Audio-Visual LLM for Video Understanding

Fangxun Shu[1*] Lei Zhang[2*] Hao Jiang[1†] Cihang Xie [3†]
[1] Alibaba Group [2] University of California, San Diego [3] University of California, Santa Cruz

Abstract

This paper introduces Audio-Visual LLM, a novel Multimodal Large Language Model designed for holistic video understanding through integrated visual and auditory inputs. Our work innovates with a modality-augmented training approach, using uniquely designed modality-specific tokens to selectively activate the corresponding visual and auditory encoders. This mechanism is pivotal in efficient end-to-end training across diverse video data modalities, encompassing visual-only, audio-only, and combined audio-visual content. Additionally, we introduce a high-quality video instruction dataset, characterized by its robust temporal audio-visual correlations, which facilitates the model's adept handling of a wide range of audio-visual tasks, from nuanced audio-visual narratives to intricate reasoning. Extensive experiments demonstrate impressive zero-shot performance in various video understanding tasks, such as question answering, captioning, and complex reasoning, underscoring its potential in video understanding.

1. Introduction

Videos are inherently multimodal, encapsulating both auditory and visual information. This multi-modality is not just an inherent characteristic of videos but also a fundamental aspect of how humans perceive and interact with visual media. For example, in a cinematic context, simultaneous engagement with visual imagery and auditory cues significantly enriches the viewing experience, enhancing both comprehension and enjoyment. Recent developments in multimodal models [20, 22, 28, 39] highlight this by focusing on the integration of visual and auditory inputs, thus capturing a more comprehensive representation of video content.

Meanwhile, large language models (LLMs) [3, 5, 43] have shown remarkable capabilities in intent understanding and instruction following. They can interact well with human intentions and offer tailored responses. Building upon this, subsequent research [2, 17, 25] have augmented

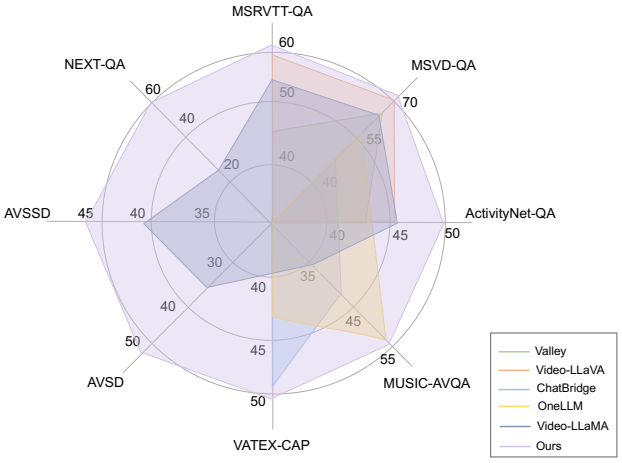


Figure 1. Audio-Visual LLM achieves advantageous performance on various video understanding tasks and consistently outperforms existing methods.

LLMs with visual perception abilities, introducing alignment mechanisms and curating datasets geared towards instruction-following in visual domains, including both image (e.g., MiniGPT-4 [55], LLaVA [30]) and video understanding (e.g., Qwen-VL [4], VALLEY [31], Video-ChatGPT [33]). Despite these advancements, most endeavors have predominantly focused on visual data, overlooking the rich auditory data inherent in videos. This oversight suggests a significant opportunity for enhancing video understanding by fully harnessing the auditory component, a gap our work aims to bridge by integrating comprehensive auditory data processing within the LLM framework.

To bridge this gap, recent works [32, 40, 52] have begun to jointly incorporate visual and audio components into LLMs to improve their video understanding ability. However, these existing models still exhibit limitations. For example, Video-LLaMA's [52] reliance on the pre-trained ImageBind model [16] for audio signal representation may restrict its capacity to capture the full depth of audio-visual interactions. Similarly, MacawLLM [32] and OneLLM [18] adopt visual and audio signals extracted from videos of different sources, which may induce bias and instability in training. These observations highlight a substantial oppor-

*Equal Contribution
†Corresponding Author

tunity for improvement in aligning audio-visual modalities, both in terms of model architecture and dataset development.

To this end, we propose a multimodal LLM framework that synergistically aligns visual and audio signals for holistic video understanding. This framework presents two main innovations. Firstly, we introduce a modality augmentation technique within the Audio-Visual LLM training process. This technique employs modality-specific tokens to trigger the relevant visual or auditory encoders depending on the input type, thereby facilitating exploration of the nuanced interplay between audio and visual components in videos. This selective activation is also pivotal in enabling end-to-end joint training, promoting dynamic merging of audio and visual elements in video content. Secondly, we introduce a GPT-assisted pipeline to curate visual/audio-text pairs into the appropriate and diverse instruction-following format, using GPT-4 [40], to improve the training of Audio-Visual LLM. Our focus during curation is to maintain audio-visual temporal consistency while minimizing hallucinations. We craft audio-visual descriptions with strong temporal associations and design intricate prompts to guide GPT-4 in curating diverse instruction data with less hallucination.

