Empowering LLMs with Pseudo-Untrimmed Videos for Audio-Visual Temporal Understanding

Yunlong Tang[1], Daiki Shimada[2], Jing Bi[1], Mingqian Feng[1], Hang Hua[1], Chenliang Xu[1, *]

[1]University of Rochester
[2]Sony Group Corporation
{yunlong.tang, jing.bi, mingqian.feng, chenliang.xu}@rochester.edu,
Daiki.Shimada@sony.com, hhua2@cs.rochester.edu

Abstract

Large language models (LLMs) have demonstrated remarkable capabilities in natural language and multimodal domains. By fine-tuning multimodal LLMs with temporal annotations from well-annotated datasets, e.g., dense video captioning datasets, their temporal understanding capacity in video-language tasks can be obtained. However, there is a notable lack of untrimmed audio-visual video datasets with precise temporal annotations for events. This deficiency hinders LLMs from learning the alignment between time, audio-visual events, and text tokens, thus impairing their ability to localize audio-visual events in videos temporally. To address this gap, we introduce PU-VALOR, a comprehensive audio-visual dataset comprising over 114,081 pseudo-untrimmed videos with detailed temporal annotations. PU-VALOR is derived from the large-scale but coarse-annotated audio-visual dataset VALOR, through a subtle method involving event-based video clustering, random temporal scaling, and permutation. By fine-tuning a multimodal LLM on PU-VALOR, we developed AVicuna, a model capable of aligning audio-visual events with temporal intervals and corresponding text tokens. AVicuna excels in temporal localization and time-aware dialogue capabilities. Our experiments demonstrate that AVicuna effectively handles temporal understanding in audio-visual videos and achieves state-of-the-art performance on open-ended video QA, audio-visual QA, and audio-visual event dense localization tasks.

Introduction

Large Language Models (LLMs) have recently advanced natural language processing (NLP), evolving into Multimodal LLMs (MLLMs) capable of comprehending various modalities like text, images, audio, and videos (Chen et al. 2023a; Zhang et al. 2023b; Liu et al. 2023; Lin et al. 2023a; Xu et al. 2023; Zhang et al. 2023a). Despite the advancements, MLLMs still struggle to provide a fine-grained understanding of spatial or temporal details in multimodal contexts. To address these limitations, several works have explored (Chen et al. 2023a; Xuan et al. 2023) incorporating object bounding box coordinates in natural language format into image-text data, enabling MLLMs to directly identify the location of objects in images using natural language and respond accurately to user-provided bounding box coordinates, thereby enhancing fine-grained region-level understanding (Chen et al. 2023a; Xuan et al. 2023; Peng et al. 2023; Bai et al. 2023). While MLLMs have demonstrated potential in fine-grained image understanding, it's equally crucial to extend their capabilities to the video domain to achieve detailed video comprehension. Some recent advancements (Yang et al. 2023; Li et al. 2023b; Wang et al. 2023a; Huang et al. 2024; Zhang et al. 2024) have leveraged natural language for temporal predictions in video understanding tasks, such as dense video captioning, video temporal grounding, etc. These approaches have demonstrated competitive performance compared to traditional regression-based methods while retaining general capabilities, such as video question answering (Video QA). However, these methods predominantly concentrate solely on visual content, overlooking the dynamic and multimodal nature of the real world. For example, audio-visual content in videos represents a common form of such data. By integrating various modalities, including sound, we can achieve a more comprehensive analysis of content. As we delve deeper into this integration, it faces two significant challenges:

(1) In contrast to the abundance of dense video caption datasets, the audio-visual domain faces a significant bottleneck due to the lack of datasets providing detailed audio-visual event captions with accurate timestamp annotations.

(2) Developing audio-visual learning methods that effectively capture the intricate blend of auditory and visual cues, enabling them to interpret complex, intertwined information across various events within videos.

To tackle the challenge (1), we propose a practical yet straightforward pipeline that leverages the VALOR-32K (Chen et al. 2023b) dataset with high-quality audio-visual captions to construct PU (**P**seudo-**U**ntrimmed) - VALOR dataset contains audio-visual videos with corresponding temporal boundary annotations. The PU-VALOR dataset is created by applying Random Temporal Scaling and Permutation to videos clustered by captions. This innovative approach, aimed at generating Pseudo-Untrimmed videos, theoretically enables the creation of an unlimited number of untrimmed videos. Consequently, the PU-VALOR dataset features over 114k video-caption pairs, each annotated with precise boundaries, thus offering a contribution toward enriching the audio-visual research domain.

*Corresponding author.

