, Aythami Morales, and Julian Fierrez. 2020. FairCVtest demo: Understanding bias in multimodal learning with a testbed in fair automatic recruitment. In *ICMI*. 760–761.
[259] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824* (2023).
[260] Juan-Manuel Pérez-Rúa, Valentin Vielzeuf, Stéphane Pateux, Moez Baccouche, and Frédéric Jurie. 2019. MFAS: Multimodal fusion architecture search. In *CVPR*. 6966–6975.
[261] Pouya Pezeshkpour, Liyan Chen, and Sameer Singh. 2018. Embedding multimodal relational data for knowledge base completion. In *EMNLP*. 3208–3218.
[262] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *AAAI*. Vol. 33, 6892–6899.
[263] Rosalind W. Picard. 2000. *Affective Computing*. MIT Press.
[264] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*.
[265] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* (2017).

[266] Shraman Pramanick, Aniket Roy, and Vishal M. Patel. 2022. Multimodal learning using optimal transport for sarcasm and humor detection. In *WACV*.
[267] Gorjan Radevski, Dusan Grujicic, Matthew Blaschko, Marie-Francine Moens, and Tinne Tuytelaars. 2023. Multimodal distillation for egocentric action recognition. In *ICCV*. 5213–5224.
[268] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, and Sandhini Agarwal. 2021. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 8748–8763.
[269] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. (2019).
[270] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67.
[271] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *ACL*. 2359–2369.
[272] Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrušaitis, and Roland Goecke. 2016. Extending long short-term memory for multi-view structured learning. In *ECCV*.
[273] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *ICML*. PMLR, 8821–8831.
[274] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *ACM Multimedia*. 251–260.
[275] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Giménez, Yury Sulsky, et al. 2022. *One Model to Learn Them All.* Deepmind Technical Report.
[276] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. In *NeurIPS*.
[277] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the Predictions of Any Classifier. In *KDD*. 1135–1144.
[278] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*. 10684–10695.
[279] Candace Ross, Boris Katz, and Andrei Barbu. 2020. Measuring social biases in grounded vision and language embeddings. *arXiv preprint arXiv:2002.08911* (2020).
[280] Yong Rui, Thomas S. Huang, and Shih-Fu Chang. 1999. Image retrieval: Current techniques, promising directions, and open issues. *Journal of Visual Communication and Image Representation* 10, 1 (1999), 39–62.
[281] Natalie Ruiz, Ronnie Taib, and Fang Chen. 2006. Examining the redundancy of multimodal input. In *Proceedings of the 18th Australia conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments*. 389–392.
[282] Shagan Sah, Sourabh Kulhare, Allison Gray, Subhashini Venugopalan, Emily Prud'Hommeaux, and Raymond Ptucha. 2017. Semantic text summarization of long videos. In *WACV*. IEEE, 989–997.
[283] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *AAAI*, Vol. 34. 8732–8740.
[284] Raeid Saqur and Karthik Narasimhan. 2020. Multimodal graph networks for compositional generalization in visual question answering. In *NeurIPS*. 3070–3081.
[285] Mehmet Emre Sargin, Yücel Yemez, Engin Erzin, and A. Murat Tekalp. 2007. Audiovisual synchronization and fusion using canonical correlation analysis. *IEEE Transactions on Multimedia* 9, 7 (2007), 1396–1403.
[286] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, and Jitendra Malik. 2019. Habitat: A platform for embodied ai research. In *ICCV*. 9339–9347.
[287] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE Transactions on Neural Networks* 20, 1 (2008), 61–80.
[288] Manos Schinas, Symeon Papadopoulos, Georgios Petkos, Yiannis Kompatsiaris, and Pericles A. Mitkas. 2015. Multimodal graph-based event detection and summarization in social media streams. In *ACM Multimedia*.
[289] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In *ICMI*. 400–408.
[290] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*. 618–626.
[291] Lucia Seminara, Paolo Gastaldo, Simon J. Watt, Kenneth F. Valyear, Fernando Zuher, and Fulvio Mastrogiovanni. 2019. Active haptic perception in robots: A review. *Frontiers in Neurorobotics* 13 (2019), 53.
[292] Luciano Serafini and Artur d'Avila Garcez. 2016. Logic tensor networks: Deep learning and logical reasoning from data and knowledge. *arXiv preprint arXiv:1606.04422* (2016).
[293] Rajiv Shah and Roger Zimmermann. 2017. *Multimodal Analysis of User-generated Multimedia Content*. Springer.

