Vision + X: A Survey on Multimodal Learning in the Light of Data

Ye Zhu, *Member, IEEE,* Yu Wu, *Member, IEEE,* Nicu Sebe, *Senior Member, IEEE,* and Yan Yan, *Senior Member, IEEE*

Abstract—We are perceiving and communicating with the world in a multisensory manner, where different information sources are sophisticatedly processed and interpreted by separate parts of the human brain to constitute a complex, yet harmonious and unified sensing system. To endow the machines with true intelligence, multimodal machine learning that incorporates data from various sources has become an increasingly popular research area with emerging technical advances in recent years. In this paper, we present a survey on multimodal machine learning from a novel perspective considering not only the purely technical aspects but also the intrinsic nature of different data modalities. We analyze the commonness and uniqueness of each data format mainly ranging from vision, audio, text, and motions, and then present the methodological advancements categorized by the combination of data modalities, such as *Vision+Text*, with slightly inclined emphasis on the visual data. We investigate the existing literature on multimodal learning from both the representation learning and downstream application levels, and provide an additional comparison in the light of their technical connections with the data nature, *e.g.*, the semantic consistency between image objects and textual descriptions, and the rhythm correspondence between video dance moves and musical beats. We hope that the exploitation of the alignment as well as the existing gap between the intrinsic nature of data modality and the technical designs, will benefit future research studies to better address a specific challenge related to the concrete multimodal task, prompting a unified multimodal machine learning framework closer to a real human intelligence system.

Index Terms—Multimodal representation learning, discriminative and generative multimodal tasks, data characteristics, survey

1 Introduction

We perceive and communicate with the world through a multisensory human system, by seeing objects, hearing sounds, speaking languages, as well as writing and reading texts. The information from these varied sources is processed by different parts of the human brain, as indicated by [5], [20], [252]. For instance, the occipital lobe acts as the primary center for visual processing, interpreting the distance and locations of objects, while the temporal lobe processes auditory information, helping us understand sounds. Language comprehension, facilitated by Wernicke's area in the posterior superior temporal lobe, is crucial for decoding both written and spoken words. Other sensory information, such as touch and movement, is processed by distinct brain areas. These integrated yet distinct functions form a complex and harmonious human sensing system. The specialized divisions in human neural processing, which highlight both unique and shared characteristics across different modalities, inspire us to think about the multimodal machine learning problem in the light of data in this paper.

Historically, the vision, audio and textual data were usually studied in separate research fields (*i.e.*, computer vision, digital signal processing, and natural language processing). With the ultimate objective to bring true intelligence to machines, the research in Artificial Intelligence (AI) nowadays has gone far beyond the exploitation of a single perception perspective but entered an era where the interplay of multiple sensing systems is studied in a collaborative way just as in human brain systems. As the research of multimodal learning has become increasingly popular in recent years, we present a survey that not only studies the technical development of recent literature but also elaborates on the data characteristics, as well as examines the connections between the logic of such technical designs and their respective data natures.

To better structure the paper, we adopt a taxonomy centered on computer vision literature, using vision as the primary data modality while incorporating others, including *audio*, *text*, and others. These modalities share commonalities but also have unique characteristics in terms of nature, format, and assessment criteria. For example, audio data can be categorized into music, speech, or ambient sounds, with speech closely linked to text and music often associated with motion in a more subjective manner. We then discuss multimodal representation learning, differentiating between *supervised* and *unsupervised* settings, alongside popular network architectures for processing various modalities. This categorization underscores a shift in research focus from traditional supervised learning with manually annotated data to large-scale pre-training on unlabeled data.

Subsequently, we delve into the downstream aspects of multimodal learning, categorizing multimodal applications into two primary directions: discriminative and generative applications. For each direction, we group the existing literature in the form of **vision+X**, where *X* primarily represents non-visual data modalities. This framework highlights the

- *Y. Zhu is with the Department of Computer Science, Princeton University, Princeton, NJ 08540 USA. E-mail: yezhu@princeton.edu*
- *Y. Wu is with the School of Computer Science, Wuhan University, China.*
- *N. Sebe is with the Department of Information Engineering and Computer Science, University of Trento, Italy.*
- *Y. Yan is with the Department of Computer Science, Illinois Institute of Technology, Chicago, IL 60616, USA.*

Manuscript received October, 2022.

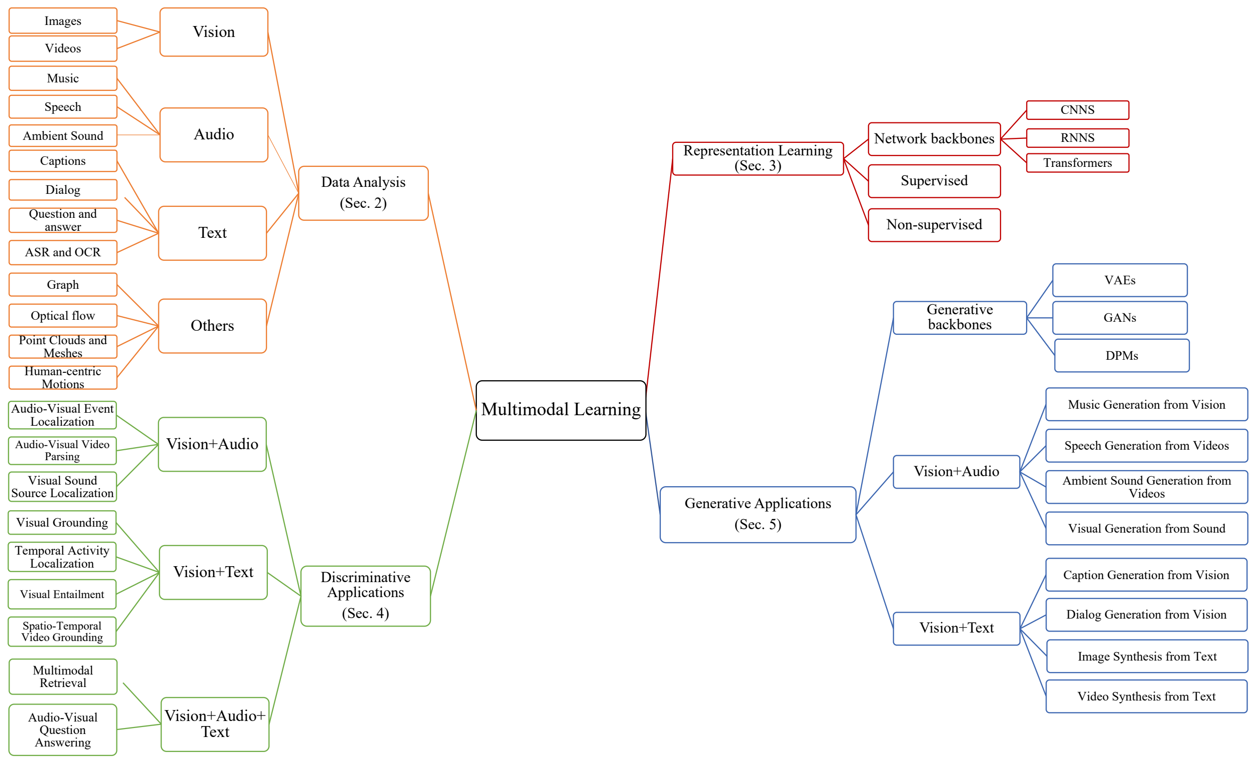


Fig. 1. **Overall structure of our survey.** We first present different data modalities and their characteristics, along with examples of multimodal datasets. We then introduce the representation learning area categorized with learning settings. Next, we mainly divide the application area into the discriminative and the generative directions, with more detailed classifications following the combination of data modalities.

adaptability and practicality of multimodal learning across diverse scenarios. For instance, akin to human multisensory perception, combining vision and language is essential for tasks like captioning, which provides textual descriptions of visual content, or imagining sounds based on visual cues. Reviewing various prominent multimodal tasks reveals that, despite differing data modalities and objectives, shared technical approaches emerge. Our detailed analysis examines these technical intricacies and their reflection on underlying data properties, strengthening the link between data types and model strategies. This exploration also addresses prevailing challenges and future directions in multimodal learning.

Compared to other surveys on multimodal learning [15], [18], [91], [115], [143], [283], we approach the problem from the unique perspective of data itself. This novel perspective allows us to establish connections between the inherent characteristics of multimodal data and the design of methodologies, leading to a profound discussion on the future of multimodal research in two major aspects. On the one hand, we believe that emphasizing and exploiting the unique characteristics of specific data modalities will contribute to solving concrete application problems associated with those modalities. On the other hand, recognizing the commonalities among different modalities will enable researchers to construct a more unified and collaborative framework that mirrors the capabilities of a real human intelligence system.

The overall structure of the paper, illustrated in Fig. 1, is as follows: In Sec. 2, we first provide an analysis in terms of data characteristics for different modalities with a focus on vision, audio, and text. Next, we explore the multimodal representation learning in Sec. 3, sub-categorized by the current popular model architectures and different learning settings. In Sec. 4 and 5, we present concrete multimodal applications with discriminative and generative tasks, respectively. In addition to the task-wise and technical introduction, we also make extra efforts by connecting the existing literature with their data characteristics as mentioned in Sec. 2, revealing which data attribute is handled and addressed in specific methods and models. The above revisit forms the basis for our discussions on the existing challenges and possible future directions in Sec. 6. Sec. 7 includes final remarks and conclusions.

2 Data Analysis

In this section, we elaborate on the intrinsic nature of multiple data modalities by analyzing their characteristics and commonalities. A list of commonly used multimodal datasets is included in Appendix A with detailed descriptions.

2.1 Vision

We categorize visual data into images and videos. As a primary information source in both human sensory systems and computer vision literature, visual data is often considered "raw data" with its high dimensionality. It encompasses a wealth of features and details, representing rich visual content. However, the redundancy in continuous spatial and temporal aspects poses challenges for processing, analysis, and efficient utilization in multimodal learning tasks.

Images. Images are fundamental to computer vision research, characterized by their inherent invariance to transformations. This key attribute drives the development of classic image processing methods and deep learning techniques like CNNs to extract meaningful visual features.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2024.3420239

In the pre-deep learning time, image processing and computer vision research primarily aimed at deciphering image content and patterns through a manual feature extraction and analysis pipeline using machine learning techniques. For example, the scale-invariant feature transform (SIFT) [156] and the histogram of oriented gradients (HOG) [54], and Speeded-Up Robust Feature (SURF) [17] are three examples of popular image feature descriptors largely used for computer vision and image processing. After having extracted those descriptive features, some machine learning algorithms such as Support-Vector Machines (SVM) [51] and Principal Component Analysis (PCA) [267] are used to further analyze the feature data. With the rapid development in the deep neural network architectures [95], [131], [216] and the availability of large-scale image datasets such as ImageNet [62], [202], computer vision entered a new era where the classic procedure of feature extraction and analysis has been automatically integrated into neural network designs.

Moreover, the applications of computer vision in the image domain have been extensively extended and enriched from simple image classifications [216], [216] to various task scenarios like object detection [310] and segmentation [154] within the image. In addition to the above-mentioned discriminative task applications that aim to mine the data patterns from existing images, there is another branch of applications that aim to synthesize the image data using generative neural networks.

Videos. Video is another form of common visual data that has been extensively studied in the computer vision community [177], [183], [281]. Unlike static images, videos encapsulate information across the temporal dimension. For instance, human actions in videos are usually defined by a series of specific movements depicted in consecutive video frames over time, as such consistency and transformation in the visual context can only be presented in the format of videos. This temporal characteristic of video data also influences video-based applications, which often require additional understanding and analysis of temporal elements (*e.g.*, actions, motions, optical flow) [46], [230]. While the conventional image representation encoded by neural networks can be applied to individual frames, extracting video representations necessitates addressing the connections between temporally related frames. An intuitive and classic approach to learning video data representation is to extend the conventional 2D convolutional neural network into 3D architectures with an additional temporal dimension, a notable example is the I3D model [30] proposed for action recognition in videos.