Extensive experiments demonstrate that our Audio-Visual LLM achieves superior zero-shot performance on a range of video understanding tasks. Notably, our model achieves an accuracy of 60.4% in MSRVTT-QA, 50.6% in ActivityNet-QA, and 47.9% in MUSIC-AVQA in question-answering tasks. In captioning task, it attains a CIDEr score of 51.4% on VATEX-CAP and 30.3% on VALOR-CAP. Furthermore, in complex reasoning tasks, Audio-Visual LLM demonstrates its prowess with a accuracy of 59. 7% in NExT-QA and a score of 26. 1% BLEU-1 in FAVDBench. These results not only surpass the performance of existing LLM-based models, such as Video-LLaMA [52] but also outperform conventional models like InterVideo [46], underscoring the superior effectiveness and innovation of our method in video understanding.

2. Related Works

2.1. Large Language Models for Video

Inspired by the strong instruction-following ability of LLMs [3, 35, 43] recent researches have extended these models to understand multimodal content, focusing particularly on images [4, 30] and videos [29, 31, 33]. They typically focus on designing multimodal projection layers to align with LLMs. LLaVA [30] and MiniGPT4 [55] connect images to LLM with learnable projectors, using image instruction datasets curated via GPT-4. Valley [31] and Video-LLaVA [29] extend it to unify images and videos into LLMs. Recently, some studies such as Video-LLaMA [52], MacawLLM [32], and OneLLM [18] have understood visual and audio content within videos. Despite these efforts, they usually concentrate on single-modal interpretation in videos, with the intricate alignment between visual and auditory modalities being relatively unexplored. To address this gap, we introduce a modality-augmented training paradigm aimed at thoroughly exploring the alignment between visual and audio modalities.

2.2. Video Instruction Datasets

A high-quality instruction dataset [3, 14, 42, 43] is crucial for LLMs and has been extended to the vision domain. Most works [10, 21, 30, 53] leverage GPT to curate the instruction-response pair based on the textural description of vision content. For instance, LLaVA [30] and GPT4RoI [53] use text annotations such as captions and detection boxes of images to generate multi-turn conversations, Complex VQA, and detail descriptions for visual instruction tuning. Recently, some works have explored the GPT-based curation of video instruction datasets. Valley [31] and VideoChat [26] sequentially input dense captions to GPT in the temporal order. Despite the inherent challenges in comprehending individual frames, the audio information is also lost. We propose an automatic audio-visual instruction curation pipeline that incorporates both visual and audio signals within videos.

3. Method

In this section, we first introduce the overall architecture of our Audio-Visual LLM as shown in Figure 2. Secondly, we propose a modality-augmented training strategy to improve audio-visual alignment. Lastly, we curate a high-quality instruction dataset of visual-only, audio-only, and joint audio-visual content as shown in Figure 3, facilitating the modality-augmented training.

3.1. Model Architecture

Our model consists of three components: the multimodal encoders, the linear projectors, and the large language model, as illustrated in Figure 2.

Multimodal Encoders. Given a video, we first decompose it into individual video frames $F \in \mathbb{R}^{T\times H\times W\times 3}$ and audio segments $A \in \mathbb{R}^{K\times M}$, where T denotes the number of frames and K denotes the number of audio segments. Each video frame is further segmented into N non-overlapping patches, creating spatio-temporal patches $V \in \mathbb{R}^{T\times N\times P\times P\times C}$, with P indicating the patch size. For frame-level processing, we employ CLIP [36] to encode each frame into embeddings $E_v \in \mathbb{R}^{T\times N\times D}$. To adapt these embeddings for longer video sequences, we introduce a novel approach that enhances flexibility. We aggregate the `[CLS]` token from each frame's embedding to compile the temporal tokens $E_t \in \mathbb{R}^{T\times D_t}$, where D_t is the dimension

