Video Understanding with Large Language Models: A Survey

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, *Graduate Student Member, IEEE*, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, *Graduate Student Member, IEEE*, Chao Huang, Zeliang Zhang, Pinxin Liu, Mingqian Feng, Feng Zheng, *Member, IEEE*, Jianguo Zhang, *Senior Member, IEEE*, Ping Luo, *Member, IEEE*, Jiebo Luo, *Fellow, IEEE*, and Chenliang Xu, *Member, IEEE*

***Abstract*—With the rapid growth of online video platforms and the escalating volume of video content, the need for proficient video understanding tools has increased significantly. Given the remarkable capabilities of large language models (LLMs) in language and multimodal tasks, this survey provides a detailed overview of recent advances in video understanding that harness the power of LLMs (Vid-LLMs). The emergent capabilities of Vid-LLMs are surprisingly advanced, particularly their ability for open-ended multi-granularity (abstract, temporal, and spatiotemporal) reasoning combined with common-sense knowledge, suggesting a promising path for future video understanding. We examine the unique characteristics and capabilities of Vid-LLMs, categorizing the approaches into three main types: *Video Analyzer* × *LLM*, *Video Embedder* × *LLM*, and *(Analyzer + Embedder)* × *LLM*. We identify five subtypes based on the functions of LLMs in Vid-LLMs: *LLM as Summarizer*, *LLM as Manager*, *LLM as Text Decoder*, *LLM as Regressor*, and *LLM as Hidden Layer*. This survey also presents a comprehensive study of the tasks, datasets, benchmarks, and evaluation methods for Vid-LLMs. Additionally, it explores the extensive applications of Vid-LLMs in various domains, highlighting their remarkable scalability and versatility in real-world video understanding challenges. Additionally, it summarizes the limitations of existing Vid-LLMs and outlines directions for future research. For more information, readers are encouraged to visit the repository at https://github.com/yunlong10/Awesome-LLMs-for-Video-Understanding.**

***Index Terms*—Video Understanding, Large Language Model, Vision-Language Model, Multimodality Learning**

I. Introduction

We live in a multimodal world where video has become the predominant form of media. With the rapid expansion of online video platforms and the growing prevalence of cameras in surveillance, entertainment, and autonomous driving, video content has risen to prominence as a highly engaging and rich medium, outshining traditional text and image-text combinations in both depth and appeal. This advancement has fueled an exponential increase in video production, with millions of videos being created every day. However, manually processing such a sheer volume of video content is labor-intensive and time-consuming. As a result, there is a growing need for tools to effectively manage, analyze, and process this abundance of video content. To meet this need, video understanding methods have emerged that use intelligent analysis techniques to automatically recognize and interpret video content, significantly reducing the workload on human operators. In addition, the ongoing development of these methods is improving their task-solving capabilities, enabling them to handle a wide range of video understanding tasks with increasing proficiency.

A. Development of Video Understanding Methods

The evolution of video understanding methods can be divided into four stages, as shown in Figure 1:

1) Conventional Methods: In the early stages of video understanding, handcrafted feature extraction techniques such as Scale-Invariant Feature Transform (SIFT) [1], Speeded-Up Robust Features (SURF) [2], and Histogram of Oriented Gradients (HOG) [3] were used to capture key information in videos. Background Subtraction [4], optical flow methods [5], and Improved Dense Trajectories (IDT) [6], [7] were used to model the motion information for tracking. Since videos can be viewed as time series data, temporal analysis techniques such as Hidden Markov Models (HMM) [8] have also been used to understand video content. Before the popularity of deep learning, basic machine learning algorithms such as Support Vector Machines (SVM) [9], Decision Trees [10], and Random Forests were also used in video classification and recognition tasks. Cluster analysis [11] for classifying video segments, or Principal Component Analysis (PCA) [12], [13] for data dimensionality reduction have also been commonly used methods for video analysis.

2) Early Neural Video Models: Compared with classical methods, deep learning methods for video understanding possess superior task-solving capabilities. DeepVideo [14] and [15] were early methods introducing a deep neural network, specifically a Convolutional Neural Network (CNN), for video understanding. However, the performance was not superior to the best handcrafted feature method due to the inadequate use of motion information. Two-stream networks [16] combined both CNN and IDT to capture the motion information to improve the performance, which verified the capability of deep neural networks for video understanding. To handle long-form video understanding, Long Short-Term Memory (LSTM) was adopted [17]. Temporal Segment Network (TSN) [18] was

Y. Tang, J. Bi, L. Song, S. Liang, D. Zhang, J. An, J. Lin, R. Zhu, A. Vosoughi, C. Huang, Z. Zhang, P. Liu, M. Feng, J. Luo, and C. Xu are with University of Rochester

T. Wang and P. Luo are with The University of Hong Kong

S. Xu, T. Wang, F. Zheng, and J. Zhang are with Southern University of Science and Technology

Corresponding to Y. Tang, J. Luo, and C. Xu ({yunlong.tang@, jluo@cs., chenliang.xu@}rochester.edu)

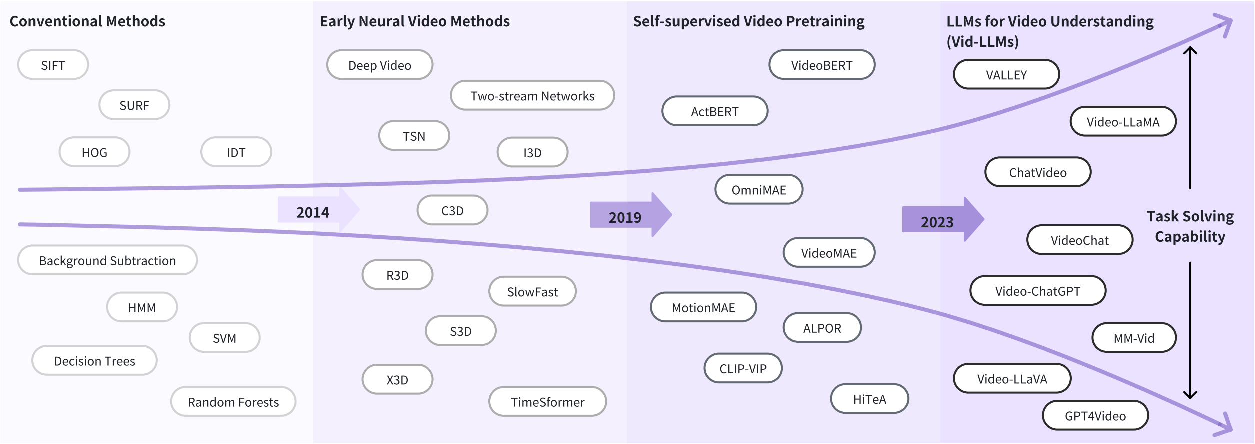

Fig. 1. The development of video understanding methods can be summarized into four stages: (1) Conventional Methods, (2) Early Neural Video Models, (3) Self-supervised Video Pretraining, and (4) Large Language Models for Video Understanding, i.e., Vid-LLMs. Their task-solving capability is continuously improving, and they possess the potential for further enhancement.

also designed for long-form video understanding by analyzing and aggregating video segments. Besides TSN, Fisher Vectors (FV) encoding [19], Bi-Linear encoding [20], and Vector of Locally Aggregated Descriptors (VLAD) [21] encoding were introduced [22]. These methods improved performance on the UCF-101 [23] and HMDB51 [24] datasets. Unlike two-stream networks, 3D networks started another branch by introducing 3D CNN to video understanding (C3D) [25]. Inflated 3D ConvNets (I3D) [26] utilizes the initialization and the architecture of 2D CNN, Inception [27], to gain a huge improvement on the UCF-101 and HMDB51 datasets. Subsequently, people began employing the Kinetics-400 (K-400) [28] and Something-Something [29] datasets to evaluate the model's performance in more challenging scenarios. ResNet [30], ResNeXt [31], and SENet [32] were also adapted from 2D to 3D, resulting in the emergence of R3D [33], MFNet [34], and STC [35]. To improve the efficiency, the 3D convolution has been decomposed into cascade 2D and 1D convolution in various studies (e.g., S3D [36], ECO [37], P3D [38]). LTC [39], T3D [40], Non-local [41], and V4D [42] focus on long-form temporal modeling, while CSN [43], SlowFast [44], and X3D [45] tend to attain high efficiency. The introduction of Vision Transformers (ViT) [46] promotes a series of prominent models (e.g., TimeSformer [47], VidTr [48], ViViT [49], MViT [50]).

3) Self-supervised Video Pretraining: Transferability [51], [52] in self-supervised pretraining models [53] for video understanding allows them to generalize across diverse tasks with minimal additional labeling, overcoming the early deep learning models' requirements for extensive task-specific data. VideoBERT [54] is an early attempt to perform video pretraining. Based on the bidirectional language model BERT [55], pertaining tasks are designed for self-supervised learning from video-text data. It tokenizes video features with hierarchical k-means. The pretrained model can be fine-tuned to handle multiple downstream tasks, including action classification and video captioning. Following the *"pretraining-finetuning"* paradigm, many studies on pretrained models for video understanding, especially video-language models, have emerged. They either use different architectures (ActBERT [56], SpatiotemporalMAE [57], OmniMAE [58], VideoMAE [59], MotionMAE [60]) or training strategies (MaskFeat [61], VLM [62], ALPRO [63], All-in-One transformer [64], MaskViT [65], CLIP-ViP [66], Singularity [67], LF-VILA [68], EMCL [69], HiTeA [70], CHAMPAGNE [71]).

4) Large Language Models for Video Understanding: Recently, large language models (LLMs) have advanced rapidly [72]. The emergence of large language models pretrained on extensive datasets has introduced a novel in-context learning capability [73]. This allows them to handle various tasks using prompts without the need for fine-tuning. ChatGPT [74] is the first groundbreaking application built on this foundation. This includes capabilities like generating code and invoking tools or APIs of other models for their use. Many studies are exploring using LLMs like ChatGPT to call vision models APIs to solve the problems in the computer vision field, including Visual-ChatGPT [75]. The advent of instruct-tuning has further enhanced these models' ability to respond effectively to user requests and perform specific tasks [76]. LLMs integrated with video understanding capabilities offer the advantage of more sophisticated multimodal understanding, enabling them to process and interpret complex interactions between visual and textual data. Similar to their impact in Natural Language Processing (NLP) [77], these models act as more general-purpose task solvers, adept at handling a broader range of tasks by leveraging their extensive knowledge base and contextual understanding acquired from vast amounts of multimodal data. This allows them to not only understand visual content but also reason about it in a way that is more aligned with human-like understanding. Many works also explore using LLMs in video understanding tasks, namely, Vid-LLMs.

B. Related Surveys

Previous survey papers either study specific sub-tasks in the area of video understanding or focus on methodologies beyond video understanding. For example, [78] surveys multimodal foundation models for general vision-language tasks, which

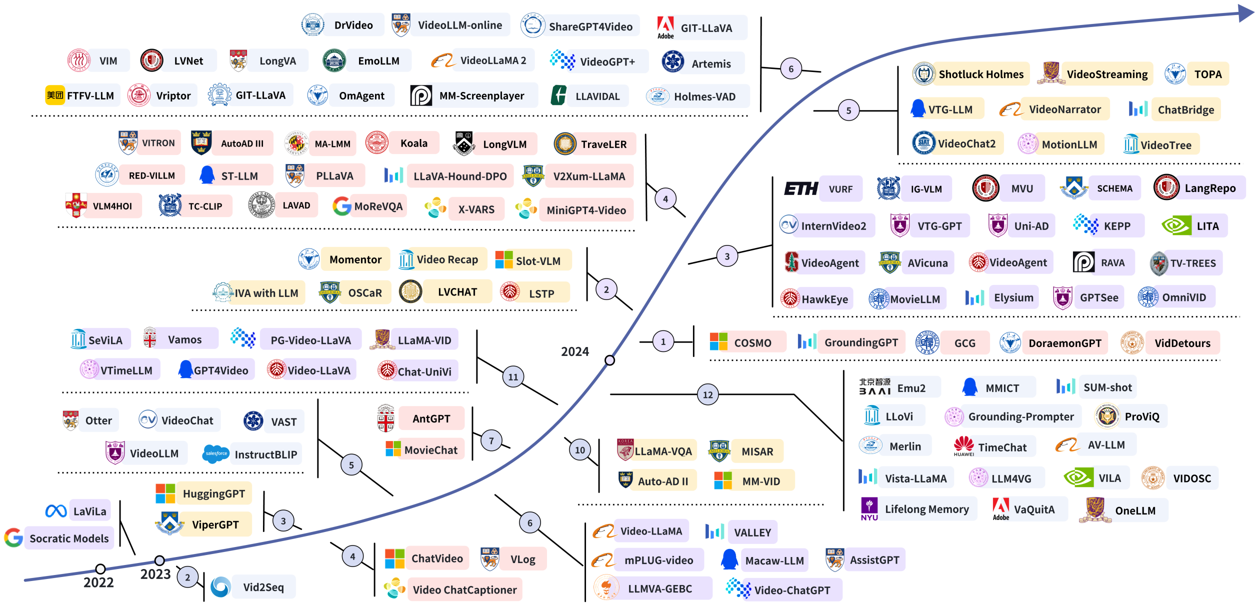

Fig. 2. A comprehensive timeline depicting the development of video understanding methods with large language models (Vid-LLMs). This survey is based on advancements up to the end of June 2024.

includes both image and video applications. [79] and [80] focus on surveying video captioning and action recognition tasks, respectively. Other video understanding tasks, such as the video question answering and grounding, are not considered. Moreover, [81], [82], and [77] survey video-related methodologies, such as video diffusion models and LLMs, lacking the concentration on video understanding. [83] centers primarily on video foundation models, with insufficient attention given to language model-based approaches. Despite the significant value to the community, previous survey papers leave a gap in surveying the general video understanding task based on large language models. This paper fills this gap by comprehensively surveying the video understanding task using large language models.

C. Survey Structure

This survey is structured as follows: Section II offers preliminaries for video understanding with LLMs, including a summary of various video understanding tasks that require handling different levels of granularity, their associated datasets, and evaluation metrics. The background of LLMs is also introduced in this section. In Section III, we delve into details of recent research leveraging LLMs for video understanding, presenting their unique approaches and impact in the field, where we divide these Vid-LLMs into three main categories, *Video Analyzer* × *LLM* and *Video Embedder* × *LLM*, and *(Analyzer + Embedder)* × *LLM*; and five sub-categories, *LLM as Summarizer/Manager/Text Decoder/Regressor/Hidden Layer*, shown as Figure 5. This section also includes the training strategies of Vid-LLMs. Section IV adds more information about popular ways to evaluate Vid-LLMs, together with some benchmarks and performances of some Vid-LLMs on the most commonly used benchmarks. Section V explores the application of Vid-LLMs across multiple significant fields and identifies unresolved challenges and potential areas for future research.

In addition to this survey, we have established a GitHub repository that aggregates various supporting resources for video understanding with large language models (Vid-LLMs). This repository, dedicated to enhancing video understanding through Vid-LLMs, can be accessed at *Awesome-LLMs-for-Video-Understanding*.

II. Preliminaries

In this section, we introduce the background of video understanding and Large Language Models (LLMs).

A. Video Understanding Tasks

Video understanding is a fundamental yet challenging task that has inspired the emergence of numerous tasks in a similar discipline aiming at interpreting complicated video content. The pioneering work for video understanding includes video classification and action recognition approaches, which classify videos into class labels and action categories, respectively. With the development of visual foundation models and expanding public datasets, current video understanding approaches can capture, analyze, and reason for more complicated video content. For instance, video captioning, as a specific task of video understanding, not only requires the model to generate detailed descriptions of the video content, but the generated video captions should be logical and follow commonsense about the scenes depicted. Additionally, the Video Question-Answering (VQA) task requires that the model understand the content and refer to external information to provide an accurate answer. The development path of video understanding from simple classification to natural language comprehension and reasoning highlights a clear trend of the video understanding model towards near-human levels of

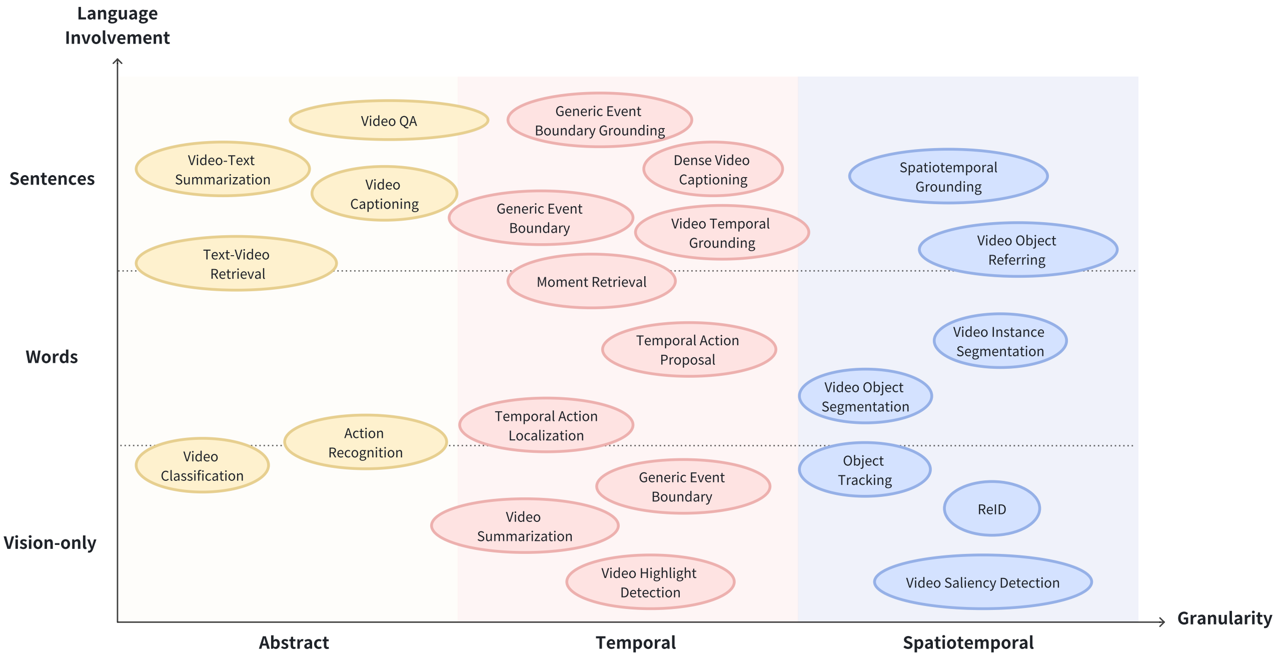

Fig. 3. This figure categorizes tasks in video understanding, delineating the granularity required and the language involvement necessary for models to perform these tasks effectively. This diagram excludes tasks involving special modalities or specific, such as audio-visual and egocentric video understanding. Notably, the tasks presented could be unified into a question-answering paradigm, and all solved by generative large models, akin to recent advances in NLP.

video interpretation capability. We summarize the main tasks in video understanding as follows.

1) Abstract Understanding Tasks: Video Classification, Action Recognition, Text-Video Retrieval, Video-to-Text Summarization, and Video Captioning.