Regarding video-based applications, the tasks are similar to those in the image domain, where the most popular discriminative tasks include video classification (sometimes referred to as action recognition) [30] and segmentation [241], and the generative tasks that seek to directly synthesize videos [243]. For the latter, Sora from OpenAI [24] stands out as a most recent large video generator.

2.2 Audio

Traditionally, the study of audio processing has predominately resided within the research field of digital signal processing. In this survey, we focus on introducing three primary types of audio data: speech, music, and ambient sound.

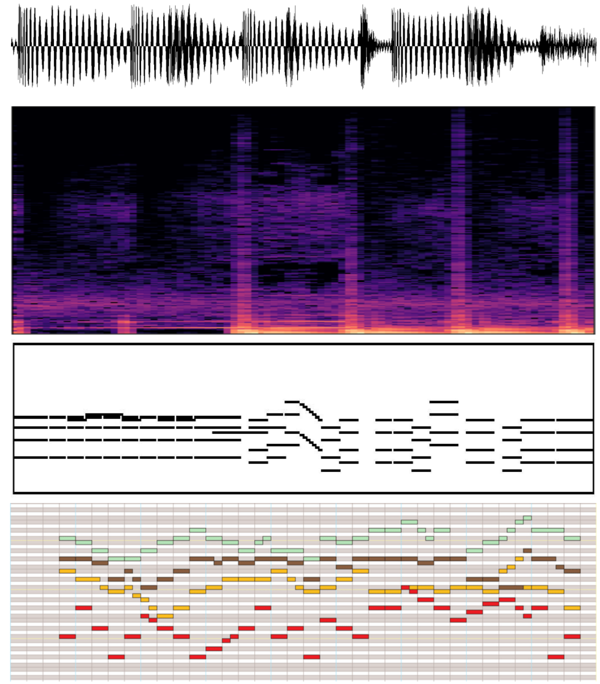


Fig. 2. **Illustrations of different audio data representations.** From top to bottom: (a) raw audio data in waveform; (b) audio data in mel-spectrogram; (c) music piece in 1D piano-roll from [72], where the horizontal and vertical axis represent the timestamps and the audio pitch, respectively; (d) music piece in MIDI from [184], where colors represent different instrument types.

Each of these audio types holds relevance and applicability in various multimodal task applications, further emphasizing the diverse nature of audio data within the context of multimodal learning. Similar to visual data, audio signals are a form of "raw data" that can be directly captured from the environment. However, unlike static images, audio signals possess inherent continuity in the temporal dimension.

Music. Music is a specific type of audio data that plays a significant role in our daily lives. As a form of expressive art, music is considered as the carrier and reflection of one's inner world. Generally speaking, the music itself has various genres such as traditional classical music, symphony, modern pop, country music, etc. Furthermore, music can also be classified into diegetic and incidental categories. Diegetic music is integral to the narrative, existing within the story's universe and perceived by its characters. In contrast, incidental music is intended solely for the audience's experience, underscoring emotions and scenes without being part of the story's world. The music categorization can often be subjective, with less strict and rigorous taxonomy for a specific genre. One common feature of those genres is that a music piece with high auditory quality has a relatively large sampling rate. For example, for the music with CD quality, the sampling rate is 44.1kHz [67], which leads to over 2 million data points for a one-minute musical piece.

From the perspective of scientific research, the high dimensionality of musical audio waveform imposes difficulties in data processing, therefore, researchers have developed different forms of musical data representations. In this survey, we classify the musical data representations into *"non learning-based"* and *"learning-based"* depending on whether the representations are obtained via deep learning techniques. For the *"non learning-based"* musical representations, we can further divide them into continuous and discrete subcategories. The most general non-learning based continuous data format for audio, including the music, is the waveform as shown in Fig. 2(a). The waveform is a two-dimensional data that depicts the sound pressure variation measured by the air vibration in the time domain. Another popular and general type of audio representation is the spectrogram. Compared to the waveform that emphasizes the temporal variations of the audio signals, spectrogram also reflects the frequency content of sounds over time, as illustrated in Fig. 2(b). In most cases, we refer to the waveform as the raw audio data. 1D piano-roll [72] and 2D Musical Instrument Digital Interface (MIDI) are classic discrete representations [22]. As illustrated in Fig. 2(c), piano-roll is a sparse data representation format, where the horizontal axis is the time stamps, and the vertical axis represents the acoustic pitch. 2D MIDI can be interpreted as a composed piano-roll format with instrument types, as represented by different colors in Fig. 2(d). Both 1D piano-roll and 2D MIDI discrete forms can be decoded back to raw audio space by pre-defined music synthesizers. On the other hand, the *"learning-based"* representations can be similarly specified by their discrete and continuous natures. Recent progress in deep learning has introduced novel learning-based discrete representations, namely the *vector quantization (VQ)*, to further reduce the high dimensional data into a discrete token space [179], [197]. Continuous learning-based musical representations share similar attributes as those in vision data, where we usually adopt neural networks such as CNNs to encode the raw audio signals to an embedding feature with desired dimensions.

Compared to other data modalities, music audio signals have several unique features to consider while applied to specific downstream tasks. Firstly, the musical data is a sequence where the temporal coherence within a complete music piece should be emphasized. Also, in addition to the temporal dimension, the audio data are often characterized by their frequency features in the form of spectrogram representation. Other than temporal coherence, rhythm is another important unique music characteristic to consider while assessing the music quality.

Speech. Speech primarily refers to audio signals of spoken languages, closely related to natural languages with their intrinsic correspondence. The data representations for speech are similar to music, where the waveform and the spectrogram are commonly used types in the non-learning based category. However, one noticeable uniqueness of speech audio lies in its natural correlation to language, where discrete representations of speech audio align with language tokens. As a result, the discrete language token-driven representation of speech features a more unified format compared to the learning-based VQ representation used in music audio. This characteristic also influences the methodology design when it comes to architecture selection, as detailed in the following of the paper.

In the application areas of speech, classic tasks like speech separation, which aims to isolate individual speech tracks from a composite audio mix, are well-studied [14]. Another pivotal task is automatic speech recognition (ASR) [295], which focuses on converting spoken language into text. ASR systems are designed to accurately transcribe human speech, making them essential for voice-activated interfaces and transcription services. At the same time, due to the intrinsic correspondence nature between speech and language natures, speech data are often applied in multilingual translation problems [70] or cross-modal translation between the audio and text [49]. More recent works in the multimodal generative area also look into the speech generation from visual input with talking lips [124].

There exists another special type of speech-language for the hard-of-hearing and speech-impaired community, which is the "sign language". Unlike audible speech, sign language necessitates the interpretation of visual signals from gestures, thereby exhibiting natural connections with visual and motion data. Research focused on sign languages explores tasks such as sign language recognition and generation [21], [196]. Essentially, sign language recognition aims to translate specific hand gestures into textual data, whereas generation addresses the reverse process. Although sign language is discussed within the "speech" category, the datasets [74], [172] typically comprise visual data, such as images and videos, accompanied by language annotations.

Ambient Sound. Beyond speech and music, there are other types of audio signals, such as the sound accompanying certain events, which we refer to as *"ambient sound"* in this survey. Compared to music, which is subjective, and speech, which is closely related to natural languages, ambient sound is more frequently encountered in conjunction with videos to characterize specific actions and events. For instance, we naturally relate the sound of a crying baby with a video showing the corresponding visual scenes. This unique correspondence enables ambient sound to provide additional information to conventional video action recognition tasks, leveraging the audio modality [113], [239]. The aforementioned audio representations are also applicable to ambient sound.

However, in contrast to music and speech audio signals, ambient sound exhibits a more noisy nature with less preprocessing. Unlike music, which can be represented using highly processed data formats like musical MIDI, and speech, which benefit from natural correspondence to text, the representation of ambient sound is more ambiguous. It lacks explicit formats like discrete token representations in speech or specific features like rhythm in music. These characteristics contribute to the inherent ambiguity and challenges in representing ambient audio.

2.3 Text

Text has been studied within the Natural Language Processing (NLP) communities from the early years. While there exist diverse formats of textual data, in this survey of multimodal learning, we mainly focus on introducing several types of textual data that exhibit close connections with other data modalities. The NLP community has witnessed significant attention in recent years, especially with the remarkable success in developing Large Language Models (LLMs) such as GPT-3 [193]. The tremendous achievements in NLP are closely linked to the nature of textual data and language. Unlike visual and audio information, which can be considered as "raw data", textual data undergoes substantial processing. More specifically, it is a data type that has evolved through human civilizations, characterized by a highly unified format and precise semantics despite linguistic differences. It signifies the fact that text is highly informative and compact, while visual and audio signals usually contain abundant information redundancy. Another unique characteristic of text on the application side is that the problem formulation of most NLP tasks can be unified under the notion of "next word token prediction". This formulation represents a common underlying structure in various NLP tasks, which contributes to the coherence and consistency within the field and its potential to solve multiple tasks via a large foundation model [19].

Captions. Captions provide sentence descriptions that summarize either the entirety or a portion of the visual content in vision and text-related multimodal works. They may consist of single sentences or extend into longer paragraphs composed of multiple sentences. Bag-of-Words (BOW) [168] is a classic form of text representation, representing a text corpus as a multiset (*i.e.*, a bag) of its words. Following the development in deep learning, the captions are also frequently processed by Recurrent Neural Networks (RNNs) (*e.g.*, LSTM [105]) with an internal memory state to obtain the learning-based representations. Compared to the visual data processed by CNNs, the memory state design in RNNs allows for establishing the recurrent connections among consecutive words in a given sentence to better interpret the overall textual features. A more recent breakthrough in the NLP community is the success of BERT (Bidirectional Encoder Representations from Transformers) [66], which is a large-scale pre-trained model for word embedding. This unique characteristic has been widely addressed in the subsequent research studies involved with captions.

Dialogue. Dialogue is another common form of textual data in multimodal machine learning, distinct from captions due to its inherently interactive nature, which involves conversation among participants with logical coherence rather than a unilateral description of visual content. Therefore, when handling dialogue data, special attention should be paid not only to the words within a sentence like plain captions but also to the connections between different sentences within a complete dialogue. In the literature on multimodal learning for vision and language, these unique characteristics of dialogue are often addressed by constituting an additional data component - history - in the design of the framework. This component typically captures the flow of the conversation, including prior exchanges, and is processed by dedicated mechanisms that operate alongside the processing module for individual sentences.

Question and Answer. A more specific categorization of textual data is *question and answer*. While its representation is overall similar to its other textual counterparts (tokens), they are often leveraged in vision-language tasks as an approach to study the visual reasoning ability of networks or to evaluate specific task performance. Visual question answering (VQA) [9] is a representative task that uses question and answer to reason the visual context. Question and answer are often closely related to the dialogue, as the interactions within the dialogue can take the form of questions and answers.

ASR and OCR Text. While captions and dialogue are correlated with the visual context in a more high-level and semantic manner, automatic speech recognition (ASR) and optical character recognition (OCR) based texts present a slightly different connection to the audio and visual information. Specifically, ASR and OCR are foundational multimodal research topics that have matured over decades, they exhibit precise correspondence between text and other data modalities [80], [117]. In addition, OCR also serves as a technique to obtain textual data from textual corpus.

2.4 Other Modalities

Multimodal learning encompasses a wide array of data modalities beyond vision, audio, and text. For instance, 3D data represents a significant category, including subcategories like point clouds and meshes. This survey focuses on exploring data modalities with cognitive significance that mirror the human perceptual system. Therefore, we categorize data modalities other than vision, audio, and text together, highlighting their relationships with these primary modalities for a more integrated understanding and presentation structure.

Graph. Graph data offers structured representations of relational information via nodes and edges which captures the connections and interactions between elements. While it may not be a data modality that naturally exists through a human perception system, it plays a significant role in machine learning when connected to other data modalities. For example, the scene graph establishes a graphical representation from the images to interpret the connections among objects. A typical example of a multimodal application that leverages both visual and graph data is the scene graph generation from the visual context [142], [157]. The non-Euclidean nature of graph data also inspires the design of graph neural networks [274], [299], which serve as a powerful model architecture to process graph data.

