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

PAUL PU LIANG, AMIR ZADEH, and LOUIS-PHILIPPE MORENCY,
Machine Learning Department and Language Technologies Institute, Carnegie Mellon University, USA

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 paper 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. 2022. Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. *Preprint* 1, 1, Article 1 (October 2022), 36 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

It has always been a grand goal of artificial intelligence to develop computer agents with intelligent capabilities such as understanding, reasoning, and learning through multimodal experiences and data, similar to how humans perceive and interact with our world using multiple sensory modalities. With recent advances in embodied autonomous agents [37, 222], self-driving cars [295], image and video understanding [11, 243], image and video generation [210, 234], and multisensor fusion in application domain such as robotics [136, 170] and healthcare [119, 151], we are now closer than ever to intelligent agents that can integrate and learn from many sensory modalities. This

Authors' address: Paul Pu Liang, pliang@cs.cmu.edu; Amir Zadeh, abagherz@cs.cmu.edu; Louis-Philippe Morency, morency@cs.cmu.edu,
Machine Learning Department and Language Technologies Institute, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, USA, 15213.

0360-0300/2022/10-ART1
https://doi.org/XXXXXXX.XXXXXXX

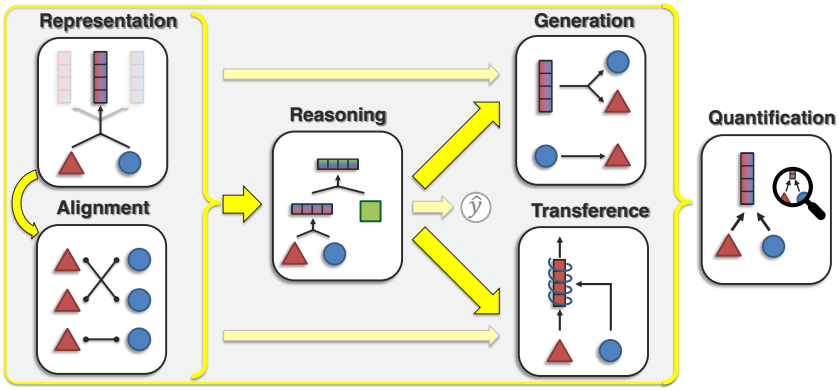

Fig. 1. Core research challenges in multimodal learning: (1) *Representation* studies how to represent and summarize multimodal data to reflect the heterogeneity and interconnections between individual modality elements. (2) *Alignment* aims to identify the connections and interactions across all elements. (3) *Reasoning* aims to compose knowledge from multimodal evidence usually through multiple inferential steps for a task. (4) *Generation* involves learning a generative process to produce raw modalities that reflect cross-modal interactions, structure, and coherence. (5) *Transference* aims to transfer knowledge between modalities and their representations. (6) *Quantification* involves empirical and theoretical studies to better understand the multimodal learning process.

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 [184], affective computing [204], robotics [127, 136], human-computer interaction [190, 228], and healthcare [40, 180].

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 multimodal research, this paper is designed to provide an overview of the methodological, computational, and theoretical foundations of multimodal machine learning, which complements recent application-oriented surveys in vision and language [269], language and reinforcement learning [161], multimedia analysis [19], and human-computer interaction [114].

To better understand the foundations of multimodal machine learning, we begin by defining (in §2) three key principles that have driven subsequent technical challenges and innovations: (1) modalities are *heterogeneous* because the information present often shows diverse qualities, structures, and representations, (2) modalities are *connected* since they are often related and share commonalities, and (3) modalities *interact* to give rise to new information when used for task inference. Building upon these definitions, we propose a new taxonomy of six core challenges in multimodal learning: *representation*, *alignment*, *reasoning*, *generation*, *transference*, and *quantification* (see Figure 1). These constitute core multimodal technical challenges that are understudied in conventional unimodal machine learning, and need to be tackled in order to progress the field forward:

(1) **Representation (§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 with the goal of keeping the same number of representations but improving 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 (§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 (§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 (§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 (§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 (§8):** The sixth and final challenge involves empirical and theoretical studies to better understand (1) the dimensions of *heterogeneity* in multimodal datasets and how they subsequently influence modeling and learning, (2) the presence and type of modality *connections and interactions* in multimodal datasets and captured by trained models, and (3) the *learning* and optimization challenges involved with heterogeneous data.

Finally, we conclude this paper with a long-term perspective in 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 check out these publicly available video recordings, additional reading materials, and discussion probes motivating open research questions in multimodal learning.

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,

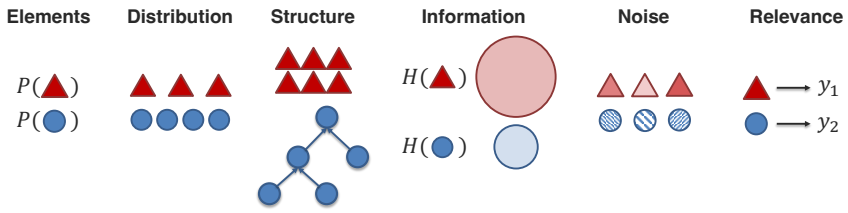

Fig. 2. The information present in different modalities will often show diverse qualities, structures, and representations. **Dimensions of heterogeneity** can be measured via differences in individual elements and their distribution, the structure of elements, as well as modality information, noise,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

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 which capture the same view modulo camera wear and tear are closer to homogeneous, two different languages which 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 [26, 147]. 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 dimensions 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 as a classic 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 [38]. 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 [227].

(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 [151]. 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 [78]. Task relevance describes how modalities can be used for inference, while context relevance describes how modalities are contextualized with other modalities.

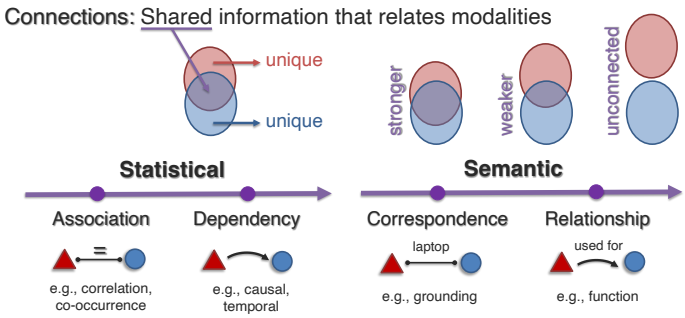

Fig. 3. **Modality connections** describe how modalities are related and share commonalities, such as correspondences between the same concept in language and images or dependencies across spatial and temporal dimensions. Connections can be studied through both statistical and semantic perspectives.

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 [38]. In the multimodal case, modeling heterogeneity is useful when learning representations and capturing alignment [314], and is a key subchallenge in quantifying multimodal models [150].

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 [290]. Modality connections describe the extent and dimensions in 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 in 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 [257].

(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 [188, 267].

(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 [192]. Identifying correspondences is fundamental in many problems related to language grounding [46], translation and retrieval [203], and cross-modal alignment [248].

(4) **Semantic relations**: Finally, semantic relations generalize semantic correspondences: instead of modality elements sharing the same exact meaning, semantic relations includes an attribute describing the exact nature of the relationship between two modality elements, such as semantic,

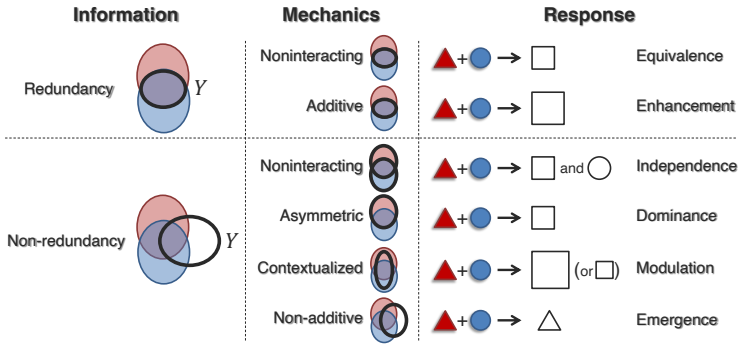

Fig. 4. **Several dimensions of modality interactions**: (1) Interaction information studies whether common redundant information or unique non-redundant information is involved in interactions; (2) interaction mechanics study the manner in which interaction occurs, and (3) interaction response studies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

logical, causal, or functional relations. Identifying these semantically related connections is important for higher-order reasoning [26, 172].

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 [290].

(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 [117], as well as from a semantic perspective where two elements interact through a logical, causal, or temporal operation [268].

(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 [197].

2.4 Core Technical Challenges

Building on these three core principles and on 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.

Table 1. This table summarizes our taxonomy of 6 core challenges in multimodal machine learning, their subchallenges, categories of corresponding approaches, and representative examples. We believe that this taxonomy can help to catalog rapid progress in this field and better identify the open research questions.

