LongVALE: Vision-Audio-Language-Event Benchmark Towards Time-Aware Omni-Modal Perception of Long Videos

Supplementary Material

1. More Details of LongVALE Benchmark

1.1. Quantitative Analysis of Event Boundaries

To quantitatively verify the semantic coherence of segmented events of different modalities, we introduce Max Running Semantic Difference (MRSD), inspired by [2]. For a n-second event clip, we compute the embedding for each second as $\{f_1, \ldots, f_n\}$, and get the most significant semantic change within the clip, denoted as:

$$\max(\{\text{Diff}(f_i, f_{i+1}) | i \in [1, n-1]\}). \quad (1)$$

We apply ImageBind [5] and CLAP [13] to extract embeddings for visual and audio clips, respectively. As in Tab. 1, for single-modal events, the clips after the second stitching stage effectively avoid being overly fragmentary while maintaining strong semantic coherence. Further, although semantic shifts may occur between single-modal events within an omni-modal event, no event is truncated, ensuring the semantic integrity of all events from various modalities.

Method	MRSD-V↓	MRSD-A↓	Avg.len
Visual event boundary (splitting)	0.531	-	3.0s
Visual event boundary (stitching)	0.532	-	10.7s
Audio event boundary (splitting)	-	0.676	1.5s
Audio event boundary (stitching)	-	0.703	5.8s
Omni-modal event boundary	0.601	0.784	16.7s

Table 1. Semantic coherence and event length analysis. We randomly sample 1K long videos in our LongVALE.

1.2. More Statistics

Based on YouTube metadata, we further analyze the distribution of video categories, as shown in Fig. 1. It reflects that our LongVALE covers a wide range of video topics. Besides, since our focus is on long-form videos with rich, event-driven storylines, the diversity of their content cannot be easily summarized by just a few simple categories. Moreover, as shown in Fig. 2, we also illustrate the distribution of the lengths of our omni-modal event captions and visualize their word cloud to highlight the rich omni-modality content within the captions.

1.3. Manual Check and Correction

During the manual check process, annotators are asked to check each omni-modal event and verify whether the caption and the corresponding temporal boundaries are accurate. Besides, videos containing only monotonous background music and speech are filtered out to ensure the dataset includes rich sound types. Afterward, during the manual correction process, another group of annotators correct all inaccurate annotations and submit the revised versions. Totally, we checked 2K videos with each taking 3 minutes, and corrected about 300 errors, totally 115 human hours. We show the interfaces in Fig. 3

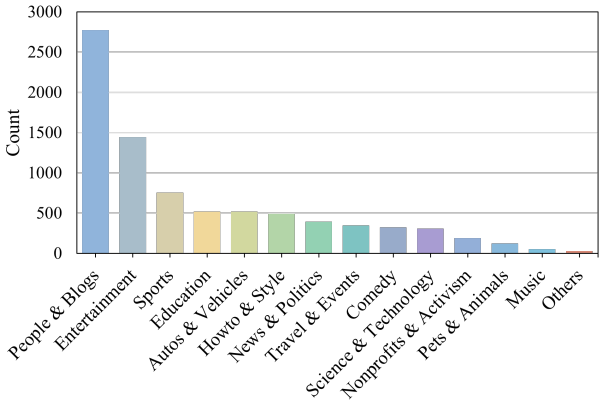
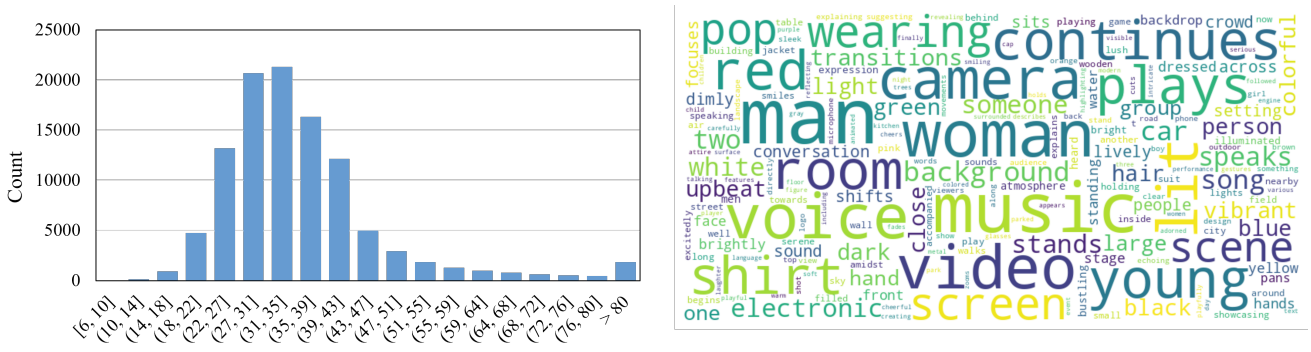


