RECENT ADVANCES IN AUDIO-VISUAL-LANGUAGE MODELING

Kairui Zhang
Intelligent Systems Laboratory
University of Bristol
pu22650@bristol.ac.uk

Zahraa S. Abdallah
Intelligent Systems Laboratory
University of Bristol
zahraa.abdallah@bristol.ac.uk

Martha Lewis
University of Amsterdam
m.a.f.lewis@uva.nl

ABSTRACT

Multimodal data plays a crucial role in our daily lives. It comprises heterogeneous data, such as audio, visual, language, etc., that surrounds us in the world. Machine learning for single modality data can result in limitations to performance owing to a lack of information, while many applications are more suitable to be modeled as multimodal learning tasks. Audiovisual learning and vision-language modeling have been extensively studied, and recent research is starting to focus on audio-visual-language joint modeling. Currently, there is no review focused on audio-visual-language joint modeling, even though audio, visual, and language modalities frequently co-occur. Therefore, we survey audio, visual, and language trimodal learning to understand the ability of multimodal learners. Problem formulations and benchmark datasets are introduced. We summarize the state-of-the-art methods for each task with corresponding evaluation criteria, and present current trends of multitask learning, larger model size, as well as pretrain-finetune training paradigm.

***K*eywords** Deep Learning · Multimodal Learning · Audio-Visual-Language Modeling

1 Introduction

Multimodal machine learning has appeared as a crucial paradigm for modeling real-world scenarios that inherently contain heterogeneous and interconnected signals across multiple modalities [1]. Applications, ranging from technology to healthcare, as depicted in Figure 1c, inherently interact with various types of information. These applications require an understanding of multiple modalities, especially audio, visual, and language data, which provides rich semantic contexts for more accurate perception, reasoning, and decision-making across diverse tasks.

Although there is no review specifically on the topic of trimodal modeling, extensive reviews have been published on the general topic of multimodal learning, driven by the rapid growth of the field.

Overly General Reviews Some reviews primarily emphasize general methodologies, with limited discussion of specific downstream tasks. For instance, Liang et al. review six technical challenges, including representation, alignment, reasoning, generation, transference, and quantification [1]. In practice, however, some of these aspects are deeply intertwined, especially when using deep learning methods, making them difficult to treat separately. Additionally, some literature includes an excessive number of modalities. For example, Jabeen et al. cover various physical signals and heterogeneous data types [2]. Several signals lack semantic information and exist in lower dimensions. Integrating such modalities into a unified framework remains challenging. A common approach involves hard-coded systems or separate models within an engineering pipeline, limiting their generalizability. An overly general survey might not truly bring us specific insights on real-world applications with a requirement for semantic-rich information. Despite Jiao et al. providing a comprehensive discussion on multimodal fusion, it does not address engineering challenges in real-world scenarios, such as lightweight fusion on edge devices with computational constraints and cross-modal synchronization issues [3].

Overly Specific Reviews Other reviews were limited to a specific task. For instance, Zhu et al. primarily analyze visual modalities while treating others as auxiliary [4]. Sura reviews multimodal clustering [5]. Similarly, Yu et al. focus solely on continual learning [6]. Furthermore, some reviews consider only two modalities, which limits their

applicability. For instance, Ahmed explores audiovisual learning without considering language [7]. Liu et al. focus on vision-language applications in medicine [8], and Lu reviews question-answering tasks [9]. Both neglect the role of audio information, which is prevalent in real-world scenarios and provides essential semantic information.

Pre-Transformer Reviews Review papers are time-dependent. In the past five years, Transformer models have been widely adopted in various modalities beyond text, such as vision [10] and audio [11], allowing the construction of models under a unified framework. In the last two years, large models have demonstrated remarkable performance in language tasks and advances in reasoning. On the other hand, they require increased computational resources. Multimodal large models have emerged as a new research direction, meaning that older reviews do not capture these recent developments [2, 12, 13, 14, 15].

Since vision-language learning and audiovisual learning have been extensively researched, and these modalities are frequently found together, it is natural to study audio-visual-language modeling jointly. Given these limitations in existing reviews, a novel survey is needed to address the recent advancements, integrate audio-visual-language triple modalities effectively, and provide insights into various downstream tasks with new challenges. Thus, in this survey, we concentrate on audio-visual-language modeling approaches within the range from 2020 to 2025, covering methods, trends, and challenges in terms of feature extraction, fusion, tasks, datasets, and models.

Through organizing and analyzing landmark research and recent literature, we answer the following questions:
Q1 (Tasks): What representative downstream tasks motivated by practical demands are used to evaluate joint audio-visual-language models across understanding and generation?
Q2 (Data): What benchmark datasets and evaluation metrics are commonly used for these tasks, and how suitable are they for joint audio-visual-language modeling?
Q3 (Methods): What state-of-the-art approaches have been proposed for joint audio-visual-language modeling? How do they compare in terms of performance, efficiency, and modality integration?
Q4 (Trends): What are the recent trends in model structures and training paradigms for joint audio-visual-language learning?
Q5 (Challenges): What are the major obstacles to understanding and achieving practically deployable joint modeling across audio, visual, and language modalities in real-world environments?

To this end, we conduct a review of recent research in audio-visual-language joint modeling. We collect and analyze research from 2020 to 2025, categorizing the keywords by their statistics. As shown in Figure 1, within multimodal learning studies involving audio, visual, and language modalities, the proportions of papers for each modality are roughly comparable. Two core techniques underpinning multimodal learning are alignment and fusion. Alignment establishes correspondences across modalities, such as feature-space alignment, semantic alignment, or temporal alignment, whereas fusion integrates information from multiple modalities to produce more accurate representations or outputs. Among these, fusion has been discussed more extensively in the literature. We also analyze application studies and find that multimodal learning applications are distributed across various industries. Meanwhile, the benchmark tasks used in research are quite diverse, with Multimedia Event Recognition (MER), Cross-Modal Retrieval (CMR), and Emotion Learning (EL) being the most extensively studied. These observations provide a broad context for our subsequent survey of representative tasks, benchmarks, and methodologies.

The selection of studies for this review followed systematic inclusion and exclusion criteria for audio-visual-language modeling research published between 2020 and 2025. Included papers were required to present deep learning methodologies utilizing all three modalities (audio, visual, and textual) as inputs or outputs, published in top-tier venues such as CVPR, ICCV, ECCV, NeurIPS, ICML, ICLR, ACL, EMNLP, ICASSP, INTERSPEECH, ACM MM, and leading AI journals. Studies must demonstrate empirical contributions with quantitative evaluation across established multimodal tasks such as action recognition, emotion recognition, question answering, cross-modal retrieval, localization, captioning, and generation. Only English-language peer-reviewed articles were considered. Exclusion criteria eliminated purely theoretical work without experimental validation, and studies using fewer than three modalities. The selection process employed iterative expansion from seminal works through systematic citation tracking to ensure comprehensive coverage of the field.

Our analysis reveals several key trends: audio-visual-language modeling has benefited from multitask learning and multi-stage training [16, 17, 18]; The integration of large language models into audio-visual-language reasoning has gained significant traction. Pretraining and finetuning have emerged as prevailing paradigms in recent research.

Our key contributions are summarized as follows:

- We present a comprehensive and up-to-date survey on audio-visual-language multimodal tasks and representative models, filling a gap in prior work.
- We discuss several trends from bimodal to trimodal modeling.

(a) **Modalities**.

(b) **Multimodal core techniques**.

(c) **Applications**.

(d) **Tasks**.

Figure 1: **Statistical analysis of recent literatures.** The figures show the frequency of the papers that are collected by audio, visual, and language multimodal learning over the past 5 years. In Figure 1c, from top to bottom: Advanced industries; Business, legal, and professional services; Consumer goods and retail, Energy and materials, Financial services; Health care, pharma, and medical products; Media and telecom, Technology. In Figure 1d, from top to bottom: Multimodal Action Recognition, Multimodal Emotion Recognition, Question Answering, Cross-Modal Retrieval, Object Localization, Event Localization, Speech Recognition, Video Captioning, Video Generation, Audio Generation.

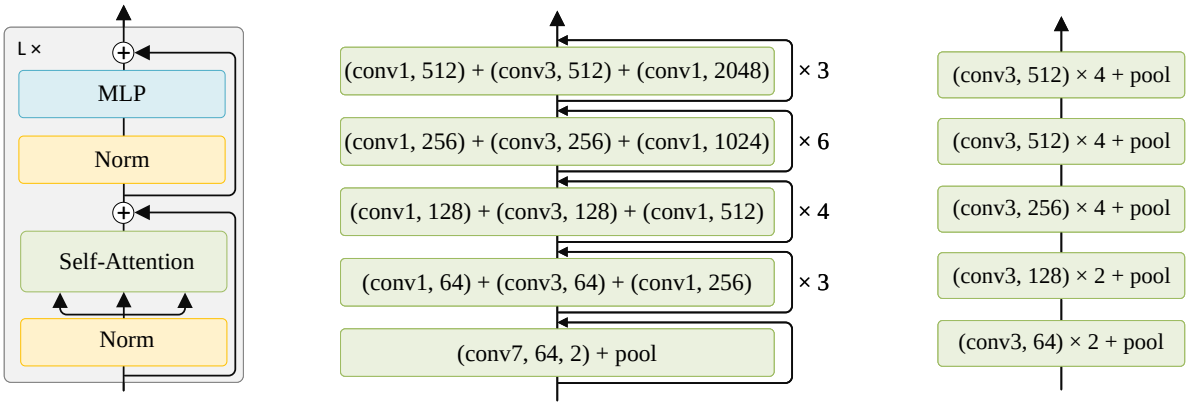
Table 1: Encoders in recent audio-visual-language models.

Year	Method	Audio	Visual	Language
2024	Amuse [19]	^{T}HTS [20]	^{T}Swin Transformer-V2 [21]	^{T}Transformer [22]
2024	TSPM [23]	^{C}VGGish [24]	^{T}CLIP [25]	^{T}CLIP [25]
2024	TeSO [26]	^{C}VGGish [24]	^{T}Swin Transformer [27]	^{T}ImageBind [28]
2024	TFGCN [29]	^{C}Wav2Vec [30]	^{T}CLIP [25]	GloVe [31]
2024	OmniVec [32]	^{T}AST [33]	^{T}ViT [10]+^{T}ViViT [34]	^{T}BERT [35]
2023	CORECT [36]	FC	FC	^{T}Transformer [22]
2023	VALOR [37]	^{C}VGGish [24]	^{C}ResNet [38]+^{C}R(2+1)D [39]	^{T}CLIP [25]+^{T}CLAP [40]
2023	ImageBind [28]	^{T}AST [33]	^{T}CLIP [25]	^{T}CLIP [25]

FC = Fully Connected Layer; HTS = HTS-Audio Transformer; AST = Audio Spectrogram Transformer;
^{T}: architecture based on Transformer; ^{C}: architecture based on CNN;
Certain models, such as ImageBind, are composed of multiple encoders. When referred to in the specific modality column, it means using the corresponding modality encoder (e.g., TeSO uses ImageBind language stream as the language encoder).

- We present key challenges such as limited long-term interactive capabilities, high computational demands, and lack of interpretability in deep models, and future research directions.

The rest of this paper is organized as follows. We review feature extraction in Section 2 and feature fusion in Section 3. Section 4 presents diverse multimodal tasks. We discuss current trends in Section 5, summarize challenges and future directions in Section 6, and conclude in Section 7.



(a) **Transformer**. (b) **Resnet50**. (c) **VGGish**.

Figure 2: **Common encoders used for feature extraction**. Figure 2b and Figure 2c remove the final several fully connected layers, as these layers are often used for classification.

2 Feature Extraction

Recent advances in audio-visual-language modeling hinge on deep learning's ability to extract rich vector representations from raw multimodal inputs. We survey recent literature and summarize audio, visual, and language encoders that produce these vector representations in Table 1.

2.1 Language Representation

Language can be represented as a sequence of discrete tokens of length L, denoted as $\mathbf{x}^L = [x_1, x_2, \ldots, x_L]$, where each token corresponds to a word or subword. Unlike visual or audio signals, these tokens are symbolic and lack intrinsic meaning. Their semantics arise from context, which enables models to learn vector representations, known as embeddings, that capture their contextual meaning. In this paper, features, latent vectors, and embeddings are used interchangeably.

GloVe GloVe is a classic method for learning word embeddings based on global word co-occurrence statistics [31]. It constructs a word-word co-occurrence matrix from a large corpus, where each entry captures how frequently two words appear together within a given context window. GloVe then optimizes word vectors so that their dot product approximates the logarithm of their co-occurrence count. The result is a static embedding for each word: its vector representation is fixed regardless of context.

Transformer-Based Encoder The Transformer architecture was originally introduced to capture long-range dependencies in machine translation through a global self-attention mechanism [22]. Building on this foundation, BERT adapted the Transformer encoder by introducing a special `[CLS]` token to aggregate global sequence information for classification tasks [35]. It also proposed a pivotal pretraining objective—masked language modeling—which we discuss in detail in Section 5.3. Notably, both the original Transformer encoder and BERT are pretrained exclusively on language modality. A commonly used structure is shown in Figure 2a.

Multimodal Encoder CLIP [25] and CLAP [40] adopt dual-encoder contrastive learning frameworks to align language with vision and audio, respectively. CLIP uses a Vision Transformer (ViT) image encoder and a BERT-based text encoder, while CLAP follows a similar paradigm for audio–text alignment. In both models, the language encoder maps text inputs into a shared multimodal embedding space, aligned with image or audio features to enable effective multimodal understanding. ImageBind [28] extends this approach by aligning additional modalities, using AST for audio and directly adopting CLIP's pretrained image and text encoders for visual and textual representations. Notably, TeSO [26] uses the text encoder from ImageBind for language encoding (See Method TeSO in Table 1), though it is structurally identical to CLIP's text encoder.