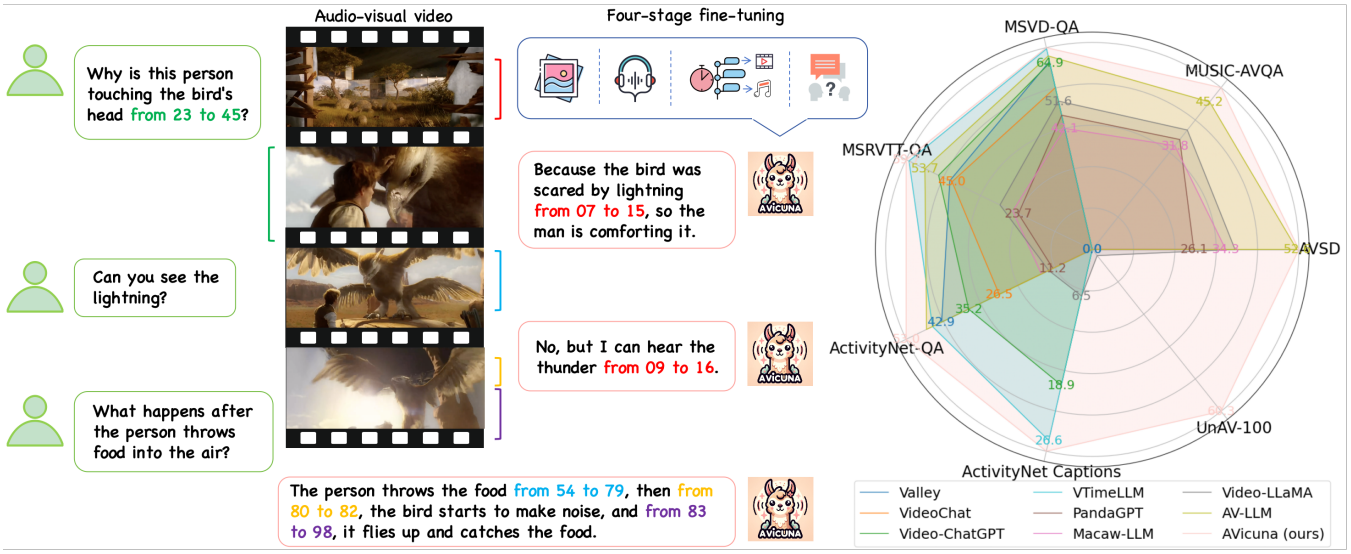

Figure 1: **Left**: AVicuna's four-stage fine-tuning aligns natural language with exact time segments in audio-visual videos, highlighting its adeptness in dynamic content analysis. **Right**: AVicuna's superior performance across various video and audio-visual understanding tasks compared to other models.

Moving forward from the proposed dataset, we recognize the second challenge in audio-visual learning: accurately modeling the temporal dynamics in untrimmed audio-visual content. Previous methods (Shu et al. 2023; Zhang et al. 2023b; Su et al. 2023; Lyu et al. 2023) have often combined embeddings from different modalities into single embeddings without adequately considering their intrinsic temporal relation. To tackle this critical issue, we introduce AVicuna, which comprises Multimodal Encoders, two Connective Adapters, an Audio-Visual Token Interleaver (AVTI), and an LLM. Multimodal Encoders extract embeddings from vision and audio modalities, which are aligned with the LLM's token space through Connective Adapters. The AVTI orchestrates the temporal relation of tokens from audio and video by creating interleaved audio-visual token sequences as inputs for the LLM. We employ a multi-stage fine-tuning approach to enhance AVicuna's capabilities, focusing on four critical stages: Vision-Text Alignment, Audio-Text Alignment, Time-Event Alignment, and Instruction Tuning. To foster effective alignment between multimodal tokens and LLM's token space, we have also aggregated several audio datasets, including AudioSet (Gemmeke et al. 2017), AudioCap (Kim et al. 2019), and Auto-ACD (Sun et al. 2023), to form a comprehensive audio-text dataset with 222K pairs, termed A5-222K (**A**udio-text **A**lignment with **A**udioSet, **A**udioCap, and **A**uto-CAD).

Our experiments demonstrate that the AVicuna fine-tuned on PU-VALOR achieves outstanding performance in both coarse-grained QA tasks and fine-grained temporal understanding tasks, as Figure 1 shown. It surpasses most LLM-based video understanding models and sets a new benchmark in the Audio-Visual Event Dense Localization (AVEDL) task.

In summary, our contributions are three-fold:

- We propose a novel approach to synthesize pseudo-untrimmed audio-visual videos and corresponding temporal boundary annotations using high-quality captions from the VALOR dataset, resulting in the PU-VALOR dataset.
- We introduce AVicuna, an audio-visual LLM with an Audio-Visual Token Interleaver and Time-Event Alignment Tuning on the PU-VALOR dataset, which achieves temporal synchronism and fine-grained understanding in audio-visual videos.
- Our experiments demonstrate that AVicuna significantly advances the state-of-the-art in the AVEDL task and exhibits strong performance in both coarse-grained QA and fine-grained temporal understanding tasks.

Related Work

Untrimmed Video Understanding. Temporal localization is key for understanding long-form videos by linking specific segments to their semantics. Key tasks include temporal video grounding (Luo et al. 2023a; Wang et al. 2022b), dense video captioning (Wang et al. 2021a; Yang et al. 2023), and video highlight detection (Lei et al. 2021; Jiang et al. 2023). However, action localization and highlight detection models often rely on predefined labels (Zhang et al. 2022a), limiting the scope. Recent work on event boundary detection (Shou et al. 2021; Wang et al. 2022a; Tang et al. 2023) and dense captioning (Krishna et al. 2017) aims to address these constraints using event description datasets (Lin et al. 2023b). Despite this, most models still rely on regression for temporal predictions, requiring extra heads for captioning and regression (Wang et al. 2021b; Zhang et al. 2022b; Tang et al. 2022). Recent LLM advancements offer a shift by using natural language to directly specify temporal locations, offering a more intuitive approach.