Optical Flow. The concept of optical flow has been first proposed in the last century as a measurement to characterize the movement of objects in a visual scene caused by the relative motion between an observer and a scene [108]. With the advancement of computer vision, especially with the deep learning techniques, the optical flow has also been studied together with visual data [236], [251], [257], [264]. Compared to other motion data formats, optical flow is usually defined in a more precise manner via the pixel-level change within the consecutive image sequences. However, the calculation of optical flow itself has been a quite challenging research problem, due to the fact that environmental lighting can also impose large effects on the pixel values of images. Overall speaking, optical flow can be considered as a specific motion presentation explicitly derived from visual information.

Point Clouds and Meshes. Point clouds and meshes are both important forms of 3D data, providing spatial and structural information that enriches our understanding of physical environments. While point clouds are collections of vertices in a three-dimensional coordinate system, meshes further build on this by connecting points with edges and faces, creating a comprehensive model that represents the shape and topology of 3D objects. Like the other data modalities discussed in this section, point clouds and meshes are not directly captured by our sensory systems but are constructed through processes that often incorporate human insights.

Human-Centric Motions. Human motions are often used to define various daily activities. 2D skeleton data of human body is a common representation form for human motions, which captures the keypoints of the human body and represents them as axis coordinates within an image. They can be used to define various daily activities, which can be practically applied in real-life applications. For instance, automatic detection of human actions is particularly valuable in human-centered assistant systems, such as the health assistant to detect falls for elderly people. Typically, we can extract 2D skeleton data for each frame via pre-trained networks such as OpenPose [28], [29]. 3D human motion data can usually provide richer information with the extra data dimension. A classic form of 3D motion data involves incorporating depth information alongside the conventional 2D keypoints data. Besides the RGB camera-based methods mentioned for keypoint acquisition, keypoints can also be derived through alternative approaches within the keypoint detection domain, such as geometric reasoning from 3D data [231] and SLAM techniques applied to lidar sensor data [226]. Furthermore, there are other forms of 3D motion representations that are more frequently adopted in the Computer Graphics (CG) field, such as the Skinned Multi-Person Linear Model (SMPL) [155]. SMPL integrates skinning and blends shapes to represent human bodies. Compared to optical flow, which captures the motion of all pixels between two frames, keypoint movement, on the other hand, tracks specific points of interest across frames, allowing for a more focused analysis of object or feature dynamics. Meanwhile, 3D video features extend this analysis into the spatial domain, integrating depth information with motion to provide a richer, more detailed representation of the visual structure and movement patterns.

3 Multimodal Representation Learning

In this section, we focus on multimodal representation learning studies. We structure this section into three parts: the introduction to several popular network architectures and evaluation, the supervised learning setting, and the non-supervised setting. The rationale behind this categorization is based on the fact that the multimodal representation learning field has gone through a shift from conventional supervised representation to large-scale pretraining. Classic methods under the supervised learning setting usually require fully annotated data to train the networks, thus imposing limitations on the size of available training datasets due to the tedious human labor work for labeling.

To overcome the bottleneck, the research trend in the multimodal representation learning has turned to *"non-supervised"* setting, using data that do not necessarily require human annotations. These datasets are often collected directly from the Internet and consist of paired data from different modalities. It is important to note that while these datasets possess intrinsic correspondence between modalities, they are considered non-supervised in this survey due to the lack of manual labeling. Notably, these non-supervised approaches benefit from larger dataset sizes and have witnessed an increase in model scale. Therefore, in Section 3.3 for non-supervised representation learning works, we mainly focus on introducing the large-scale pre-training studies, which have been attracting much research attention in recent years. The primary research objective in representation learning under the multimodal context is to learn an effective and discriminative mapping between the corresponding data representations from multiple modalities.

3.1 Network Architectures

We introduce several popular network architecture backbones for learning data representations of the above-mentioned principal data modalities (*i.e.*, visual, audio and text). Nevertheless, we note that there exists other popular network designs for specified data modality, *e.g.*, Graph Neural Networks (GNNs) [205] for graph data and PointNet [188] for point clouds.

Convolutional Neural Networks (CNNs). As one of the most classic network architectures in the computer vision field, CNNs [95], [131], [216] have been widely adopted as the backbone architecture in representation learning for visual data. The core idea of CNNs is to extract high level data representations from the raw data via complex functions composed of convolutional layers and activation functions. Similarly, the same idea has also been adapted in learning data representations for audio signals [96]. In the context of representation learning for classic visual and audio data, the training of CNNs often leverages the multi-class cross-entropy loss for classification tasks using:

$$l = -\sum_{c=1}^{N} y_c \log(p_c), \tag{1}$$

where y_c is the class label, and p_c denotes the predicted probability. Then features extracted from the last layer of the CNNs are further used as the actual data representation.

Recurrent Neural Networks (RNNs). A specific demand for learning data representations for natural languages is to consider its temporal correlations with sequential order for words. Therefore, NLP community follows a different vein of network architectures to address this challenge using recurrent neural networks (RNNs) and LSTM [85], [105]. Efforts have also been made to learn audio data representations via RNNs [77].

Transformers. Transformers [245] have gained great popularity in machine learning community in both computer vision and natural language processing areas. The core technical design of Transformers is the self-attention mechanism, which operates on sequential data to learn the overall information. Compared to CNNs and RNNs, Transformers have several distinctive advantages in model designs: the flexibility to deal with sequential data with variant lengths; the efficiency to allow for parallel computing instead of following the sequential processing as for RNNs. Despite the fact that the Transformers are initially designed for NLP tasks [25], it has been successfully applied in representation learning for visual [94] and audio data [240].

Recently, Mamba [87] comes out as a new popular model that shows promising downstream performance on long language and audio sequence processing compared to Transformers [245]. One of its key strengths is to address the computational challenges by incorporating a selective mechanism into the state space models.

For the representation learning of textual data, commonly referred to as Language Model Pre-training, a widely used problem formulation involves "next token prediction", which frames learning as a joint conditional probability challenge.

As a result, the foundational approach for optimizing neural networks in this context frequently involves maximizing the likelihood, typically achieved through the use of cross-entropy loss. However, it is worth noting that various other objective functions have been introduced to further enhance the modeling ability of neural networks.

Overall, data representation learning has always been an important research direction in machine learning, and it is a topic that lies within the upstream of the research pipeline. Consequently, the evaluation of multimodal representation learning methods usually relies on concrete downstream tasks, which we will present in detail in Section 4.

3.2 Supervised Learning

Supervised setting requires the annotations from multimodal sources to guide the learning process, which is also the most classic representation learning setting [65], [300].

Generally speaking, there are two approaches that are widely adopted for the supervised representation learning. One possible way is to establish mapping after having obtained the data representations from their respective feature space, which can be considered as a two-stage method denoted " representation learning in individual modality domain + mapping among modalities", usually with fixed backbone models for the first feature extraction stage [151], [167], [263]. Alternatively, another way to tackle the multimodal representation problem is to learn a unified representation for a given data pair in an end-to-end manner, freely optimizing the feature extraction backbones [10], [258].

For the first approach, Liu *et al.* [151] exploit the existing pre-trained semantic embeddings from visual content, and propose a collaborative experts model to aggregate the multimodal information. [167] learns the text-video embedding from heterogeneous data with a Mixture-of-Embedding-Experts (MEE) model. Wang *et al.* [263] focus on global-local sequence alignment of video and textual representations. For the second approach, [258] proposes to learn two-branch neural networks for matching the text and image data. In the audio-visual domain, [10] learns the mutual representation via a "audio-visual correspondence" learning task.

3.3 Non-supervised Learning

In contrast to supervised learning where exhaustive manual annotations are required in training, there are other paradigms to learn data representation in the multimodal context. In the existing literature, "unsupervised", "weakly-supervised", and "self-supervised" are representative terminologies to describe the setting with slight nuance. Specifically, "unsupervised" often refers to the network training without human supervision, "weakly-supervised" describes the case where the supervisions may be noisy, limited, or imprecise; "self-supervised" is used to describe the model trains itself to learn one part of the input from another part of the input. In this section, we refer to them as "non-supervised" only for easy structure and presentation purposes.

The fundamental idea for non-supervised learning setting relies on the premise of the intrinsic synchronization nature among paired data from multiple modalities [192], [290], [319]. For example, certain video actions are naturally accompanied with characteristic sounds as described in the ambient sound section in Section 2.2. The images and captions are also paired to train the vision and language models.

Nowadays, there are several large-scale pre-trained models in multimodal learning research area, especially in the text-image domain, that have been attracting much attention, due to their impressive performance as well as the wide downstream applications [160], [173], [195]. We can consider the large-scale pre-training as a specific type of multimodal representation learning, since the primary goal of pre-training is to learn a joint and unified cross-modal representation that can be flexibly transferred to other domain or downstream tasks.

There are in general two popular methods for the pre-training field, which are contrastive learning based [4], [192] and mask reconstruction based [45], [135], [158]. One of the most popular examples of such models include CLIP (Contrastive Language-Image Pre-Training) [192] for the vision and language pre-training. Most of those models are developed following the BERT [66] and GPT (Generative Pre-trained Transformer) models for natural language [25] and images [41], whose the core design consists of transformer architectures [245] pre-trained for text and image generation tasks. Inspired by the success of GPT models in showing the potential to use the language or image to guide a large neural network to accomplish a variety of generation tasks in their respective domains, researchers naturally proceed to the multimodal area to bridge these modalities. The CLIP model is trained on 400 millions text-image pairs, and is considered as one of the first large-scale pre-trained models for the multimodal learning area to bridge the text and image data space. Another example is VATT (Video-Audio-Text Transformer) [4], which is a transformer-based self-supervised large-scale model for learning representations from raw video, audio and text. It first processes raw data from different modalities via linear projection, and trains the model to learn a semantically latent space via the Noise Contrastive Estimation (NCE). One commonness among those pre-training works is that the proposed models are trained with enormous amount of data using extensive computational resources. From the technical point of view, CLIP follows the general idea to align the embedding space of paired images and the corresponding textual descriptions. It adopts the batch construction technique [218] by encoding the entire sentence description as an entirety, instead of processing the textual word by word. CLIP jointly trains a text and an image encoder by optimizing the similarity scores of a given pair. During the inference time, the model can be used for zero-shot prediction by embedding the names or the descriptions of the target dataset's classes in textual form via the learned text encoder.

It is worth noting that while those large-scale models [4], [192], [193] are able to achieve very impressive results, there are few radical novelties in the model architecture and training techniques. Therefore, while they have received large attention, there are also controversies regarding those works. One of the discussions over such large-scale pre-trained models is that the impressive results are largely due to the diverse and enormous data that have been carefully designed to train the model, as well as their scale on the existing models on large. Other concerns over the privacy and ethic issues have also been raised against these works. Overall speaking, despite the controversial voices over the topic, those models do help with building a more unified toolkit in connecting the visual and textual space in the multimodal learning area, which also promotes large number of following works that are developed based on the aligned feature space for various downstream tasks.

3.4 Trend in Representation Learning

The machine learning and computer vision research community is rapidly advancing, with a trend towards scaling up data representation learning using emerging foundation models, empowered by the upgrade on the dataset scale and computational resources. The multimodal representations learned by large pre-trained models such as CLIP [192] have been successfully applied in various multimodal downstream tasks, boosting the performance, especially in the axes of generalization ability of models.

However, we also want to emphasize that scaling up is not a panacea. Despite the benefits, fundamental issues persist, such as out-of-distribution challenges and amplified model bias [253]. While large pre-trained models are powerful for many multimodal tasks, future research needs to focus on real-world scenarios with more edge cases and complex data formats for safe and responsible deployment.

4 Discriminative Applications

In this section, we discuss the multimodal learning works for discriminative task applications, with subsections categorized with specific data modality combination in form of "Vision+X", where X stands for the additional data modalities.