2.2 Vision Representation

Vision plays a crucial role in human perception [41]. Visual information is often represented computationally in the form of RGB images or videos. An RGB image is denoted as $\mathbf{x}_V \in \mathbb{R}^{C\times H\times W}$, where $C = 3$ represents the number of channels, H and W denote the image height and width of the image, respectively. Videos, as a common visual modality, are represented as sequences of RGB images ordered in time. We categorize visual encoders into CNN-based and Transformer-based, and multimodal encoder architectures. As a multimodal encoder, the CLIP visual encoder is also commonly used for visual feature extraction, as seen in methods such as TSPM, TFGCN, and ImageBind in Table 1. Since the CLIP model is revisited in the context of multimodal encoders for language feature extraction, we do not repeat its details here.

CNN-Based Encoder Convolutional Neural Networks (CNNs) have long been the state-of-the-art in visual tasks. Their core component, the convolutional layer, introduces an inductive bias that effectively captures local spatial patterns. Since visual information is predominantly localized in images, CNNs are naturally well-suited for grid-like data structures such as images. In contrast to Transformer-based models, CNNs typically reduce the feature dimensionality progressively, making them strong candidates for lightweight model design. Numerous architectural variations have been proposed to improve CNN performance across diverse application needs [42, 43]. ResNet introduced the residual block, enabling the training of deeper CNNs capable of learning richer image representations. It has since become the default backbone in CNN-based vision models [38]. For video tasks, R(1+2)D models are a popular CNN-based choice [39]. To better exploit both temporal and spatial cues in video data, some approaches combine features extracted from both ResNet and R(2+1)D networks (e.g., VALOR in Table 1). A commonly used structure, Resnet50, is shown in Figure 2b.

Vision Transformer-Based Encoder Transformers have demonstrated exceptional performance in various natural language processing tasks, motivating their application in computer vision. As noted earlier, Transformers are inherently designed for sequential token processing. To adapt them for images without altering the architecture, ViT [10] partitions images into patches and encodes each patch similarly to language token embeddings. Since images contain more complex spatial information than text, extending ViTs beyond image classification requires finer-grained modeling. Swin Transformer addresses this by enabling hierarchical feature extraction and local attention [27, 21]. The success of ViT highlights the potential for unified vision-language models. Compared to CNNs, ViTs exhibit better scalability; although they typically require larger training datasets, they can surpass CNNs under appropriate conditions. Analogous to the R(2+1)D CNN architecture for video tasks, ViViT [34] adapts ViT for video inputs, offering enhanced model capacity and improved performance on temporal visual tasks.

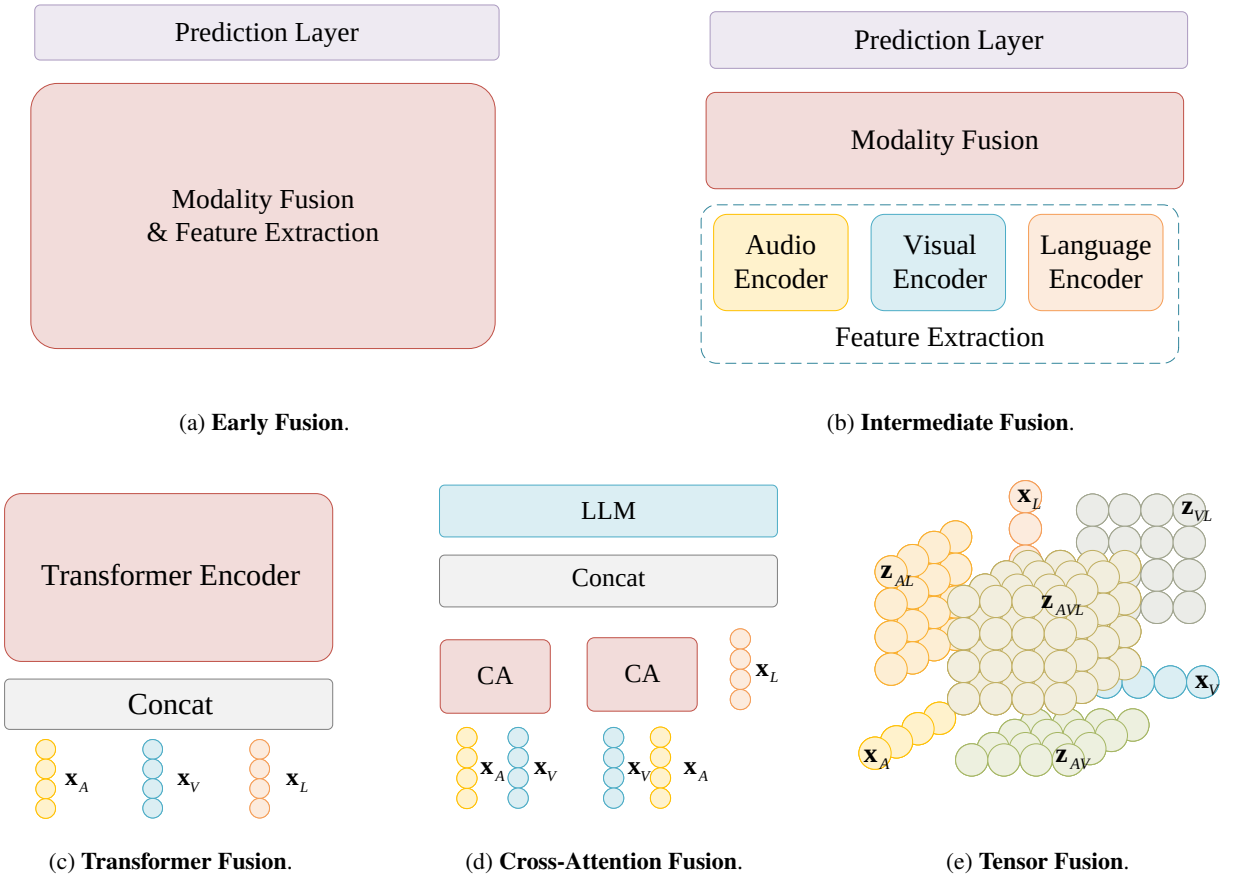
2.3 Audio Representation

The raw mono audio is commonly sampled as a waveform. For a better process, many methods transform it into a mel spectrogram via the short-time Fourier transform. Specifically, a mel spectrogram is a matrix $\mathbf{x}_A \in \mathbb{R}^{T\times F}$, where T denotes the number of time frames and F denotes the number of frequency bins. We classify methods from the past five years into CNN-based encoders and Transformer-based encoders.

CNN-based Encoder Beyond images, audio is also a continuous signal, making CNNs well-suited for audio processing. Wav2Vec is a self-supervised model that learns audio representations directly from raw waveforms [30]. Similar to unsupervised pretrained models in natural language processing, Wav2Vec encodes features from context; however, unlike discrete language tokens, audio is continuous, so convolutional layers are used to capture local features. VGGish [24], inspired by the VGG architecture in computer vision, processes mel spectrograms, which can be treated as 2D grayscale images with channel $C = 1$ (though not natural images). Thus, adapting image-based CNNs for mel spectrograms is a natural choice for audio feature extraction. The structure of VGGish is shown in Figure 2c.

Audio Transformer-Based Encoder As Transformers have proven effective in language and vision, audio encoding has been inspired by them. Motivated by ViT in computer vision, the mel spectrogram is split into patches and fed into the AST Transformer model [33]. HTS [20], another approach based on Transformer, is inspired by the Swin Transformer, employing a hierarchical structure to reduce model size and training time.

In summary, Transformer Encoders have been widely adopted across modalities due to their strong modeling capacity and the potential to unify architectures for diverse input types. Despite this trend, convolutional neural networks remain prevalent in audio and visual domains, particularly in scenarios where lightweight models are preferred. A key challenge in audio-visual-language modeling lies in the fundamentally different raw data formats and characteristics. As described in the previous three subsections, images are spatial pixel arrays, audio is temporal waveforms or mel spectrogram



(a) **Early Fusion**.

(b) **Intermediate Fusion**.

(c) **Transformer Fusion**.

(d) **Cross-Attention Fusion**.

(e) **Tensor Fusion**.

Figure 3: **Fusion.** Figure 3a and Figure 3b are 2 common fusion stages. Figure 3c to Figure 3e are 3 common fusion operations. Concat represents concatenation. CA represents CrossAttention. $\boldsymbol{x}_A$, $\boldsymbol{x}_V$, and $\boldsymbol{x}_L$ denote audio, visual, and language input features, respectively. $\boldsymbol{z}$ denotes the fused feature.

matrices, and language consists of discrete symbolic sequences. Effective alignment of the data often leads to notable performance gains. While two-modal alignment, such as the vision-language pairing exemplified by CLIP, has seen significant progress [16], aligning three modalities (e.g., vision, audio, and language) remains relatively less explored. In particular, audio and video data require not only semantic alignment but also precise temporal synchronization, which introduces additional complexity. Currently, there is a lack of alignment methods that are both practical and scalable for such settings. Nevertheless, developing such techniques is crucial and arguably even more pressing than for traditional bimodal tasks.

3 Feature Fusion

Multimodal fusion is a promising technique for sophisticated tasks [44]. Many studies categorize fusion methods by the stage at which modalities are integrated, typically classifying them as early (feature-level), late (decision-level), or intermediate (hybrid/model-level) fusion [45, 46]. Early fusion combines features at the input stage, followed by joint processing; late fusion aggregates decision scores from modality-specific branches. Intermediate fusion strikes a balance, extracting unimodal features via separate encoders and merging latent representations for joint inference. However, in contemporary audio-visual-language deep learning models, these boundaries blur, often resembling intermediate fusion in practice. Hence, rather than adhering to theoretical distinctions, we adopt a categorization grounded in fusion mechanisms from recent practical implementations in audio-visual-language modeling (grouped in Table 2), namely, MLP fusion, Cross-attention fusion, Transformer fusion, and tensor fusion. Figure 3 illustrates some commonly used fusion techniques. Note that the mechanisms discussed in this paper are not exhaustive but represent a useful review of widely adopted methods in audio-visual-language modeling. For example, although graph-based methods share

Table 2: Commonly used fusion operations

Operation	Example	Aligned	Task	Stage
MLP Fusion	TTE [46]	✗	DD	Late Fusion
Transformer Fusion	VATT [47]	✗	AVU	Early Fusion
Transformer Fusion	WCMA [52]	✗	MER	Intermediate Fusion
Transformer Fusion	BTV [50]	✗	AVQA	Intermediate Fusion
Cross-attention Fusion	TFGCN [29]	✗	MER	Intermediate Fusion
Cross-attention Fusion	MEERKAT [48]	✓	AVU	Intermediate Fusion
Cross-attention Fusion	MSFT [53]	✓	AVVC	Intermediate Fusion
Tensor Fusion	TF-BERT [54]	✗	MER	Intermediate Fusion
Tensor Fusion	InTense [55]	✗	MER	Intermediate Fusion

DD = Depression Detection; AVU = Audiovisual Video Understanding (which contains multiple subtasks); MER = Multimodal Emotion Recognition; AVQA = Audiovisual Question Answering; AVVC = Audiovisual Video Captioning;
The Aligned column indicates whether latent features from different modalities are mapped into a shared representation space such that semantically corresponding elements are aligned before fusion.

some structural similarities with modern attention-based approaches, they are not included here because, in practice, attention-based methods have largely supplanted them in audio-visual-language modeling [47, 48].

Simple Fusion Before introducing modern fusion mechanisms, we first review several fundamental techniques. Basic fusion methods include additive fusion $\boldsymbol{z} = w_1\boldsymbol{x}_1 + w_2\boldsymbol{x}_2$ and element-wise multiplicative fusion $\boldsymbol{z} = \boldsymbol{w} \odot (\boldsymbol{x}_1 \odot \boldsymbol{x}_2)$. These techniques can be applied independently or integrated with nonlinear neural networks to enhance representational capacity. They are widely employed in feature fusion scenarios, for instance, using modality embeddings to indicate the source modality of each segment, analogous to position encoding in sequence models [49, 50, 51]. While modern audio-visual-language systems rarely rely solely on these methods for modality fusion, they serve as alternatives to concatenation in more advanced fusion architectures in the rest of this section.

MLP Fusion A straightforward approach to feature fusion in neural networks is to concatenate modality-specific features and feed them into a multilayer perceptron (MLP), typically consisting of one or two hidden layers. While this method does not align with traditional late fusion, where final decision scores are combined, it is sometimes referred to as late fusion when applied near the end of a network [46].

Transformer Fusion Transformer encoder models are frequently used for fusion. Similar to early fusion, where raw features are combined and fed into a nonlinear processor such as a Transformer encoder, latent features can also be fused by combining them and inputting them into a Transformer encoder model [52, 50]. The main advantage of using Transformers at the fusion stage lies in their ability to capture long-range dependencies and complex cross-modal interactions via self-attention, resulting in more accurate and robust multimodal representations.

Cross-Attention Fusion Cross-attention is a key component in Transformer decoders, enabling the fusion of latent representations from the encoder with prior decoder outputs. In multimodal learning, a common strategy for leveraging pretrained large language models is to integrate modality-specific features, such as visual and audio, via cross-attention [48]. Similar to self-attention, cross-attention also employs the query-key-value (QKV) mechanism in Equation 1, but differs in that it operates across distinct modalities. For instance, visual features can serve as queries while language features act as keys and values.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{1}$$

Typically, cross-attention fusion is followed by additional layers to further process the combined representations. For instance, TFGCN [29] constructs a graph from the cross-attention fused features and subsequently applies graph convolutional networks. Since standard cross-attention handles two modalities, recent work such as MSFT [53] proposes multi-stage cross-attention to handle three (audio, visual, and language).