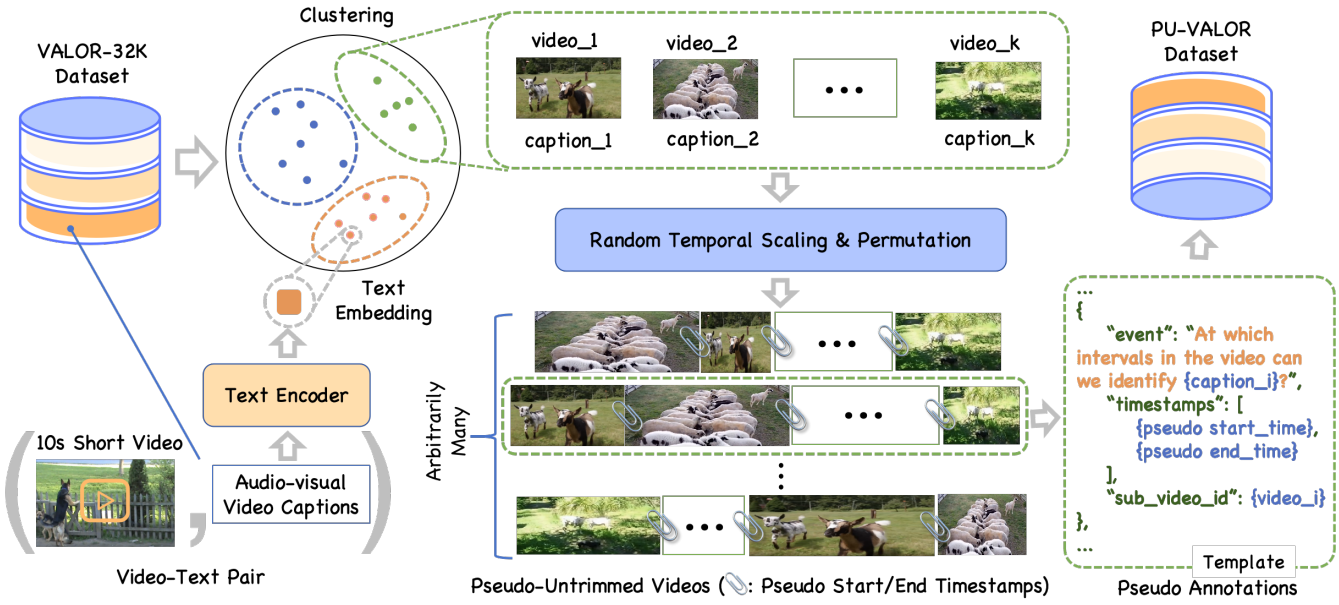

Figure 2: Pipeline for creating the PU-VALOR dataset, which involves extracting text embeddings from high-quality audio-visual captions of the original trimmed VALOR-32K dataset, clustering these embeddings, and then applying Random Temporal Scaling & Permutation to generate pseudo-untrimmed videos. These synthesized videos are then annotated with temporal boundaries using a template-based approach to facilitate the following audio-visual time-event alignment.

Multimodal LLMs. The development of Multimodal LLMs (MLLMs) has been driven by advancements in LLMs, enabling the integration of multimodal inputs (Alayrac et al. 2022; Liu et al. 2023; Hu et al. 2022b; Wang et al. 2023b; Hua et al. 2024a; Bi et al. 2023, 2024; Tang et al. 2024a). Flamingo (Alayrac et al. 2022) utilizes visual in-context learning for visual question answering, while LLaVA (Liu et al. 2023) introduces visual instruction tuning to enhance visual understanding. Models like VisionLLM (Wang et al. 2023c), KOSMOS-2 (Peng et al. 2023), (Chen et al. 2023a), and Qwen-VL (Bai et al. 2023) further advance MLLMs with visual grounding. Recent advancements such as VideoChat (Li et al. 2023b), ChatVideo (Wang et al. 2023a), V2Xum-LLM (Hua et al. 2024b), Valley (Luo et al. 2023b), and VTimeLLM (Huang et al. 2024) extend this fine-grained understanding to dynamic video content using natural language to define temporal boundaries without special tokens. We integrate audio cues to offer a more comprehensive approach to audio-visual temporal understanding tasks.

Audio-Visual Video Datasets. Increasing attention is being directed toward LLMs that support audio-visual inputs, such as Video-LLaMA (Zhang et al. 2023b), PandaGPT (Su et al. 2023), Macaw-LLM (Lyu et al. 2023), and AV-LLM (Shu et al. 2023), which are trained on audio-visual video datasets to enhance understanding of audio-visual content. However, these models struggle with fine-grained understanding of long or untrimmed videos due to the lack of detailed annotations in existing datasets. The VALOR dataset (Chen et al. 2023b) offers high-quality audio-visual captions but consists of trimmed 10-second clips. Other datasets like VGG-Sound-AVEL100K (Zhou et al. 2023), AVVP (Tian et al. 2020), UnAV-100 (Geng et al. 2023), and LFAV (Hou et al. 2023) provide temporal annotations but lack rich captioning. While AVSD (Alamri et al. 2019) and MUSIC-AVQA (Li et al. 2022) offer quality question-answer pairs, their temporal questions lack precise timestamps. This gap in datasets limits the models' ability to learn the relationship between audio-visual context and temporal boundaries. Table 1 compares important features across different video datasets, including untrimmed videos, audio-visual modalities, captions, and timestamps.

Dataset	Un-trimmed	Audio-Visual	Captions	Time-stamps
ActivityNetCaps	✓	×	✓	✓
InternVid	✓	×	✓	✓
VGGSound-AVEL	×	✓	×	✓
AVVP	×	✓	×	✓
LFAV	✓	✓	×	✓
UnAV-100	✓	✓	×	✓
VALOR	×	✓	✓	×
PU-VALOR (ours)	✓	✓	✓	✓

Table 1: Comparison of datasets based on untrimmed videos, audio-visual modalities, captions, and timestamps, showcasing the full coverage of all attributes by our PU-VALOR dataset.

Methodology

PU-VALOR Dataset

One of the primary challenges in untrimmed audio-visual video understanding is the scarcity of datasets with fine-grained annotations for temporal audio-visual events. To tackle this issue, we propose a practical yet straightforward pipeline, as illustrated in Figure 2, to utilize the existing VALOR-32K audio-visual dataset (Chen et al. 2023b), which comprises exclusively trimmed videos. By synthesizing untrimmed videos with precise temporal labels, we have created the PU-VALOR dataset that enables LLMs to learn the alignment among temporal cues, audio-visual events, and text tokens.

Clustering Videos with Similar Events. When creating untrimmed videos from trimmed clips, it is crucial to maintain semantic coherence within the untrimmed video to ensure a natural flow of content. This means that transitions between different segments should not be abrupt or disjointed, as such sudden shifts can disrupt the viewer's understanding. To ensure the content within an untrimmed video is semantically related, we group similar video segments based on the semantic similarity of their captions for follow-up untrimmed video generation.