For discriminative applications, popular approaches usually inherit the neural networks from general representation learning in Sec. 3.1, with additional modules to adapt to task-specific objectives. A general methodological design in multimodal learning follows the ideas of *"separate processing"* and *"unified fusion"*. To be more specific, data of different modalities are first processed with respective network branches, and then the inter-modality learning is further performed by extra mutual modules before outputting the final results for different tasks. Since the exact objectives depend on the task scenarios, we leave the detailed introduction in the following subsections. In terms of the evaluations, different multimodal tasks have their corresponding evaluation protocols. Similarly to the specific methodology design, we detail the evaluations in the subsections below.

4.1 Vision+Audio

Audio-Visual Event Localization (AVEL). The Audio-Visual Event (AVE) is defined as an event that is both audible and visible in a video segment [239], and the AVEL task aims to localize the AVE within an unconstrained video [73], [147], [239], [273], [285]. This task was first proposed in [239], together with the AVE video dataset (details in Table 1 and Appendix A). The overall task objective resembles the action recognition with ambient audio data and the requirement for temporal localization under supervised or weakly-supervised settings. To tackle the additional ambient sound accompanying an event, a common method is to achieve the cross-modal interactions via different attention modules [73], [239], [273], [296]. Many of the existing works follow the framework that processes audio and visual data with separate encoders, and fuses the processed information for temporal localization and activity classification. The temporal connection from the video stream is often addressed via the model backbone such as LSTM [105]. Evaluation of the AVEL task usually utilizes the prediction accuracy metric.

Audio-Visual Video Parsing (AVVP). The AVVP problem aims to parse a video into temporal segments and label them as either audible, visible or both [146], [238], [271]. It is initially developed from the AVEL task, with its task emphasis on the recognition, while the AVEL focuses more the temporal localization. As a variant of the AVEL task, several works have been developed with the common core idea that seeks to learn an effective audio-visual feature as the foundation, and then to incorporate further refined technical designs to address specific task requirements. For instance, Lin *et al.* [146] introduced a sequence-to-sequence manner integration of audio and visual features. Yu *et al.* [296] explore the AVVP task by taking the potential audio-visual asynchrony into account.

Visual Sound Source Localization (VSSL). Visual Sound Source Localization (VSSL) task aims to locate the corresponding visual locations within an image given the sound [180], [190], [191], [209], [210], [223]. While the original sound source localization task (SSL) has been widely studied in the signal processing field [86], the deep learning based visual localization was first proposed in [209]. The high-level idea also focuses on learning the correlations between paired audio and visual data, except for the visual part, the VSSL task tends to switch the regions of interest within the visual data given different ambient audio signals. Overall pipeline often consists of separate encoders for visual and audio input, and then fuses the audio-visual information for learning a localization module during training. More technical details may differ in terms of how to perform the fusion via attention mechanisms [190], [209], the training technique using various localization or contrastive loss [190], [191], [209]. To assess the performance of the VSSL task, metrics such as the cIoU (Complete IOU) and AUC (Area under the ROC Curve) scores are often used to quantify the precision of the predicted areas for sound sources.

4.2 Vision+Text

Visual Grounding. As a popular discriminative vision and language task, visual grounding aims to localize the object within an image given a text description as input [61], [63], [78], [106], [112], [140], [148], [153], [214], [276], [291], [292], [298]. The idea of realizing cross-reference between sentences and visual context was first proposed and studied in [122], where the task was also known as *"referring expression comprehension"*. Pioneering works on the referring usually only require the grounding of a single object from the description sentence input [110], [161], [297], working on the premise that the region of interest is expected to achieve the maximum posterior probability for the given textual description. More recent works tackle a more challenging visual grounding setting, where the task is refined into two sub-objectives: phrasing and grounding. For the first sub-objective, models are expected to localize all the objects mentioned in the given textual description, and then to individually detect their corresponding boxes in the image [152], [186].

In terms of methodology designs for the visual grounding task, most works [61], [63], [78], [106], [110], [112], [140], [148],

[153], [161], [214], [276], [291], [292], [297], [298] can be categorized into supervised, weakly-supervised and unsupervised settings. The supervised setting refers to the condition where the annotation of phrase-object pairs is provided, the weakly supervised removes the phrase annotations for the textual description input, and the unsupervised setting completely removes annotations for both data modalities. As for general pipelines, most methods follow either a two-stage or one-stage framework. For the two-stage frameworks, models first extract region proposals for potential objects within the image, and then rank and match the proposals with language phrases. For the one-stage frameworks, visual objects and textual phrases are aligned and connected during the learning process to avoid redundant region proposals as in two-stage designs. In the case of weakly- or unsupervised settings, some extra regularization losses such as structural loss and discriminative loss [229], [276] are usually needed to better learn the correlations between corresponding object regions and textual phrases. The evaluation for the visual grounding resembles other visual localization tasks, which often use the IoU (intersection over union) between the predicted and ground truth boxes as a quantitative measurement, with a threshold value of 0.5. Another unique metric for the visual grounding task is *PointIt* (pointing game metric) [276], which computes the pixel location with maximum predicted attention weight, if the selected hit point lies within the ground truth box area, the prediction is counted as valid.

Temporal Activity Localization (TAL). Activity localization task (TAL) is also known as the video grounding, which seeks to locate the temporal segment of a video clip given the language description of a certain activity as query [8], [43], [44], [81], [261], [305], [305]. Compared to the visual grounding within the image, TAL requires additional reasoning and matching along the temporal direction as indicated by its name. For this task, the models are expected not only to capture the correlations between the visual activity and the language, but also able to temporally localize the segment among consecutive video frames. While the high-level framework structure remains similar to previous multimodal discriminative tasks with separate encoders, a multimodal processing module for fusing the features, a decoder module adapted for specific task objectives, and early representative works for TAL task introduce different techniques to emphasize the network ability for temporal reasoning. Gao *et al.* [81] propose a Cross-modal Temporal Regression Localizer (CTRL) with a Temporal localization regression network to align the fused visual-textual information with video temporal locations. The popular quantitative metrics used for evaluation include the *mean IoU* and *IoU@a*, with *a* standing for the percentage of overlap between the predicated segment and the ground truth annotations.

Visual Entailment (VE). Visual entailment (VE) seeks to predict the logical relationship of a piece of text to an image [237], [278], [279]. It is developed from the textual entailment task [52], whose initial objective is to decide if a hypothesis can be logically deduced from the premise. Xie *et al.* [278] extends the textual entailment to the multimodal context, which replaces the textual premise with an image. Thomas *et al.* [237] further refines the task by introducing different levels of granularity. The emphasis of the VE task lies within the multimodal reasoning ability of networks. To achieve the reasoning between the image and textual hypothesis, early methods [278], [279] adopt separate network branches to process visual and textual data and leverage the attention interaction for interactions. A refined framework [237] further disentangles the textual hypothesis into its constituent parts and proposes to enhance the reasoning by introducing an abstract meaning representation (AMR) graph for the decomposed textual components. The performance is often evaluated via prediction accuracy given the premise as input.

Spatio-Temporal Video Grounding (STVG). Spatio-Temporal Video Grounding (STVG) is a recent multimodal task that lies at the intersection of visual grounding and temporal localization, integrating reasoning among space, time, and language within the visual context of videos [119], [228], [287], [307]. Specifically, given an untrimmed video and a textual description of an object, the task seeks to localize a spatio-temporal tube (*i.e.*, a sequence of bounding boxes) for the target object described. Most existing methods for STVG either build upon the idea from visual grounding or focus on temporal localization designs. One popular paradigm for tackling the task adopts the two-stage design, which leverages pre-extracted object proposals and then integrates the temporal localization via attention mechanisms [260], [289]. In the meantime, another thread of works proposes one-stage frameworks and does not rely on prior for object proposals [121], [287]. In terms of network architectures, Transformers are widely adopted as the backbone for such methodology designs [228], [287], [307]. STVG is usually evaluated via IoU metrics by comparing the frame overlap between the ground truth and the predicted timestamps.

4.3 Vision+Audio+Text

Multimodal Retrieval. Another multimodal discriminative task that has been widely studied is the retrieval [47], [88], [255], [256], [266], [311]. Most retrieval works operate on the representation space by measuring the similarities among learned representations from different modalities. As consequences, retrieval task is also one of the most frequently used downstream tasks in representation learning works.

While the retrieval task can be conducted within a single modality of data, multimodal retrieval seeks to extend its original setting to cross-modality scenarios, where we want to retrieve items that match the input from a different data modality, *e.g.,* text based vision retrieval, audio based vision retrieval. For example, CAMP [265] learns the text and image embedding via message passing across modalities. Gu *et al.* [88] propose to improve the text-visual retrieval with auxiliary generative models. [263] looks into the task of text based video retrieval by additionally looking into the local details via a global-local alignment method in the learned representation space. [319] learns a mutual audio-visual latent space via VAE-based framework for audio-visual cross-modality retrieval. Oncescu *et al.* [178] propose to retrieve audio signals given natural language queries.

Audio-Visual Question Answering. Audio-Visual Question Answering, as indicated by its name, is an extension based on visual question answering with integrated audio modality [136], [137], [288], [301]. Specifically, AVQA often involves questions regarding different visual objects, sounds, and their associations in videos. Existing methodology designs are often extended from VQA frameworks with extra interactions with audio data. For instance, an intuitive framework [137] expands the two-branch encoder design into three-branch and separately processes video, audio, and textual data before introducing interactions via the attention mechanism. Answer prediction accuracy is often used for evaluation.

5 Generative Applications

In this section, our focus is on cross-modality synthesis tasks for generative applications. These tasks involve generating data from a specific modality or multiple modalities as input.

There are usually typically two high-level approaches to generating data in cross-modality synthesis tasks: retrieving an item from a given database, or directly synthesizing and decoding data via the neural networks. For the retrieval-based generations, the core idea follows the logic to search for an item or several items that mostly resemble the "generated" data. A large percentage of the retrieval-based works perform the similarity measurement on the data representation level without actually considering the decoding part. Technically, we argue that such works are categorized in the *Representation Learning* section. As consequence, we mainly focus on introducing the works that "truly generate" data instead of retrieving items in this section for generative applications.

5.1 Generative Networks

Before diving into the concrete application tasks, we first introduce three popular backbone models for general generative tasks that have been widely adopted in multimodal generative literature.

VAE-Based Models. Variational Auto Encoders (VAEs) [126] are classic generative models proposed based on the deep neural autoencoders [100] under the unsupervised learning setting. The core of autoencoders relies on the premise that an effectively trained encoder should learn the data representation in a way that the encoded representations can be decoded to reconstruct the original data input by a decoder. Compared to conventional autoencoders, VAEs introduce regularizations on the bettleneck level by re-parametering the latent space using the Gaussian priors, where the learned Gaussian parameters allow for sampling new data, as illustrated in Fig. **??**(a). Typical training of VAEs often includes two types of losses, which are the variational loss (ELBO) [126] that consists of a regularization loss on the latent representation space (*e.g.*, Kullback–Leibler divergence) and the reconstruction losses on the output data (*e.g.*, Mean Square Errors (MSE)). The classic variational objective can be formulated and derived from the following equation:

$$\log p(x) = \mathbb{E}_{q(x|z)}[\log p(x|z)] - D_{KL}[q(z|x)||q(z)], \tag{2}$$

where p represents the decoder, q is the encoder, x and z denote the original raw data and the learned latent embedding, respectively. In the actual implementation, the re-parameterization technique, which assumes $z \sim \mathcal{N}(\mu, \sigma)$, allows us to minimize the KL divergence based on sampling from N samples:

$$\sum_{i=1}^{n} \sigma_i^2 + \mu_i^2 - \log(\sigma_i) - 1. \tag{3}$$

VAEs have been widely exploited in various generative tasks in audio and images [99], [126] as well as in the multimodal context for cross-modality generations [224], [319].