[294] Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27, 3 (1948), 379–423.
[295] Chhavi Sharma, William Paka, Scott, Deepesh Bhageria, Amitava Das, and Soujanya Poria. 2020. Task report: Memotion analysis 1.0 @SemEval 2020: The visuo-lingual metaphor!. In *SemEval*.
[296] Rajeev Sharma, Vladimir I. Pavlovic, and Thomas S. Huang. 1998. Toward multimodal human-computer interface. *Proc. IEEE* 86, 5 (1998), 853–869.
[297] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *EMNLP-IJCNLP*. 3398–3403.
[298] Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, and S. Yu Philip. 2016. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering* 29, 1 (2016), 17–37.
[299] Yuge Shi, Brooks Paige, and Philip Torr. 2019. Variational mixture-of-experts autoencoders for multimodal deep generative models. In *NeurIPS*.
[300] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).
[301] Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin. 2005. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of ICML Workshop on Learning with Multiple Views*. Vol. 2005, Citeseer, 74–79.
[302] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, and Qiyuan Hu. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792* (2022).
[303] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, and Guillaume Couairon. 2021. FLAVA: A foundational language and vision alignment model. *arXiv preprint arXiv:2112.04482* (2021).
[304] Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *NeurIPS*.
[305] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing* 65 (2017), 3–14.
[306] Dandan Song, Siyi Ma, Zhanchen Sun, Sicheng Yang, and Lejian Liao. 2021. KVL-BERT: Knowledge enhanced visual-and-linguistic BERT for visual commonsense reasoning. *Knowledge-Based Systems* 230 (2021), 107408.
[307] Daria Sorokina, Rich Caruana, Mirek Riedewald, and Daniel Fink. 2008. Detecting statistical interactions with additive groves of trees. In *ICML*. 1000–1007.
[308] Karthik Sridharan and Sham M. Kakade. 2008. An information theoretic framework for multi-view learning. (2008).
[309] Tejas Srinivasan and Yonatan Bisk. 2021. Worst of both worlds: Biases compound in pre-trained vision-and-language models. *arXiv preprint arXiv:2104.08666* (2021).
[310] Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, and Anyi Rao. 2022. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv preprint arXiv:2209.05481* (2022).
[311] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: A core of semantic knowledge. In *WWW*.
[312] Alane Suhr and Yoav Artzi. 2019. NLVR2 visual bias analysis. *arXiv preprint arXiv:1909.10411* (2019).
[313] Ömer Sümer, Patricia Goldberg, Sidney D'Mello, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2021. Multimodal engagement analysis from facial videos in the classroom. *IEEE Trans. on Affective Computing* 14, 2 (2021), 1012–1027.
[314] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. VideoBERT: A joint model for video and language representation learning. In *ICCV*. 7464–7473.
[315] Shiliang Sun. 2013. A survey of multi-view machine learning. *Neural Computing and Applications* 23 (2013), 2031–2038.
[316] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. MIT Press.
[317] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing Obama: Learning lip sync from audio. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 1–13.
[318] Riko Suzuki, Hitomi Yanaka, Masashi Yoshikawa, Koji Mineshima, and Daisuke Bekki. 2019. Multimodal logical inference system for visual-textual entailment. *arXiv preprint arXiv:1906.03952* (2019).
[319] Fuwen Tan, Song Feng, and Vicente Ordonez. 2019. Text2scene: Generating compositional scenes from textual descriptions. In *CVPR*. 6710–6719.
[320] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*. 5100–5111.
[321] Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *EMNLP*. 2066–2080.
[322] Zhulin Tao, Yinwei Wei, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat-Seng Chua. 2020. MGAT: Multimodal graph attention network for recommendation. *Information Processing & Management* 57, 5 (2020), 102277.
[323] Makarand Tapaswi, Martin Bauml, and Rainer Stiefelhagen. 2015. Book2movie: Aligning video scenes with book chapters. In *CVPR*. 1827–1835.