We utilize a text encoder E_{txt} to embed captions from the video-caption pairs $\{(v_1, c_1), (v_2, c_2), \ldots, (v_n, c_n)\}$ sourced from the VALOR-32K dataset. For each caption c_i, the embedding is given by: $\bar{c}_i = E_{\text{txt}}(c_i)$ Next, we apply a clustering algorithm to the set of embeddings $\{\bar{c}_1, \bar{c}_2, \ldots, \bar{c}_n\}$ to identify and group videos with similar events, resulting in clusters denoted as $K = \{k_1, k_2, \ldots, k_m\}$. From each video cluster k, we randomly select one and then identify the $m-1$ most similar clips within the same cluster. This selection is defined by:

$$S_k = \{v_i \mid v_i \in V_k, \text{ where } i \in \{1, 2, \ldots, m\}\}. \quad (1)$$

We repeat this process for each cluster until no more than m clips remain in any cluster. This process ensures that each set of m clips is closely related in content.

Random Temporal Scaling & Permutation. To ensure diverse temporal relationships between the clips, for each selected video $v_i \in S_k$, we randomly scale its duration within the range $[T_{\text{min}}, T_{\text{max}}]$ as

$$T_{v_i}^{\text{new}} = T_{v_i} \times \text{random}(T_{\text{min}},\ T_{\text{max}}). \quad (2)$$

Then, we shuffle the order of the selected videos S_k and concatenate the videos in S_k to form a new untrimmed video U_k:

$$U_k = v_{\pi(1)} || v_{\pi(2)} || \ldots || v_{\pi(m)}, \quad (3)$$

where π is a random permutation of the indices $\{1, 2, \ldots, m\}$.

Annotation. As we know the original duration T_{v_i} and the scaled duration $T_{v_i}^{\text{new}}$ for each video, we can map captions to specific temporal intervals in U_k. If the original duration for a caption in v_i is d_i, the new timestamp in U_k becomes:

$$[t_{\text{start},i}^{\text{new}},\ t_{\text{end},i}^{\text{new}}] = [T_{\text{offset,i}},\ T_{\text{offset,i}} + d_i \cdot \epsilon] \quad (4)$$

where T_{offset} is the cumulative duration of all preceding videos in U_k after scaling, and the scaling factor $\epsilon = T_{v_i}^{\text{new}} / T_{v_i}$. We annotate temporal intervals $[t_{\text{start},i}^{\text{new}}, t_{\text{end},i}^{\text{new}}]$ within U_k to correspond with the content described by the captions c_i.

AVicuna Model

Overview. Figure 3 illustrates AVicuna's architecture, comprising Multimodal Encoders, Connective Adapters, an Audio-Visual Token Interleaver (AVTI), and an LLM. The encoders extract embeddings aligned to the LLM's token space, and the AVTI interleaves them, with an Audio-Interleaving Rate (AIR) enhancing temporal synchronism.

Multimodal Encoders. Multimodal Encoder includes Vision Encoder and Audio Encoder. For visual input, we employ the CLIP ViT-14/L (Radford et al. 2021) as Vision Encoder to extract visual embeddings $F = \{f_i\}_{i=1}^{M}$, where M denotes the number of visual embeddings. When the input is image, $M = 1$. For audio input, we utilize CLAP (Elizalde et al. 2023) as Audio Encoder to obtain audio embeddings $A = \{a_i\}_{i=1}^{N}$, with N representing the number of audio embeddings.

Connective Adapters. To avoid interference between the different modalities during the alignment, we adopt two MLPs as Vision Adapter and Audio Adapter for visual embeddings and audio embeddings, respectively, to get visual tokens $\bar{F} = \{\bar{f}_i\}_{i=1}^{M}$ and audio tokens $\bar{A} = \{\bar{a}_i\}_{i=1}^{N}$ that are aligned with LLM's token space.

Audio-Visual Tokens Interleaver. Unlike simply adding positional embeddings to audio and video embeddings in previous work (Lyu et al. 2023; Shu et al. 2023), the Audio-Visual Token Interleaver (AVTI) rearranges video and audio embeddings without altering their sequence order, keeping the overall token length constant. The audio-interleaving rate (AIR), denoted as $\rho \in [0, 1]$, controls the ratio of video to audio tokens. Audio tokens and video tokens are interpolated or downsampled to $\tilde{A} = \{\tilde{a}_i\}_{i=1}^{\tilde{N}}$ and $\tilde{F} = \{\tilde{f}_i\}_{i=1}^{\tilde{M}}$, where $\tilde{N} = \rho T$ and $\tilde{M} = (1-\rho)T$, with T being the total sequence length generated by AVTI. As shown in Figure 3, AVTI systematically interleaves audio and video tokens, preserving their original order to maintain temporal alignment. The output, represented by the audio-visual context $\Psi = \{\psi_t\}_{t=1}^{T}$, is given by:

$$\psi_t = \begin{cases} \tilde{a}_{\lceil t/(\omega_\rho+1) \rceil} & \text{if } t \bmod (\omega_\rho + 1) \equiv 0, \\ \tilde{f}_{\lceil t/(\omega_\rho+1) \rceil} & \text{otherwise,} \end{cases} \quad (5)$$

where $\omega_\rho = \lfloor \frac{1-\rho}{\rho} \rfloor$. Each context token represents both the audio-visual content and its temporal position within the video.

Large Language Model. We use the fine-tuned Vicuna-7B-v1.5 (Touvron et al. 2023) as our LLM to process interleaved audio-visual tokens and user queries Q, generating responses R:

$$R = \text{LLM}(I, Q), \quad (6)$$

where $I \in \{\bar{F}, \bar{A}, \Psi\}$ represents vision, audio, or both.