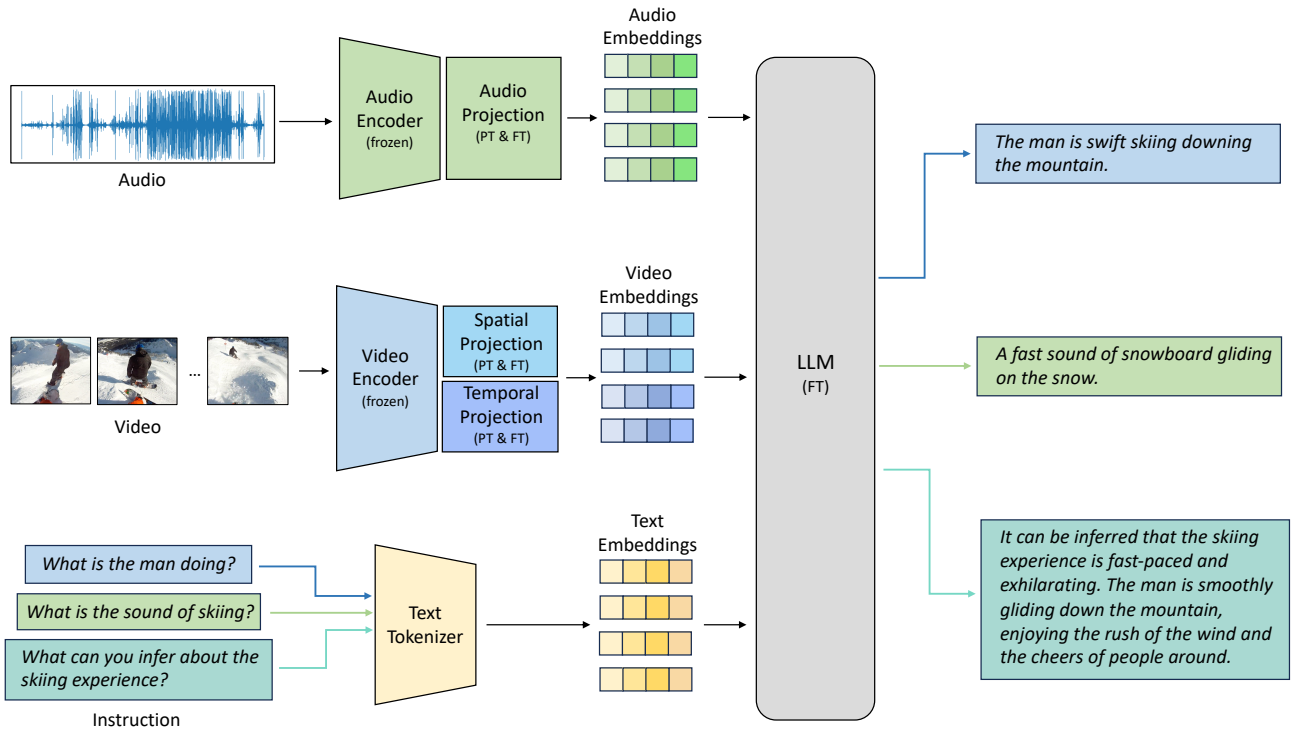


Figure 2. The framework of our method consists of three components: multimodal encoders, linear projections, and LLM. We pre-train the projections in the pre-training stage and fine-tune both the projections and LLM in the supervised-finetuning stage. Both stages freeze the multimodal encoders.

of these tokens. Concurrently, we average the patch embeddings across the temporal axis to generate spatial tokens $E_s \in \mathbb{R}^{N \times D_s}$, with D_s as their dimensional size. This strategy reduces the overall sequence length fed into the LLMs to $T + N$, streamlining the input while preserving essential spatiotemporal information, as shown below:

$$E_t = \{v_{cls}^1, v_{cls}^2, ..., v_{cls}^T\}, \tag{1}$$

$$E_s = \{\bar{v}^1, \bar{v}^2, \bar{v}^3, ..., \bar{v}^N\}. \tag{2}$$

For audio segments, we employ CLAP [47] to derive the audio embeddings by capturing the last hidden state, which reflects the semantic content of the audio. This process distills the essential auditory information into a compact form, represented as $E_a \in \mathbb{R}^{K \times D_a}$, where D_a denotes the dimension of the embeddings for each audio segment, as shown below:

$$E_a = \{a^1, a^2, a^3, ..., a^K\}. \tag{3}$$

Linear Projectors. To align with the LLM's embedding dimension D_l, we incorporate linear projection layers to normalize the dimensions of temporal tokens $E_t \in \mathbb{R}^{T \times D_t}$, spatial tokens $E_s \in \mathbb{R}^{N \times D_s}$, and auditory tokens $E_a \in \mathbb{R}^{K \times D_a}$. Linear transformations are applied to each token type through specific weight matrices W and bias vectors b, facilitating their integration into the LLM framework as follows:

$$\begin{aligned} H_t &= W_t E_t + b_t, \\ H_s &= W_s E_s + b_s, \\ H_a &= W_a E_a + b_a. \end{aligned} \tag{4}$$

These transformations result in the final video token representation $H_v \in \mathbb{R}^{(T+N+K) \times D_l}$, comprising T temporal tokens (H_t), N spatial tokens (H_s), and K auditory tokens (H_a).

Large Language Model. Our methodology utilizes the Vicuna model, an open-source LLM that has been fine-tuned on the LLaMA [43], using a diverse dataset of approximately 70,000 dialogues from ShareGPT. To interpret video content. We integrate instruction tokens I with the video tokens H_v, and this composite input is processed by Vicuna. This setup facilitates the generation of textual responses R by the LLM, ensuring relevance and coherence with the video's context, as depicted in the following equation:

$$R = \text{LLM}(H_v, I). \tag{5}$$

3.2. GPT-Assisted Data Curation

In this section, we introduce a GPT-assisted data curation approach as illustrated in 3. The primary focus in the