Some studies adopt paired cross-attention, where each modality attends to the other symmetrically: one modality uses the other as key/value while acting as the query, and vice versa [53]. This results in two updated modality-

specific representations that capture cross-modal interactions bidirectionally. While these operations may be viewed as coordinated rather than fused in conventional multimodal learning since they retain the same number of outputs as inputs [44], they still facilitate interaction across modalities, and from the level of the entire neural network, the modality-specific information is gradually integrated into the decision space. Thus, we argue that complete separation of these operations is unnecessary. In this work, we focus on the fusion of modality information, where both coordinated and conventional fused outputs capture meaningful multimodal interactions.

Tensor Fusion Tensor fusion was first introduced to fuse audio, visual, and language modalities for sentiment analysis [56]. It provides a general and expressive approach to model both unimodal and multimodal interactions through an outer product operation.

Let the modality-specific feature vectors be $\tilde{\boldsymbol{x}}_A = [\boldsymbol{x}_A; 1]$, $\tilde{\boldsymbol{x}}_V = [\boldsymbol{x}_V; 1]$, $\tilde{\boldsymbol{x}}_L = [\boldsymbol{x}_L; 1]$, where $\boldsymbol{x}_A$, $\boldsymbol{x}_V$, and $\boldsymbol{x}_L$ represent the feature vectors for audio, visual, and language modalities respectively, and appending 1 enables bias terms and interactions with the other modalities. The tensor fusion is computed as in Equation 2.

$$\boldsymbol{z} = \tilde{\boldsymbol{x}}_A \otimes \tilde{\boldsymbol{x}}_V \otimes \tilde{\boldsymbol{x}}_L \tag{2}$$

Where $\otimes$ denotes the outer product. The resulting tensor $\boldsymbol{z}$ captures:

- Unimodal features (e.g., $\boldsymbol{x}_A$, $\boldsymbol{x}_V$, $\boldsymbol{x}_L$)
- Bimodal interactions (e.g., $\boldsymbol{x}_A \otimes \boldsymbol{x}_V$, $\boldsymbol{x}_V \otimes \boldsymbol{x}_L$, $\boldsymbol{x}_A \otimes \boldsymbol{x}_L$)
- Trimodal interactions (e.g., $\boldsymbol{x}_A \otimes \boldsymbol{x}_V \otimes \boldsymbol{x}_L$)

While tensor fusion can be computationally expensive due to the high-dimensional output, it has been commonly used as a baseline or as a submodule within larger architectures for multimodal sentiment analysis [54]. Much recent work has been devoted to improving its efficiency and interpretability for audio-visual-language tasks [55].

In summary, intermediate fusion has become the mainstream approach in audio-visual-language modeling, where a wide variety of fusion operations are commonly employed to effectively integrate information across modalities. Among fusion mechanisms, cross-attention and Transformer encoder-based architectures are currently more widely adopted than alternative approaches. The choice mainly depends on computational efficiency and practical convenience. Cross-attention modules tend to be more lightweight and easier to integrate into pretrained Transformer models compared to concatenation followed by Transformer encoding. In contrast, MLPs remain viable for simple fusion tasks. Tensor fusion explicitly captures intra- and inter-modal interactions through high-order representations, but their high memory overhead restricts practical usage. Improved variants of tensor fusion show promising potential for broader adoption in future work.

4 Understanding and Generation Tasks

Having reviewed the core components of audio-visual-language models in Section 2 and Section 3, we now turn to their downstream applications. Figure 1d provides an overview of the capabilities of these models by illustrating the proportion of publications per task over the past five years. This section surveys these tasks in detail and organizes them into two broad categories: understanding tasks and generation tasks. While this taxonomy is not exhaustive, it provides a meaningful framework that covers the majority of existing work, representing relatively mature tasks with established benchmarks for evaluation and comparison.

4.1 Understanding Tasks

Scene understanding is a fundamental general goal of multimodal learning [2]. This part of the survey focuses on audio-visual-language understanding tasks, which we define as tasks requiring models to produce structured outputs based on audio, visual, and language inputs, such as classifications, localizations, retrievals, or other predictions. Specifically, we review key audio-visual-language understanding tasks, including multimodal action recognition [47], multimodal emotion recognition [36], audiovisual question answering [19], audiovisual object localization [68], audiovisual event localization [37], and cross-modal retrieval [74]. For each task, we survey representative datasets, model architectures, and recent research advances. Table 3 summarizes benchmark comparisons and highlights state-of-the-art (SOTA) methods across different audio-visual-language understanding tasks.

Table 3: Benchmark datasets for multimodal learning understanding tasks.

Task	Dataset	Size	Audio	Visual	Text	SOTA	Metric
MAR	Kinetics-400 [57]	306K videos	✓ / ✓	✓ / ✓	✗ / ✓	VATT [47]	Acc=82.1
	EPIC-KITCHENS [58]	39K action segment	✓ / ✓	✓ / ✓	✓ / ✓	MTCN [59]	Acc=45.5
MER	IEMOCAP [60]	10K turns	✓ / ✓	✓ / ✓	✓ / ✓	CORECT [36]	Acc=84.7
	CMU-MOSEI [61]	23K datapoints	✓ / ✓	✓ / ✓	✓ / ✓	RMA [62]	F1=88.1
	MELD [63]	1K dialogues	✓ / ✓	✓ / ✓	✓ / ✓	TelME [64]	F1=67.4
AVQA	Music-AVQA [65]	9K videos	✓ / ✓	✓ / ✓	✓ / ✓	Amuse [19]	Acc=82.4
	AVQA [66]	57K videos	✓ / ✓	✓ / ✓	✓ / ✓	TSPM [23]	Acc=90.8
AVOL	RefCOCO [67]	19k images	✗ / ✓	✓ / ✓	✓ / ✓	GroundingGPT [68]	Acc=91.6
	Flickr-SoundNet [69]	3K image-sound pairs	✗ / ✓	✓ / ✓	✓ / ✓	MEERKAT [48]	AUC=67.9
AVEL	AVE [70]	4K videos	✓ / ✓	✓ / ✓	✗ / ✓	OV-AVELBaseline [71]	Acc=61.9
	LLP [72]	11K videos	✓ / ✓	✓ / ✓	✗ / ✓	VALOR++ [37]	F1=59.0
CMR	MSR-VTT [73]	10K videos	✓ / ✓	✓ / ✓	✓ / ✓	GRAM [74]	R@1=64.8
	AudioCaps [75]	46K audio caption	✓ / ✓	✓ / ✓	✓ / ✓	CoAVT [76]	R@1=44.9

MAR = Multimodal Action Recognition; MER = Multimodal Emotion Recognition; AVQA = Audiovisual Question Answering; AVOL = Audiovisual Object Localization; AVEL = Audiovisual Event Localization; CMR = Cross-Modal Retrieval. The marks under audio, visual, and text represent whether that modality is presented in the dataset (left) or method (right).

4.1.1 Multimodal Action Recognition

The surge of user-generated videos, driven by online platforms and low upload barriers, has led to massive, diverse audiovisual data. This growth poses challenges in retrieval, recommendation, and regulation, highlighting the need for automated video understanding. A core task in this domain is understanding human actions in audiovisual videos [59], which has propelled the advancement of Multimodal Action Recognition (MAR).

Datasets To facilitate research, several benchmark datasets have been developed, providing standardized evaluation platforms:

- Kinetics400, introduced by Kay in 2017, includes 400 human action classes, with each class containing at least 400 video segments of approximately 10 seconds in length [57]. Due to the dataset's large size, many studies use its subsets. For example, the SEAR [77] utilized the Kinetics-Sounds subset, and AVTeST [78] worked with the MiniKinetics subset. Kinetics400 is typically evaluated using Top-1 and Top-5 accuracy metrics.
- EPIC-KITCHENS, a popular MAR dataset, is based on daily activities performed by individuals in home kitchens, captured from a first-person perspective [58]. It uses an innovative "pause and speak" annotation interface. The dataset contains 100 hours of video with 90K action segments.

Models VATT [47] proposes a unified Transformer architecture capable of processing audio, visual, and textual modalities. Trained via contrastive learning, it demonstrates versatility across multiple downstream tasks, including multimodal action recognition (MAR) on the Kinetics-400 dataset. In contrast, MTCN [59] adopts a dual-encoder framework, separately modeling audiovisual and textual inputs. The audiovisual encoder generates candidate action sequences, from which the language model selects the most probable outcome. These approaches exemplify two distinct paradigms in multimodal modeling: unified versus modular architectures and end-to-end versus staged training strategies. The results of these two methods are presented in Table 3. Although VATT is trained using three modalities, it does not utilize textual labels as MTCN does during inference on the Kinetics-400 dataset. Instead, it follows the standard MAR protocol for Kinetics-400, which involves using only audiovisual inputs to predict action labels.

Advances Recent developments in MAR extend beyond conventional settings. ActorShift [79] introduces domain adaptation techniques that incorporate non-human (e.g., animal) actions using audio cues, thus broadening MAR's applicability. Charades-Ego [80] presents a cross-view benchmark combining first- and third-person perspectives, encouraging the development of viewpoint-agnostic models. Additionally, the field is witnessing a shift toward fine-grained annotation and improved multimodal alignment. Fine-grained refers to detailed, precise, and localized characteristics or annotations.

4.1.2 Multimodal Emotion Recognition

Multimodal emotion recognition aims to predict an individual's emotional state by integrating signals from multiple modalities, such as speech, facial expressions, body language, and physiological cues. Tasks like sentiment analysis [45] and facial expression recognition [81] also fall under this category. Applications span various domains, including assistive technologies for individuals with affective disorders, emotionally intelligent human-computer interaction, personalized learning experiences, medical diagnostics, and immersive entertainment.

Datasets Several benchmark datasets have been established to facilitate the evaluation and comparison of multimodal emotion recognition methods:

- IEMOCAP [60]. Comprising 302 video segments from dyadic conversations between five speaker pairs, this dataset includes annotations for nine emotions, along with valence, arousal, and dominance ratings.
- MELD [63]. An extension of the EmotionLines dataset, MELD features audio, visual, and textual modalities derived from Friends TV show dialogues. It includes over 1,400 dialogues and 13,000 utterances, each labeled with one of seven basic emotions and a sentiment polarity.
- CMU-MOSEI [61]. Currently, the largest multimodal emotion recognition dataset, CMU-MOSEI, focuses on sentence-level annotations in online videos. It spans over 12 hours of labeled content from more than 1,000 speakers and 250 topics.

Models Emotion modeling is typically categorized into two paradigms: discrete and dimensional. Discrete models treat emotions as categorical labels (e.g., happiness, sadness, anger, fear), framing the task as classification [36, 64]. Dimensional models represent emotions as continuous values along axes such as valence (positivity) and arousal (intensity), aligning with regression objectives [62]. Due to their interpretability and ease of evaluation, discrete models dominate standard benchmarks.

Advances Recent efforts have addressed persistent challenges in modality fusion and dynamic modeling within multimodal emotion recognition. To mitigate the impact of data corruption, such as facial occlusion or audio noise, and temporal misalignment, new approaches aim to model cross-modal consistency under asynchronous conditions [82]. For instance, techniques have been developed to resolve conflicts when visual, audio, and textual cues express divergent emotions, improving robustness to signal discrepancies.

Another line of research focuses on addressing modality dominance. Models often exhibit biased reliance on modalities that show stronger statistical correlation with the target labels during training, which can lead to misclassification [83]. Recent approaches introduce adaptive weighting mechanisms or attention-based fusion to dynamically balance modality contributions, improving generalization across diverse scenarios.

Furthermore, recognizing the dynamic nature of emotions, recent models increasingly incorporate temporal context, such as conversational history or preceding utterances [83]. This shift toward temporal-aware architectures enables more accurate modeling of evolving affective states, which is critical for real-world applications involving continuous interaction.

4.1.3 Audiovisual Question Answering

Question Answering (QA) serves as a fundamental task for evaluating a model's ability to comprehend and reason over contextual information [84]. Traditional textual QA involves a question paired with a textual context [85], whereas Audiovisual Question Answering (AVQA) leverages acoustic and visual information from videos as context [65, 66].

Datasets Two commonly used benchmark datasets for AVQA are:

- **Music-AVQA** [65]: This dataset comprises over 45K QA pairs grounded in musical audio, visual content, and their interplay within videos. It includes five question types: existential, location, counting, comparative, and temporal. The dataset contains 9.3K videos, each averaging 60 seconds in length.
- **VAQA** [66]: Composed of 57,015 videos depicting daily activities and 57,335 QA pairs, this dataset spans diverse categories such as animals, vehicles, and sports. It introduces two additional question types, causal and purpose, compared to Music-AVQA, but does not include comparative questions.

While both datasets incorporate audio, they differ in focus: VAQA centers on everyday scenarios, whereas Music-AVQA targets musically rich content with higher information density. However, questions in both datasets often require only coarse-grained understanding, limiting their applicability in complex scenarios and broader multimodal evaluations.

Models To address the complexity of dense audio signals, **Amuse** [19] integrates music, visual, and language modalities. It leverages annotated rhythmic and musical sources from Music-AVQA and aligns modalities along the temporal axis to capture music-specific characteristics. Another approach, **TSPM** [23] emphasizes temporal-spatial information perception. Acknowledging that only a portion of audiovisual components may be pertinent to a given question, TSPM introduces dedicated temporal and spatial perception modules, which are subsequently fused through cross-modal interactions. Together, these models exemplify different strategies for enhancing AVQA performance by leveraging temporal alignment and selective attention across modalities.