Figure 1. Distribution of video categories of LongVALE dataset.

1.4. Captioning and AV correlation Prompts

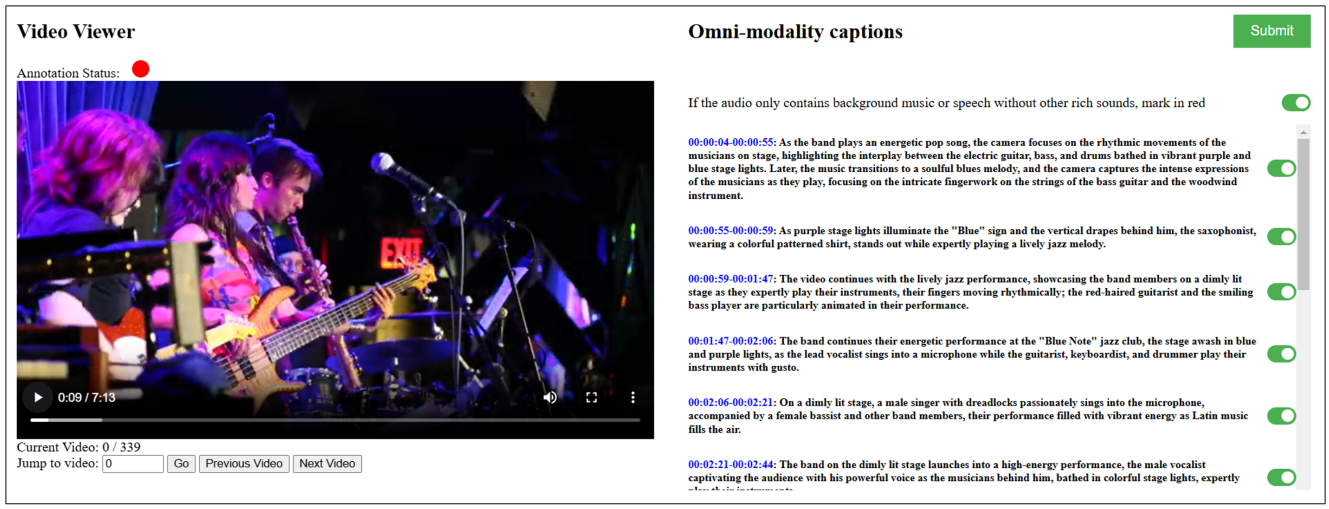
In Sec.3.3, for each segmented video clip, we apply LLaVA-NeXT-Video (34B) [14] to generate a video caption emphasizing dynamic information and apply GPT-4o [10] to generate keyframe caption emphasizing spatial details. For each segmented audio clip, we apply Qwen-Audio-Chat (7B) [4] to generate an audio caption, and utilize Whisper-Large-V3 [12] to get accurate subtitles. Note that we found that the performance of the audio captioner lags significantly behind that of visual models, leading to more hallucination issues, such as generating repetitive sentences or incorrect ASR. To address this, we cleaned up these generations, retaining only general descriptions for each audio event (*e.g.*, "this is a man speaking") while removing the specific speech content. Accurate ASR outputs generated by the advanced speech recognition model [12] were used as replacements. After obtaining modality-specific captions, we instruct Gemini-1.5-Pro [6] to integrate and correlate them explicitly. The detailed prompts are shown in Fig. 4. In Sec.3.5, we quantitatively identify the characteristics of our omni-modal event captions, including audio-visual correlations and fine-grained temporal dynamics us-