[324] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. MovieQA: Understanding stories in movies through question-answering. In *CVPR*. 4631–4640.
[325] Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond J. Mooney. 2016. Learning multi-modal grounded linguistic semantics by playing "I Spy". In *IJCAI*. 3477–3483.
[326] Bruce Thompson. 2000. Canonical correlation analysis. (2000).
[327] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*. 5238–5248.
[328] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *ECCV*. Springer, 776–794.
[329] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. 2020. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *ECCV 2020*. Springer, 436–454.
[330] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. 2018. Audio-visual event localization in unconstrained videos. In *ECCV*. 247–263.
[331] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning? In *NeurIPS*. 6827–6839.
[332] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. 2021. Contrastive learning, multi-view redundancy, and linear models. In *ALT*.
[333] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. 2017. Missing modalities imputation via cascaded residual autoencoder. In *CVPR*.
[334] Nhat C. Tran and Jean X. Gao. 2021. OpenOmics: A bioinformatics API to integrate multi-omics datasets and interface with public databases. *Journal of Open Source Software* 6, 61 (2021), 3249.
[335] George Trigeorgis, Mihalis A. Nicolaou, Björn W. Schuller, and Stefanos Zafeiriou. 2017. Deep canonical time warping for simultaneous alignment and representation learning of sequences. *IEEE TPAMI* 40, 5 (2017), 1128–1138.
[336] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *ACL*. 6558–6569.
[337] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Learning factorized multimodal representations. In *ICLR* (2019).
[338] Yao-Hung Hubert Tsai, Martin Ma, Muqiao Yang, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Multimodal routing: Improving local and global interpretability of multimodal language analysis. In *EMNLP*. 1823–1833.
[339] Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Self-supervised learning from a multi-view perspective. In *ICLR*.
[340] Michael Tsang, Dehua Cheng, and Yan Liu. 2018. Detecting statistical interactions from neural network weights. In *ICLR*.
[341] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *NeurIPS*.
[342] Matthew Turk. 2014. Multimodal interaction: A review. *Pattern Recognition Letters* 36 (2014), 189–195.
[343] Peter D. Turney and Michael L. Littman. 2005. Corpus-based learning of analogies and semantic relations. *Machine Learning* 60 (2005), 251–278.
[344] Len Unsworth and Chris Cléirigh. 2014. Multimodality and Reading: The Construction of Meaning through Image-Text Interaction. Routledge.
[345] Shagun Uppal, Sarthak Bhagat, Devamanyu Hazarika, and Navonil Majumder. 2022. Multimodal research in vision and language: A review of current and emerging trends. *Information Fusion* 77, 1 (2022), 149–171.
[346] Naushad UzZaman, Jeffrey P. Bigham, and James F. Allen. 2011. Multimodal summarization of complex sentences. In *IUI*. ACM, 43–52.
[347] Aaron Van Den Oord and Oriol Vinyals. 2017. Neural discrete representation learning. *NeurIPS* 30 (2017).
[348] Emile van Krieken, Erman Acar, and Frank van Harmelen. 2022. Analyzing differentiable fuzzy logic operators. *Artificial Intelligence* (2022).
[349] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*. 5998–6008.
[350] Ramakrishna Vedantam, Karan Desai, Stefan Lee, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2019. Probabilistic neural symbolic models for interpretable visual question answering. In *ICML*. PMLR, 6428–6437.
[351] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *ICLR*.
[352] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2015. Order-embeddings of images and language. *arXivpreprint arXiv:1511.06361* (2015).
[353] René Vidal. 2011. Subspace clustering. *IEEE Signal Processing Magazine* 28, 2 (2011), 52–68.
[354] Cédric Villani. 2009. *Optimal Transport: Old and New*. Vol. 338. Springer.

[355] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2016. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 4 (2016), 652–663.

[356] Alvin Wan, Lisa Dunlap, Daniel Ho, Jihan Yin, Scott Lee, Suzanne Petryk, Sarah Adel Bargal, and Joseph E. Gonzalez. 2020. NBDT: Neural-backed decision tree. In *ICLR*.

[357] Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. 2021. Screen2words: Automatic mobile UI summarization with multimodal learning. In *ACM UIST*. 498–510.

