ChronusOmni: Improving Time Awareness of Omni-Modal Large Language Models

Yijing Chen[1⋆], Yihan Wu[1⋆], Kaisi Guan[1], Wenhui Tan[1], Yuchen Ren[1], Yuyue Wang[1], Ruihua Song[1⋆⋆], and Liyun Ru[2]

[1] Renmin University of China, Beijing, China
[2] Baichuan Inc., Beijing, China

Abstract. Time awareness is a fundamental ability of omni-modal large language models, especially for understanding long videos and answering complex questions. Previous approaches mainly target vision-language scenarios and focus on the explicit temporal grounding questions, such as identifying when a visual event occurs or determining what event happens at a specific time. However, they often make insufficient use of the audio modality, and overlook implicit temporal grounding across modalities (for example, identifying what is visually present when a character speaks, or determining what is said when a visual event occurs), despite such cross-modal temporal relations being prevalent in real-world scenarios. In this paper, we first formally define the **audiovisual temporal grounding task** to systematically encompass both explicit and implicit cross-modal temporal relations. To tackle this task, we propose **ChronusOmni**, an omni-modal large language model designed to enhance temporal awareness for both explicit and implicit audiovisual temporal grounding. Specifically, we interleave text-based timestamp tokens with visual and audio representations at each time unit, enabling unified temporal modeling across modalities. Furthermore, to enforce correct temporal ordering and strengthen temporal reasoning, we incorporate reinforcement learning with specially designed reward functions. Moreover, we construct ChronusAV, a temporally-accurate, modality-complete, and cross-modal-aligned dataset to support the training and evaluation of the audiovisual temporal grounding task. Experimental results demonstrate that ChronusOmni achieves state-of-the-art performance on audiovisual temporal grounding datasets such as ChronusAV and LongVALE, outperforming the second-best results by 17.3% and 28.8%, respectively, and delivers top-tier performance on visual-only temporal grounding datasets including Charades-STA and ActivityNet. These achievements highlight the strong temporal awareness of our model across modalities, while preserving its robust capabilities in general video and audio understanding.

Keywords: Omni-modal LLMs · Temporal grounding · Audiovisual understanding

⋆ Equal contributions.
⋆⋆ Corresponding author.

1 Introduction

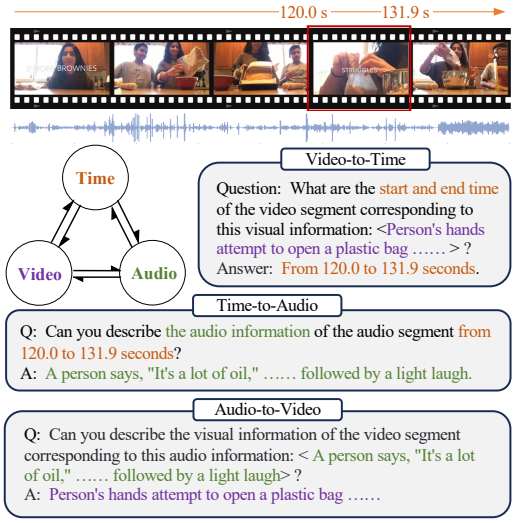
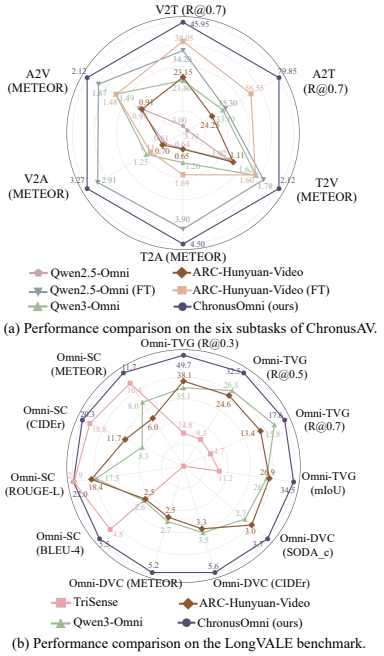


Fig. 1: Illustration of the audiovisual temporal grounding task. Three primary elements Video, Audio and Time are connected through six directional basic temporal grounding subtasks: Video-to-Time, Time-to-Video, Audio-to-Time, Time-to-Audio, Video-to-Audio, and Audio-to-Video.



(a) Performance comparison on the six subtasks of ChronusAV.

(b) Performance comparison on the LongVALE benchmark.

Fig. 2: Performance Comparison on ChronusAV and LongVALE.