- *Video Classification & Action Recognition:* Video classification and action recognition classify videos based on class labels or activities and events categories within a video sequence. Datasets specifically introduced for these tasks include UCF-101 [23], HMDB51 [24], Hollywood [84], ActivityNet [85], Charades [86], Kinetics-400 [28], Kinetics-600 [87], Kinetics-700 [88], SomethingSomethingV2 [29], HACS [89], YouTube8M [90], and PortraitMode-400 [91]. Usually, Top-K accuracy is adopted as the main metric for these tasks.
- *Text-Video Retrieval:* Text-video retrieval task matches and retrieves relevant video clips based on the similarity between video clips and the input textual descriptions. Datasets like Kinetic-GEB [92], MSRVTT [93], DiDeMo [94], YouCook2 [95], and Youku-mPLUG [96] are relevant to this task. The standard evaluation metric for this task is Recall at K (R@K), which measures the accuracy of the first K retrieved results.
- *Video-to-Text Summarization:* Video-to-text summarization is a task that generates concise textual summaries of videos. The video summarization approaches are trained to extract and interpret key visual and audio content to produce coherent and informative summaries. ViTT [97], VideoXum [98], VideoInstruct-100K [99], and Instrcut-V2Xum [100] are datasets that related to the task. Metrics of BLEU, METEOR, CIDEr, and ROUGE-L often evaluate this task.
- *Video Captioning:* Video captioning generates descriptive and coherent textual captions of given videos. The video caption models usually use visual and auditory information from video to produce accurate and contextually relevant descriptions. Notable datasets for this tasks are MSVD [101], MSR-VTT [102], TGIF [103], Charades [86], Charades-Ego [104], YouCook2 [95], Youku-mPLUG [96], VAST-27M [105], and VideoInstruct-100K [99]. This task is often evaluated by metrics of BLEU, METEOR, CIDEr, and ROUGE-L.
- *Video QA:* Video Question-Answering (VQA) aims to answer textual questions based on a given video, where the model analyzes visual and auditory information, understands the context, and eventually generates accurate responses. Datasets involved in the QA task are VCR [106], MSVD-QA [93], MSRVTT-QA [93], TGIF-QA [107], Pororo-QA [108], TVQA [109], ActivityNet-QA [110], and NExT-QA [111]. This task is evaluated using Top-1, Top-K accuracy.

2) Temporal Understanding Tasks:

- *Video Summarization:* Video Summarization aims at condensing a long video into a shorter version while preserving essential content. F1-score, Spearman, and Kendall usually evaluate this task as metrics. Commonly-used datasets include SumMe [112], TVSum [113], Ads-1k [114], VideoXum [98], Instrcut-V2Xum [100].
- *Video Highlight Detection:* Video highlight detection aims at identifying and extracting the most important and interesting segments from a video. Commonly used datasets on this task include the YouTube Highlights [115], the TV-Sum [113], and the Videos Titles in the Wild (VTW) [116].

- *Temporal Action/Event Localization:* This task aims at identifying the precise temporal segments of actions or events within a video. By analyzing sequential frames, models trained for this task must indicate when specific activities start and end. Datasets for Temporal Action/Event Localization involve THUMOS'14 [117], ActivityNet-1.3 [85], and UnAV-100 [118].
- *Temporal Action Proposal Generation:* Temporal action proposal generation involves generating candidate segments within a video that are likely to contain actions or events. Relevant datasets like THUMOS'14 [117], ActivityNet [85], and Charades [86] are used for the training and evaluation for this task.
- *Video Temporal Grounding:* Video temporal grounding is the task of locating specific moments or intervals within a video that correspond to a given textual query. This process involves aligning linguistic descriptions with visual content, enabling precise identification of relevant segments for applications in video search and content analysis. Common benchmarks are Charades-STA [119], ViTT [97], DiDeMo [94], and PU-VALOR [120]. The metrics of R1@0.5 and R1@0.7 often evaluate this task.
- *Moment Retrieval:* Moment retrieval is the task of identifying and extracting precise video segments that correspond to a given textual or visual query, which aligns semantic content between queries and video frames. DiDeMo [94] is a dataset for this task.
- *Generic Event Boundary Detection:* Generic event boundary detection involves identifying certain frames in a video where significant changes occur and splitting videos based on different events or activities. Kinetics-GEBD [121] is a widely-used dataset for this task.
- *Generic Event Boundary Captioning & Grounding:* Generic event boundary captioning and grounding involve identifying and describing the transition points between significant events in a video, where Kinetics-GEB+ [92] is the dataset for this task.
- *Dense Video Captioning:* Dense video captioning [122]–[126] aims at generating detailed and continuous textual descriptions for multiple events and actions occurring throughout a video. Evaluation metrics like BLEU, METEOR, CIDEr, and ROUGE-L are used to evaluate this task. Relevant datasets are ActivityNet Captions [127], VidChapters-7M [128], YouCook2 [95], and ViTT [97].

3) Spatiotemporal Understanding Tasks:

- *Object Tracking:* Object tracking aims at continuously identifying and following the position of specific objects within a video over time. A good tracking model should maintain accurate and consistent trajectories of objects, even for videos with occlusions, changes in appearance, and motions. Benchmarks like OTB [129], UAV [130], and VOT [131] are commonly used for this task.
- *Re-Identification:* Re-Identification (ReID) is the task of recognizing and matching individuals or objects across different video frames or camera views. Common datasets in ReID are Market-1501 [132], CUHK03 [133], MSMT17 [134], and DukeMTMC-reID [135].
- *Video Saliency Detection:* Video saliency detection aims at identifying the most visually important and attention-grabbing regions in a video [136]. This task highlights areas that stand out due to factors like motion, contrast, and unique features. Relevant datasets to this task are DHF1K [137], Hollywood-[86], UCF-Sports [138], AVAD [139], Coutrot1 [140], Coutrot2 [141], ETMD [142], and SumMe [112].
- *Video Object Segmentation:* Video object segmentation aims at partitioning a video into segments that correspond to individual objects, accurately delineating their boundaries over time. YouTube-VOS [143] and DAVIS [144] are datasets related to this task.
- *Video Instance Segmentation:* Video instance segmentation is the task of identifying, segmenting, and tracking each unique object instance within a video. YouTube-VIS [145] and Cityscapes-Seq [146] are two common benchmarks for this task.
- *Video Object Referring Segmentation:* Video object referring segmentation involves segmenting specific objects in a video based on language descriptions. It identifies and isolates the referred objects accurately across frames, where MeViS [147] is a common benchmark for this task.
- *Spatiotemporal Grounding:* Spatiotemporal grounding aims to identify and localize specific objects or events within a video's spatial and temporal dimensions based on a given query. Datasets like Vid-STG [148], HC-STVG [149], Ego4D-MQ and Ego4D-NLQ [150] are proposed to aid the training and testing for this task.

B. Background for LLMs

Language models are trained to learn a joint probability distribution $p(x_{1:L})$ with a sequence of text tokens $x_{1:L}$. This joint distribution is usually equivalent to a product of the conditional probabilities conditioned on each token with the chain rule:

$$p(x_{1:L}) = \prod_{i=1}^{L} p(x_i | x_{1:i-1}), \tag{1}$$

where L is the sequence length.

Large Language Models (LLMs) refer to language models with a large number of parameters, *e.g.*, billions. The architecture of LLMs incorporates a text tokenizer and multiple self-attention layers. LLMs are trained in a teacher-forcing manner to predict the next token's probability, where the generation process utilizes the autoregressive paradigm:

$$\mathcal{M}(x_{1:i-1}) = p(x_i \mid x_{1:i-1}), \tag{2}$$

where $\mathcal{M}$ represents an LLM.

Decoding strategies dictate how to harness the next token probability and select the next token y_t from the set S of all possible tokens in the vocabulary, which includes special tokens such as <SOS>, <EOS>, and <PAD>. Greedy decoding, the simplest strategy, selects the token with the highest probability, formalized as:

$$x_t = \arg\max_{s \in S} \log p_{\mathcal{M}}(s \mid \boldsymbol{x}_{1:t-1}). \tag{3}$$

Besides deterministic strategies, sampling strategies that randomly select the next tokens using model probability are also popular in real applications. These strategies provide diverse outputs and enable self-consistency methods.

Large language models typically exhibit the following characteristics:

- *Scaling Laws* [151]: With a significant extension of the model size (number of parameters), the pertaining data size, and computational resources, the performance of the model exhibits a pattern of regular growth, which can help researchers and engineers predict performance improvements or make effective decisions on model design and training.
- *Emergent Abilities* [77]: When the parameter size and volume of training data for a large language model exceed a certain magnitude, some novel capabilities emerge, such as in-context learning, instruction following, and step-by-step reasoning. *In-context learning* allows the model to learn and make predictions based on the context provided within the input text without requiring explicit retraining. *Instruction following* enables the model to perform tasks based on natural language instructions. *Step-by-step reasoning*, e.g. chain-of-thought (CoT), allows the model to follow a logical sequence of steps to arrive at a conclusion, which is particularly useful for solving complex problems.

LLMs possess extensive generalization abilities that can be applied to various downstream tasks, including multi-modal tasks. Multimodal Large Language Models (MLLMs) [76], [152]–[155] typically incorporate multimodal encoders, cross-modality aligners, and an LLM core structure. By combining multimodal encoders with the LLM, MLLMs excel at integrating visual and linguistic contexts to produce detailed content.

III. Vid-LLMs

In this section, we introduce a novel taxonomy of Vid-LLMs, providing a comprehensive overview of their classification. Following this, we explore the diverse training strategies that empower Vid-LLMs to achieve their capabilities.

A. Taxonomy

Based on the method of processing input video, we categorize Vid-LLMs into three primary types: *Video Analyzer* × *LLM*, *Video Embedder* × *LLM*, and *(Analyzer + Embedder)* × *LLM*. Each category represents a unique approach to integrating video processing with LLMs, as illustrated in Figure 4.

1) Video Analyzer × *LLM:* Video Analyzer is defined as a module that takes video input and outputs an analysis of the video, typically in text form, which facilitates LLM processing. This text may include video captions, dense video captions (detailed descriptions of all events in the video with timestamps), object tracking results (labels, IDs, and bounding boxes of objects), as well as transcripts of other modalities present in the video, such as speech recognition results from ASR or subtitle recognition results from OCR. The text generated by the Video Analyzer can be directly fed into the subsequent LLM, inserted into pre-prepared templates before being fed into the LLM, or converted into a temporary database format for the LLM to retrieve later.

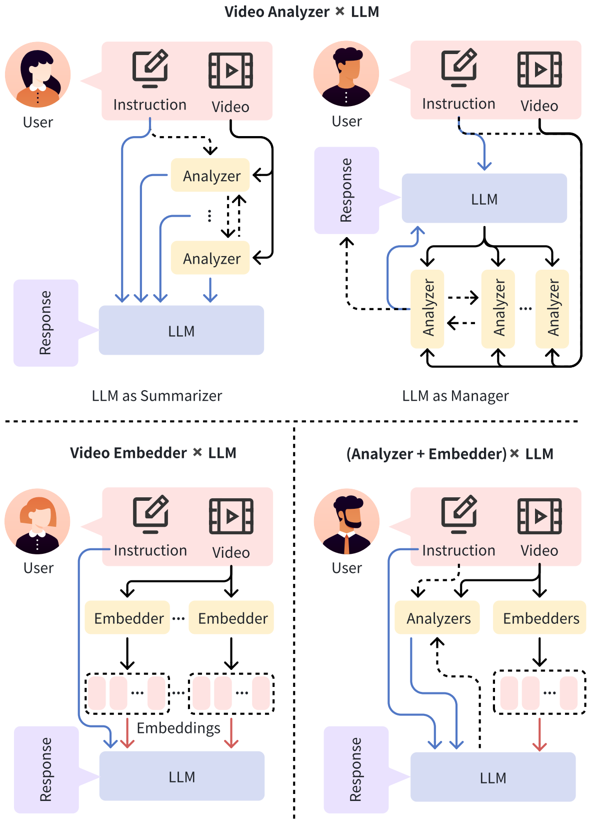

Fig. 4. The figure illustrates three primary frameworks for Vid-LLMs: (1) *Video Analyzer* × *LLM*, where video analyzers convert video inputs to textual analysis for the LLM; (2) *Video Embedder* × *LLM*, where video embedders generate vector representations (embeddings) for the LLM to process; (3) *(Analyzer + Embedder)* × *LLM*, a hybrid approach combining both analyzers and embedders to provide the LLM with textual analysis and embeddings. The arrows indicate the direction of information flow, with dashed arrows representing optional paths. Blue arrows denote textual information flow, while red arrows denote embeddings.

For the *Video Analyzer* × *LLM* category, we further create two subcategories, *LLM as Summarizer* and *LLM as Manager*, based on the function of the LLM within the Vid-LLM system:

- *LLM as Summarizer:* In this subcategory, the primary function of the LLM is to summarize the analysis obtained from the Video Analyzers. The summarization approach varies based on the prompts provided to the LLM, ranging from highly condensed summary texts and captions to comprehensive summaries for answering specific questions. Notably, in *Video Analyzer* × *LLM as Summarizer* systems, the information flow is usually unidirectional (see Figure 4), with data flowing from the video to the Video Analyzer and then to the LLM, without any reverse process. Examples of Vid-LLMs in the *Video Analyzer* × *LLM as Summarizer* category include: LaViLa [168], VLog [167], VAST [105], AntGPT [166], VIDOSC [165], Grounding-Prompter [164], LLoVi [163], Video ReCap [162], MVU [161], LangRepo [160], IG-VLM [159], MoReVQA [158], MM-Screenplayer [157],

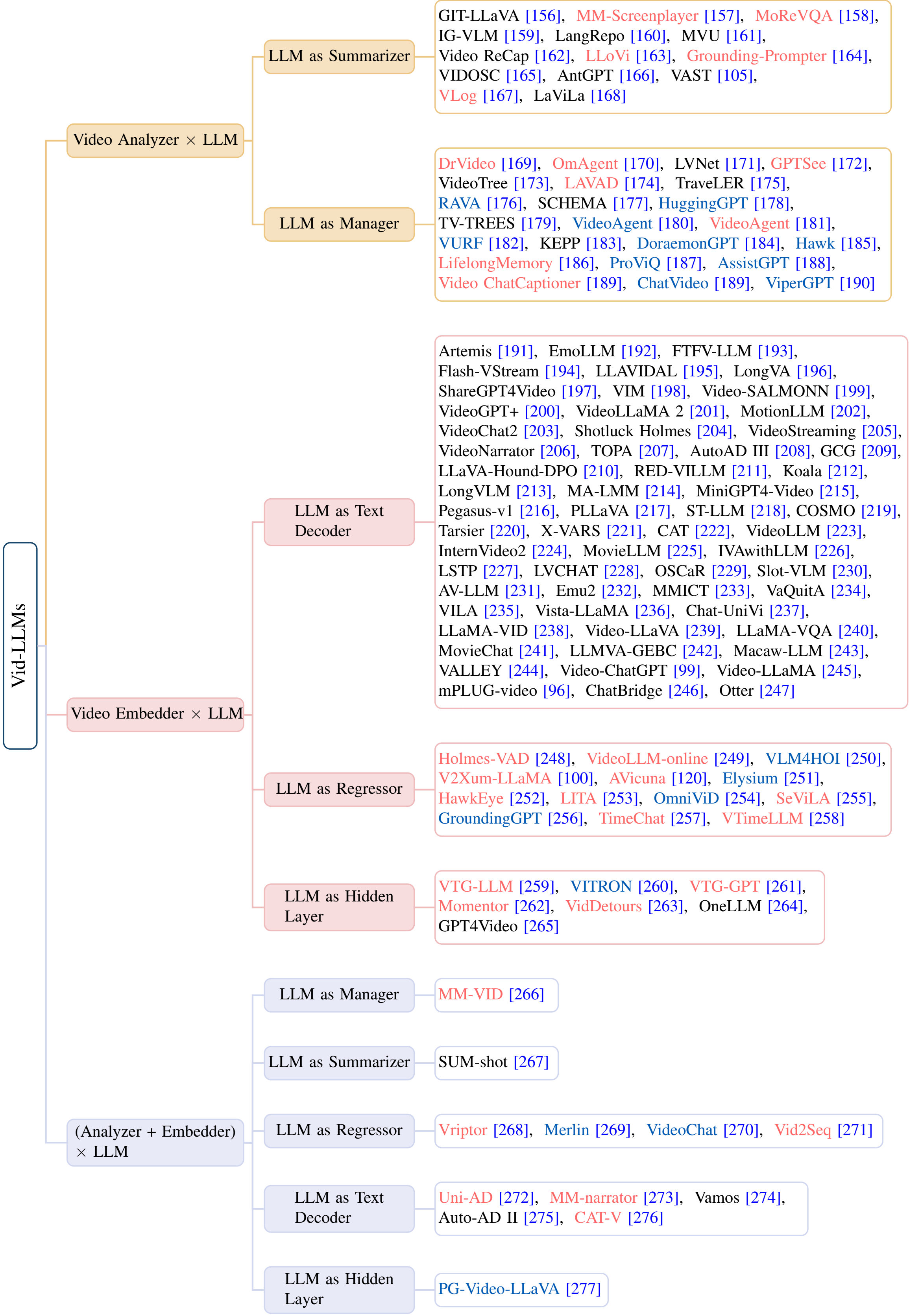
Vid-LLMs
Video Analyzer × LLM
LLM as Summarizer
GIT-LLaVA [156], MM-Screenplayer [157], MoReVQA [158], IG-VLM [159], LangRepo [160], MVU [161], Video ReCap [162], LLoVi [163], Grounding-Prompter [164], VIDOSC [165], AntGPT [166], VAST [105], VLog [167], LaViLa [168]
LLM as Manager
DrVideo [169], OmAgent [170], LVNet [171], GPTSee [172], VideoTree [173], LAVAD [174], TraveLER [175], RAVA [176], SCHEMA [177], HuggingGPT [178], TV-TREES [179], VideoAgent [180], VideoAgent [181], VURF [182], KEPP [183], DoraemonGPT [184], Hawk [185], LifelongMemory [186], ProViQ [187], AssistGPT [188], Video ChatCaptioner [189], ChatVideo [189], ViperGPT [190]
Video Embedder × LLM
LLM as Text Decoder
Artemis [191], EmoLLM [192], FTFV-LLM [193], Flash-VStream [194], LLAVIDAL [195], LongVA [196], ShareGPT4Video [197], VIM [198], Video-SALMONN [199], VideoGPT+ [200], VideoLLaMA 2 [201], MotionLLM [202], VideoChat2 [203], Shotluck Holmes [204], VideoStreaming [205], VideoNarrator [206], TOPA [207], AutoAD III [208], GCG [209], LLaVA-Hound-DPO [210], RED-VILLM [211], Koala [212], LongVLM [213], MA-LMM [214], MiniGPT4-Video [215], Pegasus-v1 [216], PLLaVA [217], ST-LLM [218], COSMO [219], Tarsier [220], X-VARS [221], CAT [222], VideoLLM [223], InternVideo2 [224], MovieLLM [225], IVAwithLLM [226], LSTP [227], LVCHAT [228], OSCaR [229], Slot-VLM [230], AV-LLM [231], Emu2 [232], MMICT [233], VaQuitA [234], VILA [235], Vista-LLaMA [236], Chat-UniVi [237], LLaMA-VID [238], Video-LLaVA [239], LLaMA-VQA [240], MovieChat [241], LLMVA-GEBC [242], Macaw-LLM [243], VALLEY [244], Video-ChatGPT [99], Video-LLaMA [245], mPLUG-video [96], ChatBridge [246], Otter [247]
LLM as Regressor
Holmes-VAD [248], VideoLLM-online [249], VLM4HOI [250], V2Xum-LLaMA [100], AVicuna [120], Elysium [251], HawkEye [252], LITA [253], OmniViD [254], SeViLA [255], GroundingGPT [256], TimeChat [257], VTimeLLM [258]
LLM as Hidden Layer
VTG-LLM [259], VITRON [260], VTG-GPT [261], Momentor [262], VidDetours [263], OneLLM [264], GPT4Video [265]
(Analyzer + Embedder) × LLM
LLM as Manager
MM-VID [266]
LLM as Summarizer
SUM-shot [267]
LLM as Regressor
Vriptor [268], Merlin [269], VideoChat [270], Vid2Seq [271]
LLM as Text Decoder
Uni-AD [272], MM-narrator [273], Vamos [274], Auto-AD II [275], CAT-V [276]
LLM as Hidden Layer
PG-Video-LLaVA [277]

Fig. 5. Taxonomy of Video Understanding with Large Language Models (Vid-LLMs), consists of *Video Analyzer × LLM*, *Video Embedder × LLM* and *(Analyzer + Embedder) × LLM*, and the sub-categories are *LLM as Summarizer/Manager/Text Decoder/Regressor/Hidden Layer*. Font color indicates the granularity of video understanding supported by the Vid-LLMs: black for abstract understanding, red for temporal understanding, and blue for spatiotemporal understanding.