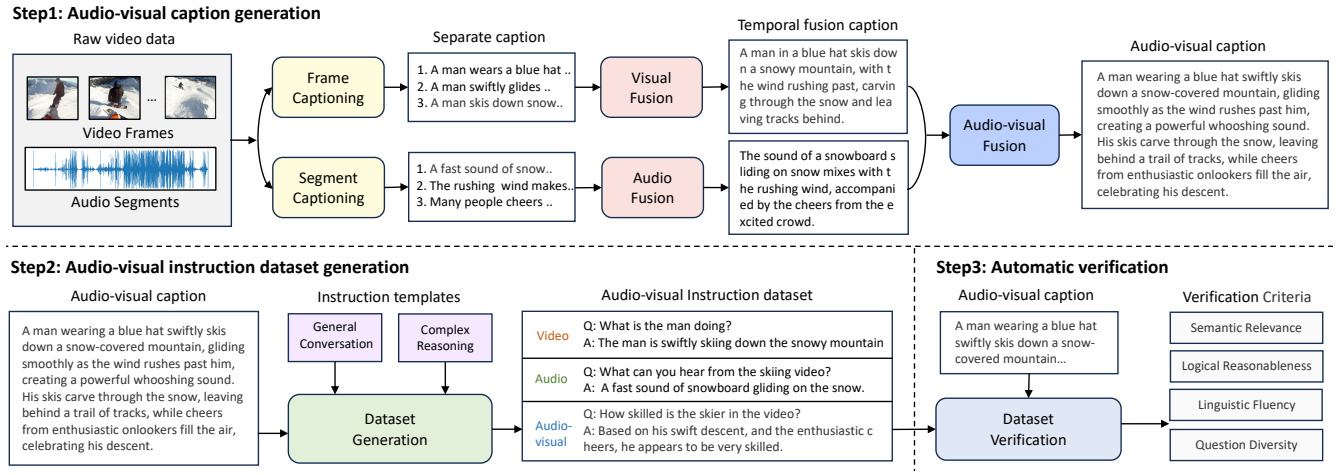


Figure 3. The framework for GPT-assisted automated data curation consists of three key steps. 1) we employ pre-trained visual and audio models to extract frame-level and segment-level captions, followed by using GPT-4 to fuse them into temporally coherent video and audio captions, ultimately combining them into a comprehensive audiovisual caption; 2) based on the audio-visual captions, we construct data generation templates, prompting GPT-4 to create a diverse dataset of multimodal instruction data. 3) we design various verification criteria to filter the high-quality instruction data.

curation is to maintain audio-visual temporal consistency while minimizing hallucinations. We utilize GPT-4 [35] and pre-trained models [24, 34] to generate audio-visual captions with strong temporal associations and design intricate prompts to guide GPT-4 to curate diverse instruction data with less hallucination based on the captions. Detailed explanations are shown below.

Audio-visual captions generation. For our selection of the datasets, we use ACAV100M [19] and VGGSound [8], recognized for their high-quality synchronization of audio and visual content. This choice guarantees that the audio and visual elements in each video align correctly over time, allowing us to produce sequential captions for both video frames and audio slices. We use BLIP to create captions at the frame level, denoted $V_c = \{v_c^1, v_c^2, ..., v_c^T\}$, and we use HTSAT [34] to generate captions for audio segments, denoted $A_c = \{a_c^1, a_c^2, ..., a_c^K\}$. Following this, we employ GPT-4 to fusion them into video and audio captions that are temporally coherent, and ultimately combine them into a comprehensive and consistent audio-visual caption.

Instruction Type	Audio	Visual	Aud-Vis	Total
Audio-visual Descriptions	20k	50k	30k	100k
Multi-Turn Conversations	20k	60k	40k	120k
Complex Reasoning	5k	20k	15k	40k

Table 1. The distribution of the instruction dataset. We generate various instruction types including multi-turn conversations, audio-visual descriptions, and complex reasoning. We also cover various video data types, including audio, visual, and audio-visual.

Audio-visual instruction dataset generation. We introduce a systematic prompting framework designed to derive varied instruction-following data from audio-visual captions. Leveraging GPT-4 for data creation, our framework comprises four main elements: ROLE, defining the task at hand; REQUIREMENT, specifying the guidelines for formulating instructions with a focus on high-quality and diverse questions; EXAMPLE, offering an in-context example; and CONTEXT, where the generated audio-visual captions serve as the foundation for data production. Through this comprehensive approach, we aim to produce a rich and diverse dataset of instruction data that enhances the performance of models in understanding and responding to complex queries related to audio-visual content.

Automatic verification. To ensure the generated instruction question-answer pairs are relevant, coherent, and useful, we employ a systematic filter process based on several key criteria. First, we assess semantic relevance, ensuring that the generated data aligns closely with the original audio-visual captions. Next, we evaluate the completeness of the information, confirming that the instruction data comprehensively covers the essential aspects of the audio-visual content. We also consider logical soundness, verifying that the instructions provide coherent and meaningful guidance. Additionally, we emphasize linguistic fluency, ensuring that the text is grammatically correct and clearly articulated. Finally, we prioritize diversity, filtering out redundant or repetitive entries to promote a rich variety of instruction data. Through this systematic filtering approach, we successfully identify and retain only high-quality instruction data that meet our established standards. As illus-

trated in Table 1, we generate a total of 260,000 instruction data pairs from a combination of audio, visual, and audio-visual instances, including 100,000 comprehensive audio-visual descriptions, 120,000 multi-turn conversations, and 40,000 complex reasoning cases.

3.3. Modality-Augmented Training

Video inherently contains visual signals, audio signals, and combined audio-visual signals. We propose a novel training paradigm known as Modality-Augmented Training (MAT), which enables the simultaneous training of three types of modalities: visual-only, audio-only, and joint audio-visual samples within a single batch. This approach allows our model to simultaneously consider multiple perspectives of the video, facilitating a more comprehensive understanding.

System: $\texttt{X}_{\texttt{prompt}}$\n
Human: $\texttt{X}^1_{\texttt{instruction}}$<VIS>\n Assistant: $\texttt{X}^1_{\texttt{response}}$\n
Human: $\texttt{X}^2_{\texttt{instruction}}$<AUD>\n Assistant: $\texttt{X}^2_{\texttt{response}}$\n
Human: $\texttt{X}^3_{\texttt{instruction}}$<AUD_VIS>\n Assistant: $\texttt{X}^3_{\texttt{response}}$\n

Figure 4. Data format for modality-augmented training.