Video temporal grounding [17, 18, 52] aims to align video semantics with temporal segments, serving as a critical capability for video understanding. Early studies in this field [6, 17, 18, 23, 24, 42, 50, 51] primarily rely solely on visual information to align video content with textual descriptions. However, in-the-wild videos inherently contain rich audio cues (e.g., dialogues, environmental sounds, and background music) that have largely been neglected by previous works. This motivates a shift toward audiovisual temporal grounding, which incorporates both visual and auditory modalities to capture complementary cues and improve time awareness in complex real-world scenarios.

Temporal grounding serves as a prerequisite for sophisticated multimodal reasoning. Consider a typical audiovisual scenario (e.g., in a movie dialogue scene, the camera first stays on Character A's expression; off-screen, Character B delivers the line "Let's go now." Then the shot cuts to B, and B's lip movements align with the dialogue.), where the semantic meaning of an event can only be accurately understood when both visual and auditory cues are jointly analyzed and temporally aligned. In such cases, the model is required to understand not only the **localization information in video streams** and the **localization**

information in audio streams (explicit temporal grounding), but also **the synchronization of audiovisual interactions** (implicit temporal grounding). However, most existing works focus solely on explicit temporal grounding [17, 23, 24, 48, 60], relying on either visual frames or audio features without capturing temporal dependencies across audio and video. As a result, they struggle in complex real-world scenarios where auditory and visual cues complement each other.

To address the above challenge, in this work, we improve the audiovisual temporal grounding task in a systematic way. First, we formally define the audiovisual temporal grounding task to systematically encompass both explicit and implicit cross-modal temporal relations. As illustrated in Figure 1, a model with audiovisual temporal grounding capability should support six directions of prediction across time, video, and audio: **video-to-time**, **time-to-video**, **audio-to-time**, **time-to-audio** (which constitute explicit temporal grounding), **audio-to-video**, and **video-to-audio** (which constitute implicit temporal grounding). To tackle this task, we propose ChronusOmni, an omni-modal large language model designed to enhance fine-grained temporal awareness. To explicitly model cross-modal temporal dependencies, we design an audiovisual temporal-synchronized representation that interleaves text-based timestamp tokens with visual and audio representations at each time unit. Furthermore, to enforce correct temporal ordering and strengthen fine-grained temporal reasoning, we incorporate a progressive refinement optimization strategy, culminating in reinforcement learning equipped with specially designed reward functions. Moreover, existing temporal grounding datasets often lack comprehensive audio annotations or implicit cross-modal alignment. Therefore, we construct ChronusAV, a temporally-accurate, modality-complete, and cross-modal-aligned dataset specifically tailored to support the training and evaluation of the audiovisual temporal grounding task.

The main contributions can be summarized as follows.

- We first formally define the **audiovisual temporal grounding task** to systematically encompass both explicit and implicit cross-modal temporal relations, comprehensively covering six distinct directions of prediction across time, video, and audio.
- We propose **ChronusOmni**, an omni-modal LLM that leverages an interleaved cross-modal representation and reinforcement learning with specially designed reward functions to explicitly enhance fine-grained temporal reasoning. Moreover, we construct ChronusAV, a large-scale, temporally-accurate, modality-complete, and cross-modal-aligned dataset to support its training and evaluation.
- Experiments demonstrate that ChronusOmni sets new state-of-the-art results on audiovisual temporal grounding datasets like ChronusAV and LongVALE, outperforming the second-best results by 17.3% and 28.8% respectively. It also delivers top-tier performance on visual-only datasets including Charades-STA and ActivityNet. These achievements highlight its strong

temporal awareness across modalities, while preserving robust capabilities in general video and audio understanding.

2 Related Work

Temporal Grounding MLLMs. Traditional video temporal grounding [52] aims to precisely align video semantics with corresponding time, including various tasks such as moment retrieval [4, 13, 21, 59], dense video caption [25, 58], and video highlight detection [54]. With the development of video large language models (LLMs), some recent works leverage the reasoning capabilities of LLMs to enhance temporal grounding. Some models directly infer frame indices relying on their ability to reason over frame order [23, 30], while others introduce absolute timestamp encoding to strengthen temporal localization [6,17,18,40]. Beyond the video modality, several audiovisual LLMs [14, 16, 32, 47, 55, 56, 61] extend temporal perception capabilities to audiovisual scenarios. Many use learnable temporal positional embeddings [55,56,61] or specially designed time encoders [32], which require substantial amounts of temporal training data to develop fine-grained temporal sensitivity from scratch. ARC-Hunyuan-Video [14] instead embeds a watermark representing absolute time on each frame. It introduces additional complexity because the model must perform OCR-like operations to recover temporal information. To address these limitations, we propose an audiovisual temporal-synchronized representation. We encode time directly as a text modality and interleave audio, visual, and time tokens along the timeline, enabling explicit, fine-grained alignment across modalities. With our temporal-aware training strategy, the model acquires strong audiovisual temporal grounding ability using only a relatively small amount of training data.