Challenge	Subchallenge	Approaches & key examples
Representation (§3)	Fusion (§3.1)	Abstract [117, 310] & raw [24, 209] fusion
	Coordination (§3.2)	Strong [75, 206] & partial [276, 319] coordination
	Fission (§3.3)	Modality-level [94, 262] & fine-grained [1, 48] fission
Alignment (§4)	Discrete connections (§4.1)	Local [60, 100] & global [142] alignment
	Continuous alignment (§4.2)	Warping [90, 103] & segmentation [243]
	Contextualization (§4.3)	Joint [140], cross-modal [93, 159] & graphical [301]
Reasoning (§5)	Structure modeling (§5.1)	Hierarchical [15], temporal [297], interactive [161] & discovery [200]
	Intermediate concepts (§5.2)	Attention [299], discrete symbols [13, 274] & language [109, 317]
	Inference paradigm (§5.3)	Logical [82, 246] & causal [4, 189, 304]
	External knowledge (§5.4)	Knowledge graphs [86, 324] & commonsense [196, 315]
Generation (§6)	Summarization (§6.1)	Extractive [52, 270] & abstractive [139, 193]
	Translation (§6.2)	Exemplar-based [122, 135] & generative [6, 115, 210]
	Creation (§6.3)	Conditional decoding [63, 191, 321]
Transference (§7)	Cross-modal transfer (§7.1)	Tuning [208, 266], multitask [150, 235] & transfer [160]
	Co-learning (§7.2)	Representation [118, 312] & generation [202, 249]
	Model Induction (§7.3)	Co-training [33, 68] & co-regularization [239, 302]
Quantification (§8)	Heterogenity (§8.1)	Importance [78, 195], bias [92, 199] & noise [163]
	Interconnections (§8.2)	Connections [3, 42, 255] & interactions [94, 149, 285]
	Learning (§8.3)	Generalization [150, 212], optimization [284, 293] & tradeoffs [151]

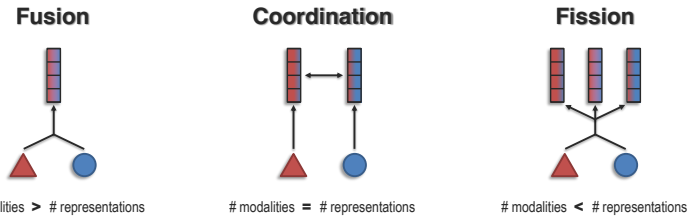

Fig. 5. Challenge 1 aims to learn **representations** that reflect cross-modal interactions between individual modality elements, through (1) *fusion*: integrating information to reduce the number of separate representations, (2) *coordination*: interchanging cross-modal information with the goal of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

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

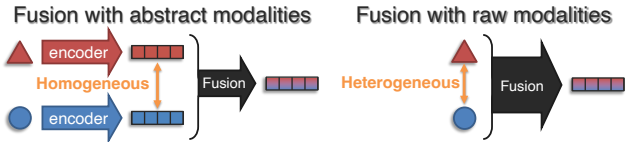

Fig. 6. We categorize **representation fusion** approaches into (1) *fusion with abstract modalities*, where unimodal encoders first capture a holistic representation of each element before fusion at relatively homogeneous representations, and (2) *fusion with raw modalities* which entails representation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

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 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 [74]. Otherwise, the additive representation $\mathbf{z}_{\text{mm}}$ can also undergo subsequent unimodal or multimodal processing [23]. Multiplicative interactions extend additive interactions to include a cross term $w_3(\mathbf{x}_1 \times \mathbf{x}_2)$. These models have been used extensively in statistics, where it can be interpreted as a *moderation* effect of $\mathbf{x}_1$ affecting the linear relationship between $\mathbf{x}_2$ and y [25]. 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 [310]. 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 [28, 76]. Tensor products of higher order represent polynomial interactions of higher order between elements [98]. 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 [98, 158]. Finally, *Multiplicative Interactions (MI)* generalize additive and multiplicative operators to include learnable parameters that capture second-order interactions [117]. 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 [47, 284]. 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 [261], fully-connected neural network layers [18, 47], or even hard gated units for sharper attention [55].

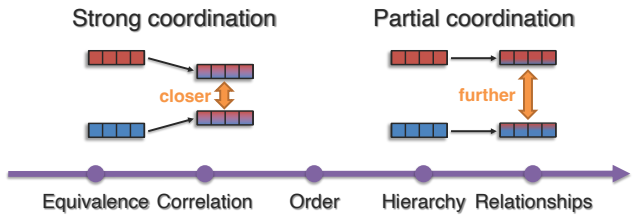

Fig. 7. There is a spectrum of **representation coordination** functions: *strong coordination* aims to enforce strong equivalence in all dimensions, whereas in *partial coordination* only certain dimensions may be coordinated to capture more general connections such as correlation, order, hierarchies,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

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 [23], 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. [24] demonstrated robustness benefits of fusion at early stages, while Gadzicki et al. [77] 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 [24, 77] and Transformers [150, 153]. However, do these complex non-additive fusion models actually learn non-additive interactions between modality elements? Not necessarily, according to Hessel and Lee [94]. We cover these fundamental analysis questions and more in the quantification challenge (§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 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) [75]. The coordination distance is typically cosine distance [174, 287] or max-margin losses [102]. Recent work has explored large-scale representation coordination by scaling up contrastive learning of image and text pairs [206], and also found that contrastive learning provably captures redundant information across the two views [256, 258] (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 [69]. For example, Socher et al. [236] maps image embeddings into word embedding spaces for zero-shot image classification. Similar ideas were used to learn coordinated representations between text, video, and audio [202], as well as between pretrained language models and image features [249].

Partial coordination: Instead of strictly capturing equivalence via strong coordination, partial coordination instead captures more general modality connections such as correlation, order, hierarchies, or relationships. To achieve these goals, partially coordinated models enforce different types of constraints on the representation space beyond semantic similarity, and perhaps only on certain dimensions of the representation.

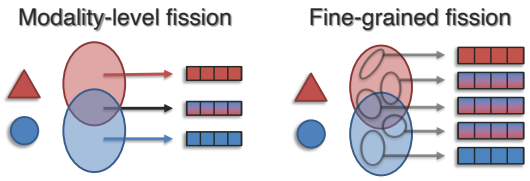

Fig. 8. **Representation fission** creates a larger set of decoupled representations that reflects knowledge about internal structure. (1) *Modality-level fission* factorizes into modality-specific information primarily in each modality, and multimodal information redundant in both modalities, while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

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 [254]. CCA models have been used extensively for cross-modal retrieval [211] audio-visual signal analysis [221], and emotion recognition [186]. To increase the expressiveness of CCA, several nonlinear extensions have been proposed including Kernel CCA [134], Deep CCA [16], and CCA Autoencoders [283].

Ordered and hierarchical spaces: Another example of representation coordination comes from order-embeddings of images and language [276], 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. [306] 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. [319] use structured representations of text and images to create multimodal concept taxonomies. Delaherche and Chetouani [61] learn coordinated representations capturing hierarchical relationships, while Alviar et al. [12] apply multiscale coordination of speech and music using partial correlation measures. Finally, Xu et al. [298] learn coordinated representations using a Cauchy loss to strengthen robustness to outliers.

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 [101, 262]. *Disentangled representation learning* aims to learn mutually independent latent variables that each explain a particular variation of the data [30, 95], and has been useful for modality-level fission by enforcing independence constraints on modality-specific and multimodal latent variables [101, 262]. Tsai et al. [262] and Hsu and Glass [101] study factorized multimodal representations and demonstrate the importance of modality-specific and multimodal factors towards generation and prediction. Shi et al. [231] study modality-level fission in multimodal variational autoencoders using a mixture-of-experts layer, while Wu and Goodman [292] 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) [94] 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

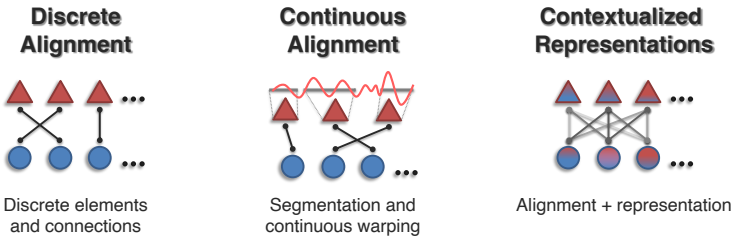

Fig. 9. **Alignment** aims to identify cross-modal connections and interactions between modality elements. Recent work has involved (1) *discrete alignment* to identify connections among discrete elements, (2) *continuous alignment* of continuous signals with ambiguous segmentation, and (3) *contextualized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

disentanglement and interpretation [176], 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 [277]. *Clustering* approaches that group data based on semantic similarity [165] have been integrated with multimodal networks for end-to-end representation fission and prediction. For example, Hu et al. [102] combine k-means clustering in representations with unsupervised audiovisual learning. Chen et al. [48] combine k-means clustering with self-supervised contrastive learning on videos. Subspace clustering [1], approximate graph Laplacians [125], conjugate mixture models [124], and dictionary learning [126] have also been integrated with multimodal models. Motivated by similar goals of representation fission, *matrix factorization* techniques have also seen several applications in multimodal prediction [10] and image retrieval [41].