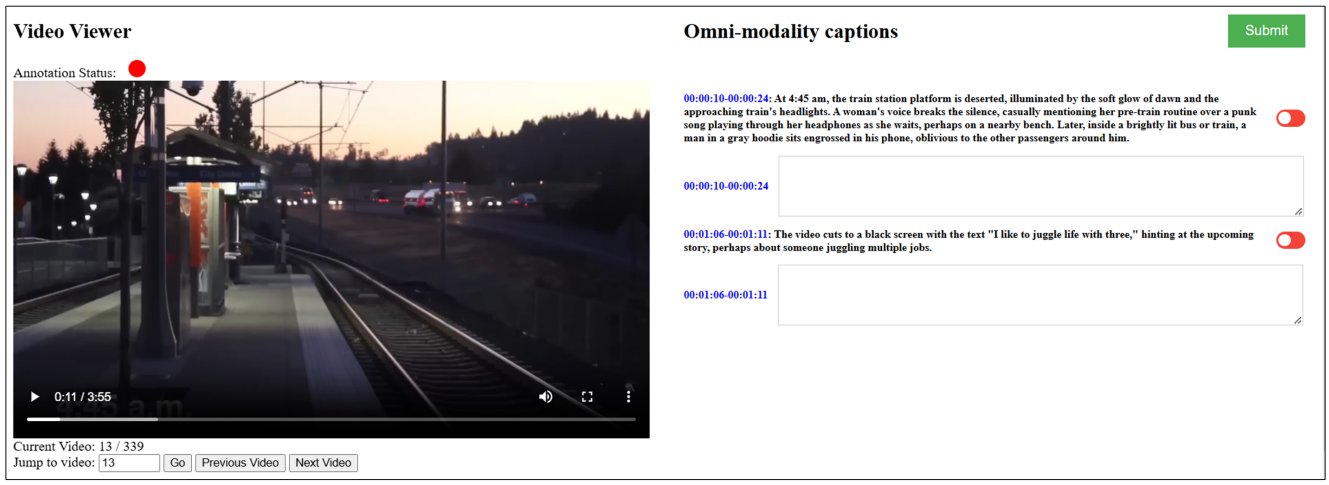
(a) Distribution of omni-modality caption length (# words)

(b) Word cloud of omni-modality captions

Figure 2. Distribution of omni-modality caption length and word cloud.



(a) Screenshot of the manual check interface



(b) Screenshot of the manual correction interface

Figure 3. Screenshots of our manual check and correction interfaces.

ing Gemini-1.5-Pro [6]. Here, we provide the detailed prompt as shown in Fig 5.

2. Task, Model and Training Data Details

2.1. Detailed Task Definition

We extend three fine-grained video tasks to the novel omni-modality domain towards omni-perception of long videos. These tasks emphasize cross-modal reasoning and fine-grained temporal understanding at the same time. Here, we provide detailed definitions for these tasks.

Omni-modal temporal video grounding. Given a textual query describing a specific omni-modal event, the model is required to identify the start and end timestamps of the corresponding video segment.

Omni-modal dense video captioning. The task is more intricate, requiring the model to perform both temporal localization and captioning for all omni-modal events occurring in a given untrimmed video.

Omni-modal segment captioning. Given a temporal boundary, the task demands the model to generate a caption summarizing the content of the corresponding omni-modal event within the untrimmed video.

2.2. Detailed Model Architecture

Multimodal encoders. Given a video, we utilize a frozen CLIP ViT-L/14 [11] as the Visual Encoder to extract visual embeddings $F_V = \{v_i\}_{i=1}^{N_v}$, where N_v denotes the number of input video frames. Since both non-speech audio (*i.e.*, natural sound and music) and speech provide crucial information for multi-modal video understanding, we employ BEATs [1] and Whisper [12] to extract non-speech audio embeddings $F_A = \{a_i\}_{i=1}^{N_a}$ and speech embeddings $F_S = \{s_i\}_{i=1}^{N_s}$, where N_a and N_s represent the number of audio and speech embeddings, respectively. Therefore, the resulting auditory features of these two encoders are complementary and suitable for general audio input.