GAN-Based Models. The generative adversarial networks (GANs) [83] are another mainstream backbone type for various generative models. From a high-level perspective, GANs involve two agents (*i.e.*, the generator $\mathcal{G}$ and the discriminator $\mathcal{D}$) playing an adversarial game. The generator aims to synthesize realistic data that resemble real data for fooling the discriminator, while the goal of the discriminator is to distinguish between the synthesized data by $\mathcal{G}$ from the real data. Similar to VAEs, the training of GAN-based models does not require external annotations but only the real raw data, and therefore is frequently used in unsupervised or weakly-supervised settings. The standard training of GANs also minimizes losses from two aspects with latent space regularization (also known as adversarial losses) and reconstruction optimizations [83], [204]. Following the original work, multiple variants of the GAN models and adversarial losses have been proposed such as the Wasserstein GANs with the Wasserstein loss [11], [90] and conditional GANs [170]. The classic GAN loss is formulated as follows:

$$\begin{aligned} &\min_G \max_D V(D,G) = \\ &\mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_Z(z)}[\log(1 - D(G(z)))], \end{aligned} \tag{4}$$

where G and D are the generator and discriminator, respectively. x and z denotes the original raw data and the learned latent embedding. On the application level, GANs are first widely applied in image generations [23], [114], later studies also explore the GAN-based models for audio synthesis [127], [132] or cross-modality areas [302], [316], [317].

DPM-Based Models. Compare to VAEs and GANs, the diffusion probabilistic models (DPMs) [217] are another type of generative backbones that have been very popular in more recent years. In principal, the DPMs include a Markov chain of a finite steps in two opposite directions. The forward direction, also known as "diffusion" process seeks to gradually add noises to a given data at each diffusion step, while the inverse denoising process aims to remove the noises added in the forward steps and recover the actual data from a non-informative noisy distribution. There are two variants of conventional DPMs that differ in the state space formulations of the Markov chain. Classic DPMs suppose the state space to be continuous, and parameterize the diffusion process with Gaussian noises [68], [101], [125], [128], [175], [219], [220], [221], while another variant of DPMs consider the discrete state space and formulate the diffusion process with state transition matrix [12], [89], [320]. The variational lower bound [101] is the classic loss function used for effective DPMs learning, other practical losses such as the auxiliary loss [12], [89], classifier-free guidance [103] and contrastive diffusion losses [320] are also proposed to further improve the generation performance. Vanilla DPMs are trained on a variational lower bound defined as follows:

$$\begin{aligned} \mathcal{L}_{\mathrm{vb}} = \mathbb{E}_q[&\underbrace{D_{\mathrm{KL}}(q(x_T|x_0)||p(x_T))}_{\mathcal{L}_T} + \\ &\sum_{t>1}\underbrace{D_{\mathrm{KL}}(q(x_{t-1}|x_t,x_0)||p_\theta(x_{t-1}|x_t))}_{\mathcal{L}_{t-1}} - \underbrace{\log p_\theta(x_0|x_1)}_{\mathcal{L}_0}], \end{aligned} \tag{5}$$

where q and p represent the diffusion and denoising processes, respectively. x_i denotes the data at diffusion step t. DPMs have received much research attention due to the competitive performance in generative tasks for images [68], [101], [102], [109], [175], audio [128], [134], [171], as well as in the cross-modality scenarios such as text-to-image [89], [174], [320] and dance-to-music generations [320].

The evaluations for the generative tasks have always been an important aspect to consider. Typically, the assessment of synthesized data in multimodal setting considers both unimodal and multimodal criteria. The unimodal metrics are utilized not only for multimodal scenarios but also for generative tasks in general, such as the fidelity in image generation. In addition to the general quality, the multimodal generations also take the cross-modality correspondence into account, such as the beat correspondence between video and music. We summarize the common evaluation metrics for various synthesized data in generative tasks in Appendix B.

5.2 Vision+Audio

Music Generation from Vision. Recent studies that seek to generate music from visual data (usually from videos) can be categorized by their adopted music representations. One branch of the music generation works [2], [69], [79], [227] rely on the symbolic audio representations such as 1D piano-roll and 2D midi and as we have introduced in Section 2.2. The symbolic musical representations can be decoded back to the raw audio waveform by pre-defined synthesizers that introduces no additional noises, thus enduring the high-quality of the generated music. This is especially true when compared to the learning-based musical representations and decoders [67], [317], [320], where the synthesized music usually has relatively high noise levels. Secondly, the computational cost of symbolic representation based works is generally lower than the pure learning-based methods, since the symbolic musical representations are very sparse and low dimensional, which facilitate the learning and inference process. However, such symbolic based music generation methods are also restricted in terms of the music diversity and flexibility. Especially in current research works, the generated music are usually limited to a certain specific pre-defined instrumental sound [2], [69], [79]. It is worth noting that most symbolic based music generation works, despite the fact that the framework output is raw music, do not directly use the generative backbones as we have introduced in Sec. 5.1. Technically, most of them are trained based on the ground truth MIDI annotations in form of cross-entropy loss. In contrast, another branch of research works deploy the learning-based musical representations either in form of continuous or discrete. However, despite the continuous music representations have been exploited in the music synthesis field [132], most recent cross-modal music generations adopt the discrete form of learned musical feature - vector quantization (VQ) - as the intermediate representations [317], [320], leveraging the large-scale pretrained music synthesis model JukeBox [67]. For example, D2M-GAN [317] proposes a GAN-based framework that takes the human body motion data and dance video frames as input, and generates musical VQ representation. CDCD [320] builds upon the diffusion probabilistic model with a discrete state space represented by the VQ, and incorporates a contrastive diffusion loss to train the network to improve the input-output correspondence for cross-modality applications.

Speech Generation from Videos. In addition to the music audio, another specific generation task seeks to synthesize speech audio from videos of human speaking [75], [76], [124], [166], [169], [187], [203], [248], [286]. One unique aspect about this audio generation task is that the speech largely relies on the movement of lips while speaking. Based on this characteristic, many works on this direction focus on reading and interpreting visual lip movement from video input and then convert it to audio waveform, which also explains the reason why this "video-to-speech" synthesis task is also known as "lip-to-speech" generation. Therefore, despite the topic of audio generation from videos, large percentage of the works in this area rather focus on the "motions in the videos", instead of the raw videos. To enhance the correlations between lip movement and speech audio, the audio-visual cross-modal attention mechanisms are further adopted to improve the generation quality. Kim *et al.* [124] propose an attentional GAN with visual context to read lips for speech synthesis. Yadav *et al.* [286] use the VAE generative backbone with the stochastic modelling approach. At the same time, more refined variant of this problem with disentangled speech features such as the individual speaking styles have also been studied [187].

Ambient Sound Generation from Videos. Research works that seek to generate sound from natural videos [36], [42], [314] put special emphasis on the alignment between the generated sound and the visual context, which consists of both the semantic and temporal alignments. Chen *et al.* [36] address the semantic alignment problem by adopting a perceptual loss and considering the sound categories in the optimization process. Zhou *et al.* [314] follows a rather classic encoder-decoder framework for video input and audio decoder, propose three methods including the frame-to-frame, sequence-to-sequence, and flow-based variants. In [42], the authors tackle both semantic and temporal alignment with the proposed *REGNET* framework, whose core technical design includes a visual encoder and an audio forwarding regularizer. Generally speaking, compared to speech and music, the ambient sound has relatively fewer unique attributes other than its correspondence with certain activity. Summarizing the above works, we notice the high-level technical ideas are rather general and resemble to the standard pipeline designs.

Visual Generation from Sound. As the inverse direction of the sound generation from vision, to directly generate the pixel-wise natural images or videos solely from the audio modality is a challenging problem. However, as a specific type of visual generation from sound tasks, synthesizing talking faces from speech audio [38], [222], [306], [312] is a relatively well-studied sub-field. Similar to the *"video-to-speech"* task, the visual information in this reverse direction rather emphasizes the motions of lips in a video clip. In most cases, the input for the talking faces generation task includes a reference image and a driving audio track. Early works [37], [222], [249] adopt a general pipeline with two separate encoders for the input and a decoder for synthesizing the talking videos mostly via the GANs-based generative backbones. More recent works seeks to refine and improve the synthesis results by splitting the previous architecture into hierarchical structures [38], [58]. In addition to the raw videos, more specific motion data such as the flow is also used to

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2024.3420239

further enable high-resolution generations [306]. Some works also seek to generate natural videos by reformulating the video generation problem into an motion generation task in the form of optical flow [33], [64].

5.3 Vision+Text

Caption Generation from Vision. One of the classic vision and language generative tasks is the image and video captioning [35], [40], [159], [254], [259], [270], [282], [294], which aims to generate a language textual description of the given visual data. In the early stage of deep learning based visual captioning [162], [247], the general pipeline usually consists of encoder-decoder framework with the CNNs and RNNs as the backbone architectures for the image encoder and text decoder. Later on, the attention mechanism becomes a popular technique [159], [272], [282] to enhance the correlations between the sentence descriptions and the corresponding visual concept. A large amount of research works follows the updated general encoder-decoder pipeline with an additional attention modules. Other than the general encoder-decoder framework, there are some other works that tackle the captioning task using techniques such as adversarial learning [53], or reinforcement learning [150].

In addition to the technical progress, there are also some works that bring novel insights on the task setting of caption generations. For example, [6], [107], [208] incorporates ambient audio from the video as an additional input data modalities to further assist the video captioning. Following the above works, [321], [322] propose a new setting where part of the visual data becomes inaccessible as input, they further propose a dialog process between two agents as a supplement to the missing visual input, whose final goal remains to be generating a precise and complete textual description for the video.

Dialog Generation from Vision. In addition to the captioning tasks, another form of text generation from visual input focus on the dialog texts instead of plain descriptions. We can further classify the dialog generation from vision into visual question answering (VQA) [1], [9], [39], [56], [213], [277], [280] and visual dialog [57], [60], [116], [189], [211], [268], [321], [322]. As described in Sec. 2, the main difference between the two subcategories lies within the fact that the former VQA task aims to answer a single question related to the visual input, while the latter expect to maintain the question-answer interactions for multiple rounds with internal logic.

VQA task was first introduced in [9], whose task objective is to answer a language question related to the visual content. Similarly to the captioning task, the mainstream frameworks also follow the encoder-decoder pipeline, usually equipped with separate encoders for the visual and textual data, and a decoder for generating the words in language. Attention mechanisms have also been widely used in literature for the similar purpose to enhance the correlations between corresponding features from visual and language domains. One uniqueness of the VQA task compared to the previous visual captioning is the potential bias problem in the task setting. Specifically, one (models) may not truly rely on the visual context to answer the raised questions. For example, given a question such as " what is the color of the sky ?", the answer is likely to be "blue" in most cases, or given a question that the answer is expected to be "yes" or "no", one (models) can simply guess between two options. Therefore, to address such spurious pattern and bias problem, recent works have been seeking to migrate the issue by analyzing the causal relations [39], [176].

As another popular vision and language task, visual dialog follows a similar development path after being proposed in [57], where the general encoder-decoder framework is further extended to include an additional encoder for the dialog history data. From a high-level point of view, the challenge of visual dialog compared to the captioning and VQA is that the multiple rounds of question-answer interactions may refer to different parts of visual context with an image or video as the dialog goes on, leading to a higher requirement for precisely referencing the key information between two visual and textual data. In the meanwhile, the bias issue for VQA remains in the visual dialog.

Image Synthesis from Text. The inverse direction that seeks to generate image data from given text conditioning has also been a very active research topic in multimodal learning field within recent years since 2016 [89], [139], [149], [199], [284], [303], [304], [320]. It is a more challenging task compared to the text generation from visions, due to the reason that the visual data usually are richer in context, with demanding requirement for pixel-level synthesis.

Following the preliminary exploitation in the text-to-image task from [199], the recent literature in the area can be divided into different stages following the chronological order. Before the year of 2020, the mainstream approaches use the GAN based architectures to tackle the problem [284], [293], [304]. Later with the development of multimodal learning, researchers start to get inspiration from the NLP society by adapting techniques originated from language processing, one concrete example is the Auto-Regressive models for image generation [71], [195]. Starting from recent several years since 2020, DPMs have gradually become one of the most active methods for vision generation tasks [89], [101], [102], [194], [220], [320], due to its impressive performance and better tractability compared to GANs.