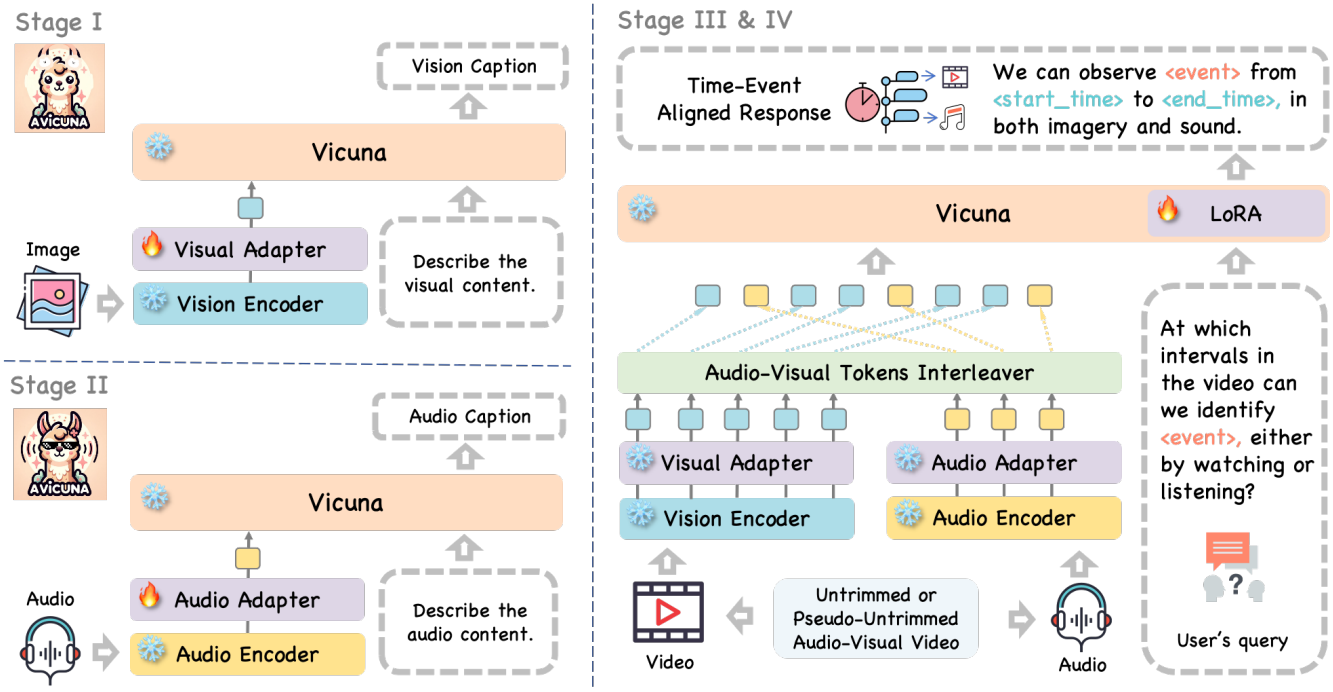

Figure 3: AVicuna model architecture and fine-tuning process. Vision and Audio Adapters are MLPs that align modalities with LLM. The Audio-Visual Tokens Interleaver ensures temporal synchronization. LoRA fine-tuning aligns temporal boundaries with events and enhances instruction-following capabilities.

Multi-stage Fine-tuning

In AVicuna training, aligning embeddings from other modalities to the LLM's token space is preferred, using Video-Text and Audio-Text Alignment. To enable specific temporal patterns, we fine-tune LoRA (Hu et al. 2022a) during Time-Event Alignment. Recent studies (Huang et al. 2024) indicate this may affect question-answering, mitigated by Instruction Tuning. We thus employ a four-stage fine-tuning process, with datasets detailed in our technical appendices (Tang et al. 2024b).

Stage I & II: Multimodal-Text Alignment. In Vision-Text Alignment, we fix the Vision Encoder and LLM, updating only the Vision Adapter, using LCS-558K (Liu et al. 2023), a 558K image-text pair subset from LAION-CC-SBU with BLIP (Li et al. 2023a) captions. Here, $I = \bar{F}$, Q is a visual content query, and R is the image caption. In Audio-Text Alignment, we fix the Audio Encoder and LLM, updating only the Audio Adapter, using A5-222K, a 222K audio-text dataset compiled from AudioSet (Gemmeke et al. 2017), AudioCap (Kim et al. 2019), and Auto-ACD (Sun et al. 2023). Here, $I = \bar{A}$, Q is an audio content query, and R is the audio caption.

Stage III: Time-Event Alignment. We freeze the fine-tuned Audio and Vision Adapters, updating only the LoRA (Hu et al. 2022a) parameters in the LLM. We create Q-R pairs with time-related information from the PU-VALOR datasets, including (1) Time-referenced Query, Time-agnostic Response: $\{(Q, R)|\tau \sqsubseteq Q, \tau \not\sqsubseteq R)\}$, and (2) Time-agnostic Query, Time-referenced Response: $\{(Q, R)|\tau \not\sqsubseteq Q, \tau \sqsubseteq R)\}$, where τ :="*from* τ_s *to* τ_e" and τ_s, τ_e are event time points. Inputs (I) can be visual, audio, or combined; we use the InternVid (Wang et al. 2023d) dataset to enrich visual event alignment training.

Stage IV: Instruction Tuning. Finally, we fine-tune AVicuna on instruction-following datasets, including UnAV-100 (Geng et al. 2023) for event localization, and other instruction datasets such as VideoInstruct100K (Maaz et al. 2023), ActivityNet Captions (Krishna et al. 2017), and DiDeMo (Anne Hendricks et al. 2017), to improve question-answering and mitigate previous tuning effects (Huang et al. 2024). At this stage, I is also Ψ or $\bar{F}$, Q is a general instruction, and R is the corresponding response.

Experimental Results

Experiment Setups

Metrics. We evaluate temporal understanding using tasks across various domains: Video Question Answering (Video QA), Audio-visual Video Question Answering (AVQA), and Audio-Visual Event Dense Localization (AVEDL). For General Video QA, zero-shot evaluation is performed on the MSVD-QA (Chen and Dolan 2011), MSRVTT-QA (Xu et al. 2016), and ActivityNet-QA (Yu et al. 2019) datasets, with open-ended QA tasks evaluated using GPT scoring (Maaz et al. 2023). AVQA tasks are assessed on