Multimodal adapters. We apply linear layers to project the obtained embeddings from different modalities to get visual tokens $\hat{F_V} = \{\hat{v_i}\}_{i=1}^{N_v}$, audio tokens $\hat{F_A} = \{\hat{a_i}\}_{i=1}^{N_a}$, and speech tokens $\hat{F_S} = \{\hat{s_i}\}_{i=1}^{N_s}$ that are aligned with LLM's token space. Subsequently, the obtained token sequences are simply concatenated as:

$$\mathbf{Z} = \text{Concat}(\hat{F_V}, \hat{F_A}, \hat{F_s}), \tag{2}$$

where $\mathbf{Z} \in \mathbb{R}^{N \times d}$, $N = N_v + N_a + N_s$, and d is the hidden dimension of LLM. Note that our model also supports single-modal and dual-modal inputs, allowing for flexible handling of video data with missing modalities.

Large language model. We use Vicuna-7B-v1.5 [3] as our LLM to process concatenated multi-modal tokens $\mathbf{Z}$ and user queries for response generation.

2.3. Training Data Details

For boundary perception, we adopted the same template-based data generation strategy as [7] with the same templates, where 20% of the data is randomly sampled to generate single-turn dialogues (omni-modal dense video captioning), and 80% is used to generate multi-turn dialogues, *i.e.*, each event is randomly assigned to one of the two tasks (omni-modal temporal video grounding and segment captioning). For instruction tuning, the prompt used to generate omni-modality dialogues is shown in Fig. 6.

3. Experimental Details

3.1. More Implementation Details

We train our model for 2 epochs with a batch size of 128 throughout the two training stages. The AdamW [9] optimizer is applied with a cosine learning rate decay and a warm-up period. The learning rate is 1×10^{-4}. The rank in LoRA is 64 with $alpha = 128$. Following [7], we merge the LoRA module trained in the boundary perception stage with the LLM parameters, and then additionally incorporate a new LoRA module for instruction tuning. This ensures the temporal understanding capabilities acquired during the boundary perception stage are effectively preserved within the model. We complete the training of our 7B model within 30 hours with 1 RTX-A100 (40G) GPU.

3.2. Evaluation Details

Evaluation of our LongVALE-LLM. For LongVALE-LLM that only undergoes boundary perception tuning without instruction tuning, we directly use the templates as queries. Specifically, for the omni-modal dense captioning task, we employ "*Could you please detail the events that took place during different time segments in the video?*" as the query. For the omni-modal temporal grounding task, we employ "*During which frames does* $<$ *event* $>$ *occur in the video?*" as the query. For the omni-modal segment captioning task, we employ "*Could you tell me what happened from* $<$ *start* $>$ *to* $<$ *end* $>$ *in the video?*" as the query. LongVALE-LLM that undergoes instruction tuning demonstrates strong instruction-following ability. For omni-modal dense captioning, we utilize the following query: "*Could you please detail the events that took place during different time segments in the video? List the events in the format: From xx to xx, event1. From xx to xx, event2...*". For the omni-modal temporal grounding task, we employ the query "*During which frames does* $<$ *event* $>$ *occur in the video? Give the timestamps in the format: From xx to xx.*" or the omni-modal segment captioning task, we employ the query "*Can you describe what occurred from* $<$ *start* $>$ *to* $<$ *end* $>$ *in the video? Please give the event description directly.*". We also adopt other similar queries such as "*Provide details about the events from* $<$ *start* $>$ *to* $<$ *end* $>$

in the video...", the results remain consistently close.

Evaluation of other video LLMs. For other Video LLMs including VideoLLaMA, PandaGPT, NExT-GPT, VideoChat, Video-ChatGPT, TimeChat, and VTimeLLM, we tried our best to assess their optimal performance, recognizing that some were not specifically trained for these tasks. For models that have been trained on tasks such as dense video captioning or grounding, we employ the queries provided in their original studies. For instance, for TimeChat, we use the original query for dense captioning: "*Localize a series of activity events in the video, output the start and end timestamp for each event, and describe each event with sentences. List the events in the format: From x1 second to y1 second: event1.*" Similarly, for temporal grounding, we use the query: "*Detect and report the start and end timestamps of the video segment that semantically matches the* {*sentence*}*. Give the timestamps in the format: From xx to xx.*" For segment captioning, we identified the most effective prompt to be the one described below.