Advances Recent progress in AVQA has shifted toward generative approaches powered by large pretrained language models [86]. In contrast to classification-based models, which are limited by fixed answer vocabularies [35, 65], generative models are capable of producing fluent, free-form responses, thereby aligning more closely with real-world usage scenarios. For example, the Music-AVQA dataset defines 42 fixed answer classes, framing the task as a closed-set classification problem [65]. While effective within the dataset's scope, this setup hinders generalization to open-ended or unseen answers in practical applications. Generative models, by design, can support multi-turn dialogues [87], enabling more dynamic reasoning and making them better suited for interactive human-computer communication [48].

Despite these advantages, evaluating generative QA remains a significant challenge. Traditional metrics such as accuracy are insufficient to capture the semantic quality or contextual relevance of open-ended responses. A common workaround involves model-based evaluation, where GPT-style models are used to compute the conditional likelihood of candidate answers given the question and context, with the highest-probability response selected as the most appropriate [88].

4.1.4 Audiovisual Object Localization

Understanding the spatial properties of objects, including their absolute positions and relative arrangements, is widely regarded as a fundamental aspect of intelligent behavior. Traditional computer vision tasks such as object detection [89], segmentation [90], and tracking [91] have achieved significant progress in modeling such spatial information. Audiovisual object localization extends this capability by incorporating sound as an informative cue, which is particularly beneficial in real-world scenarios where identifying the location of sound sources can guide attention and enable more efficient data processing [92]. Furthermore, language can serve as a powerful interface, allowing users to specify targets or attributes that should be grounded in the visual data [67].

Datasets Two related datasets are collected to localize objects in visual data.

- **RefCOCO** [67]: The term RefCOCO commonly refers to a suite of three related visual grounding benchmark datasets: RefCOCO, RefCOCO+, and RefCOCOg. RefCOCO and RefCOCO+ were collected through a two-player referential game [93]. RefCOCO contains 142,209 referring expressions across 19,994 images, while RefCOCO+ includes 141,564 expressions paired with 19,992 images. In contrast, RefCOCOg was constructed in a non-interactive setting and comprises 85,474 referring expressions over 26,711 images.
- **Flickr-SoundNet** [69]: This dataset is designed to evaluate a model's ability to learn the correspondence between visual scenes and associated sounds. Each sample is annotated with the image coordinates of the sound source and its category, either object-based or ambient. The dataset includes 2,786 aligned image–sound pairs, with 250 randomly selected for testing and the remainder used for training.

Models GroundingGPT [68] is an end-to-end model designed to perform object localization with audio, visual, and language signals. GroundingGPT adopts a three-stage training strategy, gradually enhancing the model's local semantic awareness. Based on large language models, GroundingGPT leverages the remarkable language understanding and extends to multimodal tasks. MEERKAT [48] is another audiovisual large language model equipped with a fine-grained understanding of image and audio information. It adopts a modality alignment module and a cross-attention module to enforce audiovisual consistency, thereby achieving remarkable results in multiple multimodal tasks.

Advances Recent advances leverage the semantic understanding capabilities of pretrained language models and extend them to encompass spatial reasoning. While visual input remains central to spatial tasks, relying solely on visual cues limits model flexibility in complex scenarios. For instance, in pedestrian or traffic object detection, users may wish to query based on natural language descriptions [91] (e.g., "pedestrians wearing red jackets near a bus"). Traditional approaches depend heavily on large-scale annotated datasets and extensive fine-tuning, which can be inefficient and brittle, especially when queried attributes are underrepresented in training data.

To address these limitations, Um et al. [94] propose two loss functions for sound source localization. The object-aware contrastive alignment loss leverages detailed object information from audiovisual scene understanding, while the object

region isolation loss enhances performance in multi-source scenarios, improving localization and identification accuracy in dense environments.

LLaVA-ST [95] introduces Language-Aligned Positional Embedding, which integrates coordinate-based textual tokens into the visual space, facilitating alignment of fine-grained spatial-temporal correspondences. Additionally, the Spatial-Temporal Packer decouples temporal and spatial resolution compression into two separate point-to-region attention pathways, improving modeling efficiency and accuracy.

Tasks in audiovisual object localization are becoming increasingly fine-grained. Sounding Object Localization integrates auditory cues to infer the spatial positions of sound-emitting objects in static images [96], promoting tighter integration between auditory and visual modalities. In audiovisual segmentation tasks, the granularity increases further. Models must perform pixel-level segmentation of sound-producing objects [97], moving beyond bounding boxes to generate detailed spatial masks.

4.1.5 Audiovisual Event Localization

In dynamic real-world environments, understanding an action requires not only isolated visual frames, but also modeling of temporal dynamics over longer sequences [37]. To this end, event localization has emerged as a critical task for detecting and interpreting time-bound occurrences within videos [70, 72]. Audiovisual Event Localization (AVEL) has gained increasing attention [70]. The task aims to temporally pinpoint events in a video using synchronized visual and auditory cues.

Datasets Two representative benchmark datasets for audiovisual event localization are:

- **AVE** [70]: Collected from AudioSet [98], AVE dataset contains 4143 videos. Each video contains at least a 2-second-long audiovisual event within 28 categories.
- **LLP** [72]: This dataset contains 11,849 YouTube video clips of 25 event categories. It is also a subset of AudioSet [98]. The LLP dataset provides not only category labels but also fine-grained boundary information with start time and end time of the events.

Models To achieve open-vocabulary AVEL, OV-AVELBaseline [71] adopts a zero-shot paradigm similar to CLIP [25]. Specifically, it applies ImageBind [28] to extract audio, visual, and language features. By computing the cosine similarity between audio embeddings and language embeddings, as well as between visual embeddings and language embeddings, the method finds the audio event category and video event category of each video segment. To predict audiovisual labels, OV-AVELBaseline selects the categories when the audio event category and the video event category agree. VALOR++ [37] investigates the under-explored unaligned setting, where the goal is to recognize audio and visual events in a video with only weak labels observed. These weak labels annotate the events that happen but are not specifically attached to one modality. Similarly, VALOR++ adopts CLIP and CLAP for visual-language similarity computation and audio-language similarity computation, respectively.

Advances As with AVOL in Section 4.1.4, the leading approach for AVEL uses Multimodal Large Language Models (MLLM). By leveraging both audio and visual modalities, AVEL systems mitigate the limitations of unimodal approaches in capturing temporal context and environmental semantics. For instance, segmenting lengthy lecture videos into concise, meaningful excerpts allows for more efficient navigation and targeted content retrieval [99]. However, to the best of our knowledge, there is no existing dataset that provides fine-grained annotations across audio, visual, and language modalities. While some datasets incorporate multiple modalities, they are typically not designed with detailed temporally or semantically aligned correspondences between them [70, 72]. We argue that such fine-grained alignment is essential for developing models capable of precise cross-modal understanding and reasoning, which is critical for tasks such as audiovisual event localization, multimodal grounding, and content-based retrieval. Therefore, we believe that there is a strong need for a novel dataset specifically designed to support fine-grained multimodal alignment.

4.1.6 Cross-Modal Retrieval

Cross-modal retrieval retrieves relevant content from one modality based on a query from another, and typically operates on coarse-grained semantic matching without requiring fine-grained alignment, for example, recognising elements within an image. Such retrieval methods can be applied in practical systems like search engines and recommendation systems [100, 101], where multimodal queries and results are common.

Datasets To facilitate research in various task settings, many benchmark datasets have been proposed. Two key datasets are:

- **MSR-VTT** [73] is a large-scale video-to-text benchmark. The dataset was built by collecting 257 popular video search queries from a commercial video search engine and retrieving 118 videos per query. The release contains 10K web video clips (41.2 hours) and 200K clip-sentence pairs; each clip was annotated with about 20 human-written natural-language captions.

- **AudioCaps** [75]. This dataset provides 46K audio clips collected from AudioSet [98]. The audio clips are paired with video signals. The audio and video are separately labelled with human-written text.

Models In multimodal retrieval, contrastive learning is widely used to project inputs into a joint embedding space by maximizing similarity between positive pairs and minimizing it for negative pairs [102, 103]. CLIP [25] significantly advanced this paradigm by training on 400M image-text pairs with a dual-encoder architecture and using cosine similarity between normalized embeddings. Its efficiency and scalability have led to broad adoption in audio-visual-text retrieval tasks. However, the shared embedding may suppress modality-specific features critical for some tasks. For example, in image-text retrieval, subtle visual details like brushstroke texture or lighting direction might be downplayed because such features lack direct textual counterparts, yet are important for tasks like art style recognition. To address this, Zeng et al. [104] propose dual common subspaces: an explicit subspace that captures modality-common features by aligning paired audio-visual data while discarding modality-specific details, and an implicit subspace that preserves modality-specific features to maintain category distinctions by increasing feature separability within each modality. This design retains complementary information unique to each modality, enabling richer and more discriminative multi-modal representations. Handling incomplete modalities is another challenge. Additionally, Lee et al. propose a mismatch-aware strategy that aligns and adjusts representations to learn robust audiovisual embeddings even with missing audio [105].

GRAM [74] addresses limitations of previous approaches that align each modality to a fixed anchor by directly aligning multiple modalities in a higher-dimensional space. It minimizes the volume of the parallelotope, an extension of a parallelogram in higher dimensions, spanned by the modality vectors, thereby ensuring simultaneous geometric alignment of all modalities. CoAVT [76] extracts audio and visual features using separate encoders, followed by a joint encoder to derive audio, visual, and audiovisual representations. These are fused via cross-attention layers using learnable queries. Language is processed independently in a separate stream, reflecting the distinction between verbal (language) and non-verbal (audiovisual) information. The audiovisual and language streams are optimized with a matching loss and a language modeling loss, respectively, while both are jointly trained with a contrastive loss to align audiovisual and linguistic information.

Advances Recent developments in cross-modal retrieval have achieved significant progress in zero-shot inference capabilities, enabling models to generalize effectively to unseen data without requiring additional examples during inference. Parida et al. [102] successfully demonstrate audiovisual video retrieval through coordinated joint multimodal embeddings in a zero-shot manner, establishing a robust framework for cross-modal understanding. Advanced correlation analysis techniques have been effectively integrated into retrieval systems, with clustering and canonical correlation analysis approaches proving successful in identifying corresponding data across modalities, as shown by Zeng et al. [103, 106] and Zhang et al. [107].

The field has progressed beyond simple architectures to sophisticated models that effectively handle complex multimodal scenarios. For multiscale data representation, Chen et al. [108] have successfully developed specialized models tailored for multiscale audiovisual retrieval tasks, demonstrating superior performance compared to standard approaches. Furthermore, advanced architectures have been successfully deployed to manage retrieval complexity, with VAE and Encoder-Decoder frameworks [76] showing promising results in handling sophisticated multimodal retrieval tasks, representing significant advances in architectural innovation for the field.

4.2 Generation Tasks

Generation tasks involve creating new content from existing data, including image, video, text, audio, and multimodal generation. Compared to understanding tasks, which focus on reasoning over raw data to derive conclusions, generation tasks are generally more complex and challenging to evaluate, especially in creative domains where subjectivity plays a significant role [109]. Despite their differences, understanding and generative tasks are closely linked: understanding methods often assist in data processing, model evaluation, or serve as key components in generative systems. For instance, the CLIP model, originally designed for retrieval, is used as the text encoder in TAgVM to enhance the model's ability to interpret textual prompts and control video generation accordingly [110]. Common tasks incorporated with audio, visual, and language modalities are listed in Table 4.

Table 4: Benchmark datasets for multimodal learning generation tasks.

Task	Dataset	Size	Audio	Visual	Text	SOTA	In→Out	Metric
AVSR	LRS2 [111]	144K videos	✓	✓	✓	AUTO-AVSR [113]	AV→T	WER=1.5
	LRS3 [112]	151K videos	✓	✓	✓	AUTO-AVSR	AV→T	WER=0.9
AVVC	MSR-VTT [73]	10K videos	✓	✓	✓	CLIP4VLA [120]	AV→T	METEOR=31.1
	ActivityNet Captions [121]	20k videos	✓	✓	✓	BMHRL [122]	AV→T	METEOR=10.92
ViG	AudioSet-Cap [109]	809K videos	✓	✓	✓	SVG [109]	T→AV	FID-vid=8.10
	Landscape [123]	9K videos	✓	✓	✗	TAgVM [110]	AT→V	FVD=877.21

AVSR = Audiovisual Speech Recognition; AVVC = Audiovisual Video Captioning; ViG = Video Generation.

4.2.1 Audiovisual Speech Recognition

Automatic Speech Recognition (ASR) transcribes spoken language from audio signals, while Visual Speech Recognition (VSR), or lip reading, extracts speech content from visual cues. Both tasks have been widely studied due to their broad range of applications. In real-world scenarios, audio signals can be corrupted by multi-source noise, and visual inputs may suffer from occlusions. These challenges naturally motivate the integration of both modalities. By leveraging the complementary strengths of audio and visual information, Audiovisual Speech Recognition (AVSR) aims to improve recognition accuracy and enhance robustness under noisy conditions.

Datasets The primary datasets for audiovisual speech recognition (AVSR) are LRS2 [111] and LRS3 [112], specifically designed for visual speech recognition (VSR) and AVSR tasks.

- **LRS2**: Previous lip-reading research was confined to word- or phrase-level recognition. LRS2 extends this by enabling sentence-level acoustic-video recognition for natural language transcription. It includes 4960 hours of BBC videos and over 100K natural sentences.
- **LRS3**: This dataset comprises over 400 hours of TED talks, with corresponding subtitles and word alignment boundaries, allowing for more precise alignment.

These datasets primarily consist of news and lecture content, lacking features typical of spontaneous, fast, and indistinct spoken language found in everyday conversations. They also lack interactive dialogue scenarios. The vocabulary in news broadcasts is more limited compared to everyday language. Furthermore, the lecture format is inherently one-directional, failing to capture elements such as interruptions or conversational flow.