4 CHALLENGE 2: ALIGNMENT

A second challenge is to identify cross-modal connections and interactions between elements of multiple modalities. For example, when analyzing the speech and gestures of a human subject, how can we align specific gestures with spoken words or utterances? Alignment between modalities is challenging since it may depend on long-range dependencies, involves ambiguous segmentation (e.g., words or utterances), and could be either one-to-one, many-to-many, or not 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 [131], visual referring expression recognition [58, 59], and cross-modal retrieval [75, 203]). 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 [23]. Several objective functions for learning aligned spaces from varying quantities of paired [43, 107] and unpaired [85]

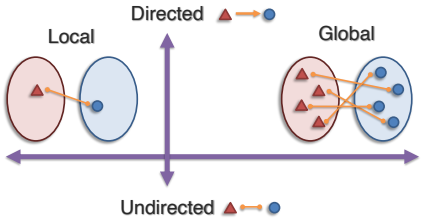

Fig. 10. **Discrete alignment** identifies connections between discrete elements, spanning (1) *local alignment* to discover connections given matching pairs, and (2) *global alignment* where alignment must be performed globally to learn both the connections and matchings between modality elements.

data have been proposed. Many of the ideas that enforce strong [75, 152] or partial [16, 276, 319] representation coordination (§3.2) are also applicable for local alignment. Several examples include aligning books with their corresponding movies/scripts [323], matching referring expressions to visual objects [169], and finding similarities between image regions and their descriptions [105]. 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 [107], sound [60, 236], and multimedia [323] 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 [278] (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 [142, 205]. To alleviate computational issues, several recent advances have integrated them with neural networks [54], approximated optimal transport with entropy regularization [288], and formulated convex relaxations for efficient learning [85].

4.2 Subchallenge 2b: Continuous Alignment

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

Continuous warping aims to align two sets of modality elements by representing them as continuous representation spaces and forming a bridge between these representation spaces. *Adversarial training* is a popular approach to warp one representation space into another. Initially used in domain adaptation [27], adversarial training learns a domain-invariant representation across domains where a domain classifier is unable to identify which domain a feature came from [8]. These ideas have been extended to align multimodal spaces [100, 103, 181]. Hsu et al. [100] use adversarial training to align images and medical reports, Hu et al. [103] design an adversarial network for cross-modal retrieval, and Munro and Damen [181] design both self-supervised alignment and adversarial alignment objectives for multimodal action recognition. *Dynamic time warping (DTW)* [133] is a related approach to segment and align multi-view time series data. DTW measures the similarity between two sequences and finds an optimal match between them by time warping (inserting frames) such that they are aligned across segmented time boundaries. For multimodal tasks, it is necessary to design similarity metrics between modalities [17, 251]. DTW was extended using CCA to map the modalities to a coordinated space, allowing for both alignment (through DTW) and coordination (through CCA) between different modality streams jointly [260].

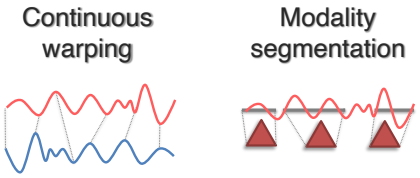

Fig. 11. **Continuous alignment** tackles the difficulty of aligning continuous signals where element segmentation is not readily available. We cover related work in (1) *continuous warping* of representation spaces and (2) *modality segmentation* of continuous signals into discrete elements at an appropriate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

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 [309]. Malmaud et al. [167] explore multimodal alignment using a factored hidden Markov model to align ASR transcripts to the ground truth. *Clustering* approaches have also been used to group continuous data based on semantic similarity [165]. Clustering-based discretization has recently emerged as an important preprocessing step for generalizing language-based pretraining (with clear word/bytepair 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 [243] perform masked pretraining on raw video and audio data. Similarly, approaches such as DALL.E [210], VQ-VAE [271], and CMCM [156] 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 [164]. 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 [47], tensor products [158, 310], and multiplicative interactions [117]. More recently, transformer models [273] have emerged as powerful encoders for sequential data by automatically aligning and capturing complementary features at different time steps. Building upon the initial text-based transformer model, multimodal transformers have been proposed that perform joint alignment using a full self-attention over modality elements concatenated across the sequence dimension (i.e., early fusion) [140, 243]. 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 [297, 318]. 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

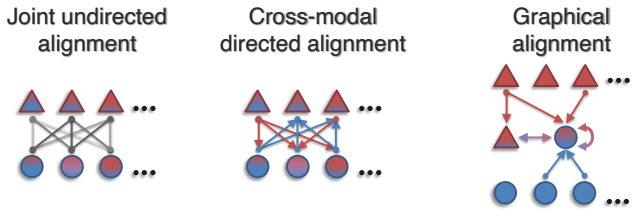

Fig. 12. **Contextualized representation** learning aims to model modality connections to learn better representations. Recent directions include (1) *joint undirected alignment* that captures undirected symmetric connections, (2) *cross-modal directed alignment* that models asymmetric connections in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

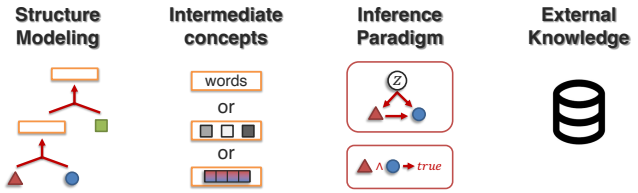

Fig. 13. **Reasoning** aims to combine knowledge, usually through multiple inferential steps, exploiting the problem structure. Reasoning involves (1) *structure modeling*: defining or learning the relationships over which reasoning occurs, (2) the *intermediate concepts* used in reasoning, (3) *inference* of increasingly abstract concepts from evidence, and (4) leveraging *external knowledge* in the study of structure, concepts, and inference.

sequence to another, before repeating in a bidirectional manner. This results in two sets of asymmetric contextualized representations to account for the possibly asymmetric connections between modalities [159, 248, 261]. These methods are useful for sequential data by automatically aligning and capturing complementary features at different time-steps [261]. Self-supervised multimodal pretraining has also emerged as an effective way to train these architectures, with the aim of learning general-purpose representations from larger-scale unlabeled multimodal data before transferring to specific downstream tasks via supervised fine-tuning [140]. These pretraining objectives typically consist of unimodal masked prediction, crossmodal masked prediction, and multimodal alignment prediction [93].

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. Solutions in this subcategory typically make use of graph neural networks [275] to recursively learn element representations contextualized with the elements in locally connected neighborhoods [223, 275]. These approaches have been applied for multimodal sequential data through MTAG [301] that captures connections in human videos, and F2F-CL [289] that additionally adds factorizes nodes along speaker turns.

5 CHALLENGE 3: REASONING

Reasoning is defined as combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and the problem structure. We categorize work towards multimodal reasoning into 4 subchallenges of structure modeling, intermediate concepts, inference paradigm, and external knowledge (Figure 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.

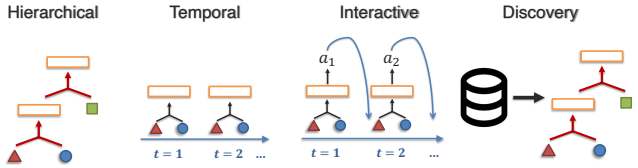

Fig. 14. **Structure modeling** aims to define the relationship over which composition occurs, which can be (1) *hierarchical* (i.e., more abstract concepts are defined as a function of less abstract ones), (2) *temporal* (i.e., organized across time), (3) *interactive* (i.e., where the state changes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

5.1 Subchallenge 3a: Structure Modeling

Structure modeling aims to capture the hierarchical relationship over which composition occurs, usually via a data structure parameterizing atoms, relations, and the reasoning process. Commonly used data structures include trees [97], graphs [308], or neural modules [15]. 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 [230], such as using language syntactic structure to guide visual modules that discover specific information in images [15, 58]. Graph-based reasoning approaches have been applied for visual commonsense reasoning [155], visual question answering [220], machine translation [305], recommendation systems [250], web image search [281], and social media analysis [224].

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. [209] explore various memory representations including multimodal fusion, coordination, and factorization. Insights from key-value memory [297] and attention-based memory [311] 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 [244]. To tackle the challenges of interactive reasoning, the growing research field of multimodal reinforcement learning (RL) has emerged from the intersection of language understanding, embodiment in the visual world, deep reinforcement learning, and robotics. We refer the reader to the extensive survey paper by Luketina et al. [161] and the position paper by Bisk et al. [32] for a full review of this field. Luketina et al. [161] separate the literature into multimodal-conditional RL (in which multimodal interaction is necessitated by the problem formulation itself, such as instruction following [47, 286]) and language-assisted RL (in which multimodal data is optionally used to facilitate learning, such as reading instruction manuals [185]).

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 [200, 300]. 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 [110]. Finally, Hu et al. [104] extend the predefined reasoning structure obtained through language parsing in Andreas et al. [15] 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 towards interpretable attention maps, discrete symbols, and language as an intermediate medium for reasoning.