As illustrated in Figure 4, we structure the data as a multi-turn conversation instruction format. `System` refers to the system message $\texttt{X}_{prompt}$ that outlines the role and expected operation of LLMs, and we set it as *"A chat between a curious human and a video assistant."*. `Human` represents the instruction, which includes a modality-specific token <MOD> $\in$ {<VIS>, <AUD>, <AUD_VIS>} that indicates the modality being employed; this token is replaced with the embedding extracted from the corresponding encoder. `Assistant` denotes the text response $\texttt{X}_{response}$ of the model.

The training process is structured into two distinct phases. In the initial phase, we focus on crafting general audio-visual descriptions for pre-training, during which only the projectors are updated. This phase allows the model to learn effective representations from audio-visual captions. The second phase shifts to generating detailed responses for multi-turn dialogues and complex reasoning tasks, where both the LLMs and projectors are jointly fine-tuned. In this phase, we incorporate modality-specific tokens <MOD> to specify the input modality within the prompts, helping the model to distinguish between various sources of modality. The training objective is formulated using a cross-entropy loss that focuses on the response sequence Y, represented as follows:

$$\mathcal{L} = -\sum_{t=1}^{T} \log P(y_t | y_{<t}, \texttt{<MOD>}, x) \tag{6}$$

where $P(y_t | y_{<t}, \texttt{<MOD>}), x$ denotes the probability of generating the token y_t given all prior tokens $y_{<t}$, the modality token <MOD> and the input x. This structured approach ensures effective learning from both general audio-visual contexts and complex instruction data, ultimately enhancing the model's understanding and performance.

4. Experiment

In this section, we first introduce the experimental setup and implementation details. We then describe the downstream tasks for which our method is evaluated and report strong results and ablation studies.

4.1. Experimental Setup

Model Settings. We build the visual encoder with ViT-L/14 [13], which is a transformer-based model composed of 24 layers of blocks, with a patch size of 14. We initialize it from the pre-trained CLIP [36] version via contrastive learning. Similarly, we build the audio encoder with HTSAT [9], which is also a transformer-based model with 4 groups of swin transformer blocks. We initialize it from the pre-trained CLAP [47] via contrastive learning. We use fully connected layers as linear projectors to convert spatial, temporal, and audio tokens to the LLM with a dimension of 4096. We build our LLM with Vicuna-v1.5 7B [12], which is an effective chat version that has been fine-tuned on LLaMA [43].

Training Datasets. We curate 260k video instruction data including 100k detailed descriptions and 160k complex instructions. We also use 650k image data from LLaVA, and 770k video data from Valley to enhance the shared spatio-temporal perception and reasoning capacity.

Implementation Details. We resize the videos to a resolution of 224×224 and uniformly sample 32 frames during training. We evenly divide the audio signals into 4 segments and sample each segment at a sampling rate of 48Khz. We treat images as 1-frame videos so that we can jointly train images and videos in a unified manner. We adopt FlashAttention and Deepspeed ZeRO for efficient training. We use the AdamW optimizer with $\beta = (0.9, 0.98)$. A cosine annealing learning rate schedule is applied with a warm-up ratio of 0.03. The training is conducted on 8×A100 GPUs with 80GB GPU memory. During the pre-training stage, we freeze the encoders and LLM, while only training the projectors. The learning rate is set to 2e-3. We use a total batch size of 256 and set the training epoch of 3, taking approximately 16 hours. During the instruction fine-tuning stage, we only freeze the encoders and jointly train the linear projectors and LLM. The learning rate is set to 2e-5. We use a total batch size of 128 and set the training epoch of 1, taking approximately 10 hours.

Method	MSRVTT-QA	MSVD-QA	ActivityNet-QA	AVSD	AVSSD	MUSIC-AVQA
JustAsk [50]	41.5	46.3	38.9	-	-	-
VALOR* [11]	46.7	56.4	44.8	-	-	76.6
FrozenBiLM [51]	47.0	54.8	43.2	-	-	-
InterVideo [46]	47.1	55.5	–	-	-	-
LLaMA-Adapter [15]	43.8	54.9	34.2	-	-	-
VideoChat [26]	45.0	56.3	26.5	-	-	-
Valley [31]	45.7	65.4	42.9	-	-	-
Video-ChatGPT [33]	49.3	64.9	35.2	-	-	-
Video-LLaVA [29]	59.2	70.7	45.3	-	-	-
MacawLLM [32]	25.5	42.1	14.5	34.3	36.2	31.8
ChatBridge [54]	-	45.3	-	-	-	43.0
OneLLM [18]	-	56.5	-	-	-	47.6
Video-LLaMA [52]	54.9	65.0	45.8	36.7	40.8	36.6
PandaGPT [40]	23.7	46.7	11.2	26.1	32.7	33.7
FAVOR* [41]	-	-	-	51.2	50.5	-
Ours	**60.4**	**72.1**	**50.6**	**53.4**	**48.3**	**47.9**

Table 2. Zero-shot evaluation of state-of-the-art methods on video question-answering (MSRVTT-QA, MSVD-QA, and ActivityNet-QA) and audio-visual question-answering (AVSD, AVSSD, and MUSIC-AVQA). VALOR* and FAVOR* are evaluated via finetuning.

Method	VATEX-CAP	VALOR-CAP	AVSD-COMP
ChatBridge	48.9	24.7	75.4
OneLLM	43.8	29.2	74.5
Ours	**51.4**	**30.3**	**76.8**

Table 3. Zero-shot evaluation of state-of-the-art methods on video captioning tasks (VATEX-CAP) and audio-visual captioning tasks (VALOR-CAP and AVSD-COMP).