Temporal Grounding Datasets. Most existing temporal grounding benchmarks [20, 26, 27, 45, 62] rely on visual information solely, neglecting the audio modality. AVEL [49] and UnAV-100 [15] incorporate both audio and video streams with coarse event-level labels without detailed captions, limiting the model's ability to fine-grained audiovisual understanding. LongVALE [16] and TriSense-2M [32] provide detailed temporal event descriptions that combine audio and visual information but lack separate audio and video captions. This entanglement makes it difficult to analyze or assess a model's temporal understanding capability within and across each modality independently. VUE-TR [48] provides separate temporal annotations for video and audio streams, but its audio annotations focus only on human speech, restricting its applicability to more diverse audiovisual scenarios. Furthermore, none of the existing temporal grounding benchmarks evaluate implicit temporal perception across audio and video modalities, such as grounding what is heard when something is seen, or what is seen when something is heard. To address these limitations, we introduce ChronusAV, a temporally-accurate, modality-complete, and cross-modal-aligned dataset that provides time-aligned triplets of (time, video segment caption, audio segment caption) for each video. This design enables flexible construction of all six multimodal temporal grounding subtasks (as shown in Figure 1) and establishes

a comprehensive foundation for advancing multimodal temporal grounding research.

3 Task Definition: Audiovisual Temporal Grounding

In this section, we formally define the audiovisual temporal grounding task and its six subtasks accordingly. Suppose we have a video v and its corresponding audio a, both with a duration t. We denote a sequence of absolute time points as $t_0, t_1, t_2, ..., t_n$, which divide the video into n temporal segments: $(t_0, t_1), (t_1, t_2), ..., (t_{n-1}, t_n)$. For each interval (t_i, t_{i+1}), we extract a video clip v_i and an audio clip a_i. The clip v_i (and similarly a_i) starts at t_i and ends at t_{i+1}. Thus, we obtain n aligned tuples:

$$D_i = \{v_i, a_i, (t_i, t_{i+1})\}, \tag{1}$$

where each tuple contains a video segment, an audio segment, and its corresponding time boundaries.

Given a tuple D_i, the audiovisual temporal grounding task consists of querying one element to infer another, forming six subtasks, as illustrated in Figure 1.

- **Video-to-Time** (V2T). Given the information of the video segment v_i, the goal is to ground its corresponding absolute time segment (t_i, t_{i+1}).
- **Time-to-Video** (T2V). Given the absolute time segment (t_i, t_{i+1}), the objective is to obtain the information of the associated video segment v_i.
- **Audio-to-Time** (A2T). Given the information of the audio segment a_i, the goal is to ground its corresponding absolute time segment (t_i, t_{i+1}).
- **Time-to-Audio** (T2A). Given the absolute time segment (t_i, t_{i+1}), the objective is to obtain the information of the associated audio segment a_i.
- **Video-to-Audio** (V2A). Given the visual content of v_i, the objective is to obtain the information of the time-synchronized audio segment a_i.
- **Audio-to-Video** (A2V). Given the audio information of a_i, the objective is to obtain the information of the time-synchronized video segment v_i.

4 Proposed Method: ChronusOmni

To equip the model with fine-grained audiovisual time awareness, we propose ChronusOmni, a unified framework that achieves precise alignment across video, audio, and time. Unlike prior MLLMs that process audio and video largely in isolation, ChronusOmni organizes tokens in an interleaved manner to explicitly model cross-modal temporal dependencies. In addition, we design a two-stage, progressive refinement temporal optimization strategy to strengthen the model's temporal reasoning ability, resulting in robust performance on multimodal temporal grounding tasks.