Attention maps are a popular choice for intermediate concepts since they are, to a certain extent, human-interpretable, while retaining differentiability. For example, Andreas et al. [15] 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. [299] explore both soft and hard attention mechanisms for reasoning in image captioning generation. Related work has also used attention maps through dual attention architectures [182] or stacked latent attention architectures [71] for multimodal reasoning. These are typically applied for problems involving complex reasoning steps such as CLEVR [120] or VQA [320].

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 [15, 168, 274] or referring expression recognition [58]. A core challenge in this approach lies in maintaining differentiability of discrete symbols, which has been tackled via logic-based differentiable reasoning [13, 226].

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. [317] explore 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 [86, 324]. Hudson and Manning [109] 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 the way in which 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 §3.1 for a full review) are also generally applicable here, the goal is 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 [13, 226]. Many of these approaches use differentiable fuzzy logic [272] which provides a probabilistic interpretation of logical predicates, functions, and constants to ensure differentiability. These logical operators have been applied for visual question answering [82] and visual reasoning [13]. Among the greatest benefits of logical reasoning lies in its

ability to perform interpretable and compositional multi-step reasoning [111]. Logical frameworks have also been useful for visual-textual entailment [246] and geometric numerical reasoning [50], fields where logical inductive biases are crucial toward strong performance.

Causal inference extends the associational level of reasoning to interventional and counterfactual levels [198], which requires extensive knowledge of the world to imagine counterfactual effects. For example, Yi et al. [304] 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 [4] and Counterfactual VQA [189] to measure the robustness of VQA models under controlled interventions to the question as a step towards mitigating language bias in VQA models. Methods inspired by integrating causal reasoning capabilities into neural network models have also been shown to improve robustness and reduce biases [282].

5.4 Subchallenge 3d: External Knowledge

The final subchallenge studies the derivation of knowledge in the study of defining composition and structure. Knowledge is typically derived from domain knowledge on task-specific datasets. As an alternative to using domain knowledge to pre-define the compositional structure, recent work has also explored reasoning automatically using data-driven methods, such as widely accessible but more weakly supervised data outside the immediate task domain.

Multimodal knowledge graphs extend classic work in language and symbolic knowledge graphs (e.g., Freebase [35], DBpedia [20], YAGO [241], WordNet [178]) to semantic networks containing multimodal concepts as nodes and multimodal relationships as edges [322]. Multimodal knowledge graphs are important because they enable the grounding of structured information in the visual and physical world. For example, Liu et al. [157] constructs multimodal knowledge graphs containing both numerical features and images for entities. Visual Genome is another example containing dense annotations of objects, attributes, and relationships in images and text [132]. These multimodal knowledge bases have been shown to benefit visual question answering [294, 324], knowledge base completion [201], and image captioning [175, 303]. Gui et al. [86] integrates knowledge into vision-and-language transformers for automatic reasoning over both knowledge sources. We refer the reader to a comprehensive survey by Zhu et al. [322] 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 [315] and VisualCOMET [196], while other works have also introduced datasets with video and text inputs to test for temporal reasoning (e.g., MovieQA [252], MovieFIB [166], TVQA [137]). Benchmarks for multimodal commonsense typically require leveraging external knowledge from knowledge bases [237] or pretraining paradigms on large-scale datasets [159, 316].

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 [62]. We will cover recent advances as well as the evaluation of generated content.

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 [51], video [141], and audio [70, 116, 139]. Recent trends in multimodal summarization include *extractive* and *abstractive* approaches. Extractive

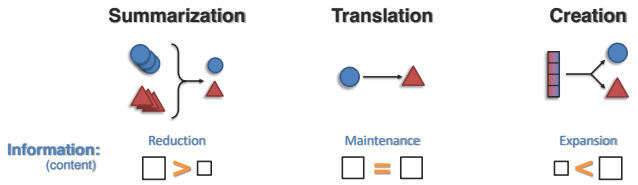

Fig. 15. How can we learn a generative process to produce raw modalities that reflect cross-modal interactions, structure, and coherence? **Generation** involves (1) *summarizing* multimodal data to highlight the most salient parts, (2) *translating* from one modality to another while being consistent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

approaches aim to filter words, phrases, and other unimodal elements from the input to create a summary [52, 116, 139]. 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 [218]. Li et al. [139] collected a dataset of news videos and articles paired with manually annotated summaries as a benchmark towards multimodal summarization. Finally, UzZaman et al. [270] 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 [51, 143]. Although most approaches only focus on generating a textual summary from multimodal data [193], several directions have also explored generating summarized images to supplement the generated textual summary [51, 141].

6.2 Subchallenge 4b: Translation

Translation aims to map one modality to another while respecting semantic connections and information content [279]. For example, generating a descriptive caption of an image can help improve the accessibility of visual content for blind people [88]. 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 [72], and *generative* models which can translate into arbitrary instances interpolating beyond the data but face challenges in quality, diversity, and evaluation [128, 210, 266]. Despite these challenges, recent progress in large-scale translation models has yielded impressive quality of generated content in text-to-image [210, 215], text-to-video [234], audio-to-image [115], text-to-speech [213], speech-to-gesture [6], speaker-to-listener [187], language to pose [7], and speech and music generation [191].

6.3 Subchallenge 4c: Creation

Creation aims to generate novel high-dimensional data (which could span text, images, audio, video, and other modalities) from small initial examples or latent conditional variables. This *conditional decoding* process is extremely challenging since it needs to be (1) conditional: preserve semantically meaningful mappings from the initial seed to a series of long-range parallel modalities, (2) synchronized: semantically coherent across modalities, (3) stochastic: capture many possible future generations given a particular state, and (4) auto-regressive across possibly long ranges. Many modalities have been considered as targets for creation. Language generation has been explored for a long time [207], and recent work has explored high-resolution speech and sound generation using neural networks [191]. Photorealistic image generation has also recently become possible due to advances in large-scale generative modeling [123]. Furthermore, there have been a number of attempts at generating abstract scenes [247], computer graphics [177], and talking heads [321]. While there has been some progress toward video generation [234], complete synchronized generation of realistic video, text, and audio remains a challenge.

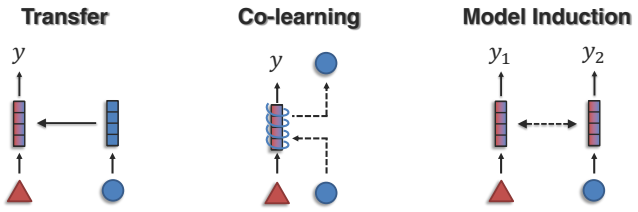

Fig. 16. **Transference** studies the transfer of knowledge between modalities, usually to help a noisy or limited primary modality, via (1) *cross-modal transfer* from models trained with abundant data in the secondary modality, (2) *multimodal co-learning* to share information across modalities by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

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 [29], hate speech [2, 79], deepfakes [89], and lip-syncing videos [245] 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 into the evaluation process [81]. Several automatic proxy metrics have been proposed [14, 53] by none are universally robust across many generation tasks.

7 CHALLENGE 5: TRANSFERENCE

Transference aims to transfer knowledge between modalities and their representations. How can knowledge learned from a secondary modality (e.g., predicted labels or representation) help a model trained on a primary modality? This challenge is particularly relevant when the primary modality has limited resources — a lack of annotated data, noisy inputs, or unreliable labels. 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 transference 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 [146] and prompt tuning [138], recent work has also studied the tuning of pretrained language models to condition on visual and other modalities. For example, Tsimpoukelli et al. [266] quickly conditions a pretrained, frozen language model on images for image captioning. Related work has also adapted prefix tuning for image captioning [49], multimodal fusion [91], and summarization [307]. 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. [325] includes additional self-attention layers between language model layers and external modalities. Rahman et al. [208] 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 [113], MultiModel [121], ViT-BERT [144], and PolyViT [153] 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) [65], images (sequence of patches) [67], video (sequence of images) [243], and other time-series data (sequence of timesteps) [154]. 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 [150], VATT [9], FLAVA [235], and Gato [212].

Transfer learning: While more research has focused on transfer within the same modality with external information [236, 296, 312], Liang et al. [152] studies transfer to new modalities using small amounts of paired but unlabeled data. Lu et al. [160] found that Transformers pretrained on language transfer to other sequential modalities as well. Liang et al. [150] builds a single multimodal model capable of transferring to completely new modalities and tasks. Recently, there has also been a line of work investigating the transfer of pretrained language models for planning [106] and interactive decision-making [145].

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 either 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 similarity space between image and text to improve image classification [75]. Marino et al. [171] use knowledge graphs for image classification via a graph-based joint representation space. Jia et al. [118] improve image classifiers with contrastive representation learning between images and noisy captions. Finally, Zadeh et al. [312] showed that implicit co-learning is also possible without explicit co-learning objectives.

Co-learning via generation instead learns a translation model from the primary to secondary modality, resulting in enriched representations of the primary modality that can predict both the label and 'hallucinate' secondary modalities containing shared information. Classic examples in this category includes language modeling by mapping contextualized text embeddings into images [249], image classification by projecting image embeddings into word embeddings [236], and language sentiment analysis by translating language into video and audio [202].