For models such as VideoLLaMA, PandaGPT, and Video-ChatGPT without training for these tasks, we found that the most effective approach involved using queries that include the video duration. For dense captioning, the query, "*This video has a duration of D seconds. From which second to which second in the video, what event happens? List the events in the format: From x1 second to y1 second: event1...*" yielded the best results. For grounding, we found that the query, "*This video lasts for D seconds. During this time, at what specific time does the event* {*sentence*}*occur? Please provide the start and end timestamps in the format: From x seconds to y seconds, the event happens.*" produced optimal performance. Moreover, we used GPT-4o mini to efficiently extract timestamps from the generated responses. Additionally, for segment captioning, we observed that using "*This video has a total duration of D seconds. Please describe in detail what happens between* $< start >$ *and* $< end >$ *in the video. Be specific about the activities of individuals, the environment, and any interactions between people or objects.*" provided the clearest and most detailed outputs. After obtaining the output, we tried to apply multiple regular expressions to format the output. For those outputs cannot be processed, we exclude the corresponding data from metric calculations.

4. More Qualitative Results

As shown in Fig. 7-10, we present more qualitative results encompassing all evaluated tasks.

Omni-modal segment captioning. In Fig. 7, VTimeLLM provides only brief descriptions of visual events within the specified moment, whereas our model offers richer information on both dynamic and auditory events, delivering a more comprehensive and vivid account.

Omni-modal temporal video grounding. In Fig. 8, given an omni-modal event caption, our model can more accurately pinpoint the time interval when the event occurs, which fully demonstrates its fine-grained temporal understanding capability in an omni-modality domain.

Omni-modal dense video captioning. In Fig. 10, given a video, our model can identify more omni-modal events and provide finer-grained descriptions, including key information from both visual and audio modalities, enabling a full understanding of the video's storyline.

General audio-visual question answering (AVQA). Our model not only excels in fine-grained omni-modal understanding but also demonstrates the ability to accurately answer more general audio-visual questions through cross-modal reasoning. For instance, in Fig. 9, it can precisely determine the location of the loudest instrument by integrating visual and auditory cues.

Overall, these examples vividly illustrate that relying solely on visual information to understand videos is far from sufficient. Integrating information from multiple modalities is both crucial and essential for comprehensive video understanding. Furthermore, thanks to our LongVALE dataset, our model is the first to combine cross-modal reasoning with fine-grained temporal understanding, setting it apart from traditional vision-only models.

5. Broader Impact

Risk mitigation. During the data generation, we used Gemini's safety mechanism to efficiently block harmful responses (*i.e.*, harassment, hate, dangerous content, *etc.*) and filter out corresponding videos. We also removed all individual names with the NLTK framework to protect privacy.

Data Licenses. We sourced our data from the open-sourced database, ACAV-100M [8] under MIT License[1]. Besides, the annotations of our LongVALE will be provided to the public under CC BY-NC-SA 4.0 license[2]. We hope our dataset can serve as a pivotal benchmark for promoting comprehensive multi-modal video understanding.

References

[1] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*, 2022. 3

[2] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024. 1

[1] https://opensource.org/license/mit

[2] https://creativecommons.org/licenses/by-nc-sa/4.0/

[3] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, march 2023. *URL https://lmsys. org/blog/2023-03-30-vicuna*, 3(5), 2023. 3

[4] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023. 1

[5] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 1

[6] Google. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf, 2024. 1, 3

[7] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280, 2024. 3

[8] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10274–10284, 2021. 4

[9] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[10] OpenAI. Gpt-4o system card. https://cdn.openai.com/gpt-4o-system-card.pdf, 2024. 1

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[12] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 1, 3

[13] 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 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1

[14] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 1

 # Prompt for video clip captioning

Please describe the video in detail, following instructions:

1. Focus on key visual details including appearance, motion, sequence of actions, objects involved, scene context and interactions between elements in the video.
2. Emphasize important points like the order of events, actions of people or objects, and any significant changes or camera movements.
3. Do not mention uncertain information or counting.
4. If there are characters, do not give specific recognition results, but explain their meaning.
5. Don't add extra sound descriptions.
6. Ensure the description is concise, clear, informative.