Method	NExT-QA	MVBench	FAVDBench
Video-LLaMA	22.5	34.1	20.8 / 15.0
FAVOR	42.5	29.2	24.9 / 14.8
Ours	**59.7**	**44.6**	**26.1 / 15.8**

Table 4. Zero-shot evaluation of state-of-the-art methods on complex reasoning tasks (NExT-QA, MVBench, and FAVDBench).

4.2. Results

Downstream Tasks and Datasets We explore our method across a range of video-related tasks, including question-answering (video QA and audio-visual QA), captioning (video captioning and audio-visual captioning), and complex audio-visual reasoning, demonstrating its capability in holistic video understanding. Unless stated otherwise, we report top-1 accuracy [33] on QA tasks and CIDEr [44] on captioning tasks.

Question-answering. As shown in Table 2, we evaluate across various video QA tasks (MSRVTT [49], MSVD [7], and ActivityNet [6]) and audio-visual QA tasks (AVSD [1], AVSSD [8], MUSIC-AVQA [23]). Our method yields superior performance across these benchmarks over previous methodologies, indicating a more nuanced audio-visual understanding.

For video QA, by integrating audio with visual information, our model achieves notable accuracy improvements: +1.2% on MSRVTT-QA, +1.5% on MSVD-QA and +4.8% on ActivityNet-QA. The significant improvement on ActivityNet-QA, with its longer video length, highlights our model's strength in processing long-duration audio-visual content. For audio-visual QA, our curation of instruction datasets with high audio-visual consistency enables more effective modality-augmented training. This approach yields substantial accuracy improvements over the state-of-the-art: +15.9% on AVSD, +7.5% on AVSSD, and +0.3% on MUSIC-AVQA, demonstrating the model's superior capability in comprehensive audio-visual alignment.

Captioning. As shown in Table 3, we evaluate across video captioning (VATEX-CAP [45]) and audio-visual captioning (VALOR-CAP [11] and AVSD-COMP [1]). By developing captions that are both contextually rich and temporally aligned with the corresponding audio-visual content, we achieve a more effective modality-augmented training process. This nuanced approach to synchronizing video and audio elements results in notable CIDEr score enhancements over the state-of-the-art: we observe increases of +2.5% on VATEX-CAP, +1.1% on VALOR-CAP, and +1.4% on AVSD-COMP. These results underscore our model's adeptness in capturing and expressing the intricate dynamics of audio-visual information.

Complex Reasoning. As shown in Table 4, we extend our evaluation to include intricate question-answering (NExT-QA [48]), comprehensive video understanding (MVBench [27]), and detailed audio-visual descriptions (FAVDBench [38]). Our model outperformed the state-of-the-art, achieving a +2.4% accuracy increase on NExT-QA [48], +10.5% on MVBench [27], and +5.3% BLEU-1 on FAVDBench [38]. These gains highlight our model's effectiveness in intricate audio-visual understanding.

4.3. Ablation Studies

In this section, we investigate the effects of various design choices including training strategy, modality integration, model size, and sequence length. Our standard configuration employs a ViT/L-14 for visual encoding, HTSAT for auditory processing, and Vicuna-7B as the foundational LLM architecture. For both training and inference phases, videos are segmented into 32 frames coupled with 4 audio segments.

Training Strategy. We explore the influence of training strategy on the understanding of video content through ablation experiments on video QA and audio-visual QA tasks. Specifically, we evaluate our Modality-Augmented Training (MAT) against two alternative strategies: PT1, which combines visual and audio data training initially and then separates them, and PT2, which starts with separate training and then merges the modalities

Table 5 shows that MAT significantly outperforms the alternatives, with improvements of +2.1% on MSRVTT-QA and +3.2% on AVSSD. This confirms the effectiveness of MAT in enhancing audio-visual comprehension. Additionally, the lower performance of PT1 suggests that isolated training of visual and audio data can lead to misalignment, emphasizing the value of integrated modality training.

Modality Integration. In assessing the impact of integrating both visual and audio modalities for video understanding, We conduct ablation studies on video QA and audio-visual QA tasks, comparing the performance when using both audio-visual inputs to visual-only inputs during inference.

As shown in Figure 5, incorporating visual and audio modalities, rather than utilizing only the visual modality, leads to a significant improvement, such as +1.7% on MSRVTT-QA and +12.1% on AVSSD. This underscores the critical role of audio-visual integration in achieving a holistic understanding of video content. Notably, even video QA tasks, traditionally reliant on visual data alone, show marked improvement with audio input, highlighting audio's essential contribution to interpreting video.

Size of Model Architecture. The size of the model architecture critically impacts performance. Our baseline setup employs ViT-L/14 (CLIP) as the visual encoder, HTSAT (CLAP) for audio input, and Vicuna-7B (LLaMA) for language processing. To explore scalability, we experimented with various sizes for these components, replacing HTSAT with different versions of Whisper [37] due to the size restrictions of HTSAT.

Method	MSRVTT-QA	ActivityNet-QA	AVSSD	MUSIC-AVQA
PT1	56.7	45.3	43.6	43.5
PT2	58.3	48.1	45.1	45.4
MAT	**60.4**	**50.6**	**48.3**	**47.9**

Table 5. Ablation experiments on the training strategy. We report comparison results between modality-augmented training **(MAT)** versus plain training of two versions: **PT1** and **PT2**.