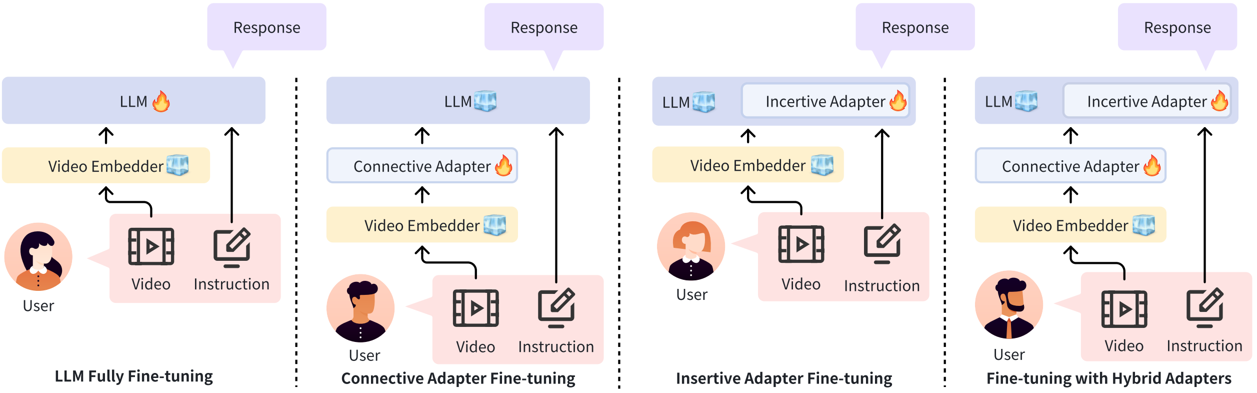

Fig. 6. Four types of Vid-LLMs fine-tuning strategies, classified according to their specific training methods: LLM fully fine-tuning, fine-tuning with connective adapters, insertive adapters, and hybrid methods.

GIT-LLaVA [156], etc.

- *LLM as Manager:* In this subcategory, the LLM primarily coordinates the overall system's operation. It may actively generate commands to invoke various Video Analyzers to produce the desired results for the user, call the Video Analyzer and further process the obtained analysis before returning it to the user, or engage in multiple rounds of interaction with the Video Analyzer. Compared to the *LLM as Summarizer category*, the *LLM as Manager* category is more flexible and can be distinguished by its information flow complexity. Examples of Vid-LLMs in the *Video Analyzer* × *LLM as Manager* category include: ViperGPT [190], Video ChatCaptioner [278], ChatVideo [189], AssistGPT [188], HuggingGPT [178], Hawk [185], ProViQ [187], LifelongMemory [186], DoraemonGPT [184], KEPP [183], VURF [182], VideoAgent (by Stanford) [181], VideoAgent (by PKU) [180], TV-TREES [179], SCHEMA [177], RAVA [176], GPTSee [172], TraveLER [175], LAVAD [174], VideoTree [173], LVNet [171], OmAgent [170], DrVideo [169], etc.

2) Video Embedder × LLM: A Video Embedder typically refers to a visual backbone/video encoder, such as ViT or CLIP, used to convert input videos into vector representations, known as video embeddings or video tokens. Some Embedders encode other modalities within the video, such as audio (e.g., CLAP [279]), which are also categorized under Video Embedder here (note that we do not consider the LLM's tokenizer as an embedder). Unlike the text generated by the Video Analyzer, the vectors generated by the Video Embedder cannot be directly utilized by the LLM and usually require an adapter to map these embeddings from the vision's (or other modalities') semantic space to the text semantic space of the LLM's input tokens.

For the *Video Embedder* × *LLM* category, we also classify them into subcategories based on the LLM's function within the Vid-LLM system:

- *LLM as Text Decoder:* In this subcategory, the LLM receives embeddings from the Video Embedder as input and decodes them into text outputs based on prompts or instructions. These tasks generally do not require fine-grained understanding or precise spatiotemporal localization, focusing mainly on general QA or captioning. Thus, the Vid-LLM behaves like a standard LLM during decoding. Examples of Vid-LLMs in the *Video Embedder* × *LLM as Text Decoder* category include: VideoLLM [223], Otter [247], Video-LLaMA [245], Video-ChatGPT [99], VALLEY [244], Macaw-LLM [243], MovieChat [241], Video-LLaVA [239], Chat-UniVi [237], Vista-LLaMA [236], VILA [235], GPT4Video [265], MovieLLM [225], InternVideo2 [224], MiniGPT4-Video [215], VideoChat2 [203], VideoLLaMA 2 [201], etc. See Figure 5 for the complete list.
- *LLM as Regressor:* In this subcategory, the LLM receives embeddings from the Video Embedder as input and, like the Text Decoder, can output text. However, unlike the Text Decoder, the *LLM as Regressor* can also predict continuous values, such as timestamp localization in videos and bounding box coordinates for object trajectories, functioning similarly to a regressor performing regression tasks, even though it is fundamentally performing classification. Examples of Vid-LLMs in the *Video Embedder* × *LLM as Regressor* category include: VTimeLLM [258], SeViLA [255], TimeChat [257], GroundingGPT [256], OmniViD [254], LITA [253], HawkEye [252], Elysium [251], AVicuna [120], V2Xum-LLaMA [100], VLM4HOI [250], VideoLLM-online [249], Holmes-VAD [248], etc.
- *LLM as Hidden Layer:* In this subcategory, the LLM also receives video embeddings as input but does not directly output text. Instead, it connects to a specially designed task-specific head to perform actual regression tasks, such as event time localization or object bounding box prediction in videos, while maintaining the LLM's text output capability. Examples of Vid-LLMs in the *Video Embedder* × *LLM as Hidden Layer* category include: GPT4Video [265], OneLLM [264], VidDetours [263], Momentor [262], VTG-GPT [261], VITRON [260], VTG-LLM [259], etc.

3) (Analyzer + Embedder) × LLM: This category of Vid-LLMs is relatively rare. As the name suggests, it involves simultaneously using a Video Analyzer to obtain textual analysis of the video and a Video Embedder to encode the video into embeddings. The LLM receives both types of inputs along with other prompts/instructions and outputs responses to complete tasks. The subcategories here can flexibly be any of the Summarizer/Manager/Text Decoder/Regressor/Hidden Layer categories. Vid-LLMs in the *(Analyzer + Embedder) × LLM* category include: Vid2Seq [271], VideoChat [270], MM-VID [266], Auto-AD II [275], Vamos [274], PG-Video-LLaVA [277], MM-Narrator [273], SUM-shot [267], Merlin [269], Uni-AD [272], Vriptor [268], etc.

B. Training Strategies for Vid-LLMs

1) Training-free Vid-LLMs: Many Vid-LLMs systems are built on powerful LLMs with strong zero-shot, in-context learning, and Chain-of-Thought capabilities. These systems do not require training in the parameters of the LLM or other modules. Most Vid-LLMs in the *Video Analyzer × LLM* category are training-free because the information from the video and other accompanying modalities has already been parsed into text. At this point, the video understanding task has been transformed into a text understanding task. Since LLMs can unify almost all NLP tasks into generation tasks, they can also handle many video understanding tasks. SlowFast-LLaVA [280] is a training-free Video LLM that uses a two-stream input design to capture both spatial semantics and temporal context without fine-tuning and demonstrates capabilities across various video understanding benchmarks.

2) Fine-tuning Vid-LLMs: In contrast to most Vid-LLMs in the *Video Analyzer × LLM* category being training-free, almost all Vid-LLMs in the *Video Embedder × LLM* category undergo fine-tuning. The common methods for fine-tuning Vid-LLMs are categorized based on the types of adapters used during fine-tuning into four main types: LLM Fully Fine-tuning, Connective Adapter Fine-tuning, Insertive Adapter Fine-tuning, and Fine-tuning with Hybrid Adapters. An adapter is a small, trainable module added to a large model for fine-tuning. By only updating the parameters of these modules, the model can adapt to specific tasks without changing the entire model's parameters, achieving efficient parameter updates and task adaptation while conserving computational resources. Illustrations of each type are shown in Figure 6.

- *LLM Fully Fine-tuning:* This fine-tuning method does not use any adapters but instead employs supervised training with a lower learning rate, updating all the parameters in the LLM. This method allows the Vid-LLM to fully adapt to the respective task and achieve good performance, especially when the target task is quite different from the pretraining tasks. For end-to-end Vid-LLMs, especially those in the *Video Embedder × LLM* category, the Video Embedder may also be fine-tuned for more comprehensive learning. However, this method consumes more computational resources than adapter-based fine-tuning methods and may potentially impair the inherent capabilities of the LLM, such as zero-shot and in-context learning. Vid-LLM adopted LLM Fully Fine-tuning include AV-LLM [231] and Vid2Seq [271]. In [231], there are both fully fine-tuning and adapter fine-tuning versions of Vid-LLMs, and the former's performance is better than the latter.
- *Connective Adapter Fine-tuning:* Here, the term "Connective" refers to adapters that bridge the Video Embedder and the LLM externally, enabling information from the video to flow into the LLM through the Connective Adapter. As illustrated in Figure 6, during training, the parameters of both the Video Embedder and the LLM are frozen, and only the parameters of the Connective Adapter are updated. Common Connective Adapters include MLP/Linear Layer and Q-former [283], their combinations, etc., whose primary function is to map video embeddings from the visual semantic space to the text semantic space of the LLM input tokens (i.e., modality alignment). Typically, fine-tuning only the Connective Adapter does not alter the LLM's inherent behavior.
- *Insertive Adapter Fine-tuning:* As the name suggests, Insertive Adapters are inserted into the LLM itself. Similar to using Connective Adapters, during training, the parameters of the Video Embedder and the LLM are frozen, and only the parameters of the Insertive Adapter are updated. Insertive Adapters, often based on LoRA, affect the LLM's behavior because they are added to the existing LLM parameters. This type of adapter is almost always present in Vid-LLMs classified as *Video Embedder × LLM as Regressor* and *Video Embedder × LLM as Hidden Layer*, as these types of Vid-LLMs require changes in the LLM's behavior, such as outputting continuous prediction values.
- *Fine-tuning with Hybrid Adapters:* Many Vid-LLMs use a combination of Connective and Insertive Adapters to achieve both modality alignment and changes in the LLM's inherent behavior. Vid-LLMs employing Hybrid Adapters typically use multi-stage fine-tuning. A common approach is to fine-tune only the Connective Adapter in the first stage for modality alignment. In the second stage, the already fine-tuned Connective Adapter is frozen, the training task (from alignment task to target task) and the training data (from data used for modality alignment to data required for the target task) are changed, and only the parameters of the Insertive Adapter are updated. There are also single-stage approaches where both Connective and Insertive Adapters are updated simultaneously.

IV. Benchmarks and Evaluation

This section provides an overview of the evaluation methods for video question-answering models and related tasks, categorized into three types: closed-ended evaluation, open-ended evaluation, and other evaluation methods, which are shown in Table III. Closed-ended evaluations rely on questions with predefined answers or formats, including multiple-choice questions and structured formats that allow for straightforward scoring. Open-ended evaluations involve questions without predefined answer options, often requiring more sophisticated

TABLE I
Comparison of video understanding models with large language models, sorted by their release dates. The table presents key details for each method, including the number of training frames, video embedders, utilization of audio information, model adaptation approaches, computational resources, the specific large language model employed, and the corresponding number of parameters. Entries marked with a hyphen ("-") indicate undisclosed details in the respective papers.

Model	#Frame	Video Embedder	Sound	Speech	Adapter	Hardware	LLM	LLM Size	Date
Socratic Models [281]	Varying	CLIP ViT-L/14	✗	✓	-	1 V100 GPU	RoBERTa, GPT-3	-	05/2022
LaViLa [168]	4	TimeSformer-B/L	✗	✗	Cross-Attention	32 V100 GPUs	GPT-2 XL frozeen	-	12/2022
Vid2Seq [271]	100	CLIP ViT-L/14	✗	✓	Transformer Encoder	64 TPU v4	T5	0.2B	02/2023
ViperGPT [190]	-	-	✗	✗	-	-	-	-	03/2023
Vid ChatCaptioner [278]	100	-	✗	✗	BLIP-2	24G GPU	ChatGPT	20B	04/2023
VLog [167]	-	CLIP ViT-G	✓	✓	-	-	ChatGPT	20B	04/2023
ChatVideo [189]	-	-	✓	✓	-	-	ChatGPT	20B	04/2023
VideoChat [270]	4-32	ViT-G	✗	✓	MLP+Q-fomer	1 A10 GPU	StableVicuna	7B	05/2023
ChatBridge [246]	4	ViT-G	✓	✗	Perceiver	8 A100 GPUs	GPT-4	-	05/2023
VAST [105]	4,8	EVA-CLIP, BEATs, BERT-B	✓	✓	-	64 V100 GPUs	Vicuna	13B	05/2023
Otter [247]	4-8	CLIP ViT-L/14	✗	✗	-	4 RTX-3090 GPUs	LLaMA	7B	05/2023
VideoLLM [223]	Varying	7 Task-Specific Video Encoders	✗	✗	Linear Layer	-	GPT-2/T5/OPT/LLaMA	1.5/6.5/6.7/7B	05/2023
AssistGPT [188]	1/3 FPS	-	✗	✓	-	4 A5000 GPUs	Vicuna	7B	06/2023
Video-LLaMA [245]	8	CLIP ViT-G	✓	✗	Video Q-former	-	Vicuna	7B	06/2023
Video-ChatGPT [99]	100	CLIP ViT-L/14	✓	✓	Linear Layer	8 A100 GPUs	Vicuna-v1.1	7B	06/2023
LLMVA-GEBC [242]	96	CLIP ViT-G	✗	✗	Video Q-former	2 A6000 GPUs	OPT	13B	06/2023
mPLUG-video [96]	8	TimeSformer	✗	✗	-	-	GPT/Blood	-	06/2023
VALLEY [244]	Varying	CLIP ViT-L/14	✗	✗	Projection Layer	8 A100 GPUs	StableVicuna	7B/13B	06/2023
Macaw-LLM [243]	Varying	CLIP VIT-B/16	✗	✓	Alignment and Integration	-	LLaMA/Vicuna/Bloom	7B	06/2023
AntGPT [166]	4	CLIP ViT-L/14	✗	✗	-	A6000 GPUs	Llama2	7B	07/2023
MovieChat [241]	2048	ViT-G/14, EVA-CLIP	✗	✗	Q-former+LSTM	-	GPT-3.5/Claude	-	07/2023
FAVOR [282]	Varying	InstructBLIP ViT-G/14	✓	✓	Causal Q-Former	-	Vicuna	7B/13B	10/2023
Auto-AD II [275]	Varying	CLIP ViT-B/32	✗	✗	-	-	GPT-2	-	10/2023
LLaMA-VQA [240]	-	CLIP ViT-L/14	✗	✗	Linear Layer	8 A6000 GPUs	LLaMA	7B	10/2023
MM-VID [266]	Varying	-	✗	✓	-	-	GPT4	-	10/2023
Video-LLaVA [239]	8-32	LanguageBind	✓	✗	MLP Projection Layer	-	Vicuna/LLaVA	7B/13B	11/2023
PG-Video-LLaVA [277]	Varying	CLIP ViT-L/14	✗	✓	MLP Projection Layer	4 A100 GPUs	Vicuna/LLaVA	7B/13B	11/2023
SeViLA [255]	32	BLIP-2 ViT	✗	✗	Q-former	4 48GB A6000 GPUs	MiniGPT4	-	11/2023
Vamos [274]	Varying	CLIP ViT-L/14, BLIP-2 ViT	✗	✗	LORA / LLaMA-Adapter	A6000 GPU	LLaMA/LLaMA2	7B	11/2023
VTimeLLM [258]	100	CLIP ViT-L/14	✗	✗	Linear Layer	1 RTX-4090 GPU	Vicuna	7B/13B	11/2023
Chat-UniVi [237]	64	ViT-L/14, ViT-G	✗	✗	Spatial Temporal Merging	-	Vicuna	7B	11/2023
MM-narrator [273]	Varying	CLIP ViT	✗	✓	-	-	GPT-4	-	11/2023
GPT4Video [265]	Varying	CLIP ViT-L/14	✗	✗	Cross-Attention	8 A100 GPUs	LLaMA	7B	11/2023
LLaMA-VID [238]	1fps	CLIP ViT-L/14	✗	✗	Projector	8 A100 GPUs	Vicuna	7B/13B	11/2023
LLoVi [163]	0.5fps	CLIP, TimeSformer, BLIP	✗	✗	-	32 RTX 3090 GPUs	GPT-4	-	12/2023
AV-LLM [231]	32	CLIP ViT-L/14	✗	✗	Linear Projectors	8×A100 GPUs	Vicuna	7B	12/2023
Emu2 [232]	Varying	EVA-02-CLIP-E-plus	✗	✗	Linear Projection	-	Emu2-Chat	38B	12/2023
Grounding-Prompter [164]	Varying	BLIP	✗	✓	-	-	GPT-3.5-turbo-16k	-	12/2023
VIDOSC [165]	1fps	CLIP	✗	✗	-	-	GPT4	-	12/2023
LifelongMemory [186]	360	CLIP ViT-L/14, TimeSformer	✗	✗	-	-	GPT-4	-	12/2023
Merlin [269]	8	CLIP ViT-L/14	✗	✗	2D Conv + Linear Layer	-	Vicuna v1.5	7B	12/2023
MMICT [233]	16	BLIP-2 ViT-G/14	✗	✗	Transformers + Linear Projection	4 V100 GPUs	FlanT5XL, OPT	2.7B	12/2023
OneLLM [264]	-	CLIP ViT-L/14	✗	✗	K Transformer Projectors + 1 MLP	16 A100 GPUs	LLaMA2	7B	12/2023
SUM-shot [267]	4	EVA-CLIP ViT-G/14	✓	✓	Q-former	16 A100-80G GPUs	Vicuna-v0	7B	12/2023
TimeChat [257]	96	BLIP-2 ViT-G/14	✗	✗	Video Q-Former+Linear Layer+LoRA	-	LLaMA-2	7B	12/2023
VaQuitA [234]	100	CLIP ViT-L/14	✗	✗	Video Perceiver + VQ-Former	8 A100 80GB GPUs	Llama2	7B	12/2023
VILA [235]	8	CLIP ViT-L/14	✗	✗	Linear Projector	16 A100 GPUs	Llama-2	7B/13B	12/2023
Vista-LLaMA [236]	-	EVA-CLIP, CLIP ViT-L/14	✗	✗	Sequential Q-Former with linear layer	8 A100 80GB GPUs	Vicuna	7B	12/2023
ProViQ [187]	-	TimeSformer	✗	✓	-	one A100 GPU	gpt-3.5-turbo	-	12/2023
COSMO [219]	3	CLIP ViT-L/14	✗	✗	Gated Cross-Attention Layers	128 V100 GPUs	OPT-IML/RedPajama/Mistral	1.8/3/7B	01/2024
VidDetours [263]	-	InternVideo	✗	✗	Linear Layer	8 A100 GPUs	LLaMA-2	13B	01/2024
DoraemonGPT [184]	-	-	✗	✓	-	-	GPT-3.5-turbo	-	01/2024
GroundingGPT [256]	64	CLIP ViT-L/14	✗	✗	Q-former, MLP	8 A100 GPUs	Vicuna-v1.5	7B	01/2024
GCG [209]	4	EVA-CLIP	✗	✗	Q-former	-	IntructBLIP-Vicuna	7B	01/2024
IVAwithLLM [226]	-	no mention	✗	✗	Linear Layer	8 A100 GPUs	Vicuna	7B	02/2024
LSTP [227]	Varying	BLIP-2 ViT-G/14	✗	✗	Q-former	1 A100 GPU	Vicuna	7B	02/2024
LVCHAT [228]	Varying	UMT-L	✗	✗	QFormer + Linear Layer	4 A6000 GPUs	Vicuna	7B	02/2024
Momentor [262]	300	CLIP ViT-L/14	✗	✗	Projection Layer	8 A1000 GPUs	LLaMA	7B	02/2024
OSCaR [229]	-	CLIP ViT-L/14	✗	✗	-	-	Vicuna	7/13B	02/2024
Slot-VLM [230]	1	CLIP ViT-L/14	✗	✗	Linear Layer	1 A100 GPU	Vicuna	7B	02/2024
Video ReCap [162]	4	TimeSformer	✗	✗	-	8 V100 GPUs	GPT-2	1.5B	02/2024
MVU [161]	16	-	✗	✗	-	2 A5000 GPUs	Mistral/LLaMA-2/Gemma	7B/13B	03/2024
VideoAgent [180]	Varying	ViCLIP	✗	✗	-	-	GPT-4	-	03/2024
VideoAgent [181]	1 FPS	EVA-CLIP-8B-plus	✗	✗	-	1 A6000 GPU	GPT-4	-	03/2024
VTG-GP [261]	0.5 FPS	EVA-CLIP	✗	✗	-	8 RTX-3090 GPUs	LLaMA-2	7B	03/2024
VURF [182]	-	-	✗	✗	-	-	GPT-3.5	-	03/2024
KEPP [183]	-	-	✗	✗	-	2 A100 GPUs	LLaMA-2	13B/70B	03/2024
GPTSee [172]	-	CLIP ViT-B/32, SlowFast	✗	✗	Cross-attention	8 RTX 3090 GPUs	-	-	03/2024
HawkEye [252]	Varying	UMT-L	✗	✓	Q-former + LORA	8 V100 GPUs	GPT-4	-	03/2024
InternVideo2 [224]	8	InternVL-6B+VideoMAE-g	✓	✓	Q-former + LoRA	256 A100 GPUs	Mixer-7B + BERT-large	7B	03/2024
LangRepo [160]	-	-	✗	✗	-	-	Mistral	7B/12B	03/2024
LITA [253]	100	CLIP ViT-L/14	✗	✗	MLP + SlowFast Token Pooling	8 A100 GPUs	Vicuna	7B/13B	03/2024
MovieLLM [225]	1 FPS	CLIP ViT-L/14	✗	✗	-	4 A100 GPUs	Vicuna	7B	03/2024
OmniViD [254]	32	VideoSwin	✗	✗	Mixed Q-former+Visual Translator	8 A100 GPUs	GPT-4	-	03/2024
RAVA [176]	30 FPS	-	✗	✗	-	-	GPT-4	-	03/2024
SCHEMA [177]	-	CLIP ViT-L/14	✗	✗	-	1 V100 GPU	GPT-3.5	-	03/2024
TV-TREES [179]	2 FPS	CLIP ViT-L/14	✗	✗	-	-	GPT-3.5	-	03/2024
IG-VLM [159]	6	CLIP ViT-L/14	✗	✗	-	-	LLaVA-v1.6, and GPT-4V	7B/13B/34B	03/2024
AVicuna [120]	100	CLIP ViT	✓	✗	MLP, LoRA	1 A6000 GPU	Vicuna-v1.5	7B	03/2024
CAT [222]	-	ImageBind	✓	✗	Linear Projector	1 A100 GPU	LLaMA-2	7B	03/2024
Uni-AD [272]	-	CLIP ViT-B/32	✓	✗	Video mapping network	8 A100 GPUs	LLaMA-2	7B	03/2024
Elysium [251]	Varying	CLIP-ViT-L	✗	✗	T-Selector	24 A100-80G	Vicuna	-	03/2024