Here is an example: <Example>

 # Prompt for video clip keyframe captioning

Please describe this image, provide comprehensive visual details, including spatial attributes, scene context, and object characteristics. Only generate highly certain information, not irrelevant association or speculation. Ensure the description is concise, clear and informative.

 # Prompt for audio clip captioning

Recognize all events in the audio and describe them in detail.

 # Prompt for omni-modal event caption generation

You are an AI assistant that can see and hear videos.

Please describe the video content within a given time range using a complete sentence of less than 50 words based on the video captions, image captions, audio captions, subtitles and the context of the previous events.

1. Pretend to see and hear the video: The description must be in a tone that you are seeing and hearing the video. Do not generate a description: "According to the description...".
2. Focus on different elements in different captions: For video captions, mainly focus on dynamic information, including actions, interactions and camera movements. For image captions, mainly focus on visual details, including appearance, foreground and background scene context. For audio captions, mainly focus on clearly heard audio events. For subtitles, if there are some weird irrelevant content, please ignore them. Note: Please do not directly quote visually recognized characters, music lyrics and specific speech content in the generated sentence. If there are multiple captions from different time spans, you need to capture the changes between them and summarize them.
3. Reason the correlations between audio and visual information: Analyze whether the source of the sound is visible. If visible, who and why makes the sound? If invisible, what complementary information is provided by sound? If there are multiple sounds, what are the occurring time order and causation? Does it reflect any emotions, atmosphere and characteristics of the scene?
4. Only use the information given and not bring in any outside knowledge, you can generate some new words by reasoning but avoid excessive speculation and irrelevant association.

Here is an example:
Video caption: [0s : 10s]: "In the living room a black dog lies on the sofa".
Image captions: [0s : 10s]: "A black dog is lying on a khaki sofa and barking. White walls and a gray door can be seen in the picture."
Audio caption: [2s : 8s]: "dog is barking with a sound of a police car".
The given time range is: [0s : 10s].
So the generated description within the given time range is: "In the living room, a black dog lies on the sofa, alertly barking at the sound of a police car siren that echoes from outside."

Video captions:
Image captions:
Audio captions and subtitles:
Previous event context:
The given time range is:

Figure 4. The prompts for the captioning of video clips, keyframes and audio clips, and integrating them for omni-modal events captions.

Prompt for analyzing characteristics of omni-modal event captions

You are an AI assistant that can see and hear videos.
Please analyze the nature of the given video depiction, considering the following two aspects. Note that please only analyze based on the information given in the depiction, avoid speculation and irrelevant information.

1. **Audio-visual correlation:** Determine which types of audio-visual correlation exist in the depiction. Answer yes or no for each type.
- Synchronicity: The audio and visual elements are aligned both semantically and temporally, such as seeing and hearing a dog bark simultaneously.
- Complementary: The audio and visual information are not semantically or time-aligned, but complement each other, providing multi-faceted information.
- Temporal association: The audio and visual events occur one after another, such as cheers of the audience after the goal/performance, the thunder is seen on the screen before it is heard.
- Corrective: The sound information corrects the visual description, e.g., the visual information shows an outdoor celebration, but the sound actually reflects a protest march or dubbing a funny video.
- Causality: An event in one modality causes an event in another modality to occur, for example, the sound of outdoor sirens causes people to run and dogs to bark.
- Enhancement: Sound information enhances the atmosphere, for example, the background sound in the movie creates a tense atmosphere of the plot and crying and laughing reflect the emotional state. Visual description alone cannot reflect emotional expression.
- Scene-aware: Sound information reflects scene context in videos, e.g., birdsong, wind, waves reflect wild environment; vehicle engine horns reflect urban environment.
- Visual-only: The depiction only contains visual elements.
- Audio-only: The depiction only contains audio elements.