Method	A&V	TU	#Pairs	LLM-size	AVSD	MUSIC-QA	MSVD-QA	MSRVTT-QA	ActivityNet-QA
Valley	×	×	1.5M	13B	-	-	65.4	45.7	26.5
VideoChat	×	✓	25M	7B	-	-	56.3	45.0	26.5
Video-ChatGPT	×	✓	0.9M	7B	-	-	64.9	49.3	35.2
VTimeLLM	×	✓	0.7M	7B	-	-	69.8	58.8	45.5
MA-LLM	×	✓	-	7B	-	-	60.6	48.5	49.8
PandaGPT	✓	×	128M	13B	26.1	33.7	46.7	23.7	11.2
Macaw-LLM	✓	×	0.3M	7B	34.3	31.8	42.1	25.5	14.5
AV-LLM	✓	×	1.6M	13B	52.6	45.2	67.3	53.7	47.2
Video-LLaMA	✓	✓	2.8M	7B	36.7	36.6	51.6	29.6	12.4
AVicuna (ours)	✓	✓	1.1M	7B	**53.1**	**49.6**	**70.2**	**59.7**	**53.0**

Table 2: Comparison with existing LLM-based methods on open-ended video QA (MSVD-QA, MSRVTT-QA, ActivityNet-QA) and AVQA (AVSD, MUSIC-AVQA) benchmarks. **A&V**: the model supports both video and audio input. **TU**: the model can perform temporal understanding task, *e.g.,* temporal grounding and localization. **#Pairs**: the instruction-response pairs for instruction tuning. **LLM-size**: the number of parameters in the LLM adopted.

the AVSD (Alamri et al. 2019) and MUSIC-AVQA (Li et al. 2022) datasets. The AVEDL task uses the UnAV-100 (Geng et al. 2023) dataset, with performance measured by mean Average Precision (mAP) at Intersection over Union (tIoU) thresholds [0.5:0.1:0.9] and the average mAP across [0.1:0.1:0.9].

Baseline Models. For AVQA tasks, we evaluate LLM-based models, including PandaGPT (Su et al. 2023), Macaw-LLM (Lyu et al. 2023), and AV-LLM (Shu et al. 2023), which support audio-visual input. For the video QA task, except for the models mentioned above, we also compare with Vally (Luo et al. 2023b), VideoChat (Li et al. 2023b), VTimeLLM (Huang et al. 2024), and MA-LLM (He et al. 2024). For AVEDL tasks, we include non-LLM baselines such as VSGN (Zhao et al. 2021), TadTR (Liu et al. 2022), ActionFormer (Zhang et al. 2022a), UnAV (Geng et al. 2023), and UniAV-AT/ST (Geng et al. 2024).

Implementation Details. We uniformly extract a minimum of 100 frames from each video to create an interleaved sequence of audio-visual tokens. More details are provided in our technical appendices (Tang et al. 2024b).

Comparison Experiments

General Video QA and AVQA. The video QA and AVQA comparison results are shown as Table 2. AVicuna supports both audio and video as input and handles temporal understanding tasks. Despite being fine-tuned with only 1.1M pairs and utilizing an LLM with 7B parameters, AVicuna surpasses all other LLM-based models on both video QA and AVQA benchmarks.

Audio-Visual Event Localization. The comparison results on the AVEDL tasks can be found in Table 3. In the AVEDL task, AVicuna's superior mAP scores, particularly at the IoU threshold of 0.5 through 0.9, indicate its enhanced precision in localizing events within a video. The results are impressive, considering they outperform other LLM-based models and specialized non-LLM methods like VSGN, TadTR, ActionFormer, and UnAV. This suggests that the AVicuna model has effectively leveraged its audio-visual capabilities to provide a more nuanced understanding of the temporal aspects of videos. We also conducted the video temporal grounding (VTG) task on the ActivityNet Captions dataset, which can be found in our technical appendices (Tang et al. 2024b).

Method	0.5	0.6	0.7	0.8	0.9	Avg.
VSGN	24.5	20.2	15.9	11.4	6.8	24.1
TadTR	30.4	27.1	23.3	19.4	14.3	29.4
ActionFormer	43.5	39.4	33.4	27.3	17.9	42.2
UnAV	50.6	45.8	39.8	32.4	21.1	47.8
UniAV-AT	54.1	48.6	42.1	34.3	20.5	50.7
UniAV-ST	54.8	49.4	43.2	35.3	22.5	51.7
AVicuna (ours)	**60.0**	**50.4**	**49.6**	**43.5**	**36.5**	**60.3**

Table 3: Comparison of the results on the UnAV-100 for the AVEDL task.

Setting	0.5	0.6	0.7	0.8	0.9	Avg.
AVicuna	**60.0**	**54.4**	**49.6**	**43.5**	**37.1**	**60.3**
w/o PU-VALOR	19.5	14.3	10.2	6.8	4.5	27.9
w/o AVTI	50.1	45.2	40.2	34.2	29.4	51.1
w/o A5-222K	22.2	16.5	11.4	6.8	2.7	30.1
w/o Audio	29.0	23.9	18.8	13.6	8.8	35.8

Table 4: Ablation study on the dataset and model components, which lead to decreases in mAP.

Ablation Study

We conduct ablation studies as shown in Table 4 to assess the impact of different components, datasets, and modalities on AVicuna's performance. Each row represents an independent experiment where a specific component or dataset is removed. Specifically, omitting the PU-VALOR or A5-222K datasets, especially PU-VALOR, leads to significant

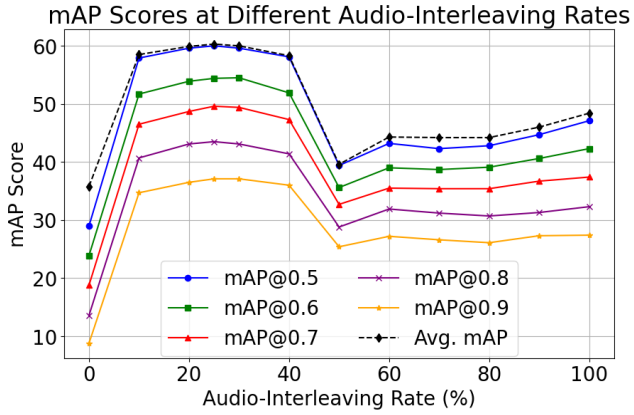

Figure 4: AVicuna's performances on UnAV-100 measured by mAP scores at different AIRs.

performance drops, emphasizing their critical roles in Time-Event Alignment and Audio-Text Alignment, respectively. Removing the AVTI module also results in a decrease in performance, further validating its necessity. Additionally, excluding audio inputs causes a marked reduction in accuracy, underscoring the value of a multimodal approach.