[358] Meng Wang, Hao Li, Dacheng Tao, Ke Lu, and Xindong Wu. 2012. Multimodal graph-based reranking for web image search. *IEEE Transactions on Image Processing* 21, 11 (2012), 4649–4661.

[359] Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou. 2020. Multimodal learning with incomplete modalities by knowledge distillation. In *KDD*. 1828–1838.

[360] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. 2020. Visual commonsense representation learning via causal inference. In *CVPR*.

[361] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2015. On deep multi-view representation learning. In *ICML*. PMLR, 1083–1092.

[362] Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard? In *CVPR*. 12695–12705.

[363] Xingbo Wang, Jianben He, Zhihua Jin, Muqiao Yang, Yong Wang, and Huamin Qu. 2021. M2Lens: Visualizing and explaining multimodal models for sentiment analysis. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 802–812.

[364] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, and Jianfeng Gao. 2019. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *CVPR*. 6629–6638.

[365] Xiaofan Wei, Huibin Li, Jian Sun, and Liming Chen. 2018. Unsupervised domain adaptation with regularized optimal transport for multimodal 2D+ 3D facial expression recognition. In *FG 2018*. IEEE, 31–37.

[366] Alex Wilf, Qianli M. Ma, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2022. Face-to-face contrastive learning for social intelligence question-answering. *arXiv preprint arXiv:2208.01036* (2022).

[367] Paul L. Williams and Randall D. Beer. 2010. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515* (2010).

[368] Eric Wong, Shibani Santurkar, and Aleksander Madry. 2021. Leveraging sparse linear layers for debuggable deep networks. In *ICML*.

[369] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, and Wynne Hsu. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*. 7623–7633.

[370] Mike Wu and Noah Goodman. 2018. Multimodal generative models for scalable weakly-supervised learning. In *NeurIPS*.

[371] Nan Wu, Stanisław Jastrzębski, Kyunghyun Cho, and Krzysztof J. Geras. 2022. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. *arXiv preprint arXiv:2202.05306* (2022).

[372] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *CVPR*. 4622–4630.

[373] Xindi Wu, Zhiwei Deng, and Olga Russakovsky. 2023. Multimodal dataset distillation for image-text retrieval. *arXivpreprint arXiv:2308.07545* (2023).

[374] Yi Xiao, Felipe Codevilla, Akhil Gurram, Onay Urfalioglu, and Antonio M. López. 2020. Multimodal end-to-end autonomous driving. *IEEE Transactions on Intelligent Transportation Systems* 23, 1 (2020), 537–547.

[375] Chen Xing, Negar Rostamzadeh, Boris Oreshkin, and Pedro O O. Pinheiro. 2019. Adaptive cross-modal few-shot learning. In *NeurIPS*.

[376] Caiming Xiong, Stephen Merity, and Richard Socher. 2016. Dynamic memory networks for visual and textual question answering. In *ICML*.

[377] Chang Xu, Dacheng Tao, and Chao Xu. 2015. Multi-view intact space learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 12 (2015), 2531–2544.

[378] Fangli Xu, Lingfei Wu, K. P. Thai, Carol Hsu, Wei Wang, and Richard Tong. 2019. MUTLA: A large-scale dataset for multimodal teaching and learning analytics. *arXiv preprint arXiv:1910.06078* (2019).

[379] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*. 2048–2057.

[380] Peng Xu, Xiatian Zhu, and David A. Clifton. 2023. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 10 (2023), 12113–12132.

[381] Zhen Xu, David R. So, and Andrew M. Dai. 2021. MUFASA: Multimodal fusion architecture search for electronic health records. *arXiv preprint arXiv:2102.02340* (2021).