2. **Temporal changes:** Determine if there are any descriptions involving temporal changes in the depiction, such as transitions between shots, or events changing over time. Answer yes or no.

The given depiction:
Please output in the form of a dictionary: {audio-visual correlation: {type1: yes or no, type2: yes or no, ...}, Temporal changes: yes or no}

Figure 5. The prompt used to analyze and identify audio-visual correlations and temporal dynamics in our omni-modal event captions.

Prompt for omni-modal instruction tuning data generation

You are an AI assistant that can see and hear with the task of analyzing a single video.
Craft a conversation between yourself and a user discussing the video's content. Develop responses that embody the persona of an active audio-visual AI assistant, capable of perceiving the video using both visual and audio information and providing insightful answers.
Include inquiries about temporal perception and reasoning, like events preceding or succeeding specific occurrences, or requesting timestamps for particular actions or events.
Ensure that the questions can be definitively answered based on the perceivable video content or confidently ascertainable absence from the video. Utilize the timestamps <s?> and <e?> to create contextual questions considering the temporal relationships between events.
The conversations should be concise.
Here's an illustrative example: <Example>

Events:
From <s1> to <e1>: Caption1.
From <s2> to <e2>: Caption2.
From <s3> to <e3>: Caption3.
…

Dialogue:

Figure 6. The prompt used to generate omni-modal instruction tuning data.

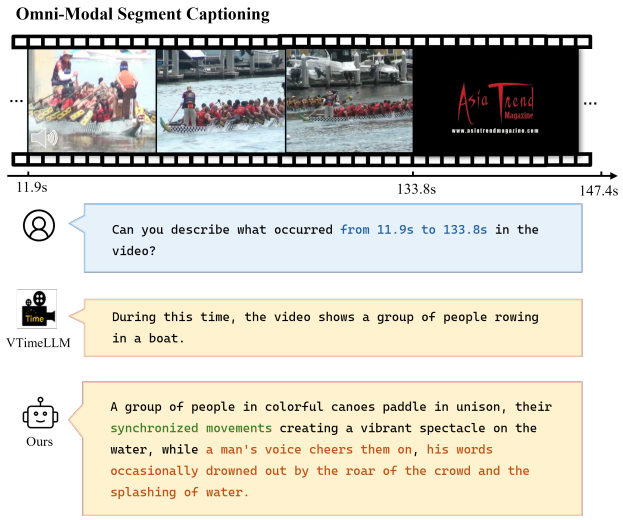


Figure 7. Additional qualitative results on omni-modal segment captioning task. The sample is from LongVALE test set.