Evaluation Metrics Both datasets are evaluated using Word Error Rate (WER). However, WER does not account for semantic equivalence or linguistic variations, as it is based on literal word matching. For example, "we'll go" and "we will go" would be considered mismatches by WER, despite being semantically identical.

Models Most current Audiovisual Speech Recognition (AVSR) methods adopt an autoregressive architecture, progressively generating text through the Transformer Decoder while leveraging cross-attention to integrate multimodal information from both audio and visual inputs. Representative methods following this design paradigm include AUTO-AVSR [113], AV-RelScore [114], MBH [115], VATLM [116], and AKVSR [117]. Among these, VATLM distinguishes itself by integrating the text modality during training, enabling joint audio-visual-text multimodal modeling that enhances language modeling capacity and cross-modal synergy of the generation module. AKVSR takes a different approach by employing a Transformer-based visual encoder and pre-storing audio features, representing one of the few methods that depart from the conventional ResNet/Conv3D paradigm in visual processing.

Some approaches explore non-Transformer-based generation structures for specific application requirements. AV-Former [118] replaces the traditional Transformer Decoder with a Conformer RNN-T architecture, specifically designed for real-time streaming recognition scenarios. The system extracts visual features using CLIP + Linear processing, while audio features are derived from a Conformer encoder. Most AVSR models adopt a joint optimization strategy that combines Connectionist Temporal Classification (CTC) loss and attention-based loss to leverage both global alignment and autoregressive sequence modeling capabilities [119].

rather than in isolated modality pairs.

Advances The field of AVSR has progressed significantly through architectural innovations that address specific challenges in multimodal speech processing. The integration of text modality during training, as demonstrated by

VATLM [116], represents a major advancement in enabling true trimodal learning that enhances cross-modal synergy beyond traditional audio-visual approaches. This development showcases the evolution from bimodal to comprehensive multimodal architectures in speech recognition systems.

Architectural diversification has emerged as another key advance, with methods like AKVSR [117] moving beyond conventional visual processing paradigms by adopting Transformer-based visual encoders. The development of streaming-oriented architectures, exemplified by AVFormer's Conformer RNN-T design [118], demonstrates the field's progression toward practical real-time applications. Furthermore, the widespread adoption of hybrid training strategies combining CTC and attention-based losses has become a standard practice, reflecting the maturation of optimization techniques that effectively balance global alignment with sequential modeling requirements.

4.2.2 Audiovisual Video Captioning

Audiovisual video captioning seeks to generate fluent, natural language descriptions of video content by leveraging both visual and audio modalities. Given that different individuals may describe the same video in various ways, the task is inherently challenging, particularly when it comes to objectively evaluating the generated captions.

Datasets Currently, the most widely used benchmarks for audiovisual video captioning are MSR-VTT [73] and Captions [121].

- **MSR-VTT** [73]: As previously mentioned in the context of visual retrieval, MSR-VTT was originally designed for video captioning, though its annotations have also been repurposed for retrieval tasks. However, not all videos in MSR-VTT contain audio. To adapt it for audiovisual captioning, silent videos must be filtered out. After filtering, 7,867 and 884 videos remain for training and testing in retrieval tasks, respectively, and 5,867, 448, and 2,617 videos are used for training, validation, and testing in captioning tasks [120].
- **ActivityNet Captions** [121]: This dataset is based on ActivityNet v1.3, comprising approximately 20,000 untrimmed YouTube videos annotated with around 100,000 captions.

Compared to ActivityNet captions, which average 13.40 words, MSR-VTT captions are shorter, averaging 9.28 words [124]. Moreover, segments in both datasets are defined based on visual cues without explicit audio alignment. Some videos even lack audio entirely, posing challenges for research that requires synchronized audiovisual inputs.

Evaluation Metrics BLEU-4 [125], METEOR [126], ROUGE [127], and CIDEr [128] are four commonly used metrics for evaluating video captioning models. While all are n-gram-based, each emphasizes different aspects of the generated text. BLEU-4 focuses on precise n-gram overlap, specifically 4-grams; METEOR accounts for stemming, synonym matching, and positional penalties; ROUGE prioritizes recall and is widely used in summarization tasks; CIDEr incorporates TF-IDF weighting with n-gram similarity to better align with human judgment. Despite their differences, these metrics primarily assess surface-level lexical similarity and do not capture the underlying semantic meaning of the generated captions.

Models CLIP4VLA extends the successful CLIP contrastive learning framework to incorporate audio modality for multimodal understanding [120]. The model adopts a similar architecture for audio processing as used for visual inputs, exploring alignment across multiple modality pairs, including audio-text, audiovisual, and original versus augmented audio. Evaluation on video captioning tasks using the MSR-VTT dataset demonstrates the effectiveness of contrastive learning for modality alignment, with the integration of a Transformer-based Multimodal Caption Generator for text output. However, the approach focuses primarily on pairwise alignment strategies, with trimodal consistency remaining a complex challenge for future optimization.

BMHRL employs reinforcement learning within a bimodal hierarchical Transformer framework that integrates visual and audio modalities for video captioning [122]. The architecture extracts visual features using an I3D network [129] and audio features via VGGish [24], with GloVe embeddings for textual encoding [31]. Cross-modal fusion is performed in two stages: initial integration of audio and visual features, followed by fusion with textual representations. The reinforcement learning implementation utilizes an actor-critic architecture where textual features contribute to the critic for reward estimation. Ablation studies demonstrate that incorporating audio significantly enhances captioning performance compared to traditional vision-only approaches, though at the cost of increased computational complexity.

Advances The integration of audio modality into traditionally vision-language frameworks represents a significant advancement in multimodal understanding. CLIP4VLA successfully demonstrates that contrastive learning principles can be effectively extended beyond vision-language pairs to encompass audio-visual-language alignment [120]. This

work establishes a foundation for trimodal contrastive learning, though challenges remain in achieving consistent alignment across all three modalities simultaneously.

Reinforcement learning approaches have proven effective for capturing long-range dependencies in multimodal content generation. The BMHRL framework demonstrates substantial improvements in video captioning performance through the incorporation of audio features, with ablation studies providing empirical evidence that multimodal integration significantly outperforms vision-only baselines [122]. This advancement highlights the importance of audio modality in comprehensive video understanding tasks, representing a shift from traditional vision-centric approaches toward truly multimodal architectures.

4.2.3 Video Generation

Sounding video generation and audio-language guided video generation are both multimodal tasks that require modeling interactions across different modalities. Sounding video generation aims to synthesize videos with synchronized audio, necessitating joint generation of both visual and auditory streams [109]. In contrast, audio-language guided video generation conditions the video synthesis process on both audio and textual inputs, requiring effective cross-modal understanding [110].

Datasets Below are some datasets designed for video generation that involve audio, visual, and language modalities.

- **AudioSet-Cap** [109] is collected based on AudioSet [98]. The dataset contains rich audio diversity videos and was annotated manually with text descriptions. There are 809,438 and 1,000 video clips of about 10 seconds in the training set and test set, respectively.
- **Landscape** [123] aims to extend existing low-resolution video datasets with high-resolution videos. 928 at least 1280×720 video are collected. They are divided into 10-second non-overlapped clips. These clips cover 9 different scenes. The scene labels are provided, but a natural language description is not constructed.

Evaluation Metrics Fréchet Inception Distance (FID) and Fréchet Video Distance (FVD) are perceptual metrics designed to quantify the quality and realism of generative models by comparing the statistical distributions of real and generated samples in a deep feature space. FID evaluates image generation by embedding samples into the feature space of a pretrained Inception network, capturing high-level semantic content rather than low-level pixel similarity. It measures the divergence between real and generated image distributions based on their mean and covariance, thus reflecting both fidelity and diversity. FVD generalizes this approach to video by utilizing the I3D network, which encodes both spatial and temporal features. Unlike metrics that assess frames independently, FVD captures temporal coherence and motion consistency across frames, making it suitable for evaluating the dynamic structure of generated videos. Both metrics assume that deep features of natural data follow approximately Gaussian distributions, and their comparison reflects perceptual alignment in a statistically grounded way.

Models Generative models for multimodal synthesis have evolved from foundational approaches including GANs [130], VAEs [131], and diffusion models [132]. Building upon these foundations, SVG-VQGAN represents a unified approach for learning both inter-modal and intra-modal representations in sounding video generation [109]. The model employs a Transformer-based autoregressive architecture to ensure semantic alignment between input text descriptions, visual frames, and audio signals while preserving temporal coherence across modalities. Similar to earlier works like MM-Diffusion that utilized dual-branch architectures for separate modality processing [133], SVG-VQGAN implements a multi-stage training pipeline. However, it advances beyond previous approaches by incorporating both contrastive losses for representation alignment and adversarial losses to enhance generation quality, demonstrating how adversarial learning principles continue to be effective in modern generative frameworks.

TAgVM introduces a comprehensive framework for text-audio guided video generation through a multi-stage architecture that synthesizes multiple generative paradigms [110]. The method first employs a 3D VQ-GAN to compress high-dimensional video data into a low-dimensional discrete latent sequence, followed by an autoregressive transformer that generates latent tokens conditioned on textual input. This approach builds upon the semantic alignment strategies explored in MM-LDM, which addressed unconditional sounding video generation through dual encoder-decoder networks with shared latent semantic spaces [134]. To enhance semantic richness and ensure alignment with both text and audio modalities, TAgVM applies a diffusion model guided by text and audio features to refine the generated video scenes. This hybrid methodology demonstrates the effectiveness of combining vector quantization, autoregressive modeling, and diffusion processes for multimodal video synthesis, representing an evolution from earlier single-paradigm approaches toward integrated multi-technique frameworks.

Advances Current trends indicate that both diffusion models and GANs are actively employed in video generation tasks. To handle complex input-output mappings, most methods operate within a low-dimensional latent space. Multi-stage training pipelines and the use of multiple loss functions, such as reconstruction, contrastive, and adversarial losses, have become standard practices to enforce consistency across modalities. However, the optimal combination and balancing of these losses remain an open research question. Moreover, existing approaches are primarily limited to short video clips, and achieving temporal coherence in long-duration video generation continues to be a significant challenge.

5 Trends: From Bimodal to Trimodal

As audiovisual and vision-language learning continue to evolve, research is increasingly shifting toward audio-visual-language trimodal modeling to address more complex real-world problems. This evolution reflects a natural progression within the broader paradigm shifts occurring across multimodal research, characterized by a continuation and amplification of key trends. The transition from bimodal to trimodal modeling brings new challenges and intensifies existing ones, particularly in task scope, model scale, and training paradigms. We discuss these developments in the following sections.

5.1 From single-task learning to multitask learning

As we can see from previous sections, due to the diversity of multimodal tasks, research in multimodal learning has developed in a fragmented manner. This fragmentation slows the transfer of advancements from one task to another, hindering cross-task communication. For example, Kosmos-2 discretizes location information into tokens [135], while LITA discretizes temporal information into tokens, two conceptually similar approaches that have evolved independently [136]. Similarly, model training can be constrained by resource limitations, as some complex tasks require large-scale models but suffer from limited training data. For instance, CLIP was trained on a dataset of 400M image-text pairs [25], whereas VisDial [87], a common benchmark for visual dialog, has only 1.2M dialogues. Since both datasets associate each sample with a single image, a model that is trained only on VisDial for that specific task effectively utilizes only 0.3% of the data volume to train the CLIP model.

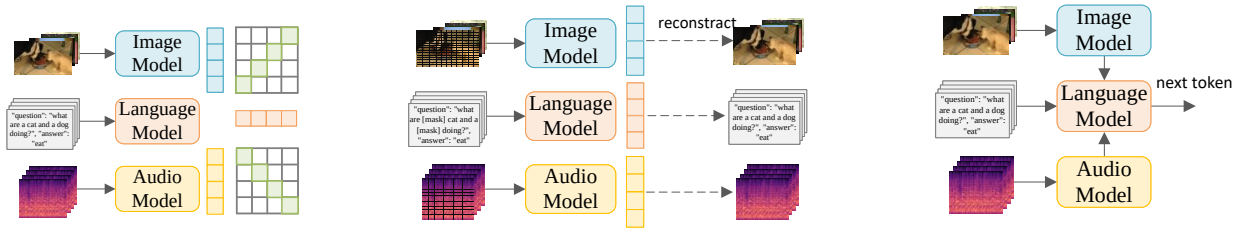
Training models on multiple tasks can leverage a larger data pool, improving model performance. Different task datasets can complement missing modalities. For example, RefCOCO contains matched text-image pairs but lacks audio signals [67], while AVE includes audiovisual information but lacks detailed textual descriptions [70]. In the absence of high-quality audio-visual-textual datasets, these datasets can still contribute to training trimodal models. In other words, low-resource tasks can benefit from high-resource tasks, reducing data requirements.

Moreover, the inherent similarities among different tasks enable models to learn more generalizable features. For example, MaPLe enhances the coupling between vision and language prompts by designing multimodal prompt learning for both branches, allowing more flexible adaptation of representation spaces across various downstream tasks [137]. Similarly, Yang et al. leverage general knowledge acquired by large language models (LLMs) to improve video-based stress detection for a specific task [138].

Furthermore, task differences help prevent overfitting to a specific dataset and mitigate spurious correlations. For example, in VQA, models may exploit shortcuts, such as always answering "tennis" when asked "what sport?" due to dataset biases rather than actual image understanding [139]. Training on diverse action recognition datasets like Sports-1M [140] and UCF101 [141], which expose models to various sports and their distinctions, can help ensure more accurate and robust reasoning. As a result, multitask learning serves as a regularization mechanism, enhancing the model's generalization ability [142].