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 aim to induce behavior in both models. Model induction is exemplified by co-training, in which two learning algorithms are trained separately on each view of the data before using each algorithm's predictions to pseudo-label new unlabeled examples to enlarge the training set of the other view [33]. Therefore, information is transferred across multiple views through model predictions instead of shared representation spaces.

Multimodal co-training extends co-training by jointly learning classifiers for multiple modalities [96]. Guillaumin et al. [87] study semi-supervised learning by using 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. [56] performs semi-supervised multimodal learning using a diversity-preserving co-training algorithm. Finally, Dunnmon et al. [68] applies ideas from data programming to the problem of cross-modal weak supervision, where weak labels derived from a secondary modality (e.g., text) are used to train models over the primary modality (e.g., images).

Co-regularization: Another set of models employs a regularizer that penalizes functions from either modality that disagree with each other. This class of models, called co-regularization, is a useful technique to control model complexity by preferring hypothesis classes containing models that predict similarly across the two views [233]. Sridharan and Kakade [239] provide guarantees for these

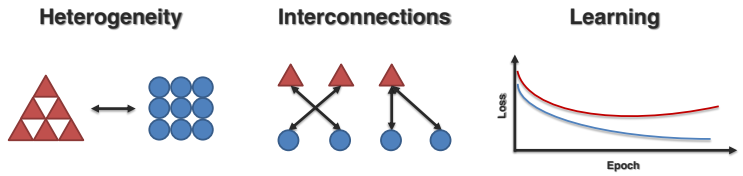

Fig. 17. **Quantification**: what are the empirical and theoretical studies we can design to better understand (1) the dimensions of *heterogeneity*, (2) the presence and type of *interconnections*, and (3) the *learning* and optimization challenges?

approaches using an information-theoretic framework. More recently, similar co-regularization approaches have also been applied for multimodal feature selection [99], semi-supervised multimodal learning [302], and video summarization [179].

8 CHALLENGE 6: QUANTIFICATION

Quantification aims to provide a deeper empirical and theoretical study of multimodal models to gain insights and improve their robustness, interpretability, and reliability in real-world applications. We break down quantification into 3 sub-challenges: (1) quantifying the *dimensions of heterogeneity* and how they subsequently influence modeling and learning, (2) quantifying the presence and type of *connections and interactions* in multimodal datasets and trained models, and (3) characterizing the *learning and optimization* challenges involved when learning from heterogeneous data (Figure 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 entire modalities and their constituents is important for determining which segment of each modality contributed to subsequent modeling. Recent work can be categorized into: (1) interpretable methods that explicitly model how each modality is used [195, 263, 313] or (2) post-hoc explanations of black-box models [45, 84]. In the former, methods such as Concept Bottleneck Models [129] and fitting sparse linear layers [291] or decision trees [280] on top of deep feature representations have emerged as promising choices. In the latter, approaches such as gradient-based visualizations [84, 225, 232]) and feature attributions (e.g., modality contribution [78], LIME [214], and Shapley values [176]) have been used to highlight regions of each modality used by the model.

Modality biases are unintended correlations between input and outputs that could be introduced during data collection [31, 36], modeling [80], or during human annotation [64]. Modality biases can lead to unexpectedly poor performance in the real world [219], or even more dangerously, potential for harm towards underrepresented groups [92, 199]. For example, Goyal et al. [83] found *unimodal biases* in the language modality of VQA tasks, resulting in mistakes due to ignoring visual information [5]. Subsequent work has developed carefully-curated diagnostic benchmarks to mitigate data collection biases, like VQA 2.0 [83], GQA [111], and NLVR2 [242]. Recent work has also found compounding *social biases* in multimodal systems [57, 216, 240] stemming from gender bias in both language and visual modalities [39, 229], which may cause danger when deployed [199].

Modality noise topologies and robustness: The study of modality noise topologies aims to benchmark and improve how multimodal models perform in the presence of real-world data imperfections. Each modality has a unique noise topology, which determines the distribution of noise and imperfections that it commonly encounters. For example, images are susceptible to blurs and shifts, typed text is susceptible to typos following keyboard positions, and multimodal time-series data is susceptible to correlated imperfections across synchronized time steps. Liang et al. [151] collect a comprehensive set of targeted noisy distributions unique to each modality. In addition

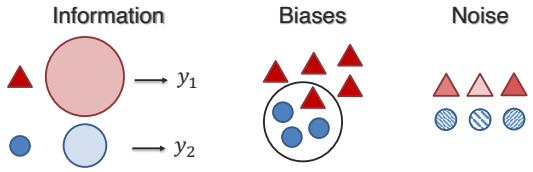

Fig. 18. The subchallenge of **heterogeneity** quantification aims to understand the dimensions of heterogeneity commonly encountered in multimodal research, such as (1) different quantities and usages of *modality information*, (2) the presence of *modality biases*, and (3) quantifying and mitigating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

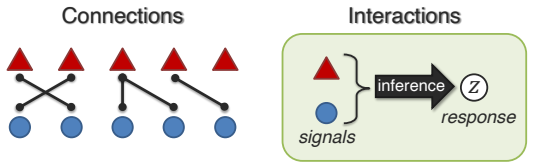

Fig. 19. Quantifying **modality interconnections** studies (1) *connections*: can we discover what modality elements are related to each other and why, and (2) *interactions*: can we understand how modality elements interact during inference?

to natural noise topologies, related work has also explored adversarial attacks [66] and distribution shifts [73] in multimodal systems. There has also been some progress in accounting for noisy or missing modalities by modality imputation using probabilistic models [163], autoencoders [259], translation models [202], or low-rank approximations [148]. However, they 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 quantification of (1) *connections*: how modalities are related and share commonality, and (2) *interactions*: how modality elements interact during inference (Figure 19).

Connections: Recent work has explored the quantification of modality connections through visualization tools on joint representation spaces [112] or attention maps [3]. Perturbation-based analysis perturbs the input and observes changes in the output to understand internal connections [149, 189]. Finally, specifically curated diagnostic datasets are also useful in understanding semantic connections: Winoground [255] probes vision and language models for visio-linguistic compositionality, and PaintSkills [57] measures the connections necessary for visual reasoning.

Interactions: One common categorization of interactions involves redundancy, uniqueness, and synergy [290]. 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 [22, 33] and contrastive learning estimators [258, 264]. Other approaches have studied these measures in isolation, such as redundancy via distance between prediction logits using either feature [173], statistical distribution tests on input features [21], or via human annotations [217]. From the semantic view, recent work in Causal VQA [4] and Counterfactual VQA [189] 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 [238, 265]. Parallel research such as EMAP [94], DIME [162], M2Lens [285], and MultiViz [149] aims to quantify the interactions in real-world multimodal datasets and models.

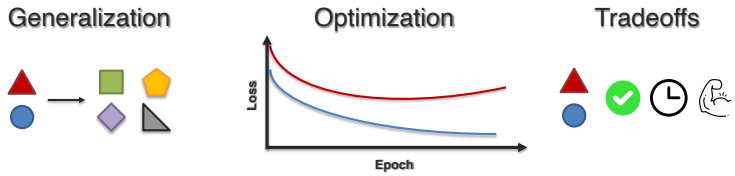

Fig. 20. Studying the multimodal **learning process** involves understanding (1) *generalization* across modalities and tasks, (2) *optimization* for balanced and efficient training, and (3) *tradeoffs* between performance, robustness, and complexity in the real-world deployment of multimodal models.

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 [108]. Recent work has studied generalization across paired modality inputs [152, 206] and in unpaired scenarios where each task is defined over only a small subset of all modalities [150, 160, 212].

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 [284]. Subsequent work has suggested both empirical and theoretical studies of why joint training of multimodal networks may be difficult and has proposed methods to improve the optimization process via weighting approaches [293].

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 [151] as well as increased susceptibility to noise and imperfection in the additional modality [202]. 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 [147, 253, 264], while information-theoretic arguments have also provided useful insights [33, 239].

9 CONCLUSION

This paper 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 in multimodal machine learning, there remain many unsolved challenges from theoretical, computational, and application perspectives:

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 the brain and human cognition, including the brain's neural architecture [34], intrinsic multimodal properties [130], mental imagery [183], and the nature of neural signals [194]. 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 several challenges and 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 in regard to 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 [255]. How can we best test for compositionality, and what reasoning approaches can enable compositional generalization?

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 [108]. 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).

Generation: *Creation and real-world ethical concerns.* The complete synchronized creation of realistic video, text, and audio remains a challenge. Furthermore, the recent success in modality generation has brought ethical concerns regarding their use. For example, large-scale pretrained language models can potentially generate text denigrating to particular social groups [229], toxic speech [79], and sensitive pretraining data [44]. 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.

Quantification: *Modality utility, tradeoffs, and selection.* How can we formalize why modalities can be useful for a task, and the potential reasons a modality can be harmful? Can we come up with formal guidelines to compare these tradeoffs and subsequently select modalities? *Explainability and interpretability.* Before models can be safely used by real-world stakeholders such as doctors, educators, or policymakers, we need to understand the taxonomy of multimodal phenomena in datasets and trained models we should aim to interpret. 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 core challenges in understanding and quantifying *modality and social biases* as well as *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.