[382] Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. 2022. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. *arXiv preprint arXiv:2206.06487* (2022).
[383] Zihui Xue, Sucheng Ren, Zhengqi Gao, and Hang Zhao. 2021. Multimodal knowledge expansion. In *ICCV*. 854–863.
[384] Xiaoqiang Yan, Shizhe Hu, Yiqiao Mao, Yangdong Ye, and Hui Yu. 2021. Deep multi-view learning methods: A review. *Neurocomputing* 448 (2021), 106–129.
[385] Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, and Amir Zadeh. 2021. MTAG: Modal-temporal attention graph for unaligned human multimodal language sequences. In *NAACL-HLT*.
[386] Yi Yang and Shawn Newsam. 2010. Bag-of-visual-words and spatial extensions for land-use classification. In *GIS*. 270–279.
[387] Yang Yang, Ke-Tao Wang, De-Chuan Zhan, Hui Xiong, and Yuan Jiang. 2019. Comprehensive semi-supervised multi-modal learning. In *IJCAI*.
[388] Yang Yang, De-Chuan Zhan, Xiang-Rong Sheng, and Yuan Jiang. 2018. Semi-supervised multi-modal learning with incomplete modalities. In *IJCAI*. 2998–3004.
[389] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. In *NeurIPS*. 20744–20757.
[390] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. 2019. CLEVRER: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442* (2019).
[391] Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. In *ACL*. 3025–3035.
[392] M. H. Peter Young, Alice Lai, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL* 2, 1 (2014), 67–68.
[393] Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. Vision guided generative pre-trained language models for multimodal abstractive summarization. In *EMNLP*. 3995–4007.
[394] Weijiang Yu, Jingwen Zhou, Weihao Yu, Xiaodan Liang, and Nong Xiao. 2019. Heterogeneous graph learning for visual commonsense reasoning. In *NeurIPS*.
[395] Jiahong Yuan, Mark Liberman, et al. 2008. Speaker identi?cation on the SCOTUS corpus. *Journal of the Acoustical Society of America* 123, 5 (2008), 3878.
[396] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250* (2017).
[397] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *AAAI*. Vol. 32.
[398] Amir Zadeh, Paul Pu Liang, and Louis-Philippe Morency. 2020. Foundations of multimodal co-learning. *Information Fusion* 64 (2020), 188–193.
[399] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* (2016).
[400] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *ACL*.
[401] Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. 2018. Taskonomy: Disentangling task transfer learning. In *CVPR*. 3712–3722.
[402] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *CVPR*.
[403] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. In *NeurIPS*.
[404] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, and Federico Tombari. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598* (2022).
[405] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. 2017. Leveraging video descriptions to learn video question answering. In *AAAI*.
[406] Da Zhang and Mansur Kabuka. 2019. Multimodal deep representation learning for protein interaction identification and protein family classification. *BMC Bioinformatics* 20, 16 (2019), 1–14.
[407] Hao Zhang, Zhiting Hu, Yuntian Deng, Mrinmaya Sachan, Zhicheng Yan, and Eric Xing. 2016. Learning concept taxonomies from multi-modal data. In *ACL*. 1791–1801.
[408] Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *Comput. Surveys* 56, 3 (2023), 1–37.
[409] Tong Zhang and C.-C. Jay Kuo. 2001. Audio content analysis for online audiovisual data segmentation and classification. *IEEE Transactions on Speech and Audio Processing* 9, 4 (2001), 441–457.
[410] Weifeng Zhang, Jing Yu, Hua Hu, Haiyang Hu, and Zengchang Qin. 2020. Multimodal feature fusion by relational reasoning and attention for visual question answering. *Information Fusion* 55 (2020), 116–126.

[411] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. 2019. Deep supervised cross-modal retrieval. In *CVPR*. 10394–10403.
[412] Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, and Xianyi Cheng. 2023. WebArena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854* (2023).
[413] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592* (2023).
[414] Hao Zhu, Huaibo Huang, Yi Li, Aihua Zheng, and Ran He. 2021. Arbitrary talking face generation via attentional audio-visual coherence learning. In *IJCAI*. 2362–2368.
[415] Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, and Xuwu Wang. 2022. Multi-modal knowledge graph construction and application: A survey. *arXiv preprint arXiv:2202.05786* (2022).
[416] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*.
[417] Yuke Zhu, Ce Zhang, Christopher Ré, and Li Fei-Fei. 2015. Building a large-scale multimodal knowledge base system for answering visual queries. *arXiv preprint arXiv:1507.05670* (2015).
[418] Zachary M. Ziegler, Luke Melas-Kyriazi, Sebastian Gehrmann, and Alexander M. Rush. 2019. Encoder-agnostic adaptation for conditional language generation. *arXiv preprint arXiv:1908.06938* (2019).
[419] George Kingsley Zipf. 2016. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Ravenio books.

Received 15 February 2023; revised 31 January 2024; accepted 2 April 2024