5.2 Emergence of Larger Multimodal Models

As the number of modalities increases from two to three, audio-visual-language models grow in size to handle the added complexity [28]. The strong performance of large language models (LLMs) in language tasks has motivated their integration into multimodal systems, enabling improved understanding of non-text modalities [48, 68]. This has led to the rise of Multimodal Large Language Models (MLLMs) as a dominant trend. Larger models, due to their higher parameter counts, exhibit stronger representational capacity and can learn more complex features [143]. For instance, in image captioning, increasing model size from a Base (B/16) to a Large (L/14) architecture resulted in consistent performance gains [143]. Beyond improved representation, large models can store and retrieve broader knowledge, enhancing both understanding and generation. MLLMs have also been applied to demanding tasks such as fake news detection, which requires rich domain knowledge; in this context, they achieved over 90% F1 scores across all evaluated datasets, outperforming prior methods [144]. Furthermore, increased parameter capacity and longer attention windows



(a) **MCL.** (b) **MDM**. (c) **NTP.**

Figure 4: **Pretraining methods commonly used in audio-visual-language modeling.** The example image and text are from the MSR-VTT [73] dataset. Figure 4a shows Multimodal contrastive learning; Figure 4b shows masked data modeling; Figure 4c shows next token prediction;

in large models support better comprehension of extended content. For example, MA-LMM achieved state-of-the-art results on long-video understanding benchmarks [145], and introduced a Memory Bank Compression (MBC) technique that leverages temporal redundancy to further improve contextual understanding.

To enable the application of LLMs in audio-visual-language settings, modality-specific features, such as visual and linguistic representations, must first be extracted using dedicated encoders (Section 2) and then fused through integration modules (Section 3). These fusion layers serve as adapters, mapping pretrained features from each modality into a shared representation space compatible with LLM processing [48, 68]. Since full fine-tuning is often impractical on resource-constrained devices, this paradigm allows for partial training while still achieving competitive performance [146].

In deep learning, the Scaling Law describes how model error (e.g., perplexity) decreases following a power-law as data, model size, and compute scale proportionally [147]. This principle suggests that larger models trained on more data can consistently yield improved performance. However, increasing model depth introduces challenges such as hindered gradient propagation. To address this, techniques like Residual Networks and normalization strategies, Batch Normalization [148] and Layer Normalization [149], have been introduced to stabilize training and enhance efficiency.

As models scale, emergent capabilities become apparent [37]. These capabilities extend beyond basic alignment or descriptive tasks, enabling advanced reasoning such as temporal grounding [37] and ambiguity resolution across modalities, e.g., detecting emotional states when audio and visual signals conflict [62]. In complex downstream tasks like audiovisual question answering, large models exhibit the ability to synchronize temporal events, model long-range dependencies [68], and infer causal relationships across modalities. VALOR exemplifies such integration, outperforming modality-specific baselines and underscoring the strength of large-scale cross-modal learning [37].

Ultimately, this trajectory leads to the development of unified, end-to-end trainable trimodal models that process audio, visual, and language data within a single framework [48]. Compared to bimodal systems, the trimodal setting introduces added complexity, such as signal alignment across three streams [32] and the design of efficient cross-modal attention mechanisms [53]. Nevertheless, large Transformer-based architectures offer the capacity and flexibility needed to address these challenges, making large trimodal models strong candidates for general-purpose multimodal agents [48].

5.3 From Training from Scratch to Pretrain-Finetune Paradigm

Since model sizes have been increasing, training from scratch becomes difficult due to the datasets and computing resources. On the other hand, models pretrained on large datasets can be easily adapted to downstream tasks with better performance than those trained from scratch. Therefore, in the audio-visual-language modeling with large models, the pretrain-finetune paradigm is becoming more essential compared to bimodal architectures [48, 68].

We discuss three useful approaches to pretrain audio-visual-language models: multimodal contrastive learning, masked data modeling, and next token prediction. They can be used separately or combined in a flexible manner. An illustration is shown in Figure 4.

Multimodal Contrastive Learning (MCL) MCL learns to associate corresponding data from different modalities (e.g., an image and its caption) while distinguishing between non-matching pairs [25]. The core idea is to project inputs from different modalities into a shared latent space, where representations of matching pairs are close together, while those of non-matching pairs are pushed apart. By leveraging this approach, multimodal contrastive learning enables models to learn cross-modal relationships and align information from diverse sources without requiring labelled

data. It has proven effective in various tasks, such as image captioning, visual question answering, and audiovisual representation learning, enabling better integration and understanding of multimodal data [28, 150].

Masked Data Modelling (MDM) MDM refers to a self-supervised learning technique where part of the input data is intentionally masked or hidden, and the model is tasked with predicting the missing information based on the unmasked portions [35, 151]. This approach leverages the model's ability to learn useful representations of the data without requiring labelled examples. This can be applied to various domains, such as text, images, and audio. For example, in text modelling, MDM could mask words or phrases, and the model would predict the masked content. In images, random patches or regions could be masked, and the model learns to predict the hidden pixels[32, 152].

Next Token Prediction (NTP) Next Token Prediction (NTP) is widely used for training LLMs [88]. By predicting the next token based on preceding ones, the model captures the underlying patterns of language data. This capability naturally extends to tasks that can be reformulated as sequence-to-sequence problems. In the case of training MLLMs, NTP is adapted by predicting the next token not only with previous ones but also conditioned on other modality inputs [68]. For example, MEERKAT uses Llama2 as their LLM backbone, which is pretrained with NTP [48]. MEERKAT then adopts audio and visual encoders and is fine-tuned for multimodal downstream tasks, such as audio-referred image grounding, image-guided audio temporal localization, etc.

6 Challenges and Future Directions

Audio-visual-language modeling is still an ongoing area. In this section, three challenges are introduced and discussed. Potential future directions are highlighted subsequently for each of them.

6.1 Limited Interpretability of audio-visual-language Models

Building on Section 4, while the audio-visual-language model consistently delivers state-of-the-art performance across a range of tasks, our understanding of its decision-making process remains limited. We typically observe only the final outputs, with little insight into the input features driving those decisions [23]. This "black-box" nature is particularly problematic in high-stakes domains. Furthermore, discrepancies between deployment and testing environments, coupled with the inherent limitations in training data quality and coverage, heighten the risk of unpredictable model failures [47]. In these settings, it is critical that every model prediction be accompanied by a clear, interpretable rationale to enable expert verification and to ensure the reliability of the underlying reasoning process.

Cause 1. Uninterpretable latent space Audio-visual-language modeling tends to create entangled representations where cross-modal relationships, intra-modal features, and task-specific adaptations become inextricably mixed within the same embedding space [153]. This entanglement prevents researchers from isolating specific types of semantic relationships or understanding how different modalities contribute to final predictions, ultimately rendering the model's decision-making process opaque.

Cause 2. Complexity of deep neural network The sophisticated neural network architectures underlying audio-visual-language models present fundamental interpretability challenges due to their nonlinear parameter interactions and their layer depth. These architectures resist straightforward analysis, as their decision-making processes cannot be decomposed into interpretable individual components or layers. Current interpretability approaches face significant limitations when applied to audio-visual-language models. Model-agnostic methods [154, 155] produce explanations that lack the granularity needed for multimodal understanding, while architecture-specific techniques struggle with the heterogeneous nature of audio-visual-language systems. Transformer-based components suffer from attention mechanism interpretability gaps [156, 157], CNN-based audiovisual processing remains difficult to visualize in complex scenarios [158, 159], and spatiotemporal video analysis components [39] present particularly intractable interpretation challenges. The fundamental issue lies in the disconnect between the model's internal representational complexity and our current analytical capabilities, leaving critical aspects of audio-visual-language decision-making opaque [160].

Direction 1. Encouraging semantically-rich latent space To make the latent space more understandable, structured representation learning might emerge as a promising approach that decomposes high-dimensional embeddings into functionally distinct subspaces while preserving model expressiveness, involving designing training objectives that explicitly encourage the separation of different types of semantic information, such as object categories or attributes, into dedicated representational components, incorporating structured regularization terms that promote orthogonality between these subspaces while maintaining semantic coherence within each component. This approach comple-

ments alignment objectives that emphasize cross-modal correspondence by organizing the internal structure of latent representations, thus facilitating effective cross-modal alignment.

Direction 2. Advancing interpretability methods To address the problem of network complexity, a promising direction would be advancing hierarchical explanation frameworks using prior knowledge of audio, visual, and language modalities. This could include establishing principled methods for interpreting cross-modal fusion mechanisms, developing spatiotemporal visualization techniques that can effectively represent the complex interactions in audio-visual-language modeling [39, 161], and creating interpretability approaches that leverage the structural properties of Transformer architectures for multimodal contexts [162, 163, 164].

6.2 Reasoning in Complex Environments

While MLLM models have demonstrated strong capabilities across various tasks in static environments, they continue to struggle with complex environmental reasoning.

Cause 1. Environments are inherently dynamic When environmental changes are substantial, the distribution of data may diverge from the real-world distribution, adding complexity to the model's reasoning process. In real-world applications, the model interacts with the environment, where its outputs influence the state of the environment, establishing causal relationships with subsequent inputs. Current models, however, often treat interactions as static, overlooking the critical role of temporal information [48].

Cause 2. Inconsistency of audio, visual, and language data Currently, datasets with fine-grained feature alignment remain scarce. By fine-grained alignment, we refer to cases where not only the general scene matches across modalities, but also detailed elements of the audio track, visual frames, and textual descriptions correspond to the same events at the same time and location. While models like GroundingGPT [68] benefit from training across various tasks, their performance improvement does not stem from the availability of finely annotated datasets with precise multimodal alignment. Additionally, multimodal data is often asynchronous, and even the causal ordering of events may differ, adding further complexity to the alignment process.

Cause 3. Information is distributed over extended time periods Current video processing systems often use frame sampling or focus only on short clips to reduce computational load, assuming key information is concentrated in these segments [59]. However, this assumption is often flawed. Critical information, especially from physical signals, is distributed over longer durations. As a result, existing models struggle to capture long-range temporal dependencies effectively [64].

Direction 1. Reinforcement learning (RL) for interactive environment To address the interactive environment, RL can be introduced. In the training of large language models, RL has already been employed to align with human reasoning, enabling models to engage in more natural conversations and exhibit improved reasoning capabilities. Future research could focus on exploring how to adapt and integrate more RL approaches to address reasoning challenges in complex environments, adding active exploration in the real-world environment [122].

Direction 2. Collecting high-quality datasets To address Cause 2, a direction is to collect datasets with fine-grained audio-visual-language alignment. Within this fine-grained alignment, ensuring causal consistency across modalities for specific objects [165] constitutes a stricter and more meaningful criterion. For example, if the audio indicates a dog barking, the visual modality should causally correspond by showing the dog actively barking, not merely the presence of a dog. Such causal alignment requires not only correlation but also causation between modalities, enabling richer multimodal correspondence and more robust cross-modal reasoning and inference.

Direction 3. Augmentation with a memory system From a modeling perspective, current methods typically rely on frame sampling, and the models themselves lack memory capabilities. While this approach addresses the issue of information being limited to key frames, it falls short when handling information distributed across extended periods [138]. Although large models are gradually increasing context length, their capacity remains finite. Research indicates that, with longer contexts, models tend to prioritize more recent information. To meet the demand for processing infinite-length inputs, future model architectures must incorporate memory mechanisms. While Neural Turing Machines [166] and memory-augmented neural networks [167] offer memory features, they cannot be easily integrated into existing multimodal frameworks. Thus, a key challenge for future research will be how to effectively incorporate memory modules into these models.

6.3 Efficiency Constraints

Audio-visual-language multimodal systems face significantly more severe deployment challenges in resource-constrained edge or embedded scenarios compared to audiovisual and vision-language systems. These challenges stem from the fundamental complexity of processing three distinct modalities simultaneously while maintaining real-time performance requirements.

Cause 1. Heterogeneous feature extraction bottlenecks Audio, visual, and language modalities each demand fundamentally different representational approaches. Audio processing requires capturing spectral-temporal features, visual processing necessitates extracting spatial textures and shape information, while language processing demands understanding symbolic embeddings and contextual relationships. As depicted in Table 1, current approaches typically deploy separate deep encoders for each modality, resulting in parameter counts that scale multiplicatively, creating substantial computational and memory overhead during both storage and inference phases [168].

Cause 2. Temporal alignment complexity Fusion of audio, visual, and language features requires precise alignment of temporal sequences operating at different rates and granularities, while simultaneously executing attention mechanisms to achieve deep inter-modal interactions. Since each modality must attend not only to its own features but also respond to information from the other two modalities, fusion computations scale beyond simple linear growth, further exacerbating performance burdens in resource-constrained environments [37]. The deployment constraints are particularly acute in applications requiring a strong data security emphasis, where companies must process proprietary data locally rather than through cloud infrastructure, and in real-time critical systems like autonomous driving and financial trading that cannot tolerate cloud-based inference latency.

Direction 1. Shared-specific hybrid encoding architecture Future solutions should focus on identifying common representational spaces across modalities while preserving modality-specific capabilities. This approach involves designing shared lower-level convolutional or transformation layers that extract similar features using common parameters, followed by lightweight specialized subnetworks at higher levels. Such architecture reduces redundancy and decreases overall model size while maintaining the understanding required for each input type.

Direction 2. Advanced model compression Current compression techniques, including quantization, pruning, and distillation, play important roles in deep learning deployment [169, 170, 171]. Future developments should leverage audio-visual-language system characteristics by implementing modality-specific mixed precision quantization and sparsification strategies based on differential precision sensitivity across modalities. Input-aware dynamic execution strategies should be implemented to adaptively balance performance and efficiency, where corresponding subnetworks can be dynamically disabled or simplified when input signal-to-noise ratios are low or certain modalities are missing.