On the other hand, it is worth mentioning that several large-scale models have been recently introduced in the text-to-image field such as DALL·E [195] and DALL·E 2 [194]. DALL·E model includes two separate stages to train for the raw image generation objective. In the first stage, DALL·E learns the visual concept from images via the discrete VAE [195], and then fuse the learned discrete image embedding with the textual tokens to train the transformer [245] in the second stage. In inference, DALL·E first obtains the fused text-image embedding for the given text description and potential image candidates, and then use the pre-trained CLIP to do image re-ranking to get the generated images with higher similarities. DALL·E 2, as an improved text-to-image generator, uses the Diffusion Probabilistic Models (DPMs) (see details in Section 5.1) on the image embedding space from CLIP to do the image synthesis.

Text-Guided Image Editing. As a step further from the image synthesis from text, another popular generative task that combines the vision and text is to perform editing based on text prompt for the given raw images [3], [13], [123], [133], [174], [318]. Compared to the text to image generation, the setup of text-guided image editing using the generative models takes not only the text prompt, but also raw real images as part of the input. The task objectives are usually two-fold: achieve the target editing effect, and preserve the remaining features of the given images.

Similar to the text-to-image (T2I) literature, as a refined task with additional raw image data as input, the community has explored several mainstream generative approaches using different architectures such as GANs [181], [315] and DPMs [13], [123], [174], [318]. As the state-of-the-art approaches, one way to categorize the DPM-based image editing methods is to see whether the proposed methods require additional learning given a pre-trained generative model. In most intuitive and straightforward cases, image editing via text prompt requires fine-tuning the parameters of given pre-trained models [123], or learning extra neural network modules [133] to achieve the target editing effect. Another category of image editing methods propose to solve the editing objective in a leaning-free manner [318], by explicitly leveraging the intrinsic abilities of DPMs in exhibiting semantics along the generation trajectories.

Video Synthesis from Text. There are also works tacking the more challenging vision generation from text that focus on the videos [104], [111], [141]. Compared to the images, videos often consist of multiple consecutive frames that are temporally and spatially correlated, in addition to the pixel-level computational extensive synthesis for a single visual frame. Existing works usually deploy VAE or GAN-based generative backbones to tackle the task [141] with language priors. Alternatively, Hu *et al.* [111] modifies the task formulation by providing a reference image and a textual description, and synthesize videos based on the given input. More recent work [104] introduces a video diffusion model for the task via modified 3D U-Net architecture equipped with an extra temporal dimension. *Make-A-Video* is one of the first large-scale video generation models. It leverages the recent advances in text-to-image field to learn the visual information, and then proposes to learn the temporal motions from unlabelled large-scale video data. In inference, the model composes the visual content with learned motions to synthesize a realistic video.

6 Further Discussion

6.1 Insights from Data and Methodology Design

Our paper aims to provide a novel perspective to understand the multimodal learning in the light of data, we revisit and discuss the correlation between the data nature and the methodology design as part of discussion from two aspects: the semantics of data modalities, and their specific formats.

For data semantics, as explained in Sec. 2, visual data and several types of audio data, such as the ambient sound, can be considered as raw information sources. These modalities include sensory information directly captured from the environment that is usually high-dimensional and can be further processed and analyzed with information redundancy. In contrast to raw information sources, text data and certain types of audio signals, such as speech, have undergone extensive processing throughout the evolution of human civilization. These data modalities already possess meaningful semantics and are more information compact with a unified representation in the form of tokens. Additionally, most NLP tasks have a rather unified problem formation under the notion of ”next word token prediction”. This distinction in data nature, particularly in terms of semantics, has played a significant role in shaping the methodological and technical advancements in their respective research domains. In the field of NLP, the highly processed nature of text data and its consistent problem formulation have paved the way for the development of large-scale foundational models. These models, such as GPT-3, have demonstrated remarkable performance across a wide range of NLP tasks. The unified nature of text data allows for the application of these models to various tasks without significant modifications, leveraging the semantic richness and the consistent problem formulation. On the other hand, the CV community faces different challenges. Visual data, being raw information sources, require extensive representation learning and specific downstream application stages to obtain effective and processed visual representations. The complexity of visual data and the diversity of visual tasks make it more challenging to develop a unified foundational model that can be applied across the board. As a result, researchers in the CV domain are constantly exploring novel representation learning techniques and task-specific approaches to address the intricacies of visual data and achieve state-of-the-art performance in complex visual tasks. As for audio data, researchers are following either community based on the specific audio types as well as the specific task requirements.

Another aspect to understand the logic between data nature and the methodology design focuses on the data format. The format, whether continuous or discrete, plays a crucial role in determining the suitable model architecture for effective processing. For continuous data, such as images and ambient sound, the continuity in the spatial or temporal dimensions usually benefit from model architectures like CNNs that are specifically designed to handle spatial and temporal dependencies and to capture the local and global correlations within the data. As for data with discrete formats, such as MIDI musical representations or textual word tokens, models like Transformers are more appropriate to model the dependencies among discrete elements.

6.2 Future Directions and Challenge

As we have extensively discussed in this survey paper, the research in multimodal machine learning is diversely spread from general representation learning to detailed downstream tasks within a specific field.

After having introduced various discriminative and generative multimodal applications that involve the vision and data of other modalities, we revisit and summarize the existing works from the perspective of their technique designs and the connections with respect to the data attributes.

For the discriminative tasks that involve the visual and audio data, we can observe from the introduced existing works that majority of them follow the general pipeline that contains separate data encoders, cross-modality attention feature fusions, as well as a decoder module designed for various task objectives. It is worth noting that all of the existing works process the ambient audio data as an entirety without specifically looking into the acoustic features of the audio signals. For example, certain type of ambient audio signals may include higher pitch and frequency than others, which can be used as a strong supplementary indicator for purely vision based recognition. In contrast, existing audio-involved generative works have explored more the disentangled features such as rhythms, pitches, and genres for either both synthesis and editing purposes. As for the vision and text (natural language) combination, representative early classic methods often use the LSTM model to deal with the textual language data with word orderings. Later, the success of Transformer model has encouraged the fast technical transition from LSTM to Transformers for the text processing branch in the multimodal learning context.

Return to current multimodal studies, while great success has been achieved within recent years, the challenges for the future research remain. From the technical perspective, we believe that we can summarize the future research directions into two direction with regard to the connections with the data modalities. On the one hand, the research community is seeking to establish a unified and general model that efficiently learns the representation of all the modalities of interest. Such a unified model, similar to the large-scale pretraining models we have introduced in Section 3.3, should greatly help with various downstream applications such as specific cross-modality generations, interactive editing, and evaluations. On the other hand, with increasing demand of more fine-grained and detailed applications in our daily life, we also expect to develop and achieve better performance for more specific and crafted tasks.

Another possible future direction for the multimodal learning could be the human intervention for ultimate multimodal perception AI systems. As our ultimate objective for multimodal learning is to bring intelligence to machines as real humans, human intervention could be a critical part to guide the general research direction in this fast developing area. A concrete example could be involving the human to provide more controls on the cross-modality generations and several downstream tasking such as editing [164], [165] .

7 Conclusions

In this paper, we present a survey on the multimodal learning domain from an unique perspective of data characteristics. We start by mainly analyzing the intrinsic natures of different data modalities for the vision, audio, and text. We proceed to introduce the multimodal representation learning, in which we mainly categorize the current literature by their learning settings. Following the general representation learning in the multimodal field, we then introduce the concrete task applications from both discriminative and generative natures, each structured into sub-classes with specific data combinations in the form of "Vision + X". For the discriminative tasks, we also include a revisit in light of data after presenting the task-specific works, where we provide our analysis to bridge the existing technique designs and their connections to the data nature from different modalities. For the generative tasks, we include an introduction to the popular generative backbone models before diving into the detailed task explanations. Finally, we provide our discussions in terms of the challenges and future directions for the multimodal learning domain.

References

[1] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *CVPR*, 2020.

[2] Gunjan Aggarwal and Devi Parikh. Dance2music: Automatic dance-driven music generation. *arXiv:2107.06252*, 2021.

[3] Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, et al. Cm3: A causal masked multimodal model of the internet. *arXiv:2201.07520*, 2022.

[4] 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. *NeurIPS*, pages 24206–24221, 2021.

[5] Claude Alain, Stephen R Arnott, Stephanie Hevenor, Simon Graham, and Cheryl L Grady. "what" and "where" in the human auditory system. *Proceedings of the national academy of sciences*, 98(21):12301–12306, 2001.

[6] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *CVPR*, 2019.

[7] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016.

[8] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pages 5803–5812, 2017.

[9] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015.

[10] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, 2017.

[11] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223. PMLR, 2017.

[12] Jacob Austin, Daniel Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *NeurIPS*, 2021.

[13] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, 2022.

[14] Fahimeh Bahmaninezhad, Jian Wu, Rongzhi Gu, Shi-Xiong Zhang, Yong Xu, Meng Yu, and Dong Yu. A comprehensive study of speech separation: spectrogram vs waveform separation. *arXiv:1905.07497*, 2019.

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

[16] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[17] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *ECCV*. Springer, 2006.

[18] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, pages 1–32, 2021.

[19] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv:2108.07258*, 2021.

[20] Ina Bornkessel-Schlesewsky, Matthias Schlesewsky, Steven L Small, and Josef P Rauschecker. Neurobiological roots of language in primate audition: common computational properties. *Trends in cognitive sciences*, 19(3):142–150, 2015.

[21] Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, 2019.

[22] Jean-Pierre Briot, Gaëtan Hadjeres, and François-David Pachet. Deep learning techniques for music generation–a survey. *arXiv preprint arXiv:1709.01620*, 2017.

[23] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv:1809.11096*, 2018.

[24] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.

[25] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are

few-shot learners. *NeurIPS*, pages 1877–1901, 2020.

[26] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015.

[27] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 2014.

[28] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE TPAMI*, 2019.

[29] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[30] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[31] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv:1512.03012*, 2015.

[32] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.

[33] Moitreya Chatterjee and Anoop Cherian. Sound2sight: Generating visual dynamics from sound and context. In *ECCV*, pages 701–719. Springer, 2020.

[34] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011.

[35] Fuhai Chen, Rongrong Ji, Jiayi Ji, Xiaoshuai Sun, Baochang Zhang, Xuri Ge, Yongjian Wu, Feiyue Huang, and Yan Wang. Variational structured semantic inference for diverse image captioning. In *NeurIPS*, 2019.

[36] Kan Chen, Chuanxi Zhang, Chen Fang, Zhaowen Wang, Trung Bui, and Ram Nevatia. Visually indicated sound generation by perceptually optimized classification. In *ECCV Workshops*, pages 0–0, 2018.

[37] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *ECCV*, pages 520–535, 2018.

[38] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *CVPR*, pages 7832–7841, 2019.

[39] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *CVPR*, 2020.

[40] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, 2017.

[41] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, pages 1691–1703. PMLR, 2020.

[42] Peihao Chen, Yang Zhang, Mingkui Tan, Hongdong Xiao, Deng Huang, and Chuang Gan. Generating visually aligned sound from videos. *IEEE TIP*, 29:8292–8302, 2020.

[43] Shaoxiang Chen and Yu-Gang Jiang. Semantic proposal for activity localization in videos via sentence query. In *AAAI*, pages 8199–8206, 2019.

[44] Shaoxiang Chen and Yu-Gang Jiang. Hierarchical visual-textual graph for temporal activity localization via language. In *ECCV*, pages 601–618. Springer, 2020.

[45] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, pages 104–120. Springer, 2020.

[46] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *ICCV*, pages 686–695, 2017.

[47] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *CVPR*, pages 8415–8424, 2021.

[48] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *ACCV*. Springer, 2017.

[49] Yu-An Chung, Wei-Hung Weng, Schrasing Tong, and James Glass. Unsupervised cross-modal alignment of speech and text embedding spaces. *NeurIPS*, 31, 2018.

[50] Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *ICML*. PMLR, 2019.

[51] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[52] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, 2005.

[53] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *ICCV*, pages 2970–2979, 2017.

[54] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[55] 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 *ECCV*, 2018.

[56] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *CVIU*, 163:90–100, 2017.

[57] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, 2017.

[58] Dipanjan Das, Sandika Biswas, Sanjana Sinha, and Brojeshwar Bhowmick. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *ECCV*, pages 408–424, 2020.