Audio-Interleaving Rates Analysis

The ablation study further sheds light on the contribution of different components to the model's performance. We conduct ablation studies on the Audio-Interleaving Rates (AIR), denoted by ρ. Figure 4 shows AVicuna's performances on UnAV-100 datset for AVEDL task. The scores generally increase with the AIR up to 25-30%, indicating optimal audio-visual integration, and then decline, suggesting that excessive audio information may be detrimental. The performance gap between different IoU thresholds narrows at higher AIRs. The highest mAP is achieved at 25% AIR across all thresholds, indicating an optimal balance between audio and video information. The complete results are shown in our technical appendices (Tang et al. 2024b).

Qualitative Analysis

Figure 5 presents three examples of AVicuna's audio-visual dialogue capabilities. In the first video, involving a Jeep with persistent engine noise, AVicuna accurately identifies the time intervals where the engine is knocking. The second video shows people live-streaming while eating pomegranates; AVicuna correctly skips the intro, focuses on the specified time interval, and describes the chewing sounds. In the third video, where two people play silent musical instruments, AVicuna accurately responds with "NO", showcasing its ability to provide correct negative responses—a common challenge for existing models. These examples demonstrate AVicuna's flexibility in handling videos of varying lengths and resolutions. More examples can be found in our technical appendices (Tang et al. 2024b).

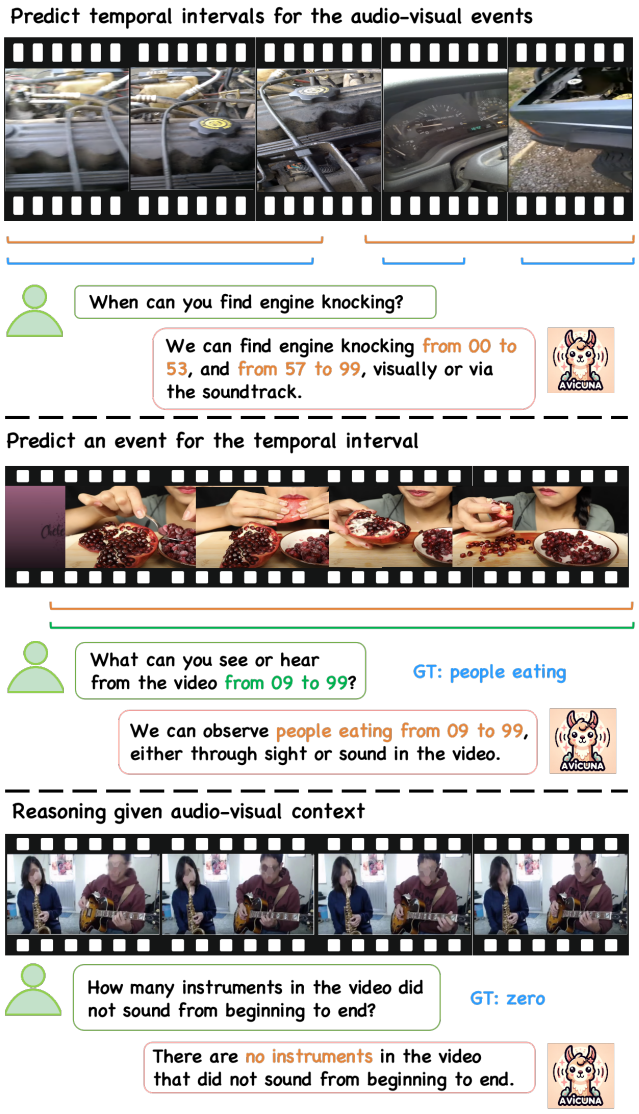

Figure 5: Qualitative results. Blue indicates ground-truth, green indicates the time intervals the user gives, and orange represents the model predictions.

Conclusion

We present a comprehensive approach to enhancing audio-visual temporal understanding in untrimmed videos using MLLMs. We introduced a novel method to construct a pseudo-untrimmed dataset, PU-VALOR, which provides pseudo-untrimmed audio-visual videos with accurate temporal boundary annotations, addressing the scarcity of datasets in the audio-visual domain. Furthermore, we developed AVicuna, an audio-visual LLM incorporating an AVTI and Time-Event Alignment to achieve fine-grained understanding and temporal synchronism in audio-visual videos. Our experiments demonstrate that AVicuna achieves state-of-the-art performance in various video and audio-visual understanding tasks, supporting both coarse-grained QA and fine-grained temporal understanding.

Acknowledgements

This work was supported by Sony Group Corporation. We would like to thank Sayaka Nakamura and Jerry Jun Yokono for insightful discussion.

References

Alamri, H.; et al. 2019. Audio Visual Scene-Aware Dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alayrac, J.-B.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35: 23716–23736.

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

Bai, J.; et al. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv:2308.12966.

Bi, J.; Tang, Y.; Song, L.; Vosoughi, A.; Nguyen, N.; and Xu, C. 2024. EAGLE: Egocentric AGgregated Language-video Engine. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 1682–1691. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706868.

Bi, J.; et al. 2023. MISAR: A Multimodal Instructional System with Augmented Reality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1–5.

Chen, D.; and Dolan, W. 2011. Collecting Highly Parallel Data for Paraphrase Evaluation. In Lin, D.; Matsumoto, Y.; and Mihalcea, R., eds., *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 190–200. Portland, Oregon, USA: Association for Computational Linguistics.

Chen, K.; et al. 2023a. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. arXiv:2306.15195.

Chen, S.; et al. 2023b. Valor: Vision-audio-language omni-perception pretraining model and dataset. arXiv:2304.08345.

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

Gemmeke, J. F.; et al. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780.

Geng, T.; et al. 2023. 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 (CVPR)*, 22942–22951.