TABLE II
CONTIUNE OF TABLE I.

Model	#Frame	Video Embedder	Sound	Speech	Adapter	Hardware	LLM	LLM Size	Date
VideoTree [173]	Varying	EVA-CLIP ViT-G/14	✗	✗	-	4 A6000 GPUs	GPT-4	-	04/2024
VTG-LLM [259]	96	EVA-CLIP ViT-G/14	✗	✗	Projector	6 ATN 910B	LLaMA-2	7B	04/2024
AutoAD III [208]	8	EVA-CLIP	✓	✓	Linear Projector	1 A40 GPU	GPT-3.5-turbo	-	04/2024
LLaVA-Hound-DPO [210]	10	LanguageBind ViT/14	✗	✗	-	-	GPT-4V	-	04/2024
RedViLLM [211]	Varying	CLIP ViT-G/14	✗	✗	Temporal Module	-	Qwen-VL	7B	04/2024
LAVAD [174]	16	BLIP-2 ViT-L/14	✗	✗	-	-	LLaMA-2-chat	13B	04/2024
VLM4HOI [250]	-	EVA CLIP	✗	✗	Projection Layer	8 V100 GPUs	LLaMA-2-Chat	13B	04/2024
Koala [212]	64	EVA-CLIP ViT-G/14	✗	✗	Video QFormer+Linear Layer	4 RTX A6000 GPUs	Vicuna-v0	7B/13B	04/2024
LongVLM [213]	100	CLIP ViT-L/14	✗	✗	Projection Layer	4 A100 GPUs	Vicuna-v1.1	7B	04/2024
MA-LMM [214]	100	EVA-CLIP ViT-G/14	✗	✗	Q-Former	4 A100 GPUs	Vicuna	7B	04/2024
MiniGPT4-Video [215]	Varying	EVA-CLIP	✗	✗	Projector + LoRA for LLM	-	LlaMA-2/Mistral	7B	04/2024
MoReVQA [158]	1 FPS	-	✗	✗	-	-	PaLM-2	-	04/2024
Pegasus-v1 [216]	-	Marengo 2.6	✗	✓	VL Alignment Module	-	GPT-4	-	04/2024
PLLaVA [217]	16	CLIP ViT-L/14	✗	✗	MM Projector	-	LLaVA	7B/13B/34B	04/2024
ST-LLM [218]	Varying	CLIP ViT-L/14	✗	✗	Projection Layer	8 A100 GPUs	Vicuna-v1.1	7B	04/2024
Tarsier [220]	Varying	CLIP ViT	✗	✗	MLP	48 A100 GPUs	Vicuna	7B/13B	04/2024
TraveLER [175]	Varying	-	✗	✗	-	8 RTX A6000 GPUs	GPT-3.5	-	04/2024
V2Xum-LLaMA [100]	1 FPS	CLIP ViT-L/14	✗	✗	Vision Adapter	8 A100 GPUs	LLaMA	7B/13B	04/2024
VITRON [260]	-	-	✗	✗	Projection Layer	10 A100 GPUs	Vicuna-v1.5	7B	04/2024
X-VARS [221]	16	CLIP ViT-L/14	✗	✗	Linear Layer	1 V100 GPU	VideoChatGPT	-	04/2024
VideoChat2 [203]	Varying	UMT-L	✗	✗	QFormer	-	Vicuna	7B	05/2024
Shotluck Holmes [204]	120	SigLip	✗	✗	MLP Projection Layer	8 H100 GPUs	Vicuna	7B	05/2024
VideoStreaming [205]	16	CLIP ViT-L/14	✗	✗	Projecter	32 A100 GPUs	Vicuna	7B	05/2024
VideoNarrator [206]	-	CLIP ViT-L/14	✗	✗	Video Projecter	4 A6000 GPUs	Baichuan	7B	05/2024
TOPA [207]	10	CLIP ViT-L/14	✗	✗	Linear Projector	4 A100 GPUs	Llama2	7/8/13B	05/2024
MotionLLM [202]	8	LanguageBind/VQ-VAE	✗	✗	Linear Layer	A100 GPUs	Vicuna	7B	05/2024
Artemis [191]	Varying	CLIP ViT-L/14	✗	✗	MLP	8 A800 GPUs	Vicuna	7B	06/2024
DrVideo [169]	2 FPS	CLIP ViT-L/14	✗	✗	-	-	LaViLa	7B	06/2024
EmoLLM [192]	8	CLIP ViT-L/14	✓	✓	Multi-Perspective Visual Projection	4 4090 GPUs	Vicuna-v1.5	7B	06/2024
FTFV-LLM [193]	8	OpenCLIP ViT-G/14	✗	✗	Vision-Language Adapter	64 A100 GPUs	Vicuna	7B	06/2024
Flash-VStream [194]	8	-	✗	✗	MLP	8 A100 GPUs	GPT-3.5	-	06/2024
Holmes-VAD [248]	Varying	CLIP ViT-L/14	✗	✗	Temporal Sampler	2 A100 GPUs	Vicuna-v1.5	7B	06/2024
LongVA [196]	1 FPS	CLIP ViT-L/14	✗	✗	MLP	8 A100 GPUs	Mistral-7B-Instruct-v0.2	7B	06/2024
OmAgent [170]	Varying	-	✗	✓	-	-	GPT-4o	-	06/2024
GIT-LLaVA [156]	6	CLIP ViT-L/14	✗	✗	MLP	4 A100 GPUs	Vicuna	7B	06/2024
ShareGPT4Video [197]	16	CLIP ViT-L/14	✗	✗	Linear Layer + LoRA for LLM	8 A100 GPUs	LLaMA-3	8B	06/2024
LVNet [171]	-	-	✗	✗	-	-	GPT-4	-	06/2024
VIM [198]	-	EVA-CLIP ViT-G/14	✗	✗	Q-former+LoRA in LLM	8 A100 GPUs	Vicuna-v1.1	13B	06/2024
Video-SALMONN [199]	2 FPS	BLIP-2 ViT-G/14	✓	✓	MRC Q-Former	-	GPT-4	-	06/2024
VideoGPT+ [200]	16/8	CLIP ViT-L/14	✗	✗	Linear Layer	8 A100 GPUs	Phi-3-Mini	3.8B	06/2024
Vriptor [268]	-	EVA-CLIP ViT-G/14	✗	✓	-	A100 GPUs	-	-	06/2024
MM-Screenplayer [157]	-	-	✗	✓	-	-	GPT-4-turbo	-	06/2024
VideoLLaMA 2 [201]	16	CLIP ViT-L/14	✓	✗	Linear Layer +STC connector	-	LLaVA-1.5	7B	06/2024
VideoLLM-online [249]	2 FPS	CLIP ViT-L/14	✗	✗	MLP projector + LoRA for LLM	-	Llama-2-Chat/Llama-3-Instruct	7B/8B	06/2024

scoring methods, including LLM-based evaluations. Other evaluation methods address specialized video understanding capabilities such as temporal/spatiotemporal reasoning.

A. Closed-ended Evaluation

Closed-ended evaluations use pre-defined answers or structured formats [284]. These include multiple-choice questions (TVQA [109], How2QA [285], STAR [286]) and questions with structured formats for direct comparison with ground truth (MSRVTT-QA [93], MSVD-QA [93], MVBench [203]). Multiple-choice performance is evaluated via accuracy percentages, while structured formats use metrics like CIDEr [287], METEOR [288], ROUGE [289], and SPICE [290] to compare predictions with ground truth. Notable benchmarks include MSRVTT-QA [93], TVQA [109], MVBench [203], EgoSchema [291], and Video-MME [292]. Each targets different video understanding aspects: TGIF-QA [107] focuses on action recognition, state transition, frame-level QA, and counting; ActivityNet-QA [110] covers motion, spatial, temporal, and free-form dimensions; VidComposition [293] emphasizes compositional reasoning; while NExT-QA [111] and MLVU [294] include causal and temporal action reasoning. These diverse question types test various reasoning abilities, though many benchmarks still exhibit domain biases toward common scenarios and lack diversity in rare events or unusual contexts.

B. Open-ended Evaluation

Open-ended evaluation involves questions without predefined options or structured formats. While ground-truth answers serve as references, scoring methods are more sophisticated than option selection or string matching. GPT-3.5/4 models often evaluate predictions by comparing them with reference answers. Notable open-ended benchmarks include MovieChat-1K [241], MLVU [294], NExT-QA [111], VELOCITI [305], and EAGLE [306]. These require more complex responses demonstrating deeper reasoning. CinePile [299] incorporates analytical tasks like character dynamics and narrative analysis. The most popular GPT-based evaluation methods, proposed in [99], are Open-end Zero-shot Video QA Evaluation and Video-based Generative Performance Benchmarking. Performance comparisons of Vid-LLMs on these metrics are shown in Table V. Originally closed-ended benchmarks like MSRVTT-QA [93], MSVD-QA [93], TGIF-QA [107], and ActivityNet-QA [110] can be repurposed as open-ended in GPT-based evaluations, as LLMs generate free-form responses that GPT models compare to references. These methods have limitations: evaluation scores change with GPT version updates, making cross-study comparisons difficult; results depend heavily on prompt engineering; and LLM evaluators may favor responses similar to their generation patterns rather than objectively assessing quality.

TABLE III
THE COMPARISON OF VARIOUS BENCHMARKS INCLUDES SEVERAL IMPORTANT ASPECTS: THE TOTAL NUMBER OF VIDEOS, THE NUMBER OF CLIPS, THE AVERAGE DURATION OF THE VIDEOS, THE NUMBER OF QA PAIRS, AND VIDEO CONTENT.

Benchmark	#Videos	#Clips	Len.(s)	Video Content	#QA Pairs	Question Type
MSRVTT-QA [93]	2,990	2,990	15.2	Open-domain	72,821	Closed-ended & open-ended (what/who/how/when/where)
MSVD-QA [93]	504	504	9.8	Open-domain	50,505	Closed-ended & open-ended (what/who/how/when/where)
TGIF-QA [107]	9,575	9,575	3.0	Open-domain	8,506	Closed-ended & open-ended, action, transition, counting
ActivityNet-QA [110]	800	800	111.4	Human activity	8,000	Closed-ended & open-ended (what/who/how/when/where/why)
TVQA [109]	2,179	15,253	11.2	TV show	15,253	Closed-ended, multiple-choice
How2QA [285]	1,166	2,852	15.3	TV episode	2,852	Multiple-choice
STAR [286]	914	7,098	11.9	Human action	7,098	Multiple-choice
NExT-QA [111]	1,000	1,000	39.5	Daily life	8,564	Multiple-choice and open-ended (causal, temporal, descriptive)
MVBench [203]	3,641	3,641	16.0	Open-domain	4,000	Closed-ended (various tasks)
Video-Bench [295]	5,917	5,917	56.0	Open-domain	17,036	Open-ended (various tasks)
EgoSchema [291]	5,063	5,063	180.0	Egocentric activity	5,063	Closed-ended, procedural understanding
AutoEval-Video [296]	327	327	14.6	Open-domain	327	Open-ended evaluation
TempCompass [297]	410	500	11.4	Open-domain	7,540	Both closed-ended & open-ended, temporal reasoning
Video-MME [292]	900	900	1,017.9	Open-domain	2,700	Closed-ended evaluation
VideoVista [298]	894	3,400	131.0	Open-domain	25,000	Open-ended (descriptive, causal, predictive)
CinePile [299]	9,396	9,396	160.0	Movie	303,828	Movie understanding, open-ended
SOK-Bench [300]	10,000	10,000	-	Open-domain	44,000	Open-ended, subject-oriented knowledge
SFD [301]	1,078	1,078	780.0	Movies	4,885	Multiple-choice, open-ended
EditVid-QA [302]	32,000	32,000	-	Entertainment	204,000	Open-ended, editing techniques
InfiniBench [303]	1,219	1,219	4,460.4	Movie/TV show	108,200	Closed-ended & open-ended, long-form understanding
MLVU [294]	1,334	1,334	180.0-7,200.0	Open-domain	2,593	Both closed-ended and open-ended, long-form understanding
MMWorld [304]	1,910	1,910	102.3	Open-domain	1,599	Open-ended, multimodal evaluation
VELOCITI [305]	900	900	10.0	Movie	-	Open-ended, visual understanding
VidComposition [293]	982	982	-	Movie/TV show	1,706	Multiple-choice, compositional reasoning
EAGLE [306]	36,000	36,000	16.0-76.0	Egocentric activity	400,000	Open-ended, procedural understanding

C. Other Evaluations

Other benchmarks evaluate fine-grained temporal and spatiotemporal understanding. Dense captioning generates [122]–[126] detailed descriptions for multiple video events/objects, using BLEU, METEOR, and CIDEr metrics that assess both temporal localization and descriptive accuracy. Vid-LLMs' performance of dense video captioning on ActivityNet Captions [127] is shown in Table IV. Several Vid-LLMs have already achieved performance comparable to traditional task-specific models in dense video captioning. Video temporal grounding localizes specific moments based on textual queries, evaluated using tIoU and Recall@K. Spatiotemporal grounding extends this to localize in both space and time, assessed via spatiotemporal IoU and mAP. Object tracking [251], [256] follows objects across frames, evaluated using precision, success rate, and tracking accuracy. Video saliency detection [307] identifies visually salient regions, evaluated with AUC-J, NSS, etc. These tasks rely on temporal or spatiotemporal annotations as ground-truth, with metrics like IoU, Recall@K, and mAP widely adopted. Human evaluation is also used for subjective aspects, though this is labor-intensive and time-consuming.