ACKNOWLEDGEMENTS

This material is based upon work partially supported by the National Science Foundation (Awards #1722822 and #1750439), National Institutes of Health (Awards #R01MH125740, #R01MH096951, and #U01MH116925), BMW of North America, and Meta. PPL is partially supported by a Facebook PhD Fellowship and a Carnegie Mellon University's Center for Machine Learning and Health Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, National Institutes of Health, BMW of North America, Facebook, or Carnegie Mellon University's Center for Machine Learning and Health, and no official endorsement should be inferred. We are extremely grateful to Alex Wilf, Arav Agarwal, Catherine Cheng, Chaitanya Ahuja, Daniel Fried, Dong Won Lee, Jack Hessel, Leena Mathur, Lenore Blum, Manuel Blum, Martin Ma, Peter Wu, Richard Chen, Ruslan Salakhutdinov, Santiago Benoit, Su Min Park, Torsten Wortwein, Victoria Lin, Volkan Cirik, Yao-Hung Hubert Tsai, Yejin Choi, Yiwei Lyu, Yonatan Bisk, and Youssouf Kebe for helpful discussions and feedback on initial versions of this paper.

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 *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 298–306.

[3] 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.

[4] 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.

[5] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the Behavior of Visual Question Answering Models. In *EMNLP*. 1955–1960.

[6] 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.

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

[8] Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. 2014. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446* (2014).

[9] Hassan Akbari, Liangzhe Yuan, Rui Qian, et al. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *arXiv preprint arXiv:2104.11178* (2021).

[10] 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.

[11] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, et al. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198* (2022).

[12] 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.

[13] 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.

[14] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*. Springer, 382–398.

[15] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *CVPR*. 39–48.

[16] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *ICML*.

[17] Xavier Anguera, Jordi Luque, and Ciro Gracia. 2014. Audio-to-text alignment for speech recognition with very limited resources. In *Fifteenth Annual Conference of the International Speech Communication Association*.

[18] 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).

[19] 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.

[20] 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.

[21] Benjamin Auffarth, Maite López, and Jesús Cerquides. 2010. Comparison of redundancy and relevance measures for feature selection in tissue classification of CT images. In *Industrial conference on data mining*. Springer, 248–262.

[22] Maria-Florina Balcan, Avrim Blum, and Ke Yang. 2004. Co-training and expansion: Towards bridging theory and practice. *NeurIPS* 17 (2004).

[23] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE TPAMI* 41, 2 (2018), 423–443.

[24] George Barnum, Sabera J Talukder, and Yisong Yue. 2020. On the Benefits of Early Fusion in Multimodal Representation Learning. In *NeurIPS 2020 Workshop SVRHM*.

[25] 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* (1986).

[26] Roland Barthes. 1977. *Image-music-text*. Macmillan.

[27] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2006. Analysis of representations for domain adaptation. *NeurIPS* 19 (2006).

[28] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *ICCV*. 2612–2620.

[29] 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.

[30] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation Learning: A Review and New Perspectives. *TPAMI* 35, 8 (Aug. 2013).

[31] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963* (2021).

[32] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. 2020. Experience Grounds Language. In *EMNLP*. 8718–8735.

[33] Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *COLT*. 92–100.

[34] Lenore Blum and Manuel Blum. 2022. A theory of consciousness from a theoretical computer science perspective: Insights from the Conscious Turing Machine. *Proceedings of the National Academy of Sciences* (2022).

[35] 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.

[36] 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.

[37] Simon Brodeur, Ethan Perez, Ankesh Anand, Florian Golemo, Luca Celotti, et al. 2017. HoME: a Household Multimodal Environment. In *NIPS 2017's Visually-Grounded Interaction and Language Workshop*.

[38] 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).

[39] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[40] 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.

[41] Juan C Caicedo and Fabio A González. 2012. Online matrix factorization for multimodal image retrieval. In *Iberoamerican Congress on Pattern Recognition*. Springer, 340–347.

[42] 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.

[43] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Qiang Yang. 2017. Transitive hashing network for heterogeneous multimedia retrieval. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. 81–87.

[44] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, et al. 2021. Extracting training data from large language models. In *USENIX Security*. 2633–2650.

[45] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. 2018. Do explanations make VQA models more predictable to a human?. In *EMNLP*.

[46] 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.

[47] 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*.

[48] Brian Chen, Andrew Rouditchenko, Kevin Duarte, Hilde Kuehne, Samuel Thomas, Angie Boggust, et al. 2021. Multimodal clustering networks for self-supervised learning from unlabeled videos. In *ICCV*. 8012–8021.

[49] 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).

[50] 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*.

[51] Jingqiang Chen and Hai Zhuge. 2018. Abstractive Text-Image Summarization Using Multi-Modal Attentional Hierarchical RNN. In *EMNLP*.

[52] Jingqiang Chen and Hai Zhuge. 2018. Extractive Text-Image Summarization Using Multi-Modal RNN. In *2018 14th International Conference on Semantics, Knowledge and Grids (SKG)*. IEEE, 245–248.

[53] 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).

[54] 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.

[55] 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.

[56] Yanhua Cheng, Xin Zhao, Rui Cai, Zhiwei Li, Kaiqi Huang, Yong Rui, et al. 2016. Semi-Supervised Multimodal Deep Learning for RGB-D Object Recognition.. In *IJCAI*. 3345–3351.

[57] 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).

[58] Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. 2018. Using syntax to ground referring expressions in natural images. In *AAAI*, Vol. 32.

[59] Volkan Cirik, Taylor Berg-Kirkpatrick, and L-P Morency. 2020. Refer360: A Referring Expression Recognition Dataset in 360 Images. In *ACL*.

[60] Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. 2018. Visual Referring Expression Recognition: What Do Systems Actually Learn?. In *NAACL*. 781–787.

[61] Emilie Delaherche and Mohamed Chetouani. 2010. Multimodal coordination: exploring relevant features and measures. In *Proceedings of the 2nd international workshop on Social signal processing*. 47–52.

[62] 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.

[63] Emily Denton and Rob Fergus. 2018. Stochastic video generation with a learned prior. In *ICML*. PMLR, 1174–1183.

[64] 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.

[65] 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)*.

[66] 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.

[67] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR* (2021).

[68] Jared A Dunnmon, Alexander J Ratner, Khaled Saab, Nishith Khandwala, Matthew Markert, Hersh Sagreiya, Roger Goldman, et al. 2020. Cross-modal data programming enables rapid medical machine learning. *Patterns* (2020).

[69] Chris Dyer. 2014. Notes on noise contrastive estimation and negative sampling. *arXiv preprint arXiv:1410.8251* (2014).

[70] Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, et al. 2013. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia* (2013).

[71] Haoqi Fan and Jiatong Zhou. 2018. Stacked latent attention for multimodal reasoning. In *CVPR*. 1072–1080.

[72] 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.

[73] Andreas Foltyn and Jessica Deuschel. 2021. Towards Reliable Multimodal Stress Detection under Distribution Shift. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*. 329–333.

[74] Jerome H Friedman and Bogdan E Popescu. 2008. Predictive learning via rule ensembles. *The annals of applied statistics* 2, 3 (2008), 916–954.

[75] 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 *Advances in neural information processing systems*. 2121–2129.

[76] 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.

[77] Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetzsche. 2020. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*. IEEE, 1–6.

[78] Itai Gat, Idan Schwartz, and Alex Schwing. 2021. Perceptual Score: What Data Modalities Does Your Model Perceive? *NeurIPS* 34 (2021).

[79] 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.

[80] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence* (2020).

[81] 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.

[82] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Vqa-lol: Visual question answering under the lens of logic. In *European conference on computer vision*. Springer, 379–396.

[83] 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.

[84] 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).

[85] Edouard Grave, Armand Joulin, and Quentin Berthet. 2019. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 1880–1890.

[86] 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).

[87] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. 2010. Multimodal semi-supervised learning for image classification. In *2010 IEEE Computer society conference on computer vision and pattern recognition*. IEEE, 902–909.

[88] 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.

[89] Jeffrey T Hancock and Jeremy N Bailenson. 2021. The social impact of deepfakes. , 149–152 pages.

[90] 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.

[91] 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*.

[92] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 771–787.

[93] 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).

[94] 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*.

[95] 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).

[96] 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).

[97] 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 TPAMI* (2019).

[98] Ming Hou, Jiajia Tang, Jianhai Zhang, Wanzeng Kong, and Qibin Zhao. 2019. Deep multimodal multilinear fusion with high-order polynomial pooling. *NeurIPS* 32 (2019), 12136–12145.

[99] Tsung-Yu Hsieh, Yiwei Sun, Suhang Wang, and Vasant Honavar. 2019. Adaptive structural co-regularization for unsupervised multi-view feature selection. In *2019 IEEE International Conference on Big Knowledge (ICBK)*. IEEE.