7 Conclusion

This paper surveys recent advances in audio-visual-language modeling, spanning feature extraction methods (See Section 2), feature fusion methods (See Section 3), and various understanding and generation tasks (See Section 4). As discussed in Section 5, most existing methods align modalities through contrastive learning, while Transformer-based architectures are adopted for more complex multimodal fusion. Large-scale pretraining has become the prevailing approach, substantially improving task performance but also increasing deployment complexity. As a result, a better understanding of model behavior and effective model compression are key challenges ahead. Furthermore, current methods primarily address coarse-grained tasks within short temporal windows, leaving fine-grained reasoning in complex scenarios an open area for future exploration.

Looking forward, several key challenges remain to be addressed for audio-visual-language models, as discussed in Section 6. First, improving model interpretability is crucial, including encouraging semantically-structured latent spaces and developing hierarchical explanation methods to better understand cross-modal fusion. Second, enhancing reasoning in complex, dynamic environments requires integrating reinforcement learning for interaction, collecting fine-grained aligned multimodal datasets, and incorporating memory mechanisms to handle long-term context. Third, efficiency constraints in real-world deployment call for hybrid shared-specific encoding architectures and advanced modality-aware compression techniques to balance performance and resource usage.

Limitation Although audio, visual, and language modalities are all semantically rich, each encompasses diverse subtypes. For example, the audio modality includes both speech, which significantly overlaps with language, and music, which differs from speech in terms of density, semantics, and structure. Similar subtypes exist within the visual and language modalities. While methods across subtypes may share common characteristics, they also exhibit important

differences. This paper focuses on general-purpose models that integrate audio, visual, and language modalities without emphasizing the specific distinctions among individual subtypes.

References

[1] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions, February 2023.

[2] Summaira Jabeen, Xi Li, Amin Muhammad Shoib, Bourahla Omar, Songyuan Li, and Abdul Jabbar. A Review on Methods and Applications in Multimodal Deep Learning. *ACM Trans. Multim. Comput. Commun. Appl.*, 19:76:1–76:41, 2023.

[3] Tianzhe Jiao, Chaopeng Guo, Xiaoyue Feng, Yuming Chen, and Jie Song. A Comprehensive Survey on Deep Learning Multi-Modal Fusion: Methods, Technologies and Applications. *Computers, Materials & Continua*, 80:1–35, 2024.

[4] Y. Zhu, Y. Wu, N. Sebe, and Y. Yan. Vision + X: A Survey on Multimodal Learning in the Light of Data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:9102–9122, December 2024.

[5] Sura Raya, Mariam Orabi, Imad Afyouni, and Zaher Al Aghbari. Multi-modal data clustering using deep learning: A systematic review. *Neurocomputing*, 607:128348, November 2024.

[6] Dianzhi Yu, Xinni Zhang, Yankai Chen, Aiwei Liu, Yifei Zhang, Philip S. Yu, and Irwin King. Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, October 2024.

[7] Ahmed Shahabaz and Sudeep Sarkar. Increasing Importance of Joint Analysis of Audio and Video in Computer Vision: A Survey. *IEEE Access*, 12:59399–59430, 2024.

[8] C. Liu, Y. Jin, Z. Guan, T. Li, Y. Qin, B. Qian, Z. Jiang, Y. Wu, X. Wang, Y.F. Zheng, and D. Zeng. Visual–language foundation models in medicine. *Visual Computer*, 2024.

[9] S. Lu, M. Liu, L. Yin, Z. Yin, X. Liu, and W. Zheng. The multi-modal fusion in visual question answering: A review of attention mechanisms. *PeerJ Computer Science*, 9, 2023.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[11] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR, 2023.

[12] Arnab Barua, Mobyen Uddin Ahmed, and Shahina Begum. A Systematic Literature Review on Multimodal Machine Learning: Applications, Challenges, Gaps and Future Directions. *IEEE Access*, 11:14804–14831, 2023.

[13] Jabeen Summaira, Xi Li, Amin Muhammad Shoib, Songyuan Li, and Abdul Jabbar. Recent Advances and Trends in Multimodal Deep Learning: A Review., May 2021.

[14] Su Mu, Meng Cui, and Xiaodi Huang. Multimodal Data Fusion in Learning Analytics: A Systematic Review. *Sensors*, 20:6856, 2020.

[15] Shashi Kant Shankar, Luis Pablo Prieto, María Jesús Rodríguez-Triana, and Adolfo Ruiz-Calleja. A Review of Multimodal Learning Analytics Architectures. In *18th IEEE International Conference on Advanced Learning Technologies, ICALT 2018, Mumbai, India, July 9-13, 2018*, pages 212–214, 2018.

[16] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, Virtual.*, pages 9694–9705, 2021.

[17] Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR, 2021.

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA.*, pages 12888–12900, 2022.

[19] X. Diao, C. Zhang, T. Wu, M. Cheng, Z. Ouyang, W. Wu, and J. Gui. Learning Musical Representations for Music Performance Question Answering. In *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Findings of EMNLP 2024*, pages 2803–2813, 2024.

[20] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 646–650. IEEE, 2022.

[21] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin Transformer V2: Scaling Up Capacity and Resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11999–12009. IEEE, 2022.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[23] G. Li, H. Du, and D. Hu. Boosting Audio Visual Question Answering via Key Semantic-Aware Cues. In *MM 2024 - Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5997–6005, 2024.

[24] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 131–135. IEEE, 2017.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, July 2021.

[26] Yaoting Wang, Peiwen Sun, Yuanchao Li, Honggang Zhang, and Di Hu. Can Textual Semantics Mitigate Sounding Object Segmentation Preference? In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXIV*, volume 15132 of *Lecture Notes in Computer Science*, pages 340–356. Springer, 2024.

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE, 2021.

[28] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind One Embedding Space to Bind Them All. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190, June 2023.

[29] Li Maoheng. Enhanced Emotion Recognition Through Multimodal Fusion Using TriModal Fusion Graph Convolutional Networks. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9, June 2024.

[30] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. Wav2vec: Unsupervised Pre-Training for Speech Recognition. In Gernot Kubin and Zdravko Kacic, editors, *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019*, pages 3465–3469. ISCA, 2019.

[31] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics, October 2014.

[32] Siddharth Srivastava and Gaurav Sharma. OmniVec: Learning robust representations with cross modal sharing. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1225–1237, January 2024.

[33] Yuan Gong, Yu-An Chung, and James R. Glass. AST: Audio Spectrogram Transformer. In Hynek Hermansky, Honza Cernocký, Lukás Burget, Lori Lamel, Odette Scharenborg, and Petr Motlícek, editors, *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pages 571–575. ISCA, 2021.

[34] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. ViViT: A Video Vision Transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 6816–6826. IEEE, 2021.

[35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[36] Cam-Van Thi Nguyen, Anh-Tuan Mai, The-Son Le, Hai-Dang Kieu, and Duc-Trong Le. Conversation Understanding using Relational Temporal Graph Neural Networks with Auxiliary Cross-Modality Interaction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 15154–15167. Association for Computational Linguistics, 2023.

[37] Y.-H. Lai, Y.-C. Chen, and Y.-C.F. Wang. Modality-Independent Teachers Meet Weakly-Supervised Audio-Visual Event Parser. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016.

[39] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459. IEEE, June 2018.

[40] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP Learning Audio Concepts from Natural Language Supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE, 2023.

[41] Taehyoung Kim and Sungkwon Park. Equivalent Data Information of Sensory and Motor Signals in the Human Body. *IEEE Access*, 8:69661–69670, 2020.

[42] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic Convolution: Attention Over Convolution Kernels. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 11027–11036. Computer Vision Foundation / IEEE, 2020.

[43] Irwan Bello. LambdaNetworks: Modeling long-range Interactions without Attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[44] T. Baltrusaitis, C. Ahuja, and L.-P. Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:423–443, 2019.

[45] Ruichen Li, Jinming Zhao, Jingwen Hu, Shuai Guo, and Qin Jin. Multi-modal Fusion for Video Sentiment Analysis. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, pages 19–25. ACM, October 2020.

[46] Sowmya Rasipuram, Junaid Hamid Bhat, Anutosh Maitra, Bishal Shaw, and Sriparna Saha. Multimodal Depression Detection Using Task-oriented Transformer-based Embedding. In *2022 IEEE Symposium on Computers and Communications (ISCC)*, pages 01–04, June 2022.

[47] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, Virtual*, pages 24206–24221, 2021.

[48] Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Jun Chen, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. MEERKAT: Audio-Visual Large Language Model for Grounding in Space and Time. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision -*

ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXIV, volume 15122 of *Lecture Notes in Computer Science*, pages 52–70. Springer, 2024.

[49] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just Ask: Learning to Answer Questions from Millions of Narrated Videos. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1666–1677. IEEE, 2021.

[50] Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, and Jie Zhou. Bridging Text and Video: A Universal Multimodal Transformer for Audio-Visual Scene-Aware Dialog. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2476–2483, 2021.

[51] Jacob Chalk, Jaesung Huh, Evangelos Kazakos, Andrew Zisserman, and Dima Damen. TIM: A Time Interval Machine for Audio-Visual Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18153–18163, 2024.

[52] Qiupu Chen, Guimin Huang, and Yabing Wang. The Weighted Cross-Modal Attention Mechanism With Sentiment Prediction Auxiliary Task for Multimodal Sentiment Analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2689–2695, 2022.

[53] Yulai Xie, Jingjing Niu, Yang Zhang, and Fang Ren. Global-Shared Text Representation Based Multi-Stage Fusion Transformer Network for Multi-Modal Dense Video Captioning. *IEEE Transactions on Multimedia*, 26:3164–3179, 2024.

[54] Jingming Hou, Nazlia Omar, Sabrina Tiun, Saidah Saad, and Qian He. TF-BERT: Tensor-based fusion BERT for multimodal sentiment analysis. *Neural Networks*, 185:107222, 2025.

[55] Saurabh Varshneya, Antoine Ledent, Philipp Liznerski, Andriy Balinskyy, Purvanshi Mehta, Waleed Mustafa, and Marius Kloft. Interpretable Tensor Fusion. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 5037–5045. ijcai.org, 2024.

[56] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor Fusion Network for Multimodal Sentiment Analysis. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1103–1114. Association for Computational Linguistics, 2017.

[57] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics Human Action Video Dataset, May 2017.

[58] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018.

[59] Evangelos Kazakos, Jaesung Huh, Arsha Nagrani, Andrew Zisserman, and Dima Damen. With a Little Help from my Temporal Context: Multimodal Egocentric Action Recognition. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, page 268. BMVA Press, 2021.

[60] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, 42:335–359, 2008.

[61] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2236–2246. Association for Computational Linguistics, 2018.

[62] Xianbing Zhao, Xuejiao Li, Ronghuan Jiang, and Buzhou Tang. Resolving multimodal ambiguity via knowledge-injection and ambiguity learning for multimodal sentiment analysis. *Inf. Fusion*, 115:102745, 2025.

[63] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 527–536. Association for Computational Linguistics, 2019.

[64] Taeyang Yun, Hyunkuk Lim, Jeonghwan Lee, and Min Song. TelME: Teacher-leading Multimodal Fusion Network for Emotion Recognition in Conversation. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024, pages 82–95. Association for Computational Linguistics, 2024.

[65] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to Answer Questions in Dynamic Audio-Visual Scenarios. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19086–19096. IEEE, June 2022.

[66] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. AVQA: A Dataset for Audio-Visual Question Answering on Videos. In João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni, editors, *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 3480–3491. ACM, 2022.

[67] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling Context in Referring Expressions. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, volume 9906 of *Lecture Notes in Computer Science*, pages 69–85. Springer, 2016.

[68] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, Zhida Huang, and Tao Wang. GroundingGPT: Language Enhanced Multi-modal Grounding Model. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 6657–6678. Association for Computational Linguistics, 2024.

[69] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to Localize Sound Source in Visual Scenes. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4358–4366. Computer Vision Foundation / IEEE Computer Society, 2018.

[70] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-Visual Event Localization in Unconstrained Videos. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II*, pages 252–268, 2018.

[71] Jinxing Zhou, Dan Guo, Ruohao Guo, Yuxin Mao, Jingjing Hu, Yiran Zhong, Xiaojun Chang, and Meng Wang. Towards Open-Vocabulary Audio-Visual Event Localization, March 2025.

[72] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified Multisensory Perception: Weakly-Supervised Audio-Visual Video Parsing. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III*, volume 12348 of *Lecture Notes in Computer Science*, pages 436–454. Springer, 2020.

[73] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296. IEEE, June 2016.

[74] Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian Multimodal Representation Learning and Alignment, December 2024.

[75] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating Captions for Audios in The Wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 119–132, 2019.

[76] Xianghu Yue, Xiaohai Tian, Lu Lu, Malu Zhang, Zhizheng Wu, and Haizhou Li. CoAVT: A Cognition-Inspired Unified Audio-Visual-Text Pre-Training Model for Multimodal Processing, February 2024.

[77] R. Hebbar, D. Bose, and S. Narayanan. SEAR: Semantically-grounded Audio Representations. In *MM 2023 - Proceedings of the 31st ACM International Conference on Multimedia*, pages 2785–2794, 2023.

[78] Muhammad Adi Nugroho, Sangmin Woo, Sumin Lee, and Changick Kim. Audio-visual glance network for efficient video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10150–10159, 2023.

[79] Yunhua Zhang, Hazel Doughty, Ling Shao, and Cees G. M. Snoek. Audio-Adaptive Activity Recognition Across Video Domains. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13781–13790. IEEE, June 2022.

[80] Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-Ego: A Large-Scale Dataset of Paired Third and First Person Videos, April 2018.

[81] H. Siqueira, S. Magg, and S. Wermter. Efficient facial feature learning with wide ensemble-based convolutional neural networks. In *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, pages 5800–5809, 2020.