As for qualitative evaluation, several approaches can effectively assess Vid-LLMs' performance in addition to numerical metrics. Error analysis [203], [308] for open-ended QA and difference comparisons [120], [258] between model outputs and ground truth annotations (*e.g.*, intervals) for temporal/spatiotemporal understanding provide insights into model limitations. Attention visualization [309] reveals what visual elements the models prioritize when generating responses. Self-explanation [293], [308], where models justify their answers for closed-ended benchmarks, offers valuable insights into reasoning processes and potential misconceptions. Human studies, though resource-intensive, remain helpful in finding models that reflect human preferences.

TABLE IV
COMPARISON OF VID-LLMS AND CONVENTIONAL MODELS (NON-LLM-BASED) ON DENSE VIDEO CAPTIONING MODELS ON ACTIVITYNET CAPTIONS DATASET.

Model	CIDEr	SODA$_c$	METEOR	F1
Non-LLM-based				
MT [310]	9.3	–	5.0	–
PDVC [123]	29.3	6.0	7.6	–
CM^2 [311]	33.1	6.2	8.5	54.2
LLM-based				
Momentor [262]	14.9	2.3	4.7	–
TimeChat [257]	19.0	4.7	5.7	36.9
VTG-LLM [259]	20.7	5.1	5.9	34.8
AVicuna [120]	22.5	5.1	6.5	45.2
TRACE [312]	25.9	6.0	6.4	39.3
VTimeLLM [313]	27.6	5.8	6.8	–
TRACE-uni [312]	29.2	6.4	6.9	40.4
GIT [314]	29.8	5.7	7.8	50.6
Vid2Seq [271]	30.2	5.9	8.5	51.8
Streaming Vid2Seq [315]	37.8	6.2	10.0	52.9
Streaming GIT [315]	41.2	6.6	9.0	50.9

D. Analysis of Model Performance

Analyzing the correlation between model attributes and benchmark performance reveals several key factors driving recent improvements in Vid-LLMs. From Tables IV and V, we observe that models built on larger and more recent foundation LLMs (e.g., IG-VLM with 34B parameters) consistently outperform their smaller counterparts, particularly in zero-shot VideoQA tasks. Models employing more powerful visual embedders such as EVA-CLIP or ViT-G architectures (notably in PLLaVA, IG-VLM, and Video LLaMA 2) demonstrate superior performance across both dense captioning and QA benchmarks. The frame sampling strategy also significantly impacts results, with high performers on temporal tasks (like VTimeLLM, AVicuna, and ST-LLM) typically processing more frames (100+) than general understanding models, while sophisticated adaptation mechanisms beyond simple projection layers (such as Q-formers or cross-attention) contribute to

TABLE V

This table comprehensively compares various Vid-LLMs across multiple open-end zero-shot video question answering and video-based generative performance benchmarks. It includes GPT-based metrics for MSVD-QA, MSRVTT-QA, and ActivityNet-QA datasets, as well as scores for Correctness of Information, Detail Orientation, Contextual Understanding, Temporal Understanding, and Consistency aspects in video-based generative performance.

Model	**Open-end Zero-shot Video QA Evaluation**			**Video-based Generative Performance Benchmarking**					
	MSVD-QA	MSRVTT-QA	ActivityNet-QA	Correctness	Detail	Context	Temporal	Consistency	Average
GPT4-V [316]	-	-	59.5	4.09	3.88	4.37	3.94	4.02	4.06
Video-LLaMA [245]	51.6	29.6	12.4	1.96	2.18	2.16	1.82	1.79	1.98
LLaMA-Adapter [317]	54.9	43.8	34.2	2.03	2.32	2.30	1.98	2.15	2.16
VideoChat [270]	56.3	45.0	26.5	2.23	2.50	2.53	1.94	2.24	2.29
VALLY [244]	60.5	51.1	45.1	2.43	2.13	2.86	2.04	2.45	2.38
MovieLLM [223]	63.2	52.1	43.3	2.64	2.61	2.92	2.03	2.43	2.53
Video-ChatGPT [99]	64.9	49.3	35.2	2.50	2.57	2.69	2.16	2.20	2.42
Chat-UniVi [237]	65.0	54.6	46.1	2.89	2.91	3.46	2.89	2.81	2.99
Vista-LLaMA [236]	65.3	60.5	48.3	2.44	2.64	3.18	2.26	2.31	2.57
AV-LLM [231]	67.3	53.7	47.2	2.56	2.47	2.93	2.17	2.47	2.52
LLaVA-NeXT-Video [318]	67.8	-	53.5	3.39	3.29	3.92	2.60	3.12	3.26
LLaMA-VID [238]	69.7	57.7	47.4	2.96	3.00	3.53	2.46	2.51	2.89
VTimeLLM [258]	69.8	58.8	45.5	2.78	3.10	3.40	2.49	2.47	2.85
VideoChat2 [203]	70.0	54.1	49.1	3.02	2.88	3.51	2.66	2.81	2.98
LongVLM [213]	70.0	59.8	47.6	2.76	2.86	3.34	2.39	3.11	2.89
AVicuna [120]	70.2	59.7	53.0	2.81	2.62	3.25	2.53	2.59	2.76
Video-LLaVA [239]	70.7	59.2	45.3	2.87	2.94	3.44	2.45	2.51	2.84
RED-VILLM [211]	71.2	53.9	44.2	2.71	2.75	3.34	2.34	2.45	2.72
Video LLaMA 2 [201]	71.7	-	49.9	3.09	3.09	3.68	2.63	3.25	3.15
Artemis [191]	72.1	56.7	39.3	2.69	2.55	3.04	2.24	2.70	2.64
VideoGPT+ [200]	72.4	60.6	50.6	2.46	2.73	2.81	1.78	2.47	2.45
MiniGPT4-video [215]	73.9	58.8	44.4	3.09	3.02	3.57	2.56	2.67	2.98
ST-LLM [218]	74.6	63.2	50.9	3.23	3.05	3.74	2.93	2.81	3.15
MovieChat [241]	75.2	52.7	45.7	2.76	2.93	3.01	2.24	2.42	2.67
PLLaVA [217]	76.6	62.0	56.3	3.21	2.86	3.62	2.33	2.93	2.99
IG-VLM [159]	76.7	62.7	57.3	3.26	2.76	3.57	2.34	3.28	3.04

better contextual understanding. Performance gains stem from a combination of stronger foundation models, better visual encoders, appropriate temporal modeling, and more sophisticated bridging architectures rather than any single innovation.

V. Applications and Future Directions

A. *Application Scenarios*

Vid-LLMs have revolutionized various sectors by enabling advanced video and language processing capabilities. We outlines their diverse applications, demonstrating the extensive and transformative impact of Vid-LLMs across industries.

1) Media and Entertainment:

- *Online Video Platforms and Multimedia Information Retrieval:* Vid-LLMs significantly enhance search algorithms [319], generate context-aware video recommendations [320], and aid in natural language tasks such as subtitle generation and translation [271], contributing to online video platforms and multimedia retrieval systems. Their capabilities in analyzing videos for specific keyword retrieval [168], [321], [322] improve intelligent recommendation systems. In the multimedia fields, it combines videos in domains like music [323], avatar [324]–[327], and scene [328], to assist with content generation.
- *Video Summarization and Editing:* Vid-LLMs are integral in generating concise summaries of video content [329], which analyzes visual and auditory elements to extract key features for context-aware summaries. They also contribute to the field of video editing, as covered in existing literature [75] and advertisement editing [114].

2) Interactive and User-Centric Systems:

- *Virtual Education, Accessibility, and Sign Language:* Vid-LLMs serve as virtual tutors in education, analyzing instructional videos for interactive learning environments [330]. They also facilitate sign language translation into spoken language or text [331], [332], improving accessibility for the deaf and hard of hearing.
- *Interactive Gaming and Virtual Environments:* In the gaming industry, Vid-LLMs play a crucial role in creating dynamic dialogues and storylines, as well as aiding in generating procedural content, such as quests and in-game texts [333], [334]. They also power customer service chatbots [335], [336]. Additionally, in AR/VR/XR, Vid-LLMs contribute to the generation of dynamic narrative content, enhancing user immersion [337]–[340].
- *State-Aware Human-Computer Interaction and Robot Planning:* In the field of human-computer interaction, Vid-LLMs analyze user videos to discern context and provide customized assistance, as highlighted in Bi et al. [341]. Interaction forms also involve video content understanding like captioning videos [155], [342], [343]. Concurrently, in autonomous robot navigation, the SayPlan method [344] integrates LLMs with 3D scene graphs to enable robots to interpret and navigate complex spaces in large buildings.

3) Healthcare and Security Applications:

- *Healthcare Innovations:* In the healthcare sector, Vid-LLMs play a crucial role in processing and interpreting medical literature, assisting in diagnostic and educational processes [345]–[348], and providing decision support for healthcare professionals. They are used in patient interaction tools, such as chatbots for symptom assessment and addressing health-related queries, thus improving patient care and access to information [349].

- *Security, Surveillance, and Cybersecurity:* Vid-LLMs are crucial in security and protection, analyzing communications for potential threats [350], [351] and detecting anomalous patterns in data [352], [353]. In surveillance video analysis, they identify suspicious behaviors, helping law enforcement [354]. Their role in cybersecurity includes identifying phishing attempts and contributing to forensic analysis by summarizing case-related texts [355]. They may also improve video crowd counting [356] for security applications.
- *Autonomous Vehicles:* In autonomous vehicles, Vid-LLMs can process language inputs for interaction [357], assist in understanding road signs and instructions [247], [358], and improve user interfaces for vehicle control systems [357], enhancing safety and user experience.

4) Other Applications: Vid-LLMs offer valuable applications beyond those previously discussed. In video generation research [359]–[361], Vid-LLMs can evaluate model performance, refine text prompts, and provide reasoning capabilities that better reflect human intentions. Additionally, Vid-LLMs show promise in resource-constrained environments through edge computing applications [362]–[364] and can enhance privacy-preserving distributed systems through federated learning frameworks [365]–[367].

B. Future Directions

Despite enhancing multiple downstream tasks, Vid-LLMs face several challenges in real-world video understanding:

1) More Fine-grained Video Understanding: Fine-grained understanding remains challenging due to limited datasets, insufficient research, and high computational demands. The frame-by-frame analysis increases computational load while capturing spatiotemporal information. Understanding deeper semantics (emotions, scene dynamics) is harder, though text-video alignment through LLMs offers promise [293].

2) Long-form Video Understanding: Long videos' extended duration complicates the analysis, especially in understanding events over time. Thus, identifying key events and maintaining attention in long videos is difficult [186], [196], [213]. Effective mechanisms are needed to detect and highlight important parts, particularly in content-rich or complex plot videos.

3) Multimodal Video Understanding: Multimodal video understanding requires integrating different types of data, such as visual, audio, and text, for a better understanding of videos [120], [368], [369]. Aligning these data, especially in terms of spatial and temporal synchronization, is particularly critical. This area lacks relevant research and suffers from a scarcity of datasets. The field lacks research and datasets, with challenges in ensuring high-quality data annotation.

4) Hallucination in Video LLMs: Hallucination occurs when models generate responses disconnected from source material [370], caused by insufficient feature extraction, influence of video context, domain gap between vision and language, and inherent LLM hallucinations. Solutions include post-training strategies [210], enhanced spatiotemporal context understanding, and visual-linguistic latent collaboration.

5) Industrial Deployment and Scalability: Effective deployment strategies [212], [213], [363], [371]–[373] include model compression, token merging, domain-specific fine-tuning, modular architectures, efficient caching, and standardized integration frameworks, balancing efficiency with performance for industrial systems.

C. Ethical Implications

The ethical implications of Vid-LLMs center on privacy, data security, and potential misuse. These models perform tasks like video engagement analysis, transcription, summarization, and captioning, requiring access to sensitive content. This raises privacy risks, as video data may contain private or confidential information that could be exposed without proper consent. Also, Vid-LLMs can be misused for surveillance or generating misleading content. Bias is another concern, especially if training data lacks diversity. Addressing these issues requires robust data governance, consent mechanisms, and ethical deployment to prioritize privacy and fairness.

VI. Conclusion

This survey has examined the integration of LLMs in video understanding, which has enabled more sophisticated and versatile processing capabilities beyond traditional methods. We categorized current approaches into three main types: *Video Analyzer* × *LLM*, *Video Embedder* × *LLM*, and *(Analyzer + Embedder)* × *LLM*, with sub-classifications based on LLM functional roles: *Summarizer*, *Manager*, *Text Decoder*, *Regressor*, and *Hidden Layer*. Vid-LLMs demonstrate capabilities in multi-granularity reasoning from abstract to spatiotemporal analysis, showing potential across video summarization, captioning, question answering, and other applications. Despite progress, limitations remain in evaluation metrics, long-form video handling, and visual-textual modality alignment. Future research will address these challenges through more efficient training strategies, improved Vid-LLM scalability, innovative architectures for multimodal integration, enhanced long-form video understanding, and methods to mitigate hallucinations. Expanding datasets and benchmarks will be critical for advancing video understanding with LLMs.

References

[1] T. Lindeberg, "Scale invariant feature transform," 2012.

[2] H. Bay *et al.*, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[4] A. Sobral and A. Vacavant, "A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos," *Computer Vision and Image Understanding*, vol. 122, pp. 4–21, 2014.

[5] Z. Tu, H. Li *et al.*, "Optical flow for video super-resolution: a survey," *Artificial Intelligence Review*, vol. 55, no. 8, pp. 6505–6546, 2022.

[6] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3551–3558.

[7] Z. Shu, K. Yun, and D. Samaras, "Action detection with improved dense trajectories and sliding window," in *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part I 13*. Springer, 2015, pp. 541–551.

[8] X. Liu and T. Cheng, “Video-based face recognition using adaptive hidden markov models,” in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, vol. 1. IEEE, 2003, pp. I–I.

[9] H. Sidenbladh, “Detecting human motion with support vector machines,” in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 2. IEEE, 2004, pp. 188–191.

[10] Y. Yuan, Q.-B. Song, and J.-Y. Shen, “Automatic video classification using decision tree method,” in *Proceedings. International Conference on Machine Learning and Cybernetics*, vol. 3. IEEE, 2002, pp. 1153–1157.

[11] A. B. Chan and N. Vasconcelos, “Modeling, clustering, and segmenting video with mixtures of dynamic textures,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 5, pp. 909–926, 2008.

[12] T. Bouwmans and E. H. Zahzah, “Robust pca via principal component pursuit: A review for a comparative evaluation in video surveillance,” *Computer Vision and Image Understanding*, vol. 122, pp. 22–34, 2014.

[13] L. Hazelhoff *et al.*, “Video-based fall detection in the home using principal component analysis,” in *Advanced Concepts for Intelligent Vision Systems: 10th International Conference, ACIVS 2008, Juan-les-Pins, France, October 20-24, 2008. Proceedings 10*. Springer, 2008, pp. 298–309.

[14] A. Karpathy *et al.*, “Large-scale video classification with convolutional neural networks,” in *CVPR*, 2014.

[15] S. Ji *et al.*, “3d convolutional neural networks for human action recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[16] C. Feichtenhofer *et al.*, “Convolutional two-stream network fusion for video action recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1933–1941.

[17] J. Yue-Hei Ng *et al.*, “Beyond short snippets: Deep networks for video classification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4694–4702.

[18] L. Wang, Y. Xiong *et al.*, “Temporal segment networks: Towards good practices for deep action recognition,” in *European conference on computer vision*. Springer, 2016, pp. 20–36.

[19] M. Sekma *et al.*, “Human action recognition based on multi-layer fisher vector encoding method,” *Pattern Recognition Letters*, vol. 65, pp. 37–43, 2015.

[20] A. Diba, V. Sharma, and L. Van Gool, “Deep temporal linear encoding networks,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2329–2338.

[21] I. Mironică *et al.*, “A modified vector of locally aggregated descriptors approach for fast video classification,” *Multimedia Tools and Applications*, vol. 75, pp. 9045–9072, 2016.

[22] H. Li, L. Zhang *et al.*, “Transvlad: Focusing on locally aggregated descriptors for few-shot learning,” in *European Conference on Computer Vision*. Springer, 2022, pp. 524–540.

[23] K. Soomro, A. R. Zamir, and M. Shah, “Ucf101: A dataset of 101 human actions classes from videos in the wild,” in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[24] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, “Hmdb: A large video database for human motion recognition,” in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

[25] D. Tran, L. Bourdev *et al.*, “Learning spatiotemporal features with 3d convolutional networks,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[26] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[27] C. Szegedy, W. Liu *et al.*, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[28] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, “The kinetics human action video dataset,” *arXiv preprint arXiv:1705.06950*, 2017.

[29] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag *et al.*, “The” something something” video database for learning and evaluating visual common sense,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5842–5850.

[30] K. He *et al.*, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[31] S. Xie, R. Girshick *et al.*, “Aggregated residual transformations for deep neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[32] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[33] K. Hara, H. Kataoka, and Y. Satoh, “Learning spatio-temporal features with 3d residual networks for action recognition,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

[34] Y. Chen, Y. Kalantidis *et al.*, “Multi-fiber networks for video recognition,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[35] A. Diba, M. Fayyaz *et al.*, “Spatio-temporal channel correlation networks for action classification,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[36] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, “Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[37] M. Zolfaghari, K. Singh, and T. Brox, “Eco: Efficient convolutional network for online video understanding,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 695–712.

[38] Z. Qiu, T. Yao, and T. Mei, “Learning spatio-temporal representation with pseudo-3d residual networks,” in *proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5533–5541.

[39] G. Varol, I. Laptev, and C. Schmid, “Long-term temporal convolutions for action recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1510–1517, 2018.

[40] A. Diba, M. Fayyaz, V. Sharma, A. H. Karami, M. M. Arzani, R. Yousefzadeh, and L. V. Gool, “Temporal 3d convnets: New architecture and transfer learning for video classification,” *CoRR*, vol. abs/1711.08200, 2017.

[41] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[42] S. Zhang, S. Guo *et al.*, “V4d: 4d convolutional neural networks for video-level representation learning,” in *International Conference on Learning Representations*, 2019.

[43] D. Tran *et al.*, “Video classification with channel-separated convolutional networks,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5552–5561.

[44] C. Feichtenhofer *et al.*, “Slowfast networks for video recognition,” in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019.

[45] C. Feichtenhofer, “X3d: Expanding architectures for efficient video recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 203–213.

[46] A. Dosovitskiy *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, 2021.

[47] G. Bertasius, H. Wang, and L. Torresani, “Is space-time attention all you need for video understanding?” in *ICML*, vol. 2, no. 3, 2021, p. 4.

[48] X. Li, Y. Zhang *et al.*, “Vidtr: Video transformer without convolutions,” *arXiv e-prints*, pp. arXiv–2104, 2021.

[49] A. Arnab, M. Dehghani *et al.*, “Vivit: A video vision transformer,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6836–6846.

[50] H. Fan, B. Xiong *et al.*, “Multiscale vision transformers,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6824–6835.

[51] C. Li, D. Zhang, W. Huang, and J. Zhang, “Cross contrasting feature perturbation for domain generalization,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1327–1337.

[52] S. Wang *et al.*, “Feature alignment and uniformity for test time adaptation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 050–20 060.

[53] D. Zhang *et al.*, “Rethinking alignment and uniformity in unsupervised image semantic segmentation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 9, 2023, pp. 11 192–11 200.

[54] C. Sun, A. Myers *et al.*, “Videobert: A joint model for video and language representation learning,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7464–7473.