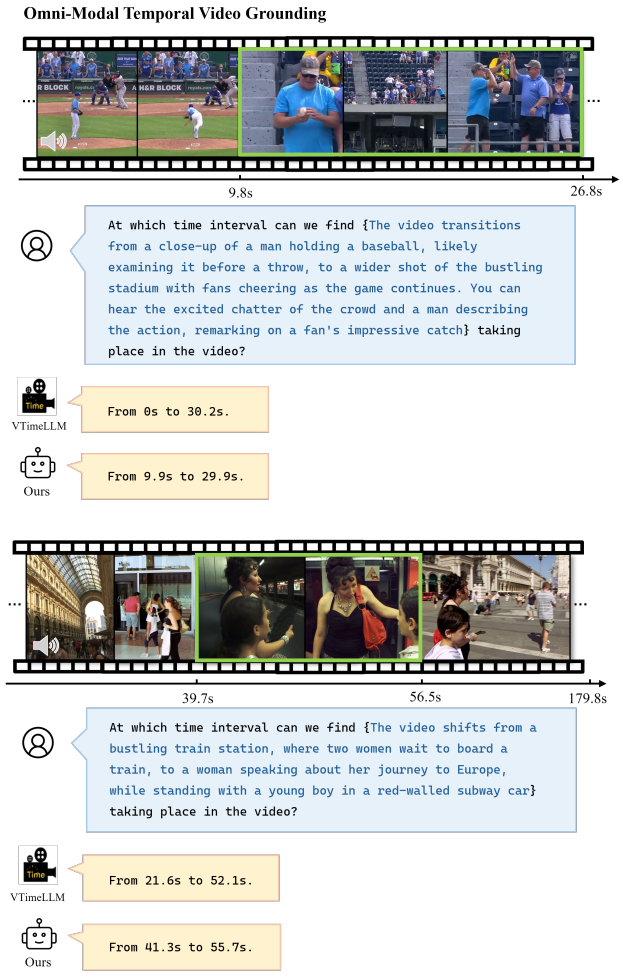


Figure 8. Qualitative results on omni-modal temporal video grounding task. The sample is from LongVALE test set. The ground-truth boundaries are displayed in green.

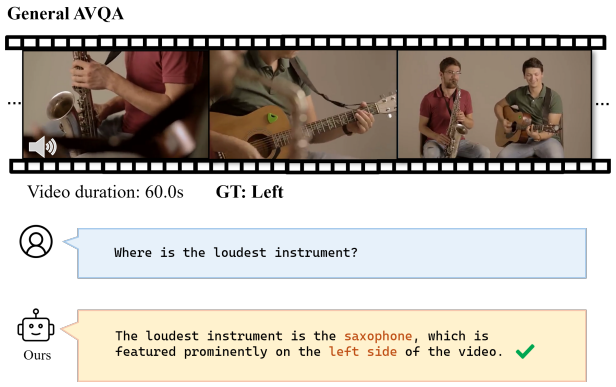


Figure 9. Additional qualitative results on general audio-visual question answering (AVQA) task. The sample is from Music-AVQA test set.

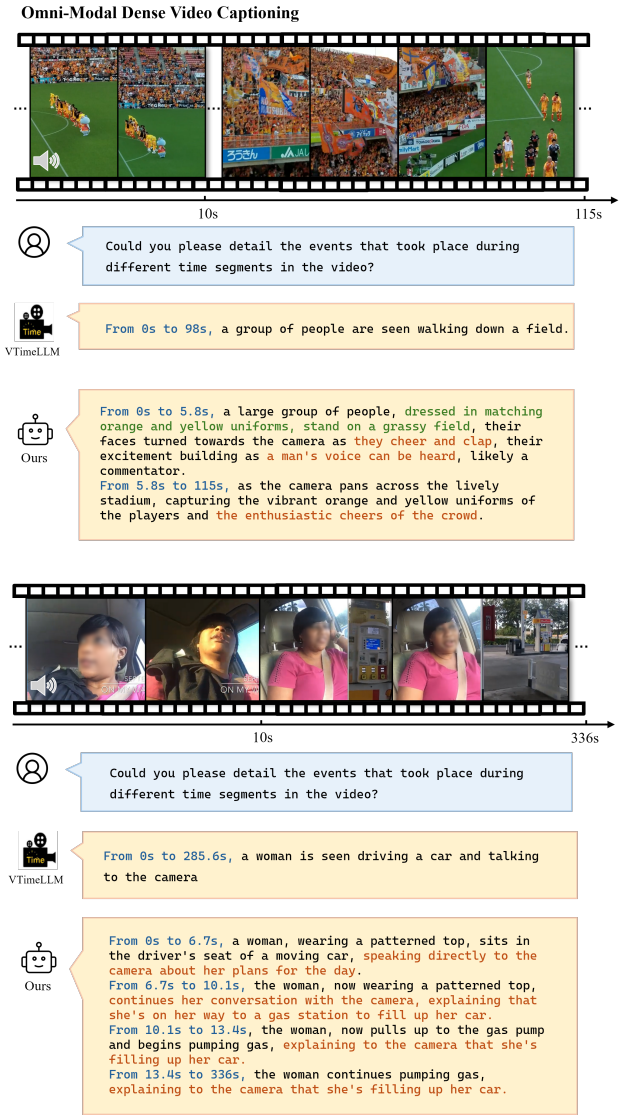


Figure 10. Qualitative results on omni-modal dense video captioning task. The sample is from LongVALE test set.