Geng, T.; et al. 2024. UniAV: Unified Audio-Visual Perception for Multi-Task Video Localization. arXiv:2404.03179.

He, B.; et al. 2024. 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*, 13504–13514.

Hou, W.; et al. 2023. Towards Long Form Audio-visual Video Understanding. arXiv:2306.09431.

Hu, E. J.; et al. 2022a. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Hu, Y.; et al. 2022b. Promptcap: Prompt-guided task-aware image captioning. arXiv:2211.09699.

Hua, H.; et al. 2024a. FINEMATCH: Aspect-based Fine-grained Image and Text Mismatch Detection and Correction. arXiv:2404.14715.

Hua, H.; et al. 2024b. V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning. arXiv:2404.12353.

Huang, B.; et al. 2024. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14271–14280.

Jiang, H.; et al. 2023. Single-stage visual query localization in egocentric videos. *Advances in Neural Information Processing Systems*, 36.

Kim, C. D.; Kim, B.; Lee, H.; and Kim, G. 2019. AudioCaps: Generating Captions for Audios in The Wild. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 119–132. Minneapolis, Minnesota: Association for Computational Linguistics.

Krishna, R.; et al. 2017. Dense-Captioning Events in Videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Lei, J.; et al. 2021. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858.

Li, G.; et al. 2022. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19108–19118.

Li, J.; et al. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv:2301.12597.

Li, K.; et al. 2023b. Videochat: Chat-centric video understanding. arXiv:2305.06355.

Lin, B.; et al. 2023a. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. arXiv:2311.10122.

Lin, J.; et al. 2023b. VideoXum: Cross-modal Visual and Textural Summarization of Videos. arXiv:2303.12060.

Liu, H.; et al. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, X.; et al. 2022. End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing*, 31: 5427–5441.

Luo, D.; et al. 2023a. Towards Generalisable Video Moment Retrieval: Visual-Dynamic Injection to Image-Text

Pre-Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23045–23055.

Luo, R.; et al. 2023b. Valley: Video Assistant with Large Language model Enhanced abilitY. arXiv:2306.07207.

Lyu, C.; et al. 2023. Macaw-LLM: Multi-Modal Language Modeling with Image, Audio, Video, and Text Integration. arXiv:2306.09093.

Maaz, M.; et al. 2023. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. arXiv:2306.05424.

Peng, Z.; et al. 2023. Kosmos-2: Grounding Multimodal Large Language Models to the World. arXiv:2306.14824.

Radford, A.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Shou, M. Z.; et al. 2021. Generic event boundary detection: A benchmark for event segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8075–8084.

Shu, F.; et al. 2023. Audio-Visual LLM for Video Understanding. arXiv:2312.06720.

Su, Y.; et al. 2023. Pandagpt: One model to instruction-follow them all. arXiv:2305.16355.

Sun, L.; et al. 2023. A Large-scale Dataset for Audio-Language Representation Learning. arXiv:2309.11500.

Tang, Y.; Zhan, G.; Yang, L.; Liao, Y.; and Xu, C. 2024a. Cardiff: Video salient object ranking chain of thought reasoning for saliency prediction with diffusion. arXiv:2408.12009.

Tang, Y.; et al. 2022. Multi-modal Segment Assemblage Network for Ad Video Editing with Importance-Coherence Reward. In *Proceedings of the Asian Conference on Computer Vision*, 3519–3535.

Tang, Y.; et al. 2023. LLMVA-GEBC: Large Language Model with Video Adapter for Generic Event Boundary Captioning. arXiv:2306.10354.

Tang, Y.; et al. 2024b. Empowering LLMs with Pseudo-Untrimmed Videos for Audio-Visual Temporal Understanding. arXiv:2403.16276.

Tian, Y.; et al. 2020. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 436–454. Springer.

Touvron, H.; et al. 2023. Llama: Open and efficient foundation language models. arXiv:2302.13971.

Wang, J.; et al. 2023a. Chatvideo: A tracklet-centric multimodal and versatile video understanding system. arXiv:2304.14407.

Wang, T.; et al. 2021a. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6847–6857.

Wang, T.; et al. 2021b. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6847–6857.

Wang, T.; et al. 2023b. Caption anything: Interactive image description with diverse multimodal controls. arXiv:2305.02677.

Wang, W.; et al. 2023c. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36.

Wang, Y.; et al. 2022a. Geb+: A benchmark for generic event boundary captioning, grounding and retrieval. In *European Conference on Computer Vision*, 709–725. Springer.

Wang, Y.; et al. 2023d. Internvid: A large-scale video-text dataset for multimodal understanding and generation. arXiv:2307.06942.

Wang, Z.; et al. 2022b. Negative sample matters: A renaissance of metric learning for temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2613–2623.

Xu, J.; et al. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu, S.; et al. 2023. Launchpadgpt: Language model as music visualization designer on launchpad. arXiv:2307.04827.

Xuan, S.; et al. 2023. Pink: Unveiling the Power of Referential Comprehension for Multi-modal LLMs. arXiv:2310.00582.

Yang, A.; et al. 2023. 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*, 10714–10726.

Yu, Z.; et al. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9127–9134.

Zhang, C.; et al. 2022a. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, 492–510. Springer.

Zhang, D.; et al. 2023a. DNAGPT: A Generalized Pre-trained Tool for Multiple DNA Sequence Analysis Tasks. bioRxiv:2023–07.

Zhang, D.; et al. 2024. CoCoT: Contrastive Chain-of-Thought Prompting for Large Multimodal Models with Multiple Image Inputs. arXiv:2401.02582.

Zhang, H.; et al. 2023b. Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv:2306.02858.

Zhang, J.; et al. 2022b. Exploiting Context Information for Generic Event Boundary Captioning. arXiv:2207.01050.

Zhao, C.; et al. 2021. Video self-stitching graph network for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13658–13667.

Zhou, J.; et al. 2023. Contrastive Positive Sample Propagation Along the Audio-Visual Event Line. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 7239–7257.