[55] J. D. M.-W. C. Kenton and L. K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[56] L. Zhu and Y. Yang, "Actbert: Learning global-local video-text representations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8746–8755.

[57] C. Feichtenhofer *et al.*, "Masked autoencoders as spatiotemporal learners," *Advances in neural information processing systems*, vol. 35, pp. 35 946–35 958, 2022.

[58] R. Girdhar *et al.*, "Omnimae: Single model masked pretraining on images and videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 406–10 417.

[59] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *Advances in neural information processing systems*, vol. 35, pp. 10 078–10 093, 2022.

[60] H. Yang, D. Huang *et al.*, "Self-supervised video representation learning with motion-aware masked autoencoders," *arXiv preprint arXiv:2210.04154*, 2022.

[61] C. Wei, H. Fan *et al.*, "Masked feature prediction for self-supervised visual pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 668–14 678.

[62] H. Xu, G. Ghosh *et al.*, "Vlm: Task-agnostic video-language model pre-training for video understanding," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 4227–4239.

[63] D. Li, J. Li *et al.*, "Align and prompt: Video-and-language pre-training with entity prompts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4953–4963.

[64] N. Moritz, G. Wichern, T. Hori, and J. Le Roux, "All-in-one transformer: Unifying speech recognition, audio tagging, and event detection." in *INTERSPEECH*, 2020, pp. 3112–3116.

[65] A. Gupta, S. Tian, Y. Zhang, J. Wu, R. Martín-Martín, and L. Fei-Fei, "Maskvit: Masked visual pre-training for video prediction," in *The Eleventh International Conference on Learning Representations*, 2022.

[66] H. Xue, Y. Sun *et al.*, "Clip-vip: Adapting pre-trained image-text model to video-language alignment," in *The Eleventh International Conference on Learning Representations*, 2022.

[67] J. Lei, T. L. Berg, and M. Bansal, "Revealing single frame bias for video-and-language learning," *arXiv preprint arXiv:2206.03428*, 2022.

[68] Y. Sun, H. Xue *et al.*, "Long-form video-language pre-training with multimodal temporal contrastive learning," *Advances in neural information processing systems*, vol. 35, pp. 38 032–38 045, 2022.

[69] P. Jin, J. Huang, F. Liu, X. Wu, S. Ge, G. Song, D. Clifton, and J. Chen, "Expectation-maximization contrastive learning for compact video-and-language representations," *Advances in Neural Information Processing Systems*, vol. 35, pp. 30 291–30 306, 2022.

[70] Q. Ye, G. Xu *et al.*, "Hitea: Hierarchical temporal-aware video-language pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 405–15 416.

[71] S. Han, J. Hessel *et al.*, "Champagne: Learning real-world conversation from large-scale web videos," *arXiv preprint arXiv:2303.09713*, 2023.

[72] H. Lyu *et al.*, "Gpt-4v (ision) as a social media analysis engine," *arXiv preprint arXiv:2311.07547*, 2023.

[73] D. Zhang, W. Zhang *et al.*, "Dnagpt: A generalized pretrained tool for multiple dna sequence analysis tasks," *bioRxiv*, pp. 2023–07, 2023.

[74] OpenAI, "Introducing chatgpt," https://openai.com/blog/chatgpt, 2022.

[75] C. Wu, S. Yin *et al.*, "Visual chatgpt: Talking, drawing and editing with visual foundation models," *arXiv preprint arXiv:2303.04671*, 2023.

[76] H. Liu *et al.*, "Visual instruction tuning," *arXiv preprint arXiv:2304.08485*, 2023.

[77] W. X. Zhao, K. Zhou *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[78] C. Li, Z. Gan, Z. Yang, J. Yang, L. Li, L. Wang, and J. Gao, "Multimodal foundation models: From specialists to general-purpose assistants," *arXiv preprint arXiv:2309.10020*, vol. 1, 2023.

[79] M. Abdar *et al.*, "A review of deep learning for video captioning," *arXiv preprint arXiv:2304.11431*, 2023.

[80] Y. Zhu *et al.*, "A comprehensive study of deep video action recognition," *arXiv preprint arXiv:2012.06567*, 2020.

[81] Z. Xing, Q. Feng *et al.*, "A survey on video diffusion models," *arXiv preprint arXiv:2310.10647*, 2023.

[82] Y. Annepaka and P. Pakray, "Large language models: A survey of their development, capabilities, and applications," *Knowledge and Information Systems*, pp. 1–56, 2024.

[83] N. Madan, A. Møgelmose, R. Modi, Y. S. Rawat, and T. B. Moeslund, "Foundation models for video understanding: A survey," *Authorea Preprints*, 2024.

[84] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2009, pp. 2929–2936.

[85] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[86] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[87] J. Carreira, E. Noland *et al.*, "A short note about kinetics-600," *arXiv preprint arXiv:1808.01340*, 2018.

[88] J. Carreira *et al.*, "A short note on the kinetics-700 human action dataset," *arXiv preprint arXiv:1907.06987*, 2019.

[89] H. Zhao *et al.*, "Hacs: Human action clips and segments dataset for recognition and temporal localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8668–8678.

[90] S. Abu-El-Haija *et al.*, "Youtube-8m: A large-scale video classification benchmark," 2016.

[91] M. Han *et al.*, "Video recognition in portrait mode," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 831–21 841.

[92] Y. Wang, D. Gao *et al.*, "Geb+: A benchmark for generic event boundary captioning, grounding and retrieval," in *European Conference on Computer Vision*. Springer, 2022, pp. 709–725.

[93] D. Xu *et al.*, "Video question answering via gradually refined attention over appearance and motion," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1645–1653.

[94] L. Anne Hendricks *et al.*, "Localizing moments in video with natural language," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5803–5812.

[95] L. Zhou, C. Xu, and J. Corso, "Towards automatic learning of procedures from web instructional videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[96] H. Xu, Q. Ye *et al.*, "Youku-mplug: A 10 million large-scale chinese video-language dataset for pre-training and benchmarks," *arXiv preprint arXiv:2306.04362*, 2023.

[97] G. Huang *et al.*, "Multimodal pretraining for dense video captioning," *arXiv preprint arXiv:2011.11760*, 2020.

[98] J. Lin, H. Hua *et al.*, "Videoxum: Cross-modal visual and textural summarization of videos," *arXiv preprint arXiv:2303.12060*, 2023.

[99] M. Maaz, H. Rasheed *et al.*, "Video-chatgpt: Towards detailed video understanding via large vision and language models," *arXiv preprint arXiv:2306.05424*, 2023.

[100] H. Hua, Y. Tang, C. Xu, and J. Luo, "V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning," *arXiv preprint arXiv:2404.12353*, 2024.

[101] D. Chen and W. 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*, 2011, pp. 190–200.

[102] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5288–5296.

[103] Y. Li, Y. Song, L. Cao, J. Tetreault, L. Goldberg, A. Jaimes, and J. Luo, "Tgif: A new dataset and benchmark on animated gif description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4641–4650.

[104] G. A. Sigurdsson *et al.*, "Charades-ego: A large-scale dataset of paired third and first person videos," *arXiv preprint arXiv:1804.09626*, 2018.

[105] S. Chen, H. Li *et al.*, "Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset," *arXiv preprint arXiv:2305.18500*, 2023.

[106] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6720–6731.

[107] Y. Jang *et al.*, "Tgif-qa: Toward spatio-temporal reasoning in visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2758–2766.

[108] K.-M. Kim *et al.*, "Deepstory: Video story qa by deep embedded memory networks," *arXiv preprint arXiv:1707.00836*, 2017.

[109] J. Lei, L. Yu *et al.*, "Tvqa: Localized, compositional video question answering," in *EMNLP*, 2018.

[110] Z. Yu, D. Xu *et al.*, "Activitynet-qa: A dataset for understanding complex web videos via question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 9127–9134.

[111] J. Xiao *et al.*, "Next-qa: Next phase of question-answering to explaining temporal actions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 9777–9786.

[112] M. Gygli *et al.*, "Creating summaries from user videos," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*. Springer, 2014, pp. 505–520.

[113] Y. Song *et al.*, "Tvsum: Summarizing web videos using titles," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5179–5187.

[114] Y. Tang, S. Xu, T. Wang, Q. Lin, Q. Lu, and F. Zheng, "Multi-modal segment assemblage network for ad video editing with importance-coherence reward," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 3519–3535.

[115] M. Sun *et al.*, "Ranking domain-specific highlights by analyzing edited videos," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer, 2014, pp. 787–802.

[116] K.-H. Zeng *et al.*, "Title generation for user generated videos," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 609–625.

[117] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar, "THUMOS challenge: Action recognition with a large number of classes," http://crcv.ucf.edu/THUMOS14/, 2014.

[118] T. Geng *et al.*, "Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 942–22 951.

[119] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "Tall: Temporal activity localization via language query," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[120] Y. Tang, D. Shimada, J. Bi, M. Feng, H. Hua, and C. Xu, "Empowering llms with pseudo-untrimmed videos for audio-visual temporal understanding," *arXiv preprint arXiv:2403.16276*, 2024.

[121] M. Z. Shou *et al.*, "Generic event boundary detection: A benchmark for event segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 8075–8084.

[122] Z. Shao *et al.*, "Region-object relation-aware dense captioning via transformer," *IEEE transactions on neural networks and learning systems*, 2022.

[123] T. Wang *et al.*, "End-to-end dense video captioning with parallel decoding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 6847–6857.

[124] Z. Shao *et al.*, "Textual context-aware dense captioning with diverse words," *IEEE Transactions on Multimedia*, vol. 25, pp. 8753–8766, 2023.

[125] Y. Long *et al.*, "Capdet: Unifying dense captioning and open-world detection pretraining," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 15 233–15 243.

[126] Z. Shao *et al.*, "Dcmstrd: end-to-end dense captioning via multi-scale transformer decoding," *IEEE Transactions on Multimedia*, 2024.

[127] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, "Dense-captioning events in videos," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 706–715.

[128] A. Yang *et al.*, "Vidchapters-7m: Video chapters at scale," *arXiv preprint arXiv:2309.13952*, 2023.

[129] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2411–2418.

[130] U. T. Benchmark, "A benchmark and simulator for uav tracking."

[131] M. Kristan *et al.*, "The tenth visual object tracking vot2022 challenge results," in *European Conference on Computer Vision*. Springer, 2022, pp. 431–460.

[132] L. Zheng, L. Shen *et al.*, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1116–1124.

[133] W. Li *et al.*, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 152–159.

[134] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 79–88.

[135] E. Ristani *et al.*, "Performance measures and a data set for multi-target, multi-camera tracking," in *European conference on computer vision*. Springer, 2016, pp. 17–35.

[136] A. Moskalenko, A. Bryncev, D. Vatolin, R. Timofte, G. Zhan, L. Yang, Y. Tang *et al.*, "Aim 2024 challenge on video saliency prediction: Methods and results," *arXiv preprint arXiv:2409.14827*, 2024.

[137] W. Wang *et al.*, "Revisiting video saliency: A large-scale benchmark and a new model," in *Proceedings of the IEEE Conference on computer vision and pattern recognition*, 2018, pp. 4894–4903.

[138] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[139] X. Min, G. Zhai, C. Hu, and K. Gu, "Fixation prediction through multimodal analysis," in *2015 Visual Communications and Image Processing (VCIP)*, 2015, pp. 1–4.

[140] A. Coutrot and N. Guyader, "How saliency, faces, and sound influence gaze in dynamic social scenes," *Journal of vision*, vol. 14, no. 8, pp. 5–5, 2014.

[141] A. Coutrot *et al.*, "Multimodal saliency models for videos," *From Human Attention to Computational Attention: A Multidisciplinary Approach*, pp. 291–304, 2016.

[142] P. Koutras and P. Maragos, "A perceptually based spatio-temporal computational framework for visual saliency estimation," *Signal Processing: Image Communication*, vol. 38, pp. 15–31, 2015.

[143] N. Xu, L. Yang, Y. Fan, D. Yue, Y. Liang, J. Yang, and T. Huang, "Youtube-vos: A large-scale video object segmentation benchmark," 2018.

[144] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 724–732.

[145] L. Yang, Y. Fan, and N. Xu, "The 4th large-scale video object segmentation challenge - video instance segmentation track," Jun. 2022.

[146] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," 2016.

[147] H. Ding, C. Liu, S. He, X. Jiang, and C. C. Loy, "Mevis: A large-scale benchmark for video segmentation with motion expressions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2694–2703.

[148] Z. Zhang, Z. Zhao, Y. Zhao, Q. Wang, H. Liu, and L. Gao, "Where does it exist: Spatio-temporal video grounding for multi-form sentences," in *CVPR*, 2020.

[149] Z. Tang *et al.*, "Human-centric spatio-temporal video grounding with visual transformers," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8238–8249, 2021.

[150] K. Grauman *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 995–19 012.

[151] J. Kaplan *et al.*, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[152] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[153] K. Chen *et al.*, "Shikra: Unleashing multimodal llm's referential dialogue magic," *arXiv preprint arXiv:2306.15195*, 2023.

[154] S. Xuan, Q. Guo, M. Yang, and S. Zhang, "Pink: Unveiling the power of referential comprehension for multi-modal llms," *arXiv preprint arXiv:2310.00582*, 2023.

[155] H. Hua, J. Shi, K. Kafle, S. Jenni, D. Zhang, J. Collomosse, S. Cohen, and J. Luo, "Finematch: Aspect-based fine-grained image and text mismatch detection and correction," *arXiv preprint arXiv:2404.14715*, 2024.

[156] A. R. Kalarani, P. Bhattacharyya, and S. Shekhar, "Seeing the unseen: Visual metaphor captioning for videos," *arXiv preprint arXiv:2406.04886*, 2024.

[157] Y. Wu, B. Li *et al.*, "Zero-shot long-form video understanding through screenplay," *arXiv preprint arXiv:2406.17309*, 2024.

[158] J. Min *et al.*, "Morevqa: Exploring modular reasoning models for video question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 235–13 245.

[159] W. Kim, C. Choi, W. Lee, and W. Rhee, "An image grid can be worth a video: Zero-shot video question answering using a vlm," 2024.

[160] K. Kahatapitiya, K. Ranasinghe, J. Park, and M. S. Ryoo, "Language repository for long video understanding," 2024.

[161] K. Ranasinghe, X. Li, K. Kahatapitiya, and M. S. Ryoo, "Understanding long videos in one multimodal language model pass," 2024.

[162] M. M. Islam, N. Ho, X. Yang, T. Nagarajan, L. Torresani, and G. Bertasius, "Video recap: Recursive captioning of hour-long videos," 2024.

[163] C. Zhang, T. Lu, M. M. Islam, Z. Wang, S. Yu, M. Bansal, and G. Bertasius, "A simple llm framework for long-range video question-answering," 2024.

[164] H. Chen, X. Wang, H. Chen, Z. Song, J. Jia, and W. Zhu, "Grounding-prompter: Prompting llm with multimodal information for temporal sentence grounding in long videos," 2023.

[165] Z. Xue, K. Ashutosh, and K. Grauman, "Learning object state changes in videos: An open-world perspective," 2024.

[166] Q. Zhao, C. Zhang *et al.*, "Antgpt: Can large language models help long-term action anticipation from videos?" *arXiv preprint arXiv:2307.16368*, 2023.

[167] showlab, "Vlog: Transform video as a document with chatgpt, clip, blip2, grit, whisper, langchain," https://github.com/showlab/VLog, accessed: 2023-12-23.

[168] Y. Zhao, I. Misra, P. Krähenbühl, and R. Girdhar, "Learning video representations from large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6586–6597.

[169] Z. Ma, C. Gou, H. Shi, B. Sun, S. Li, H. Rezatofighi, and J. Cai, "Drvideo: Document retrieval based long video understanding," *arXiv preprint arXiv:2406.12846*, 2024.

[170] L. Zhang, T. Zhao *et al.*, "Omagent: A multi-modal agent framework for complex video understanding with task divide-and-conquer," *arXiv preprint arXiv:2406.16620*, 2024.

[171] J. Park, K. Ranasinghe *et al.*, "Too many frames, not all useful: Efficient strategies for long-form video qa," *arXiv preprint arXiv:2406.09396*, 2024.

[172] Y. Sun, Y. Xu, Z. Xie, Y. Shu, and S. Du, "Gptsee: Enhancing moment retrieval and highlight detection via description-based similarity features," 2024.

[173] Z. Wang, S. Yu *et al.*, "Videotree: Adaptive tree-based video representation for llm reasoning on long videos," *arXiv preprint arXiv:2405.19209*, 2024.

[174] L. Zanella, W. Menapace *et al.*, "Harnessing large language models for training-free video anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 527–18 536.

[175] C. Shang, A. You *et al.*, "Traveler: A multi-lmm agent framework for video question-answering," *arXiv preprint arXiv:2404.01476*, 2024.

[176] J. Cao, Y. Wu, W. Chi, W. Zhu, Z. Su, and J. Wu, "Reframe anything: Llm agent for open world video reframing," 2024.

[177] Y. Niu, W. Guo, L. Chen, X. Lin, and S.-F. Chang, "Schema: State changes matter for procedure planning in instructional videos," 2024.

[178] Y. Shen *et al.*, "Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[179] K. Sanders, N. Weir, and B. V. Durme, "Tv-trees: Multimodal entailment trees for neuro-symbolic video reasoning," 2024.

[180] Y. Fan, X. Ma, R. Wu, Y. Du, J. Li, Z. Gao, and Q. Li, "Videoagent: A memory-augmented multimodal agent for video understanding," 2024.

[181] X. Wang, Y. Zhang, O. Zohar, and S. Yeung-Levy, "Videoagent: Long-form video understanding with large language model as agent," 2024.

[182] A. Mahmood, A. Vayani, M. Naseer, S. Khan, and F. S. Khan, "Vurf: A general-purpose reasoning and self-refinement framework for video understanding," 2024.

[183] K. R. Y. Nagasinghe, H. Zhou, M. Gunawardhana, M. R. Min, D. Harari, and M. H. Khan, "Why not use your textbook? knowledge-enhanced procedure planning of instructional videos," 2024.

[184] Z. Yang, G. Chen, X. Li, W. Wang, and Y. Yang, "Doraemongpt: Toward understanding dynamic scenes with large language models (exemplified as a video agent)," 2024.

[185] J. Tang *et al.*, "Hawk: Learning to understand open-world video anomalies," *arXiv preprint arXiv:2405.16886*, 2024.

[186] Y. Wang, Y. Yang, and M. Ren, "Lifelongmemory: Leveraging llms for answering queries in long-form egocentric videos," 2024.

[187] R. Choudhury, K. Niinuma, K. M. Kitani, and L. A. Jeni, "Zero-shot video question answering with procedural programs," 2023.

[188] D. Gao, L. Ji *et al.*, "Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn," *arXiv preprint arXiv:2306.08640*, 2023.

[189] J. Wang, D. Chen *et al.*, "Chatvideo: A tracklet-centric multimodal and versatile video understanding system," *arXiv preprint arXiv:2304.14407*, 2023.

[190] D. Surís, S. Menon, and C. Vondrick, "Vipergpt: Visual inference via python execution for reasoning," *arXiv preprint arXiv:2303.08128*, 2023.

[191] J. Qiu, Y. Zhang, X. Tang, L. Xie, T. Ma, P. Yan, D. Doermann, Q. Ye, and Y. Tian, "Artemis: Towards referential understanding in complex videos," *arXiv preprint arXiv:2406.00258*, 2024.