[100] 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).

[101] Wei-Ning Hsu and James Glass. 2018. Disentangling by Partitioning: A Representation Learning Framework for Multimodal Sensory Data. *arXiv preprint arXiv:1805.11264* (2018).

[102] Di Hu, Feiping Nie, and Xuelong Li. 2019. Deep multimodal clustering for unsupervised audiovisual learning. In *CVPR*. 9248–9257.

[103] Peng Hu, Dezhong Peng, Xu Wang, and Yong Xiang. 2019. Multimodal adversarial network for cross-modal retrieval. *Knowledge-Based Systems* 180 (2019), 38–50.

[104] 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.

[105] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. In *CVPR*. 4555–4564.

[106] 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).

[107] Xin Huang, Yuxin Peng, and Mingkuan Yuan. 2017. Cross-modal common representation learning by hybrid transfer network. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 1893–1900.

[108] 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.

[109] Drew Hudson and Christopher D Manning. 2019. Learning by abstraction: The neural state machine. *NeurIPS* (2019).
[110] Drew A Hudson and Christopher D Manning. 2018. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067* (2018).
[111] 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.
[112] Masha Itkina, B. Ivanovic, Ransalu Senanayake, Mykel J. Kochenderfer, and Marco Pavone. 2020. Evidential Sparsification of Multimodal Latent Spaces in Conditional Variational Autoencoders. *ArXiv* abs/2010.09164 (2020).
[113] 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).
[114] Alejandro Jaimes and Nicu Sebe. 2007. Multimodal human–computer interaction: A survey. *Computer vision and image understanding* 108, 1-2 (2007), 116–134.
[115] 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.
[116] Anubhav Jangra, Adam Jatowt, Mohammad Hasanuzzaman, and Sriparna Saha. 2020. Text-Image-Video Summary Generation Using Joint Integer Linear Programming. In *European Conference on Information Retrieval*. Springer.
[117] 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*.
[118] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, et al. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*. PMLR, 4904–4916.
[119] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, et al. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.
[120] 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.
[121] Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. 2017. One model to learn them all. *arXiv preprint arXiv:1706.05137* (2017).
[122] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. *NeurIPS* 27 (2014).
[123] 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.
[124] Vasil Khalidov, Florence Forbes, and Radu Horaud. 2011. Conjugate mixture models for clustering multimodal data. *Neural Computation* (2011).
[125] Aparajita Khan and Pradipta Maji. 2019. Approximate graph Laplacians for multimodal data clustering. *IEEE TPAMI* 43, 3 (2019), 798–813.
[126] 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.
[127] Elsa A Kirchner, Stephen H Fairclough, and Frank Kirchner. 2019. Embedded multimodal interfaces in robotics: applications, future trends, and societal implications. In *The Handbook of Multimodal-Multisensor Interfaces: Language Processing, Software, Commercialization, and Emerging Directions-Volume 3*. 523–576.
[128] Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021. Text-to-image generation grounded by fine-grained user attention. In *WACV*.
[129] 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.
[130] 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.
[131] 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.
[132] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV* 123, 1 (2017), 32–73.
[133] Joseph B Kruskal. 1983. An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM review* 25, 2 (1983), 201–237.
[134] Pei Ling Lai and Colin Fyfe. 2000. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems* (2000).
[135] Rémi Lebret, Pedro Pinheiro, and Ronan Collobert. 2015. Phrase-based image captioning. In *ICML*. PMLR, 2085–2094.
[136] 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.
[137] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, Compositional Video Question Answering. In *EMNLP*. 1369–1379.

[138] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *EMNLP*. 3045–3059.
[139] 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.
[140] 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).
[141] 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).
[142] 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.
[143] Manling Li, Lingyu Zhang, Heng Ji, and Richard J Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *ACL*. 2190–2196.
[144] Qing Li, Boqing Gong, Yin Cui, Dan Kondratyuk, Xianzhi Du, et al. 2021. Towards a Unified Foundation Model: Jointly Pre-Training Transformers on Unpaired Images and Text. *arXiv preprint arXiv:2112.07074* (2021).
[145] 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).
[146] Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *ACL-IJCNLP*.
[147] Paul Pu Liang. 2022. Brainish: Formalizing A Multimodal Language for Intelligence and Consciousness. *arXiv preprint arXiv:2205.00001* (2022).
[148] 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*.
[149] Paul Pu Liang, Yiwei Lyu, Gunjan Chhablani, Nihal Jain, Zihao Deng, et al. 2023. MultiViz: Towards Visualizing and Understanding Multimodal Models. In *ICLR*.
[150] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Shengtong Mo, Dani Yogatama, et al. 2022. HighMMT: Towards Modality and Task Generalization for High-Modality Representation Learning. *arXiv preprint arXiv:2203.01311* (2022).
[151] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Yufan Chen, et al. 2021. MultiBench: Multiscale Benchmarks for Multimodal Representation Learning. In *NeurIPS Datasets and Benchmarks Track*.
[152] 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.
[153] 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.
[154] 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* (2021).
[155] Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. In *EMNLP-IJCNLP*. 2829–2839.
[156] Alex Liu, SouYoung Jin, Cheng-I Lai, Andrew Rouditchenko, Aude Oliva, and James Glass. 2022. Cross-Modal Discrete Representation Learning. In *ACL*. 3013–3035.
[157] Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S Rosenblum. 2019. MMKG: multi-modal knowledge graphs. In *European Semantic Web Conference*. Springer, 459–474.
[158] 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.
[159] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*. 13–23.
[160] Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. 2021. Pretrained transformers as universal computation engines. *arXiv preprint arXiv:2103.05247* (2021).
[161] 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*.
[162] 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).
[163] 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).
[164] 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.
[165] T Soni Madhulatha. 2012. An overview on clustering methods. *arXiv preprint arXiv:1205.1117* (2012).
[166] 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.

[167] 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.
[168] 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*.
[169] 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.
[170] Matthew Marge, Carol Espy-Wilson, Nigel G Ward, Abeer Alwan, Yoav Artzi, Mohit Bansal, et al. 2022. Spoken language interaction with robots: Recommendations for future research. *Computer Speech & Language* (2022).
[171] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. 2017. The More You Know: Using Knowledge Graphs for Image Classification. In *CVPR*. IEEE, 20–28.
[172] Emily E Marsh and Marilyn Domas White. 2003. A taxonomy of relationships between images and text. *Journal of documentation* (2003).
[173] 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*.
[174] Dalila Mekhaldi. 2007. Multimodal document alignment: towards a fully-indexed multimedia archive. In *Proceedings of the Multimedia Information Retrieval Workshop, SIGIR, Amsterdam, the Netherlands*.
[175] Luke Melas-Kyriazi, Alexander M Rush, and George Han. 2018. Training for diversity in image paragraph captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 757–761.
[176] Luke Merrick and Ankur Taly. 2020. The explanation game: Explaining machine learning models using shapley values. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 17–38.
[177] 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 *European conference on computer vision*. Springer.
[178] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.
[179] Olivier Morere, Hanlin Goh, Antoine Veillard, Vijay Chandrasekhar, and Jie Lin. 2015. Co-regularized deep representations for video summarization. In *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 3165–3169.
[180] Ghulam Muhammad, Fatima Alshehri, Fakhri Karray, Abdulmotaleb El Saddik, et al. 2021. A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. *Information Fusion* 76 (2021), 355–375.
[181] Jonathan Munro and Dima Damen. 2020. Multi-modal domain adaptation for fine-grained action recognition. In *CVPR*. 122–132.
[182] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *CVPR*. 299–307.
[183] Bence Nanay. 2018. Multimodal mental imagery. *Cortex* 105 (2018), 125–134.
[184] 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.
[185] 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.
[186] 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.
[187] 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.
[188] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2015. A review of relational machine learning for knowledge graphs. *Proc. IEEE* 104, 1 (2015), 11–33.
[189] 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.
[190] Zeljko Obrenovic and Dusan Starcevic. 2004. Modeling multimodal human-computer interaction. *Computer* (2004).
[191] Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, et al. 2018. Parallel wavenet: Fast high-fidelity speech synthesis. In *ICML*. PMLR, 3918–3926.
[192] 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.
[193] Shruti Palaskar, Jindrich Libovickỳ, Spandana Gella, and Florian Metze. 2019. Multimodal abstractive summarization for how2 videos. *arXiv preprint arXiv:1906.07901* (2019).
[194] Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, Joseph Schmidt, and Mubarak Shah. 2020. Decoding brain representations by multimodal learning of neural activity and visual features. *IEEE TPAMI* (2020).
[195] 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.
[196] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *European Conference on Computer Vision*. Springer, 508–524.