VisEnc	MSRVTT-QA	ActivityNet-QA	AVSSD	MUSIC-AVQA
ViT-B/16	58.9	48.2	46.7	46.3
ViT-L/14	60.4	50.6	48.3	47.9
ViT-H/14	60.9	51.2	49.0	48.4

Table 6. Ablation experiments on scaling visual encoders. We report results among ViT-B/16, ViT-L/14, and ViT-H/14.

AudEnc	MSRVTT-QA	ActivityNet-QA	AVSSD	MUSIC-AVQA
W-Tiny	58.5	48.4	45.4	45.7
W-Base	59.3	49.8	47.8	48.2
W-Small	60.7	51.3	49.1	48.7

Table 7. Ablation experiments on scaling audio encoders. We report results among Whisper-Tiny, Base, and Small.

LLM	MSRVTT-QA	ActivityNet-QA	AVSSD	MUSIC-AVQA
V-7B	60.4	50.6	48.3	47.9
V-13B	61.5	51.9	50.4	49.3

Table 8. Ablation experiments on scaling LLMs. We report results between Vicuna-7B and Vicuna-13B.

The results, detailed from visual encoder scaling in Table 6 to LLM scaling in Table 8, indicate that larger models lead to better results. Notably, we obverse a +1.5% improvement in MSRVTT-QA when upgrading from ViT-B/16 to ViT-L/14, a +0.8% improvement when switching from Whisper-Tiny to Whisper-Base, and a +1.1% increase from Vicuna-7B to Vicuna-13B. These results highlight the importance of combining sophisticated multimodal encoders, which enable detailed perception, with robust LLMs, which support intricate reasoning, to improve video understanding.

Evaluation on Multiple Dimensions. The initial assessment is centered on a single aspect, such as accuracy. To comprehensively assess the effectiveness of our approach, we adopt the methodology from Video-ChatGPT and utilize GPT-4 to rate various dimensions in videos on a scale of 1 to 5: information accuracy (Correct), attention to detail (Detail), contextual grasp (Context), understanding of temporal elements (Temporal), and consistency. The notable

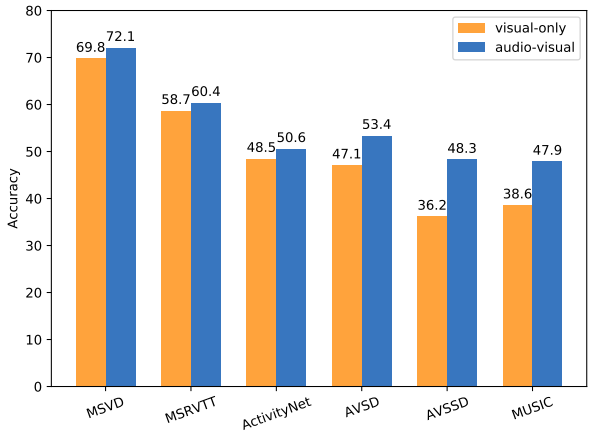


Figure 5. Ablation experiments on integrating video modalities for video understanding. We report the comparison results on video QA and audio-visual QA between joint audio-visual modalities versus only the visual modality.

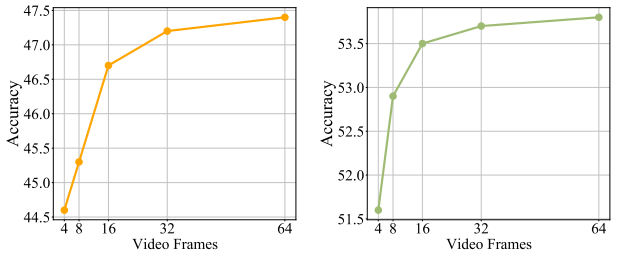
advancement shown in Table 9 highlights the comprehensive understanding of our method of video content.

Method	Correct	Detail	Context	Temporal	Consistency
LLaMA-Adapter	2.03	2.32	2.30	1.98	2.15
Video-LLaMA	1.96	2.18	2.16	1.82	1.79
VideoChat	2.23	2.50	2.53	1.94	2.24
Video-ChatGPT	2.40	2.52	2.62	1.98	2.37
Valley	2.43	2.13	2.86	2.04	2.45
Ours	**2.56**	**2.47**	**2.93**	**2.17**	**2.51**

Table 9. Evaluation on multiple dimensions for video understanding. We follow Video-ChatGPT to report score results (1∼5) on correct, detail, context, temporal, and consistency.

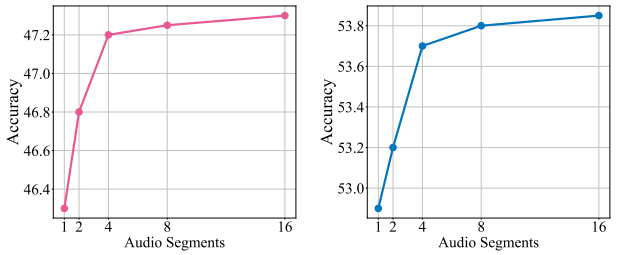
Length of Sequence. In video, each frame encapsulates instantaneous visual data while audio segments capture transient auditory information. Increasing the number of frames and audio segments enhances contextual detail extraction, thereby improving the LLM's capacity for holistic video comprehension. Our baseline configuration employs 32 video frames and 4 audio segments. To investigate the impact of sequence length, we perform ablation studies on video frames (testing lengths in 4, 8, 16, 32, 64) and audio segments (testing lengths in 1, 2, 4, 8, 16).