[82] W. Zheng, J. Yu, and R. Xia. A Unimodal Valence-Arousal Driven Contrastive Learning Framework for Multimodal Multi-Label Emotion Recognition. In *MM 2024 - Proceedings of the 32nd ACM International Conference on Multimedia*, pages 622–631, 2024.

[83] G. Tu, F. Xiong, B. Liang, H. Wang, X. Zeng, and R. Xu. Multimodal Emotion Recognition Calibration in Conversations. In *MM 2024 - Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9621–9630, 2024.

[84] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering, May 2017.

[85] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know What You Don't Know: Unanswerable Questions for SQuAD, June 2018.

[86] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, February 2023.

[87] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Stefan Lee, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41:1242–1256, 2019.

[88] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, July 2020.

[89] Yunjie Tian, Qixiang Ye, and David Doermann. YOLOv12: Attention-Centric Real-Time Object Detectors, February 2025.

[90] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment Anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3992–4003. IEEE, 2023.

[91] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixé, and Christoph Feichtenhofer. TrackFormer: Multi-Object Tracking with Transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 8834–8844. IEEE, 2022.

[92] H. Xuan, Z. Wu, J. Yang, B. Jiang, L. Luo, X. Alameda-Pineda, and Y. Yan. Robust Audio-Visual Contrastive Learning for Proposal-Based Self-Supervised Sound Source Localization in Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:4896–4907, 2024.

[93] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A Meeting of SIGDAT, a Special Interest Group of the ACL*, pages 787–798. ACL, 2014.

[94] Sung Jin Um, Dongjin Kim, Sangmin Lee, and Jung Uk Kim. Object-aware Sound Source Localization via Audio-Visual Scene Understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8342–8351, 2025.

[95] Hongyu Li, Jinyu Chen, Ziyu Wei, Shaofei Huang, Tianrui Hui, Jialin Gao, Xiaoming Wei, and Si Liu. LLaVA-ST: A Multimodal Large Language Model for Fine-Grained Spatial-Temporal Understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8592–8603, 2025.

[96] Tanvir Mahmud, Yapeng Tian, and Diana Marculescu. T-VSL: Text-Guided Visual Sound Source Localization in Mixtures. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26732–26741. IEEE, 2024.

[97] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-Visual Segmentation. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVII*, volume 13697 of *Lecture Notes in Computer Science*, pages 386–403. Springer, 2022.

[98] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 776–780. IEEE, 2017.

[99] Anchit Gupta, C. V. Jawahar, and Makarand Tapaswi. Unsupervised audio-visual lecture segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5232–5241, 2023.

[100] Yun Tie, Xiaobing Li, Tian Zhang, Cong Jin, Xin Zhao, and Jiessie Tie. Deep learning based audio and video cross-modal recommendation. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2366–2371, October 2022.

[101] F. Maxwell Harper and Joseph A. Konstan. The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems*, 5:1–19, January 2016.

[102] Kranti Parida, Neeraj Matiyali, Tanaya Guha, and Gaurav Sharma. Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3251–3260, 2020.

[103] Donghuo Zeng, Yi Yu, and Keizo Oyama. Deep Triplet Neural Networks with Cluster-CCA for Audio-Visual Cross-Modal Retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16:1–23, August 2020.

[104] Donghuo Zeng, Jianming Wu, Gen Hattori, Rong Xu, and Yi Yu. Learning Explicit and Implicit Dual Common Subspaces for Audio-visual Cross-modal Retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 19:1–23, June 2023.

[105] Sangmin Lee, Sungjune Park, and Yong Man Ro. Audio-Visual Mismatch-Aware Video Retrieval via Association and Adjustment. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, volume 13674, pages 497–514. Springer Nature Switzerland, 2022.

[106] Donghuo Zeng, Yi Yu, and Keizo Oyama. Audio-Visual Embedding for Cross-Modal Music Video Retrieval through Supervised Deep CCA. In *2018 IEEE International Symposium on Multimedia (ISM)*, pages 143–150, December 2018.

[107] Jiwei Zhang, Yi Yu, Suhua Tang, Jianming Wu, and Wei Li. Variational Autoencoder with CCA for Audio–Visual Cross-modal Retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 19:1–21, October 2023.

[108] Yaxiong Chen, Chuang Du, Yunfei Zi, Shengwu Xiong, and Xiaoqiang Lu. Scale-Aware Adaptive Refinement and Cross-Interaction for Remote Sensing Audio-Visual Cross-Modal Retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

[109] J. Liu, W. Wang, S. Chen, X. Zhu, and J. Liu. Sounding Video Generator: A Unified Framework for Text-Guided Sounding Video Generation. *IEEE Transactions on Multimedia*, 26:141–153, 2024.

[110] M. Zhao, W. Wang, T. Chen, R. Zhang, and R. Li. TA2V: Text-Audio Guided Video Generation. *IEEE Transactions on Multimedia*, 26:7250–7264, 2024.

[111] Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. Lip Reading Sentences in the Wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3444–3453, 2017.

[112] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: A large-scale dataset for visual speech recognition, October 2018.

[113] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic. Auto-AVSR: Audio-Visual Speech Recognition with Automatic Labels. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2023.

[114] J. Hong, M. Kim, J. Choi, and Y.M. Ro. Watch or Listen: Robust Audio-Visual Speech Recognition with Visual Corruption Modeling and Reliability Scoring. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2023-June, pages 18783–18794, 2023.

[115] Y. Dai, H. Chen, J. Du, R. Wang, S. Chen, H. Wang, and C.-H. Lee. A Study of Dropout-Induced Modality Bias on Robustness to Missing Video Frames for Audio-Visual Speech Recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 27435–27445, 2024.

[116] Q. Zhu, L. Zhou, Z. Zhang, S. Liu, B. Jiao, J. Zhang, L. Dai, D. Jiang, J. Li, and F. Wei. VatLM: Visual-Audio-Text Pre-Training With Unified Masked Prediction for Speech Representation Learning. *IEEE Transactions on Multimedia*, 26:1055–1064, 2024.

[117] J.H. Yeo, M. Kim, J. Choi, D.H. Kim, and Y.M. Ro. AKVSR: Audio Knowledge Empowered Visual Speech Recognition by Compressing Audio Knowledge of a Pretrained Model. *IEEE Transactions on Multimedia*, 26:6462–6474, 2024.

[118] P.H. Seo, A. Nagrani, and C. Schmid. AVFormer: Injecting Vision into Frozen Speech Models for Zero-Shot AV-ASR. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2023-June, pages 22922–22931, 2023.

[119] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tomoki Hayashi. Hybrid CTC/Attention Architecture for End-to-End Speech Recognition. *IEEE J. Sel. Top. Signal Process.*, 11:1240–1253, 2017.

[120] L. Ruan, A. Hu, Y. Song, L. Zhang, S. Zheng, and Q. Jin. Accommodating Audio Modality in CLIP for Multimodal Processing. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023*, volume 37, pages 9641–9649, 2023.

[121] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-Captioning Events in Videos. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 706–715. IEEE Computer Society, 2017.

[122] D. Rothenpieler and S. Amiriparian. METEOR Guided Divergence for Video Captioning. In *Proceedings of the International Joint Conference on Neural Networks*, volume 2023-June, 2023.

[123] Seung Hyun Lee, Gyeongrok Oh, Wonmin Byeon, Chanyoung Kim, Won Jeong Ryoo, Sang Ho Yoon, Hyunjun Cho, Jihyun Bae, Jinkyu Kim, and Sangpil Kim. Sound-Guided Semantic Video Generation. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XVII*, volume 13677 of *Lecture Notes in Computer Science*, pages 34–50. Springer, 2022.

[124] Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D. Manning. AuroraCap: Efficient, Performant Video Detailed Captioning and a New Benchmark, April 2025.

[125] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL, 2002.

[126] Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss, editors, *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics, 2005.

[127] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, 2004.

[128] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society, 2015.

[129] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, February 2018.

[130] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks, June 2014.

[131] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[132] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, Virtual*, 2020.

[133] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. MM-Diffusion: Learning Multi-Modal Diffusion Models for Joint Audio and Video Generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10219–10228. IEEE, June 2023.

[134] M. Sun, W. Wang, Y. Qiao, J. Sun, Z. Qin, L. Guo, X. Zhu, and J. Liu. MM-LDM: Multi-Modal Latent Diffusion Model for Sounding Video Generation. In *MM 2024 - Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10853–10861, 2024.

[135] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding Multimodal Large Language Models to the World, July 2023.

[136] De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. LITA: Language Instructed Temporal-Localization Assistant. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, volume 15122, pages 202–218. Springer Nature Switzerland, 2025.

[137] Muhammad Uzair Khattak, Hanoona Abdul Rasheed, Muhammad Maaz, Salman H. Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal Prompt Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19113–19122. IEEE, 2023.

[138] Yang Ding, Yi Dai, Xin Wang, Ling Feng, Lei Cao, and Huijun Zhang. Integrating Content-Semantics-World Knowledge to Detect Stress from Videos. In Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu, editors, *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 10373–10381. ACM, 2024.

[139] Qingyi Si, Fandong Meng, Mingyu Zheng, Zheng Lin, Yuanxin Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. Language Prior Is Not the Only Shortcut: A Benchmark for Shortcut Learning in VQA. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3698–3712. Association for Computational Linguistics, 2022.

[140] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-Scale Video Classification with Convolutional Neural Networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732. IEEE, June 2014.

[141] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild, December 2012.

[142] Ivonne Monter-Aldana, Adrián Pastor López-Monroy, and Fernando Sánchez-Vega. Dynamic Regularization in UDA for Transformers in Multimodal Classification. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8700–8711. Association for Computational Linguistics, 2023.

[143] M. Tschannen, M. Kumar, A. Steiner, X. Zhai, N. Houlsby, and L. Beyer. Image Captioners Are Scalable Vision Learners Too. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[144] Xiaofan Zheng, Minnan Luo, and Xinghao Wang. Unveiling Fake News with Adversarial Arguments Generated by Multimodal Large Language Models. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 7862–7869. Association for Computational Linguistics, 2025.

[145] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13504–13514. IEEE, 2024.

[146] T. Ito, S. Dan, M. Rigotti, J. Kozloski, and M. Campbell. ON THE GENERALIZATION CAPACITY OF NEURAL NETWORKS DURING GENERIC MULTIMODAL REASONING. In *12th International Conference on Learning Representations, ICLR 2024*, 2024.

[147] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models, January 2020.

[148] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 448–456. JMLR.org, 2015.

[149] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization, July 2016.

[150] Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva. Spoken Moments: Learning Joint Audio-Visual Representations from Video Descriptions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14866–14876, June 2021.

[151] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[152] Bo Yang, Lijun Wu, Jinhua Zhu, Bo Shao, Xiaola Lin, and Tie-Yan Liu. Multimodal Sentiment Analysis With Two-Phase Multi-Task Learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2015–2024, 2022.

[153] Y. Zhang, E. Sui, and S. Yeung-Levy. Connect, Collapse, Corrupt: LEARNING CROSS-MODAL TASKS WITH UNI-MODAL DATA. In *12th International Conference on Learning Representations, ICLR 2024*, 2024.

[154] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, August 2016.

[155] Scott M. Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774, 2017.

[156] Samira Abnar and Willem H. Zuidema. Quantifying Attention Flow in Transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4190–4197. Association for Computational Linguistics, 2020.

[157] Hila Chefer, Shir Gur, and Lior Wolf. Transformer Interpretability Beyond Attention Visualization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, June 19-25, 2021*, pages 782–791. Computer Vision Foundation / IEEE, 2021.

[158] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2921–2929. IEEE Computer Society, 2016.

[159] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.*, 128:336–359, 2020.

[160] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What Does BERT Look at? An Analysis of BERT's Attention. In Tal Linzen, Grzegorz Chrupala, Yonatan Belinkov, and Dieuwke Hupkes, editors, *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 276–286. Association for Computational Linguistics, 2019.

[161] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In *2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018*, pages 839–847. IEEE Computer Society, 2018.

[162] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do Vision Transformers See Like Convolutional Neural Networks? In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, Virtual*, pages 12116–12128, 2021.

[163] Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function Vectors in Large Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[164] Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. Analyzing Transformers in Embedding Space. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16124–16170. Association for Computational Linguistics, 2023.

[165] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual VQA: A Cause-Effect Look at Language Bias. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, June 19-25, 2021*, pages 12700–12710. Computer Vision Foundation / IEEE, 2021.

[166] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural Turing Machines, December 2014.

[167] Guixiang Ma, Vy A. Vo, Theodore Willke, and Nesreen K. Ahmed. Memory-Augmented Graph Neural Networks: A Brain-Inspired Review, July 2023.

[168] Z. Zhu, G. Han, G. Jia, and L. Shu. Modified DenseNet for Automatic Fabric Defect Detection with Edge Computing for Minimizing Latency. *IEEE Internet of Things Journal*, 7:9623–9636, 2020.

[169] Q. Zhang, H. Lv, J. Liu, Z. Chen, J. Duan, H. Wang, L. He, and M. Xu. An Entailment Tree Generation Approach for Multimodal Multi-Hop Question Answering with Mixture-of-Experts and Iterative Feedback Mechanism. In *MM 2024 - Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4814–4822, 2024.

[170] H. Lu, Y. Huo, G. Yang, Z. Lu, W. Zhan, M. Tomizuka, and M. Ding. UNIADAPTER: UNIFIED PARAMETER-EFFICIENT TRANSFER LEARNING FOR CROSS-MODAL MODELING. In *12th International Conference on Learning Representations, ICLR 2024*, 2024.

[171] H. He, C. Bai, K. Xu, Z. Yang, W. Zhang, D. Wang, B. Zhao, and X. Li. Diffusion Model is an Effective Planner and Data Synthesizer for Multi-Task Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.