[192] Q. Yang, M. Ye, and B. Du, "Emollm: Multimodal emotional understanding meets large language models," *arXiv preprint arXiv:2406.16442*, 2024.

[193] S. Chen, Y. Yuan, S. Chen, Z. Jie, and L. Ma, "Fewer tokens and fewer videos: Extending video understanding abilities in large vision-language models," *arXiv preprint arXiv:2406.08024*, 2024.

[194] H. Zhang, Y. Wang, Y. Tang, Y. Liu, J. Feng, J. Dai, and X. Jin, "Flash-vstream: Memory-based real-time understanding for long video streams," 2024.

[195] R. Chakraborty, A. Sinha, D. Reilly, M. K. Govind, P. Wang, F. Bremond, and S. Das, "Llavidal: Benchmarking large language vision models for daily activities of living," *arXiv preprint arXiv:2406.09390*, 2024.

[196] P. Zhang, K. Zhang *et al.*, "Long context transfer from language to vision," *arXiv preprint arXiv:2406.16852*, 2024.

[197] L. Chen, X. Wei, J. Li *et al.*, "Sharegpt4video: Improving video understanding and generation with better captions," *arXiv preprint arXiv:2406.04325*, 2024.

[198] Y. Du, K. Zhou *et al.*, "Towards event-oriented long video understanding," *arXiv preprint arXiv:2406.14129*, 2024.

[199] G. Sun, W. Yu *et al.*, "video-salmonn: Speech-enhanced audio-visual large language models," *arXiv preprint arXiv:2406.15704*, 2024.

[200] M. Maaz, H. Rasheed, S. Khan, and F. Khan, "Videogpt+: Integrating image and video encoders for enhanced video understanding," *arXiv preprint arXiv:2406.09418*, 2024.

[201] Z. Cheng, S. Leng *et al.*, "Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms," *arXiv preprint arXiv:2406.07476*, 2024.

[202] L.-H. Chen, S. Lu, A. Zeng, H. Zhang, B. Wang, R. Zhang, and L. Zhang, "Motionllm: Understanding human behaviors from human motions and videos," *arXiv preprint arXiv:2405.20340*, 2024.

[203] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, Y. Liu, Z. Wang, J. Xu, G. Chen, P. Luo *et al.*, "Mvbench: A comprehensive multi-modal video understanding benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 195–22 206.

[204] R. Luo, A. Peng, A. Vasudev, and R. Jain, "Shotluck holmes: A family of efficient small-scale large language vision models for video captioning and summarization," *arXiv preprint arXiv:2405.20648*, 2024.

[205] R. Qian, X. Dong *et al.*, "Streaming long video understanding with large language models," *arXiv preprint arXiv:2405.16009*, 2024.

[206] D. Yang, C. Zhan *et al.*, "Synchronized video storytelling: Generating video narrations with structured storyline," *arXiv preprint arXiv:2405.14040*, 2024.

[207] W. Li, H. Fan, Y. Wong, M. Kankanhalli, and Y. Yang, "Topa: Extend large language models for video understanding via text-only pre-alignment," *arXiv preprint arXiv:2405.13911*, 2024.

[208] T. Han *et al.*, "Autoad iii: The prequel-back to the pixels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 164–18 174.

[209] H. Wang, C. Lai, Y. Sun, and W. Ge, "Weakly supervised gaussian contrastive grounding with large multimodal models for video question answering," 2024.

[210] R. Zhang, L. Gui *et al.*, "Direct preference optimization of video large multimodal models from language model reward," *arXiv preprint arXiv:2404.01258*, 2024.

[211] S. Huang, H. Zhang *et al.*, "From image to video, what do we need in multimodal llms?" *arXiv preprint arXiv:2404.11865*, 2024.

[212] R. Tan, X. Sun *et al.*, "Koala: Key frame-conditioned long video-llm," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 581–13 591.

[213] Y. Weng, M. Han *et al.*, "Longvlm: Efficient long video understanding via large language models," *arXiv preprint arXiv:2404.03384*, 2024.

[214] B. He, H. Li *et al.*, "Ma-lmm: Memory-augmented large multimodal model for long-term video understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 504–13 514.

[215] K. Ataallah, X. Shen *et al.*, "Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens," *arXiv preprint arXiv:2404.03413*, 2024.

[216] R. Jung *et al.*, "Pegasus-v1 technical report," *arXiv preprint arXiv:2404.14687*, 2024.

[217] L. Xu, Y. Zhao *et al.*, "Pllava: Parameter-free llava extension from images to videos for video dense captioning," *arXiv preprint arXiv:2404.16994*, 2024.

[218] R. Liu, C. Li *et al.*, "St-llm: Large language models are effective temporal learners," *arXiv preprint arXiv:2404.00308*, 2024.

[219] A. J. Wang *et al.*, "Cosmo: Contrastive streamlined multimodal model with interleaved pre-training," *arXiv preprint arXiv:2401.00849*, 2024.

[220] J. Wang, L. Yuan, and Y. Zhang, "Tarsier: Recipes for training and evaluating large video description models," *arXiv preprint arXiv:2407.00634*, 2024.

[221] J. Held, H. Itani *et al.*, "X-vars: Introducing explainability in football refereeing with multi-modal large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3267–3279.

[222] Q. Ye, Z. Yu, R. Shao, X. Xie, P. Torr, and X. Cao, "Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios," 2024.

[223] G. Chen, Y.-D. Zheng *et al.*, "Videollm: Modeling video sequence with large language models," *arXiv preprint arXiv:2305.13292*, 2023.

[224] Y. Wang, K. Li, X. Li, J. Yu, Y. He, G. Chen, B. Pei, R. Zheng, J. Xu, Z. Wang, Y. Shi, T. Jiang, S. Li, H. Zhang, Y. Huang, Y. Qiao, Y. Wang, and L. Wang, "Internvideo2: Scaling video foundation models for multimodal video understanding," 2024.

[225] Z. Song, C. Wang, J. Sheng, C. Zhang, G. Yu, J. Fan, and T. Chen, "Moviellm: Enhancing long video understanding with ai-generated movies," 2024.

[226] Y. Li, X. Chen, B. Hu, and M. Zhang, "Llms meet long video: Advancing long video comprehension with an interactive visual adapter in llms," 2024.

[227] Y. Wang, Y. Wang, P. Wu, J. Liang, D. Zhao, and Z. Zheng, "Lstp: Language-guided spatial-temporal prompt learning for long-form video-text understanding," 2024.

[228] Y. Wang, Z. Zhang, J. McAuley, and Z. He, "Lvchat: Facilitating long video comprehension," 2024.

[229] N. Nguyen, J. Bi, A. Vosoughi, Y. Tian, P. Fazli, and C. Xu, "Oscar: Object state captioning and state change representation," 2024.

[230] J. Xu, C. Lan, W. Xie, X. Chen, and Y. Lu, "Slot-vlm: Slowfast slots for video-language modeling," 2024.

[231] F. Shu, L. Zhang, H. Jiang, and C. Xie, "Audio-visual llm for video understanding," 2023.

[232] Q. Sun, Y. Cui, X. Zhang, F. Zhang, Q. Yu, Z. Luo, Y. Wang, Y. Rao, J. Liu, T. Huang, and X. Wang, "Generative multimodal models are in-context learners," 2024.

[233] T. Chen, E. Zhang, Y. Gao, K. Li, X. Sun, Y. Zhang, and H. Li, "Mmict: Boosting multi-modal fine-tuning with in-context examples," 2023.

[234] Y. Wang, R. Zhang, H. Wang, U. Bhattacharya, Y. Fu, and G. Wu, "Vaquita: Enhancing alignment in llm-assisted video understanding," 2023.

[235] J. Lin, H. Yin, W. Ping, Y. Lu, P. Molchanov, A. Tao, H. Mao, J. Kautz, M. Shoeybi, and S. Han, "Vila: On pre-training for visual language models," 2024.

[236] F. Ma, X. Jin, H. Wang, Y. Xian, J. Feng, and Y. Yang, "Vista-llama: Reliable video narrator via equal distance to visual tokens," 2023.

[237] P. Jin, R. Takanobu, W. Zhang, X. Cao, and L. Yuan, "Chat-univi: Unified visual representation empowers large language models with image and video understanding," 2024.

[238] Y. Li, C. Wang, and J. Jia, "Llama-vid: An image is worth 2 tokens in large language models," *arXiv preprint arXiv:2311.17043*, 2023.

[239] B. Lin, B. Zhu, Y. Ye, M. Ning, P. Jin, and L. Yuan, "Video-llava: Learning united visual representation by alignment before projection," *arXiv preprint arXiv:2311.10122*, 2023.

[240] D. Ko *et al.*, "Large language models are temporal and causal reasoners for video question answering," *arXiv preprint arXiv:2310.15747*, 2023.

[241] E. Song, W. Chai, G. Wang, Y. Zhang, H. Zhou, F. Wu, X. Guo, T. Ye, Y. Lu, J.-N. Hwang *et al.*, "Moviechat: From dense token to sparse memory for long video understanding," *arXiv preprint arXiv:2307.16449*, 2023.

[242] Y. Tang, J. Zhang, X. Wang, T. Wang, and F. Zheng, "Llmva-gebc: Large language model with video adapter for generic event boundary captioning," *arXiv preprint arXiv:2306.10354*, 2023.

[243] C. Lyu, M. Wu *et al.*, "Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration," *arXiv preprint arXiv:2306.09093*, 2023.

[244] R. Luo, Z. Zhao *et al.*, "Valley: Video assistant with large language model enhanced ability," *arXiv preprint arXiv:2306.07207*, 2023.

[245] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," *arXiv preprint arXiv:2306.02858*, 2023.

[246] Z. Zhao, L. Guo, T. Yue, S. Chen, S. Shao, X. Zhu, Z. Yuan, and J. Liu, "Chatbridge: Bridging modalities with large language model as a language catalyst," *arXiv preprint arXiv:2305.16103*, 2023.

[247] B. Li, Y. Zhang *et al.*, "Otter: A multi-modal model with in-context instruction tuning," *arXiv preprint arXiv:2305.03726*, 2023.

[248] H. Zhang, X. Xu, X. Wang, J. Zuo, C. Han, X. Huang, C. Gao, Y. Wang, and N. Sang, "Holmes-vad: Towards unbiased and explainable video anomaly detection via multi-modal llm," *arXiv preprint arXiv:2406.12235*, 2024.

[249] J. Chen, Z. Lv *et al.*, "Videollm-online: Online video large language model for streaming video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 407–18 418.

[250] S. Bansal *et al.*, "Hoi-ref: Hand-object interaction referral in egocentric vision," *arXiv preprint arXiv:2404.09933*, 2024.

[251] H. Wang *et al.*, "Elysium: Exploring object-level perception in videos via mllm," *arXiv preprint arXiv:2403.16558*, 2024.

[252] Y. Wang, X. Meng, J. Liang, Y. Wang, Q. Liu, and D. Zhao, "Hawkeye: Training video-text llms for grounding text in videos," 2024.

[253] D.-A. Huang, S. Liao, S. Radhakrishnan, H. Yin, P. Molchanov, Z. Yu, and J. Kautz, "Lita: Language instructed temporal-localization assistant," 2024.

[254] J. Wang, D. Chen, C. Luo, B. He, L. Yuan, Z. Wu, and Y.-G. Jiang, "Omnivid: A generative framework for universal video understanding," 2024.

[255] S. Yu *et al.*, "Self-chained image-language model for video localization and question answering," 2023.

[256] Z. Li, Q. Xu, D. Zhang, H. Song, Y. Cai, Q. Qi, R. Zhou, J. Pan, Z. Li, V. T. Vu, Z. Huang, and T. Wang, "Groundinggpt:language enhanced multi-modal grounding model," 2024.

[257] S. Ren, L. Yao, S. Li, X. Sun, and L. Hou, "Timechat: A time-sensitive multimodal large language model for long video understanding," 2024.

[258] B. Huang, X. Wang *et al.*, "Vtimellm: Empower llm to grasp video moments," *arXiv preprint arXiv:2311.18445*, 2023.

[259] Y. Guo, J. Liu *et al.*, "Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding," *arXiv preprint arXiv:2405.13382*, 2024.

[260] H. Fei, S. Wu *et al.*, "Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing," 2024.

[261] Y. Xu, Y. Sun, Z. Xie, B. Zhai, and S. Du, "Vtg-gpt: Tuning-free zero-shot video temporal grounding with gpt," 2024.

[262] L. Qian, J. Li, Y. Wu, Y. Ye, H. Fei, T.-S. Chua, Y. Zhuang, and S. Tang, "Momentor: Advancing video large language model with fine-grained temporal reasoning," 2024.

[263] K. Ashutosh, Z. Xue, T. Nagarajan, and K. Grauman, "Detours for navigating instructional videos," 2024.

[264] J. Han, K. Gong, Y. Zhang, J. Wang, K. Zhang, D. Lin, Y. Qiao, P. Gao, and X. Yue, "Onellm: One framework to align all modalities with language," 2023.

[265] Z. Wang *et al.*, "Gpt4video: A unified multimodal large language model for lnstruction-followed understanding and safety-aware generation," *arXiv preprint arXiv:2311.16511*, 2023.

[266] K. Lin *et al.*, "Mm-vid: Advancing video understanding with gpt-4v (ision)," *arXiv preprint arXiv:2310.19773*, 2023.

[267] M. Han, L. Yang, X. Chang, and H. Wang, "Shot2story20k: A new benchmark for comprehensive understanding of multi-shot videos," 2023.

[268] D. Yang, S. Huang *et al.*, "Vript: A video is worth thousands of words," *arXiv preprint arXiv:2406.06040*, 2024.

[269] E. Yu, L. Zhao *et al.*, "Merlin:empowering multimodal llms with foresight minds," 2023.

[270] K. Li, Y. He *et al.*, "Videochat: Chat-centric video understanding," *arXiv preprint arXiv:2305.06355*, 2023.

[271] A. Yang *et al.*, "Vid2seq: Large-scale pretraining of a visual language model for dense video captioning," in *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, 2023, pp. 10 714–10 726.

[272] H. Wang, Z. Tong, K. Zheng, Y. Shen, and L. Wang, "Contextual ad narration with interleaved multimodal sequence," 2024.

[273] C. Zhang *et al.*, "Mm-narrator: Narrating long-form videos with multimodal in-context learning," *arXiv preprint arXiv:2311.17435*, 2023.

[274] S. Wang *et al.*, "Vamos: Versatile action models for video understanding," 2024.

[275] T. Han, M. Bain *et al.*, "Autoad ii: The sequel-who, when, and what in movie audio description," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 645–13 655.

[276] Y. Tang, J. Bi, C. Huang, S. Liang, D. Shimada, H. Hua, Y. Xiao, Y. Song, P. Liu, M. Feng *et al.*, "Caption anything in video: Fine-grained object-centric captioning via spatiotemporal multimodal prompting," *arXiv preprint arXiv:2504.05541*, 2025.

[277] S. Munasinghe *et al.*, "Pg-video-llava: Pixel grounding large video-language models," *arXiv preprint arXiv:2311.13435*, 2023.

[278] J. Chen, D. Zhu *et al.*, "Video chatcaptioner: Towards the enriched spatiotemporal descriptions," *arXiv preprint arXiv:2304.04227*, 2023.

[279] B. Elizalde *et al.*, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[280] M. Xu, M. Gao *et al.*, "Slowfast-llava: A strong training-free baseline for video large language models," *arXiv preprint arXiv:2407.15841*, 2024.

[281] A. Zeng *et al.*, "Socratic models: Composing zero-shot multimodal reasoning with language," *arXiv preprint arXiv:2204.00598*, 2022.

[282] G. Sun, W. Yu *et al.*, "Fine-grained audio-visual joint representations for multimodal large language models," 2023.

[283] J. Li, D. Li *et al.*, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International Conference on Machine Learning*, 2023.

[284] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *arXiv preprint arXiv:2306.13549*, 2023.

[285] L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu, "Hero: Hierarchical encoder for video+ language omni-representation pre-training," in *EMNLP*, 2020.

[286] B. Wu and S. Yu, "Star: A benchmark for situated reasoning in real-world videos," in *NeurIPS*, 2024.

[287] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[288] S. Banerjee and A. 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*, 2005, pp. 65–72.

[289] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[290] P. Anderson *et al.*, "Spice: Semantic propositional image caption evaluation," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Springer, 2016, pp. 382–398.

[291] K. Mangalam *et al.*, "Egoschema: A diagnostic benchmark for very long-form video language understanding," *arXiv preprint arXiv:2308.09126*, 2023.

[292] C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang *et al.*, "Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis," *arXiv preprint arXiv:2405.21075*, 2024.

[293] Y. Tang, J. Guo *et al.*, "Vidcomposition: Can mllms analyze compositions in compiled videos?" *arXiv preprint arXiv:2411.10979*, 2024.

[294] J. Zhou *et al.*, "Mlvu: A comprehensive benchmark for multi-task long video understanding," *arXiv preprint arXiv:2406.04264*, 2024.

[295] M. Ning, B. Zhu, Y. Xie, B. Lin, J. Cui, L. Yuan, D. Chen, and L. Yuan, "Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models," *arXiv preprint arXiv:2311.16103*, 2023.

[296] X. Chen, Y. Lin, Y. Zhang, and W. Huang, "Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering," *ArXiv preprint*, 2023.

[297] Y. Liu, S. Li, Y. Liu, Y. Wang, S. Ren, L. Li, S. Chen, X. Sun, and L. Hou, "Tempcompass: Do video llms really understand videos?" *arXiv preprint arXiv:2403.00476*, 2024.

[298] Y. Li, X. Chen, B. Hu, L. Wang, H. Shi, and M. Zhang, "Videovista: A versatile benchmark for video understanding and reasoning," *arXiv preprint arXiv:2406.11303*, 2024.

[299] R. Rawal, K. Saifullah, R. Basri, D. Jacobs, G. Somepalli, and T. Goldstein, "Cinepile: A long video question answering dataset and benchmark," *arXiv preprint arXiv:2405.08813*, 2024.

[300] A. Wang *et al.*, "Sok-bench: A situated video reasoning benchmark with aligned open-world knowledge," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 384–13 394.

[301] R. Ghermi, X. Wang, V. Kalogeiton, and I. Laptev, "Short film dataset (sfd): A benchmark for story-level video understanding," *arXiv preprint arXiv:2406.10221*, 2024.

[302] L. Xu *et al.*, "Beyond raw videos: Understanding edited videos with large multimodal model," *arXiv preprint arXiv:2406.10484*, 2024.

[303] K. Ataallah *et al.*, "Infinibench: A comprehensive benchmark for large multimodal models in very long video understanding," *arXiv preprint arXiv:2406.19875*, 2024.

[304] X. He *et al.*, "Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos," *arXiv preprint arXiv:2406.08407*, 2024.

[305] D. Saravanan *et al.*, "Velociti: Can video-language models bind semantic concepts through time?" *arXiv preprint arXiv:2406.10889*, 2024.

[306] J. Bi, Y. Tang *et al.*, "Eagle: Egocentric aggregated language-video engine," in *Proceedings of the 32nd ACM International Conference on Multimedia*, ser. MM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 1682–1691. [Online]. Available: https://doi.org/10.1145/3664647.3681618

[307] Y. Tang, G. Zhan, L. Yang, Y. Liao, and C. Xu, "Cardiff: Video salient object ranking chain of thought reasoning for saliency prediction with diffusion," *arXiv preprint arXiv:2408.12009*, 2024.