As demonstrated in Fig. 6 to Fig. 7 across video QA benchmarks with varying durations—MSRVTT-QA (15s) and ActivityNet-QA (180s)—accuracy improvements correlate with sequence length expansion, though marginal gains diminish progressively. Specifically, video frame scaling yields significant improvements up to 32 frames, while audio segments plateau near 4 units. This phenomenon suggests that while extended sequences enrich



(a) ActivityNet-QA (b) MSRVTT-QA

Figure 6. Ablation experiments on lengths of video frames. We report results on video frames of length 4, 8, 16, 32, and 64.



(a) ActivityNet-QA (b) MSRVTT-QA

Figure 7. Ablation experiments on lengths of audio segments. We report results on audio segments of length 1, 2, 4, 8, and 16.

temporal context, inherent redundancy in continuous video data ultimately constrains further performance enhancement. Future research directions should prioritize developing modules that optimize the trade-off between information sufficiency and computational efficiency, potentially through intelligent redundancy reduction mechanisms. Such advancements would enable LLMs to dynamically focus on semantically critical frames and audio segments while minimizing processing overhead from redundant content.

5. Conclusion

We introduce Audio-Visual LLM, a multimodal framework that empowers LLM with video instruction-following capability. The modality-augmented training plays a crucial role in enabling end-to-end joint training with video data across different modalities, including visual-only, audio-only, and audio-visual formats. Additionally, we present a high-quality video instruction dataset with strong audio-visual associations, derived from GPT-4, which enables our model to effectively process a wide range of task-oriented video instructions, spanning from multi-turn conversations and audio-visual narratives to complex reasoning tasks. Extensive experiments demonstrate the impressive performance of Audio-Visual LLM across diverse video understanding tasks.

References

[1] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Marks, et al. Audio scene-aware dialog. In *CVPR*, 2019. 6

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a language model for few-shot learning. *NeurIPS*, 2022. 1

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, et al. Qwen technical report. *arXiv preprint*, 2023. 1, 2

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, , et al. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint*, 2023. 1, 2

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020. 1

[6] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 6

[7] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, pages 190–200, 2011. 6

[8] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 4, 6

[9] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP*, 2022. 5

[10] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint*, 2023. 2

[11] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint*, 2023. 6

[12] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, and others. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 5

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 5

[14] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *ACL*, pages 320–335, 2022. 2

[15] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient instruction model. *arXiv preprint*, 2023. 6

[16] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 1

[17] Tanmay Gupta, Amita Kamath, Aniruddha Kembhavi, and Derek Hoiem. Towards general purpose vision systems: An end-to-end task-agnostic vision-language architecture. In *CVPR*, 2022. 1

[18] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *CVPR*, 2024. 1, 2, 6

[19] Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale Song. Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning. In *ICCV*, pages 10274–10284, 2021. 4

[20] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, 2021. 1

[21] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint*, 2023. 2

[22] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *CVPR*, 2022. 1

[23] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *CVPR*, 2022. 6

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900. PMLR, 2022. 4

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 1

[26] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, et al. Videochat: Chat-centric video understanding. *arXiv preprint*, 2023. 2, 6

[27] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, , et al. Mvbench: A comprehensive multi-modal video understanding benchmark. *arXiv preprint*, 2023. 7

[28] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *EMNLP*, pages 2046–2065, 2020. 1

[29] Bin Lin, Bin Zhu, Yang Ye, et al. Video-llava: Learning united representation by alignment before projection. *arXiv preprint*, 2023. 2, 6

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2024. 1, 2

[31] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley:

Video assistant with large language model enhanced ability. *arXiv preprint*, 2023. 1, 2, 6

[32] Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, et al. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint*, 2023. 1, 2, 6

[33] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 1, 2, 6

[34] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, et al. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *arxiv preprint*, 2024. 4

[35] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. 2, 4

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, et al. Learning transferable models from natural language supervision. In *ICML*, 2021. 2, 5

[37] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *ICML*, 2023. 7

[38] Xuyang Shen, Dong Li, Jinxing Zhou, Zhen Qin, Bowen He, Xiaodong Han, Aixuan Li, Yuchao Dai, Lingpeng Kong, Meng Wang, et al. Fine-grained audible video description. In *CVPR*, 2023. 7

[39] Fangxun Shu, Biaolong Chen, Yue Liao, Shuwen Xiao, Wenyu Sun, Xiaobo Li, Yousong Zhu, Jinqiao Wang, and Si Liu. Masked contrastive pre-training for efficient video-text retrieval. *arXiv preprint*, 2022. 1

[40] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint*, 2023. 1, 6

[41] Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, et al. Fine-grained audio-visual joint representations for multimodal large language models. *arXiv preprint*, 2023. 6

[42] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023. 2

[43] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Lacroix, et al. Llama: Open and efficient foundation language models. *arXiv preprint*, 2023. 1, 2, 3, 5

[44] R Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 6

[45] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 2019. 6

[46] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *ECCV*, 2024. 2, 6

[47] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, 2023. 3, 5

[48] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021. 7

[49] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 6

[50] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, 2021. 6

[51] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *NeurIPS*, 2022. 6

[52] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint*, 2023. 1, 2, 6

[53] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint*, 2023. 2

[54] Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. Chatbridge: Bridging modalities with large language model as a language catalyst. *arXiv preprint*, 2023. 6

[55] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint*, 2023. 1, 2