[59] Abe Davis and Maneesh Agrawala. Visual rhythm and beat. In *ACM Transactions on Graphics (TOG)*, 2018.

[60] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*, 2017.

[61] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. Visual grounding via accumulated attention. In *CVPR*, pages 7746–7755, 2018.

[62] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[63] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *ICCV*, pages 1769–1779, 2021.

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

[65] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *CVPR*, pages 11162–11173, 2021.

[66] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.

[67] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv:2005.00341*, 2020.

[68] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021.

[69] Shangzhe Di, Zeren Jiang, Si Liu, Zhaokai Wang, Leyan Zhu, Zexin He, Hongming Liu, and Shuicheng Yan. Video background music generation with controllable music transformer. In *ACMMM*, 2021.

[70] Mattia A Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. Must-c: a multilingual speech translation corpus. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2012–2017. Association for Computational Linguistics, 2019.

[71] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *NeurIPS*, pages 19822–19835, 2021.

[72] Hao-Wen Dong, Wen-Yi Hsiao, and Yi-Hsuan Yang. Pypianoroll: Open source python package for handling multitrack pianoroll. *Proc. ISMIR. Late-breaking paper*, 2018.

[73] Bin Duan, Hao Tang, Wei Wang, Ziliang Zong, Guowei Yang, and Yan Yan. Audio-visual event localization via recursive fusion by joint co-attention. In *WACV*, pages 4013–4022, 2021.

[74] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2sign: a large-scale multimodal dataset for continuous american sign language. In *CVPR*, 2021.

[75] Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. Improved speech reconstruction from silent video. In *ICCV Workshops*, pages 455–462, 2017.

[76] Ariel Ephrat and Shmuel Peleg. Vid2speech: speech reconstruction

from silent video. In *ICASSP*, pages 5095–5099. IEEE, 2017.

[77] Michael Freitag, Shahin Amiriparian, Sergey Pugachevskiy, Nicholas Cummins, and Björn Schuller. audeep: Unsupervised learning of representations from audio with deep recurrent neural networks. *The Journal of Machine Learning Research*, 2017.

[78] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv:1606.01847*, 2016.

[79] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. In *ECCV*, 2020.

[80] Sanchit Gandhi, Patrick Von Platen, and Alexander M Rush. Esb: A benchmark for multi-domain end-to-end speech recognition. *arXiv:2210.13352*, 2022.

[81] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, pages 5267–5275, 2017.

[82] Jort F Gemmeke, Daniel PW 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 *ICASSP*. IEEE, 2017.

[83] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[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. In *CVPR*, 2017.

[85] Alex Graves. Generating sequences with recurrent neural networks. *arXiv:1308.0850*, 2013.

[86] Pierre-Amaury Grumiaux, Srdjan Kitic, Laurent Girin, and Alexandre Guérin. A survey of sound source localization with deep learning methods. *The Journal of the Acoustical Society of America*, pages 107–151, 2022.

[87] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv:2312.00752*, 2023.

[88] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *CVPR*, pages 7181–7189, 2018.

[89] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, 2022.

[90] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *NeurIPS*, 30, 2017.

[91] Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019.

[92] David Harwath and James Glass. Deep multimodal semantic embeddings for speech and images. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015.

[93] Thomas Hayes, Songyang Zhang, Xi Yin, Guan Pang, Sasha Sheng, Harry Yang, Songwei Ge, Isabelle Hu, and Devi Parikh. Mugen: A playground for video-audio-text multimodal understanding and generation. *arXiv:2204.08058*, 2022.

[94] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[95] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[96] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *ICASSP*. IEEE, 2017.

[97] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

[98] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017.

[99] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

[100] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[101] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[102] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 2022.

[103] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv:2207.12598*, 2022.

[104] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022.

[105] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[106] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *IEEE TPAMI*, 2019.

[107] Chiori Hori, Huda Alamri, Jue Wang, Gordon Wichern, Takaaki Hori, Anoop Cherian, Tim K Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, et al. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *ICASSP*. IEEE, 2019.

[108] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[109] Minghui Hu, Yujie Wang, Tat-Jen Cham, Jianfei Yang, and PN Suganthan. Global context with discrete diffusion in vector quantised modelling for image generation. *arXiv:2112.01799*, 2021.

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

[111] Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make it move: controllable image-to-video generation with text descriptions. In *CVPR*, pages 18219–18228, 2022.

[112] Jianqiang Huang, Yu Qin, Jiaxin Qi, Qianru Sun, and Hanwang Zhang. Deconfounded visual grounding. In *AAAI*, pages 998–1006, 2022.

[113] Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman. Epic-sounds: A large-scale dataset of actions that sound. In *ICASSP*. IEEE, 2023.

[114] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[115] Alejandro Jaimes and Nicu Sebe. Multimodal human–computer interaction: A survey. *CVIU*, 108(1-2):116–134, 2007.

[116] Unnat Jain, Svetlana Lazebnik, and Alexander G Schwing. Two can play this game: visual dialog with discriminative question generation and answering. In *CVPR*, 2018.

[117] Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M Khapra. Towards building asr systems for the next billion users. In *AAAI*, 2022.

[118] Jesper Jensen and Cees H Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2009–2022, 2016.

[119] Yang Jin, Zehuan Yuan, Yadong Mu, et al. Embracing consistency: A one-stage approach for spatio-temporal video grounding. *NeurIPS*, 2022.

[120] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

[121] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021.

[122] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *NeurIPS*, 27, 2014.

[123] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *CVPR*, 2022.

[124] Minsu Kim, Joanna Hong, and Yong Man Ro. Lip to speech synthesis with visual context attentional gan. *NeurIPS*, pages 2758–2770, 2021.

[125] Diederik P Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *NeurIPS*, 2021.

[126] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[127] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity

speech synthesis. In *NeurIPS*, 2020.
[128] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *ICLR*, 2020.
[129] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017.
[130] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.
[131] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 2012.
[132] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. In *NeurIPS*, 2019.
[133] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *ICLR*, 2023.
[134] Junhyeok Lee and Seungu Han. Nu-wave: A diffusion probabilistic model for neural audio upsampling. *Proc. Interspeech 2021*, 2021.
[135] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, pages 7331–7341, 2021.
[136] Guangyao Li, Wenxuan Hou, and Di Hu. Progressive spatio-temporal perception for audio-visual question answering. In *ACMM*, 2023.
[137] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *CVPR*, 2022.
[138] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, 2021.
[139] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *CVPR*, pages 12174–12182, 2019.
[140] Yunzhu Li, Toru Lin, Kexin Yi, Daniel Bear, Daniel Yamins, Jiajun Wu, Joshua Tenenbaum, and Antonio Torralba. Visual grounding of learned physical models. In *ICML*, pages 5927–5936. PMLR, 2020.
[141] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *AAAI*, 2018.
[142] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *ICCV*, 2017.
[143] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv:2209.03430*, 2022.
[144] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
[145] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*. Springer, 2014.
[146] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing. *NeurIPS*, 34:11449–11461, 2021.
[147] Yan-Bo Lin and Yu-Chiang Frank Wang. Audiovisual transformer with instance attention for audio-visual event localization. In *ACCV*, 2020.
[148] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *ICCV*, pages 4673–4682, 2019.
[149] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *ECCV*, 2022.
[150] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *ICCV*, pages 873–881, 2017.
[151] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv:1907.13487*, 2019.
[152] Yongfei Liu, Bo Wan, Lin Ma, and Xuming He. Relation-aware instance refinement for weakly supervised visual grounding. In *CVPR*, pages 5612–5621, 2021.
[153] Yongfei Liu, Bo Wan, Xiaodan Zhu, and Xuming He. Learning cross-modal context graph for visual grounding. In *AAAI*, pages 11645–11652, 2020.
[154] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.
[155] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34, 2015.
[156] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.
[157] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*. Springer, 2016.
[158] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, 2019.
[159] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 2017.
[160] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv:2002.06353*, 2020.
[161] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016.
[162] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv:1412.6632*, 2014.
[163] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019.
[164] Willi Menapace, Stéphane Lathuilière, Aliaksandr Siarohin, Christian Theobalt, Sergey Tulyakov, Vladislav Golyanik, and Elisa Ricci. Playable environments: Video manipulation in space and time. In *CVPR*, pages 3584–3593, 2022.
[165] Willi Menapace, Stéphane Lathuilière, Sergey Tulyakov, Aliaksandr Siarohin, and Elisa Ricci. Playable video generation. In *CVPR*, pages 10061–10070, 2021.
[166] Daniel Michelsanti, Olga Slizovskaia, Gloria Haro, Emilia Gómez, Zheng-Hua Tan, and Jesper Jensen. Vocoder-based speech synthesis from silent videos. *arXiv:2004.02541*, 2020.
[167] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv:1804.02516*, 2018.
[168] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
[169] Rodrigo Mira, Alexandros Haliassos, Stavros Petridis, Björn W Schuller, and Maja Pantic. Svts: Scalable video-to-speech synthesis. *arXiv:2205.02058*, 2022.
[170] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv:1411.1784*, 2014.
[171] Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon. Symbolic music generation with diffusion models. *arXiv:2103.16091*, 2021.
[172] Akash Nagaraj. Asl alphabet, 2018.
[173] Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In *CVPR*, pages 3977–3986, 2021.
[174] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv:2112.10741*, 2021.
[175] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*. PMLR, 2021.
[176] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *CVPR*, pages 12700–12710, 2021.
[177] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, JK Aggarwal, Hyungtae Lee, Larry Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR*, pages 3153–3160, 2011.
[178] Andreea-Maria Oncescu, A Koepke, Joao F Henriques, Zeynep Akata, and Samuel Albanie. Audio retrieval with natural language queries. In *INTERSPEECH*, 2021.

[179] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017.
[180] Takashi Oya, Shohei Iwase, Ryota Natsume, Takahiro Itazuri, Shugo Yamaguchi, and Shigeo Morishima. Do we need sound for sound source localization? In *ACCV*, 2020.
[181] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. *arXiv:2305.10973*, 2023.
[182] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.
[183] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, pages 724–732, 2016.
[184] Sergio Picazo-Vela and Luis Hernandez. *Technology, Science, and Culture: A Global Vision*. 02 2019.
[185] Cyril Plapous, Claude Marro, and Pascal Scalart. Improved signal-to-noise ratio estimation for speech enhancement. *IEEE transactions on audio, speech, and language processing*, 2006.
[186] Bryan A Plummer, Arun Mallya, Christopher M Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. Phrase localization and visual relationship detection with comprehensive image-language cues. In *ICCV*, pages 1928–1937, 2017.
[187] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. Learning individual speaking styles for accurate lip to speech synthesis. In *CVPR*, 2020.
[188] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017.
[189] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. In *CVPR*, pages 10860–10869, 2020.
[190] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *ECCV*, pages 292–308. Springer, 2020.
[191] Kranthi Kumar Rachavarapu, Vignesh Sundaresha, AN Rajagopalan, et al. Localize to binauralize: Audio spatialization from visual sound source localization. In *ICCV*, pages 1930–1939, 2021.
[192] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021.
[193] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.
[194] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125*, 2022.
[195] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*. PMLR, 2021.
[196] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Sign language recognition: A deep survey. *Expert Systems with Applications*, 2021.
[197] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *NeurIPS*, 2019.
[198] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *CVPR*, 2016.
[199] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, pages 1060–1069. PMLR, 2016.
[200] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *ICASSP*, pages 749–752. IEEE, 2001.
[201] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.
[202] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015.
[203] Khwaja Mohd Salik, Swati Aggarwal, Yaman Kumar, Rajiv Ratn Shah, Rohit Jain, and Roger Zimmermann. Lipper: Speaker independent speech synthesis using multi-view lipreading. In *AAAI*, pages 10023–10024, 2019.
[204] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *NeurIPS*, 29, 2016.
[205] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 2008.
[206] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv:2210.08402*, 2022.
[207] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv:2111.02114*, 2021.
[208] Idan Schwartz, Alexander G Schwing, and Tamir Hazan. A simple baseline for audio-visual scene-aware dialog. In *CVPR*, 2019.
[209] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *CVPR*, pages 4358–4366, 2018.
[210] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Less can be more: Sound source localization with a classification model. In *WACV*, pages 3308–3317, 2022.
[211] Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, and Leonid Sigal. Visual reference resolution using attention memory for visual dialog. In *NeurIPS*, 2017.
[212] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.
[213] Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *CVPR*, 2016.
[214] Gunnar A Sigurdsson, Jean-Baptiste Alayrac, Aida Nematzadeh, Lucas Smaira, Mateusz Malinowski, Joao Carreira, Phil Blunsom, and Andrew Zisserman. Visual grounding in video for unsupervised word translation. In *CVPR*, pages 10850–10859, 2020.
[215] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016.
[216] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
[217] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*. PMLR, 2015.
[218] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *NeurIPS*, 29, 2016.
[219] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *NeurIPS*, 2021.
[220] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *NeurIPS*, 2020.
[221] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2020.
[222] Yang Song, Jingwen Zhu, Dawei Li, Xiaolong Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. *arXiv:1804.04786*, 2018.
[223] Zengjie Song, Yuxi Wang, Junsong Fan, Tieniu Tan, and Zhaoxiang Zhang. Self-supervised predictive learning: A negative-free method for sound source localization in visual scenes. In *CVPR*, pages 3222–3231, 2022.
[224] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *CVPR*, pages 89–98, 2018.
[225] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021.
[226] Dominc Streiff, Lukas Bernreiter, Florian Tschopp, Marius Fehr, and Roland Siegwart. 3d3l: Deep learned 3d keypoint detection and description for lidars. In *ICRA*. IEEE, 2021.
[227] Kun Su, Xiulong Liu, and Eli Shlizerman. Audeo: Audio generation for a silent performance video. In *NeurIPS*, 2020.
[228] Rui Su, Qian Yu, and Dong Xu. Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *ICCV*, 2021.
[229] Mingjie Sun, Jimin Xiao, Eng Gee Lim, Si Liu, and John Y Goulermas. Discriminative triad matching and reconstruction