[197] Sarah Partan and Peter Marler. 1999. Communication goes multimodal. *Science* 283, 5406 (1999), 1272–1273.
[198] Judea Pearl. 2009. *Causality*. Cambridge university press.
[199] 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.
[200] Juan-Manuel Pérez-Rúa, Valentin Vielzeuf, Stéphane Pateux, Moez Baccouche, and Frédéric Jurie. 2019. Mfas: Multimodal fusion architecture search. In *CVPR*. 6966–6975.
[201] Pouya Pezeshkpour, Liyan Chen, and Sameer Singh. 2018. Embedding Multimodal Relational Data for Knowledge Base Completion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 3208–3218.
[202] 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.
[203] 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*.
[204] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* (2017).
[205] Shraman Pramanick, Aniket Roy, and Vishal M Patel. 2022. Multimodal Learning using Optimal Transport for Sarcasm and Humor Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.
[206] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 8748–8763.
[207] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).
[208] 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.
[209] Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrušaitis, and Roland Goecke. 2016. Extending long short-term memory for multi-view structured learning. In *European Conference on Computer Vision*.
[210] 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.
[211] 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 *ACMMM*. 251–260.
[212] 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* (2022).
[213] 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. *NeurIPS* 32 (2019).
[214] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *KDD*. 1135–1144.
[215] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*. 10684–10695.
[216] Candace Ross, Boris Katz, and Andrei Barbu. 2020. Measuring Social Biases in Grounded Vision and Language Embeddings. *arXiv preprint arXiv:2002.08911* (2020).
[217] 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.
[218] 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.
[219] 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.
[220] Raeid Saqur and Karthik Narasimhan. 2020. Multimodal graph networks for compositional generalization in visual question answering. *NeurIPS* 33 (2020), 3070–3081.
[221] 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.
[222] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. 2019. Habitat: A platform for embodied ai research. In *ICCV*. 9339–9347.
[223] 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.
[224] 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*.
[225] 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.

[226] 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).
[227] Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal* (1948).
[228] Rajeev Sharma, Vladimir I Pavlović, and Thomas S Huang. 2002. Toward multimodal human–computer interface. In *Advances in image processing and understanding: A Festschrift for Thomas S Huang*. World Scientific, 349–365.
[229] 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.
[230] 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.
[231] Yuge Shi, Brooks Paige, and Philip Torr. 2019. Variational mixture-of-experts autoencoders for multimodal deep generative models. *NeurIPS* (2019).
[232] 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).
[233] 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.
[234] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792* (2022).
[235] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, et al. 2021. FLAVA: A Foundational Language And Vision Alignment Model. *arXiv preprint arXiv:2112.04482* (2021).
[236] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. *NeurIPS* (2013).
[237] 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.
[238] Daria Sorokina, Rich Caruana, Mirek Riedewald, and Daniel Fink. 2008. Detecting statistical interactions with additive groves of trees. In *Proceedings of the 25th international conference on Machine learning*. 1000–1007.
[239] Karthik Sridharan and Sham M Kakade. 2008. An information theoretic framework for multi-view learning. (2008).
[240] 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).
[241] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *WWW*.
[242] Alane Suhr and Yoav Artzi. 2019. Nlvr2 visual bias analysis. *arXiv preprint arXiv:1909.10411* (2019).
[243] 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.
[244] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.
[245] 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.
[246] 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).
[247] Fuwen Tan, Song Feng, and Vicente Ordonez. 2019. Text2scene: Generating compositional scenes from textual descriptions. In *CVPR*. 6710–6719.
[248] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP-IJCNLP*. 5100–5111.
[249] Hao Tan and Mohit Bansal. 2020. Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision. In *EMNLP*. 2066–2080.
[250] 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.
[251] Makarand Tapaswi, Martin Bauml, and Rainer Stiefelhagen. 2015. Book2movie: Aligning video scenes with book chapters. In *CVPR*. 1827–1835.
[252] 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.
[253] 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.
[254] Bruce Thompson. 2000. Canonical correlation analysis. (2000).
[255] 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.
[256] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *ECCV*. Springer, 776–794.
[257] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning? *NeurIPS* 33 (2020), 6827–6839.

[258] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. 2021. Contrastive learning, multi-view redundancy, and linear models. In *ALT*.
[259] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. 2017. Missing Modalities Imputation via Cascaded Residual Autoencoder. In *CVPR*.
[260] 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.
[261] 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.
[262] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Learning factorized multimodal representations. *ICLR* (2019).
[263] 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.
[264] Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Self-supervised Learning from a Multi-view Perspective. In *ICLR*.
[265] Michael Tsang, Dehua Cheng, and Yan Liu. 2018. Detecting Statistical Interactions from Neural Network Weights. In *ICLR*.
[266] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *NeurIPS* 34 (2021).
[267] Peter D Turney and Michael L Littman. 2005. Corpus-based learning of analogies and semantic relations. *Machine Learning* 60 (2005), 251–278.
[268] Len Unsworth and Chris Cléirigh. 2014. Multimodality and reading: The construction of meaning through image-text interaction. Routledge.
[269] Shagun Uppal, Sarthak Bhagat, Devamanyu Hazarika, Navonil Majumder, et al. 2022. Multimodal research in vision and language: A review of current and emerging trends. *Information Fusion* 77 (2022), 149–171.
[270] Naushad UzZaman, Jeffrey P Bigham, and James F Allen. 2011. Multimodal summarization of complex sentences. In *Proceedings of the 16th international conference on Intelligent user interfaces*. ACM, 43–52.
[271] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *NeurIPS* 30 (2017).
[272] Emile van Krieken, Erman Acar, and Frank van Harmelen. 2022. Analyzing differentiable fuzzy logic operators. *Artificial Intelligence* (2022).
[273] 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 *Advances in neural information processing systems*. 5998–6008.
[274] 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.
[275] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *ICLR*.
[276] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2015. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361* (2015).
[277] René Vidal. 2011. Subspace clustering. *IEEE Signal Processing Magazine* 28, 2 (2011), 52–68.
[278] Cédric Villani. 2009. *Optimal transport: old and new*. Vol. 338. Springer.
[279] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2016. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE TPAMI* (2016).
[280] 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*.
[281] 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.
[282] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. 2020. Visual commonsense representation learning via causal inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.
[283] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2015. On deep multi-view representation learning. In *ICML*. PMLR, 1083–1092.
[284] Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What Makes Training Multi-Modal Classification Networks Hard?. In *CVPR*. 12695–12705.
[285] Xingbo Wang, Jianben He, Zhihua Jin, et al. 2021. M2Lens: Visualizing and explaining multimodal models for sentiment analysis. *IEEE Transactions on Visualization and Computer Graphics* (2021).
[286] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, et al. 2019. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *CVPR*. 6629–6638.
[287] Jônatas Wehrmann, Anderson Mattjie, and Rodrigo C Barros. 2018. Order embeddings and character-level convolutions for multimodal alignment. *Pattern Recognition Letters* 102 (2018), 15–22.

[288] 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.
[289] 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).
[290] Paul L Williams and Randall D Beer. 2010. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515* (2010).
[291] Eric Wong, Shibani Santurkar, and Aleksander Madry. 2021. Leveraging sparse linear layers for debuggable deep networks. In *ICML*.
[292] Mike Wu and Noah Goodman. 2018. Multimodal generative models for scalable weakly-supervised learning. *NeurIPS* 31 (2018).
[293] 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).
[294] 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.
[295] 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* (2020).
[296] Chen Xing, Negar Rostamzadeh, Boris Oreshkin, and Pedro O O. Pinheiro. 2019. Adaptive Cross-Modal Few-shot Learning. In *NeurIPS*.
[297] Caiming Xiong, Stephen Merity, and Richard Socher. 2016. Dynamic memory networks for visual and textual question answering. In *ICML*.
[298] Chang Xu, Dacheng Tao, and Chao Xu. 2015. Multi-view intact space learning. *IEEE TPAMI* (2015).
[299] 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.
[300] 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).
[301] Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, et al. 2021. MTAG: Modal-Temporal Attention Graph for Unaligned Human Multimodal Language Sequences. In *NAACL-HLT*.
[302] Yang Yang, Ke-Tao Wang, De-Chuan Zhan, Hui Xiong, and Yuan Jiang. 2019. Comprehensive Semi-Supervised Multi-Modal Learning.. In *IJCAI*.
[303] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *ECCV*.
[304] 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).
[305] 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.
[306] 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 (2014), 67–68.
[307] 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.
[308] Weijiang Yu, Jingwen Zhou, Weihao Yu, Xiaodan Liang, and Nong Xiao. 2019. Heterogeneous Graph Learning for Visual Commonsense Reasoning. In *NeurIPS*.
[309] Jiahong Yuan, Mark Liberman, et al. 2008. Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America* (2008).
[310] 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).
[311] 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.
[312] Amir Zadeh, Paul Pu Liang, and Louis-Philippe Morency. 2020. Foundations of multimodal co-learning. *Information Fusion* 64 (2020), 188–193.
[313] 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*.
[314] 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.
[315] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From Recognition to Cognition: Visual Commonsense Reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[316] 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. *NeurIPS* 34 (2021).

[317] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598* (2022).
[318] 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*.
[319] 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.
[320] 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.
[321] 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.
[322] Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, et al. 2022. Multi-Modal Knowledge Graph Construction and Application: A Survey. *arXiv preprint arXiv:2202.05786* (2022).
[323] 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*.
[324] 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).
[325] 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).