[308] H. Hua, Y. Tang *et al.*, "Mmcomposition: Revisiting the compositionality of pre-trained vision-language models," *arXiv preprint arXiv:2410.09733*, 2024.

[309] J. Bi, J. Guo, Y. Tang *et al.*, "Unveiling visual perception in language models: An attention head analysis approach," *arXiv preprint arXiv:2412.18108*, 2024.

[310] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8739–8748.

[311] M. Kim *et al.*, "Do you remember? dense video captioning with cross-modal memory retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 894–13 904.

[312] Y. Guo, J. Liu, M. Li, Q. Liu, X. Chen, and X. Tang, "Trace: Temporal grounding video llm via causal event modeling," *arXiv preprint arXiv:2410.05643*, 2024.

[313] B. Huang *et al.*, "Vtimellm: Empower llm to grasp video moments," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 271–14 280.

[314] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, and L. Wang, "Git: A generative image-to-text transformer for vision and language," *arXiv preprint arXiv:2205.14100*, 2022.

[315] X. Zhou, A. Arnab, S. Buch, S. Yan, A. Myers, X. Xiong, A. Nagrani, and C. Schmid, "Streaming dense video captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 243–18 252.

[316] O. or Author Name, "Gpt-4v: An overview," https://website.com/path-to-gpt-4v, 2023, accessed: 2023-xx-xx.

[317] P. Gao, J. Han, R. Zhang, Z. Lin, S. Geng, A. Zhou, W. Zhang, P. Lu, C. He, X. Yue *et al.*, "Llama-adapter v2: Parameter-efficient visual instruction model," *arXiv preprint arXiv:2304.15010*, 2023.

[318] Y. Zhang, B. Li, h. Liu, Y. j. Lee, L. Gui, D. Fu, J. Feng, Z. Liu, and C. Li, "Llava-next: A strong zero-shot video understanding model," April 2024. [Online]. Available: https://llava-vl.github.io/blog/2024-04-30-llava-next-video/

[319] K. Mao *et al.*, "Large language models know your contextual search intent: A prompting framework for conversational search," *arXiv preprint arXiv:2303.06573*, 2023.

[320] C. Ju *et al.*, "Prompting visual-language models for efficient video understanding," in *European Conference on Computer Vision*. Springer, 2022, pp. 105–124.

[321] P. Jin, H. Li, Z. Cheng, J. Huang, Z. Wang, L. Yuan, C. Liu, and J. Chen, "Text-video retrieval with disentangled conceptualization and set-to-set alignment," *arXiv preprint arXiv:2305.12218*, 2023.

[322] P. Jin, H. Li, Z. Cheng, K. Li, X. Ji, C. Liu, L. Yuan, and J. Chen, "Diffusionret: Generative text-video retrieval with diffusion model," *arXiv preprint arXiv:2303.09867*, 2023.

[323] S. Xu, Y. Tang, and F. Zheng, "Launchpadgpt: Language model as music visualization designer on launchpad," *arXiv preprint arXiv:2307.04827*, 2023.

[324] L. Song *et al.*, "Tacr-net: editing on deep video and voice portraits," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 478–486.

[325] L. Song, G. Yin, Z. Jin, X. Dong, and C. Xu, "Emotional listener portrait: Neural listener head generation with emotion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 839–20 849.

[326] L. Song, P. Liu, L. Chen, C. Liu, and C. Xu, "Tri 2-plane: Volumetric avatar reconstruction with feature pyramid," *arXiv preprint arXiv:2401.09386*, 2024.

[327] L. Song *et al.*, "Fsft-net: face transfer video generation with few-shot views," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 3582–3586.

[328] Y. Song *et al.*, "Objectstitch: Object compositing with diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 18 310–18 319.

[329] S. Pramanick, Y. Song, S. Nag, K. Q. Lin, H. Shah, M. Z. Shou, R. Chellappa, and P. Zhang, "Egovlpv2: Egocentric video-language pre-training with fusion in the backbone," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 5285–5297.

[330] W. Gan, Z. Qi, J. Wu, and J. C.-W. Lin, "Large language models in education: Vision and opportunities," *arXiv preprint arXiv:2311.13160*, 2023.

[331] L. Liu, L. Gao *et al.*, "A survey on deep multi-modal learning for body language recognition and generation," *arXiv preprint arXiv:2308.08849*, 2023.

[332] M. De Coster *et al.*, "Machine translation from signed to spoken languages: State of the art and challenges," *Universal Access in the Information Society*, pp. 1–27, 2023.

[333] M. K. Mishra, "Generating video game quests from stories," Master's thesis, University of Twente, 2023.

[334] S. Koomen, "Text generation for quests in multiplayer role-playing video games," Master's thesis, University of Twente, 2023.

[335] V. Soni, "Large language models for enhancing customer lifecycle management," *Journal of Empirical Social Science Studies*, vol. 7, no. 1, pp. 67–89, 2023.

[336] T. Medeiros, M. Medeiros, M. Azevedo, M. Silva, I. Silva, and D. G. Costa, "Analysis of language-model-powered chatbots for query resolution in pdf-based automotive manuals," *Vehicles*, vol. 5, no. 4, pp. 1384–1399, 2023.

[337] N. Gokce Narin, "The role of artificial intelligence and robotic solution technologies in metaverse design," in *Metaverse: Technologies, Opportunities and Threats*. Springer, 2023, pp. 45–63.

[338] T. Jung and M. C. tom Dieck, *XR-Metaverse Cases: Business Application of AR, VR, XR and Metaverse*. Springer Nature, 2023.

[339] Y. Yu, Z. Zeng, H. Hua, J. Fu, and J. Luo, "Promptfix: You prompt and we fix the photo," *arXiv preprint arXiv:2405.16785*, 2024.

[340] C. Huang, Y. Tian, A. Kumar, and C. Xu, "Egocentric audio-visual object localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 910–22 921.

[341] J. Bi, N. M. Nguyen, A. Vosoughi, and C. Xu, "Misar: A multimodal instructional system with augmented reality," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 1–5.

[342] T. Wang, J. Zhang, J. Fei, Y. Ge, H. Zheng, Y. Tang, Z. Li, M. Gao, S. Zhao, Y. Shan *et al.*, "Caption anything: Interactive image description with diverse multimodal controls," *arXiv preprint arXiv:2305.02677*, 2023.

[343] Y. Hu, H. Hua, Z. Yang, W. Shi, N. A. Smith, and J. Luo, "Promptcap: Prompt-guided task-aware image captioning," *arXiv preprint arXiv:2211.09699*, 2022.

[344] K. Rana, J. Haviland *et al.*, "Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning," in *Conference on Robot Learning*. PMLR, 2023, pp. 23–72.

[345] G. Eysenbach *et al.*, "The role of chatgpt, generative language models, and artificial intelligence in medical education: a conversation with chatgpt and a call for papers," *JMIR Medical Education*, vol. 9, no. 1, p. e46885, 2023.

[346] H. Liu *et al.*, "Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis," in *European conference on computer vision*. Springer, 2022, pp. 612–630.

[347] H. Liu, N. Iwamoto, Z. Zhu, Z. Li, Y. Zhou, E. Bozkurt, and B. Zheng, "Disco: Disentangled implicit content and rhythm learning for diverse co-speech gestures synthesis," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 3764–3773.

[348] H. Liu *et al.*, "Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1144–1154.

[349] C. Li *et al.*, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," *arXiv preprint arXiv:2306.00890*, 2023.

[350] M. Al-Hawawreh, A. Aljuhani, and Y. Jararweh, "Chatgpt for cybersecurity: practical applications, challenges, and future directions," *Cluster Computing*, vol. 26, no. 6, pp. 3421–3436, 2023.

[351] H. Mouratidis *et al.*, "Modelling language for cyber security incident handling for critical infrastructures," *Computers & Security*, vol. 128, p. 103139, 2023.

[352] Y. Lee, J. Kim, and P. Kang, "Lanobert: System log anomaly detection based on bert masked language model," *Applied Soft Computing*, vol. 146, p. 110689, 2023.

[353] C. Almodovar *et al.*, "Logfit: Log anomaly detection using fine-tuned language models," 2023.

[354] I. de Zarzà, J. de Curtò, and C. T. Calafate, "Socratic video understanding on unmanned aerial vehicles," *Procedia Computer Science*, vol. 225, pp. 144–154, 2023.

[355] J. Tang, Y. Yang *et al.*, "Graphgpt: Graph instruction tuning for large language models," *arXiv preprint arXiv:2310.13023*, 2023.

[356] B. Cao *et al.*, "Efficient masked autoencoder for video object counting and a large-scale benchmark," 2025. [Online]. Available: https://arxiv.org/abs/2411.13056

[357] C. Cui, Y. Ma *et al.*, "Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 902–909.

[358] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia, "Lisa: Reasoning segmentation via large language model," *arXiv preprint arXiv:2308.00692*, 2023.

[359] P. Zhou *et al.*, "A survey on generative ai and llm for video generation, understanding, and streaming," *arXiv preprint arXiv:2404.16038*, 2024.

[360] H. Lin *et al.*, "Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning," *arXiv preprint arXiv:2309.15091*, 2023.

[361] D. Kondratyuk *et al.*, "Videopoet: A large language model for zero-shot video generation," *arXiv preprint arXiv:2312.14125*, 2023.

[362] Y. Jin *et al.*, "Efficient multimodal large language models: A survey," *arXiv preprint arXiv:2405.10739*, 2024.

[363] Z. Lu *et al.*, "B-vllm: A vision large language model with balanced spatio-temporal tokens," *arXiv preprint arXiv:2412.09919*, 2024.

[364] M. Hu *et al.*, "Edge-based video analytics: A survey," *arXiv preprint arXiv:2303.14329*, 2023.

[365] Y. Yao *et al.*, "Federated large language models: Current progress and future directions," *arXiv preprint arXiv:2409.15723*, 2024.

[366] A. Bastola *et al.*, "Fedmil: Federated-multiple instance learning for video analysis with optimized dpp scheduling," in *2024 20th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)*. IEEE, 2024, pp. 109–116.

[367] Y. Wang *et al.*, "Fedvmr: A new federated learning method for video moment retrieval," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[368] M. G. Mohammadkhani, S. Momtazi, and H. Beigy, "A survey on bridging vlms and synthetic data," 2025. [Online]. Available: https://openreview.net/pdf?id=ThjDCZOljE

[369] D. Zhang, W. Yao, X. Wang, Y. Hu, J. Luo, and D. Yu, "Multimedia-agent: A multimodal agent for multimedia content generation."

[370] J. Zhang *et al.*, "Eventhallusion: Diagnosing event hallucinations in video llms," *arXiv preprint arXiv:2409.16597*, 2024.

[371] S.-H. Lee *et al.*, "Video token merging for long-form video understanding," *arXiv preprint arXiv:2410.23782*, 2024.

[372] C. Tang *et al.*, "Enhancing multimodal llm for detailed and accurate video captioning using multi-round preference optimization," *arXiv preprint arXiv:2410.06682*, 2024.

[373] Y. Shang *et al.*, "Interpolating video-llms: Toward longer-sequence lmms in a training-free manner," *arXiv preprint arXiv:2409.12963*, 2024.

Yunlong Tang received the B.Eng. degree in Intelligence Science and Technology from the Southern University of Science and Technology (SUSTech) in 2023, supervised by Prof. Feng Zheng. She is currently pursuing a Ph.D. degree in Computer Science at the University of Rochester, advised by Prof. Chenliang Xu. Her research focuses on multimodal learning, especially video understanding and generation.

Jing Bi is currently pursuing a Ph.D. in Computer Science at the University of Rochester since 2020, advised by Prof. Chenliang Xu. He received his B.S. from Shandong University and M.S. from the University of Rochester.

Siting Xu received her B.Eng. (2019 - 2023) degree in Computer Science and Technology from Southern University of Science and Technology (SUSTech), supervised by Prof. Feng Zheng.

Luchuan Song is currently a Ph.D. candidate in Computer Science at the University of Rochester. He received his M.S. and B.S. from University of Science and Technology of China.

Susan Liang is currently a Ph.D. candidate in the Computer Science Department at the University of Rochester. His research focuses on multimodal learning.

Teng Wang is currently a Ph.D. candidate in the Department of Computer Science at the University of Hong Kong. He obtained his bachelor's and master's degrees from Sun Yat-sen University in 2017 and 2020, respectively. His research interests lie in vision-language multimodal learning and video understanding.

Daoan Zhang is currently a Ph.D. Student in Computer Science at the University of Rochester, advised by Prof. Jiebo Luo. His research focuses on generative AI.

Jie An is a Ph.D. candidate in Computer Science at the University of Rochester, advised by Prof. Jiebo Luo. Prior to that, he obtained his B.S. (2012 - 2016) and M.S. (2016 - 2019) in Applied Mathematics from Peking University, advised by Prof. Jinwen Ma. Jie's research focuses on improving the performance and extending the capability of GenAI models. He is particularly interested in image style transfer, generative models, image/video generation, and multi-modal generation/evaluation.

Jingyang Lin is a PhD student majoring in Computer Science at the University of Rochester, advised by Professor Jiebo Luo. He received his BE and MSc degrees from Sun Yat-sen University (SYSU), Guangzhou, China, in 2019 and 2022, respectively. His research interests include multimodal learning with LLMs, AI for health, and self-supervised learning.

Rongyi Zhu is currently a Ph.D. student in Computer Science at Stony Brook University. He received his MS. degree in Computer Science from the University of Rochester in 2024. His research focuses on trustworthy AI.

Ali Vosoughi is a PhD candidate in Electrical and Computer Engineering at the University of Rochester, working with Professors Chenliang Xu and Axel Wismueller. His research focuses on using AI for multimodality and complex reasoning to assist humans with challenging tasks. He holds a Bachelor's degree in Electrical Engineering from Sharif University of Technology (Iran), and two Master's degrees—one from Bogazici University (Turkey) and another from the University of Rochester (USA).

Chao Huang is currently a PhD candidate in Computer Science at the University of Rochester, advised by Prof. Chenliang Xu. Previously, he spent a year as a research assistant at the Chinese University of Hong Kong, working with Prof. Chi-Wing Fu. He received his B.Eng. from ESE Department, Nanjing University in 2019.

Zeliang Zhang received a B.Eng. degree in computer science from Huazhong University of Science and Technology in 2022. He is currently a Ph.D. candidate in Computer Science at the University of Rochester. His research focuses on trustworthy and efficient deep learning methods.

Pinxin Liu is a Ph.D. student at the University of Rochester, advised by Prof. Chenliang Xu. Before that, he was a Research Scientist Intern at Flawless AI. He received his B.S. from the Department of Computer Science at the University of Rochester. His research interest lies in human-related topics, e.g., video gesture synthesis, 3D face rendering, and text-to-motion generation.

Mingqian Feng received his B.S. degree in Physics from University of Science and Technology of China (USTC) and his M.S.E. degree in financial mathematics from Johns Hopkins University. He is currently a Ph.D. student in Computer Science at the University of Rochester. His research focuses on bias, optimization, and video understanding.

Feng Zheng is an Associate Professor at Southern University of Science and Technology (SUSTech). His research interests include machine learning (ML), computer vision (CV) and human-computer interaction (HCI). He received a Ph.D. from the University of Sheffield, UK. Before joining SUSTech, he worked as a senior researcher at Tencent YouTu Lab in Shanghai, China. Prior to this, he worked as a postdoctoral researcher at the University of Pittsburgh, USA and as an assistant research professor at the Shenzhen Institute of Advanced Technology, CAS. In terms of academic research, he has published 85 papers in top international journals and conferences, including IEEE TPAMI/TITS/TIP, AAAI, NeuIPS, CVPR, and ECCV.

Jianguo Zhang is a Professor in the Department of Computer Science and Engineering, Southern University of Science and Technology. Previously, he was a Reader in Computing, School of Science and Engineering, University of Dundee, UK. He received a PhD from the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, 2002. His research interests include object recognition, medical image analysis, machine learning, and computer vision. He is a senior member of the IEEE and an Associate Editor of IEEE Transactions on Multimedia.

Ping Luo is an Associate Professor in the Department of Computer Science at the University of Hong Kong, Associate Director of the HKU Musketeers Foundation Institute of Data Science, and Deputy Director of the Joint Research Lab of HKU and Shanghai AI Lab. He earned his Ph.D. in Information Engineering from the Chinese University of Hong Kong in 2014, supervised by Prof. Xiaoou Tang and Prof. Xiaogang Wang. Before joining HKU in 2019, he was a Research Director at SenseTime. He has published over 100 papers in top conferences and journals, with 50,000+ citations on Google Scholar. His awards include the 2015 AAAI Easily Accessible Paper, 2022 ACL Outstanding Paper, 2023 WAIC Outstanding Paper, and ICCV'23 Best Paper nomination. In 2020, he was named one of the MIT Technology Review's Innovators Under 35 in Asia-Pacific. He has mentored 30 Ph.D. students, many of whom have won prestigious awards like the Nvidia and Baidu Fellowships.

Jiebo Luo is the Albert Arendt Hopeman Professor of Engineering at the University of Rochester, which he joined in 2011 after over 15 years at Kodak Research Laboratories. He has played key roles in numerous conferences, serving as General Co-Chair of ACM Multimedia 2018 and IEEE ICME 2024, and Program Co-Chair of ACM Multimedia 2010, IEEE CVPR 2012, and IEEE ICIP 2017. He has served on editorial boards of several major journals, including IEEE Transactions on Pattern Analysis and Machine Intelligence and IEEE Transactions on Multimedia, and was the Editor-in-Chief of IEEE Transactions on Multimedia from 2020 to 2022. Prof. Luo is a Fellow of ACM, AAAI, IEEE, AIMBE, IAPR, and SPIE, and a member of Academia Europaea (AE) and the US National Academy of Inventors (NAI). His honors include the 2024 IEEE Computer Society Edward J. McClusky Technical Achievement Award, the 2021 ACM SIGMM Technical Achievement Award, the 2024 William H. Riker University Award for Excellence in Graduate Teaching, the 2024 Edmund A. Hajim Outstanding Faculty Award, and the inaugural 2024 Debra Haring Excellence in Research Award.

Chenliang Xu is an Associate Professor in the Department of Computer Science at the University of Rochester. He received his Ph.D. in Computer Science from the University of Michigan in 2016, an M.S. in Computer Science from the University at Buffalo in 2012, and a B.S. in Information and Computing Science from Nanjing University of Aeronautics and Astronautics, China, in 2010. His research originates in computer vision and tackles interdisciplinary topics, including video understanding, audio-visual learning, vision and language, and methods for trustworthy AI. Xu is a recipient of the James P. Wilmot Distinguished Professorship (2021), the University of Rochester Research Award (2021), the Best Paper Award at the 17th ACM SIGGRAPH VRCAI Conference (2019), the Best Paper Award at the 14th Sound and Music Computing Conference (2017), and the University of Rochester AR/VR Pilot Award (2017). He has authored over 100 peer-reviewed papers in computer vision, machine learning, multimedia, and AI venues. He served as an associate editor for IEEE Transactions on Multimedia and area chair/reviewer for various international conferences.