for weakly referring expression grounding. *IEEE TPAMI*, pages 4189–4195, 2021.

[230] Shuyang Sun, Zhanghui Kuang, Lu Sheng, Wanli Ouyang, and Wei Zhang. Optical flow guided feature: A fast and robust motion representation for video action recognition. In *CVPR*, pages 1390–1399, 2018.

[231] Supasorn Suwajanakorn, Noah Snavely, Jonathan J Tompson, and Mohammad Norouzi. Discovery of latent 3d keypoints via end-to-end geometric reasoning. *NeurIPS*, 2018.

[232] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.

[233] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *ICASSP*, pages 4214–4217. IEEE, 2010.

[234] Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. Naturalspeech: End-to-end text to speech synthesis with human-level quality. *arXiv:2205.04421*, 2022.

[235] Zhicong Tang, Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Improved vector quantized diffusion models. *arXiv:2205.16007*, 2022.

[236] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419. Springer, 2020.

[237] Christopher Thomas, Yipeng Zhang, and Shih-Fu Chang. Fine-grained visual entailment. In *ECCV*. Springer, 2022.

[238] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *ECCV*, 2020.

[239] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018.

[240] Thanh-Dat Truong, Chi Nhan Duong, Hoang Anh Pham, Bhiksha Raj, Ngan Le, Khoa Luu, et al. The right to talk: An audio-visual transformer approach. In *ICCV*, 2021.

[241] Yi-Hsuan Tsai, Ming-Hsuan Yang, and Michael J Black. Video segmentation via object flow. In *CVPR*, 2016.

[242] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *ISMIR*, 2019.

[243] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *CVPR*, 2018.

[244] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

[245] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[246] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.

[247] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015.

[248] Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, and Maja Pantic. Video-driven speech reconstruction using generative adversarial networks. *arXiv:1906.06301*, 2019.

[249] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *IJCV*, 128(5):1398–1413, 2020.

[250] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[251] Jacob Walker, Abhinav Gupta, and Martial Hebert. Dense optical flow prediction from a static image. In *ICCV*, pages 2443–2451, 2015.

[252] Mark N Wallace, Peter W Johnston, and Alan R Palmer. Histochemical identification of cortical areas in the auditory region of the human brain. *Experimental Brain Research*, 143(4):499–508, 2002.

[253] Angelina Wang and Olga Russakovsky. Overwriting pretrained bias with finetuning data. In *ICCV*, 2023.

[254] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *CVPR*, 2018.

[255] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *ACMM*, pages 154–162, 2017.

[256] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. *arXiv:1607.06215*, 2016.

[257] Lei Wang, Piotr Koniusz, and Du Q Huynh. Hallucinating idt descriptors and i3d optical flow features for action recognition with cnns. In *ICCV*, pages 8698–8708, 2019.

[258] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE TPAMI*, 41(2):394–407, 2018.

[259] Liwei Wang, Alexander Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *NeurIPS*, 2017.

[260] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *CVPR*, 2019.

[261] Weining Wang, Yan Huang, and Liang Wang. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *CVPR*, pages 334–343, 2019.

[262] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 2019.

[263] Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: global-local sequence alignment for text-video retrieval. In *CVPR*, pages 5079–5088, 2021.

[264] Yang Wang, Peng Wang, Zhenheng Yang, Chenxu Luo, Yi Yang, and Wei Xu. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In *CVPR*, pages 8071–8081, 2019.

[265] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. In *ICCV*, pages 5764–5773, 2019.

[266] Zeyu Wang, Yu Wu, Karthik Narasimhan, and Olga Russakovsky. Multi-query video retrieval. *arXiv:2201.03639*, 2022.

[267] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[268] Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, and Anton Van Den Hengel. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *CVPR*, 2018.

[269] Shih-Lun Wu and Yi-Hsuan Yang. The jazz transformer on the front line: Exploring the shortcomings of ai-composed music through quantitative measures. *arXiv:2008.01307*, 2020.

[270] Yu Wu, Lu Jiang, and Yi Yang. Switchable novel object captioner. *IEEE TPAMI*, pages 1–1, 2022.

[271] Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *CVPR*, 2021.

[272] Yu Wu, Linchao Zhu, Lu Jiang, and Yi Yang. Decoupled novel object captioner. In *ACMM*, 2018.

[273] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *ICCV*, 2019.

[274] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE TNNLS*, 2020.

[275] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, 2015.

[276] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *CVPR*, pages 5945–5954, 2017.

[277] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021.

[278] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment task for visually-grounded language learning. *arXiv preprint arXiv:1811.10582*, 2018.

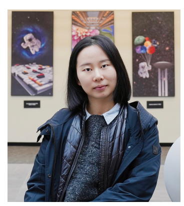
[279] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv:1901.06706*, 2019.

[280] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*. Springer, 2016.

[281] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016.

[282] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with

visual attention. In *ICML*, 2015.
[283] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE TPAMI*, 2023.
[284] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018.
[285] Hanyu Xuan, Zhenyu Zhang, Shuo Chen, Jian Yang, and Yan Yan. Cross-modal attention network for temporal inconsistent audio-visual event localization. In *AAAI*, pages 279–286, 2020.
[286] Ravindra Yadav, Ashish Sardana, Vinay P Namboodiri, and Rajesh M Hegde. Speech prediction in silent videos using variational autoencoders. In *ICASSP*, pages 7048–7052. IEEE, 2021.
[287] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *CVPR*, 2022.
[288] 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 *ACMM*, 2022.
[289] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *ICCV*, 2019.
[290] Xitong Yang, Palghat Ramesh, Radha Chitta, Sriganesh Madhvanath, Edgar A Bernal, and Jiebo Luo. Deep multimodal representation learning from temporal data. In *CVPR*, pages 5447–5455, 2017.
[291] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *ECCV*, pages 387–404. Springer, 2020.
[292] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *ICCV*, pages 4683–4693, 2019.
[293] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.
[294] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, 2016.
[295] Dong Yu and Li Deng. *Automatic speech recognition*, volume 1. Springer, 2016.
[296] Jiashuo Yu, Ying Cheng, Rui-Wei Zhao, Rui Feng, and Yuejie Zhang. Mm-pyramid: Multimodal pyramid attentional network for audio-visual event localization and video parsing. *arXiv:2111.12374*, 2021.
[297] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85. Springer, 2016.
[298] Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. Rethinking diversified and discriminative proposal generation for visual grounding. *arXiv:1805.03508*, 2018.
[299] Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. Explainability in graph neural networks: A taxonomic survey. *IEEE TPAMI*, 2022.
[300] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning. In *CVPR*, pages 6995–7004, 2021.
[301] Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *ICCV*, 2021.
[302] Fangneng Zhan, Hongyuan Zhu, and Shijian Lu. Spatial fusion gan for image synthesis. In *CVPR*, pages 3653–3662, 2019.
[303] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *CVPR*, 2017.
[304] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE TPAMI*, 2018.
[305] Mingxing Zhang, Yang Yang, Xinghan Chen, Yanli Ji, Xing Xu, Jingjing Li, and Heng Tao Shen. Multi-stage aggregated transformer network for temporal language localization in videos. In *CVPR*, pages 12669–12678, 2021.
[306] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *CVPR*, pages 3661–3670, 2021.
[307] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *CVPR*, 2020.
[308] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *ICCV*, 2019.
[309] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, 2018.
[310] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE TNNLS*, 30(11):3212–3232, 2019.
[311] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *CVPR*, pages 10394–10403, 2019.
[312] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *CVPR*, pages 4176–4186, 2021.
[313] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *CVPR*, 2019.
[314] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *CVPR*, pages 3550–3558, 2018.
[315] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *ECCV*. Springer, 2020.
[316] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *CVPR*, 2019.
[317] Ye Zhu, Kyle Olszewski, Yu Wu, Panos Achlioptas, Menglei Chai, Yan Yan, and Sergey Tulyakov. Quantized gan for complex music generation from dance videos. In *ECCV*, 2022.
[318] Ye Zhu, Yu Wu, Zhiwei Deng, Olga Russakovsky, and Yan Yan. Boundary guided learning-free semantic control with diffusion models. *NeurIPS*, 2023.
[319] Ye Zhu, Yu Wu, Hugo Latapie, Yi Yang, and Yan Yan. Learning audio-visual correlations from variational cross-modal generation. In *ICCASP*, 2021.
[320] Ye Zhu, Yu Wu, Kyle Olszewski, Jian Ren, Sergey Tulyakov, and Yan Yan. Discrete contrastive diffusion for cross-modal and conditional generation. In *ICLR*, 2023.
[321] Ye Zhu, Yu Wu, Yi Yang, and Yan Yan. Describing unseen videos via multi-modal cooperative dialog agents. In *ECCV*, 2020.
[322] Ye Zhu, Yu Wu, Yi Yang, and Yan Yan. Saying the unseen: Video descriptions via dialog agents. In *TPAMI*, 2021.



Ye Zhu is currently a postdoctoral researcher in Computer Science at Princeton University, USA. She holds the Ph.D. degree in Computer Science from Illinois Institute of Technology, USA, in 2023. She received her B.S. and M.S. degrees from Shanghai Jiao Tong University, China. Her research interests mainly include multimodal learning and generation.

Yu Wu is professor with the School of Computer Science at Wuhan University, China. His research interests are multi-modal perception and video understanding. He was the recipient of Google PhD Fellowship 2020. He was the Workshop Chair of CVPR 2023, and also served as Area Chair for CVPR 2023, NeurIPs 2023.

Nicu Sebe is professor with the University of Trento, Italy. His research covers the areas of multimedia information retrieval and human behavior understanding. He was the Program Chair of the International Conference on Image and Video Retrieval in 2007 and 2010, ACM Multimedia 2007 and 2011. He was the Program Chair of ICCV 2017 and ECCV 2016, and a General Chair of ACM ICMR 2017.

Yan Yan is currently Gladwin Development Chair Assistant Professor with the Department of Computer Science, Illinois Institute of Technology. He received the Ph.D. degree in computer science from the University of Trento. His research interests include computer vision, machine learning, and multimedia.