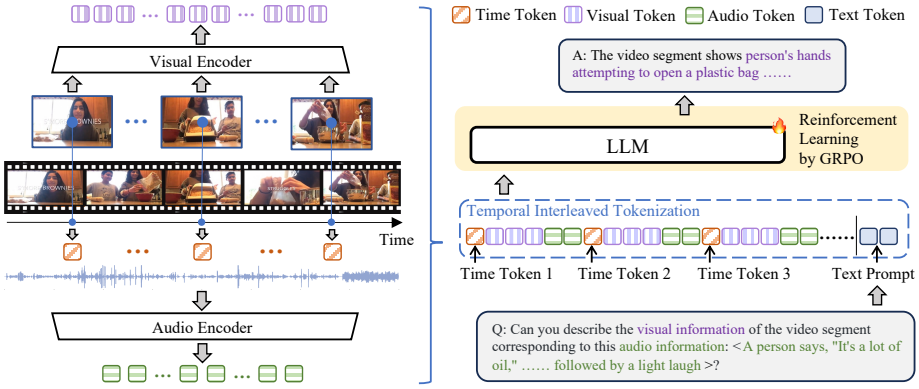


Fig. 3: The architecture of ChronusOmni. Time, video, audio are tokenized and interleaved at each time step. The token sequence along with text prompt is input into the LLM, which is supervised finetuned and further enhanced by reinforcement learning.

4.1 Audiovisual Temporal Representation

To improve cross-modal synchronization and achieve fine-grained alignment of audiovisual information and absolute time, we propose a Temporal Interleaved Tokenization (TIT) method, explicitly aligning the three elements–video, audio, and time along the timeline, as depicted in Figure 3.

Before being fed into the LLM, the video is first uniformly sampled into a *fixed number of frames*. A visual encoder and audio encoders then transform the video frames and audio into modality-specific tokens. However, uniform sampling results in *uneven time intervals between frames*. Therefore, explicit timestamps are required to convey absolute temporal information to the model: For each sampled frame, we extract its corresponding time point and convert it into text using the fixed format `second{t}` following [6]. For example, for a 126-second video sampled at 2-second intervals, we collect 64 timestamps: "second{0.0}", "second{2.0}", ..., "second{126.0}". These timestamps are encoded directly into text tokens, making them fully interpretable to the LLM.

Before feeding the tokens into the LLM, we interleave the token sequences from these three elements to achieve fine-grained temporal alignment. The interleaved tokens can be represented as:

$$I = [T_1, V_1, A_1, T_2, V_2, A_2, ..., T_i, V_i, A_i, ...], \tag{2}$$

where T_i is the absolute timestamp tokens corresponding to the specific time point t_i, V_i is the tokens encoded from the frame at t_i, and A_i is the tokens encoded from the audio between t_i and t_{i+1}. By using interleaved tokens, the model can establish a fine-grained connection among absolute time, visual information, and audio information.

4.2 Audiovisual Temporal Optimization

To enable the model to gradually acquire and robustly exploit temporal information, we further design a progressive refinement training strategy, including temporal-aware finetuning and temporal-aware reinforcement learning.

Stage 1: Temporal-aware Supervised Finetuning. In the first stage, we optimize the ChronusOmni via supervised fine-tuning (SFT). The model is trained on the video and audio dense captioning task, where for each video-audio pair input, the model learns to localize events and output time boundaries, visual caption, and audio caption for each event.

This training objective, together with our temporal-synchronized representation introduced in Section 4.1, guides the model to understand the alignment relationships among video, audio, and time.

However, SFT alone presents inherent limitations for temporal grounding [50]. First, although time is a continuous variable, SFT treats it as a categorical label, ignoring the metric structure, i.e., the magnitude and distance between time points. Second, as SFT exposes the model to gold prefixes throughout training, the model tends to rely on pattern memorization rather than learning to precisely localize boundaries. Consequently, SFT struggles to establish a robust mapping between audiovisual features and accurate temporal intervals.

Stage 2: Reinforcement Learning with GRPO. To overcome the limitations of SFT, we employ reinforcement learning in the second stage. Specifically, we adopt Group Relative Policy Optimization (GRPO) [10, 44] to train the model on the audiovisual temporal grounding task directly. This paradigm replaces the restrictive maximum likelihood objective with task-specific, outcome-driven reward functions, enabling the model to explore better strategies for aligning and integrating detailed audiovisual information with temporal boundaries.

For moment retrieval subtasks (V2T, A2T), the model outputs a temporal interval. We adopt the Intersection over Union (IoU) between the predicted interval and the ground-truth interval as the reward function:

$$R_{\text{IoU}} = \frac{|I_{\text{pred}} \cap I_{\text{gt}}|}{|I_{\text{pred}} \cup I_{\text{gt}}|}, \tag{3}$$

where I_{pred} is the predicted time interval, and I_{gt} is the ground-truth interval. To encourage consistent formatting that aligns with our timestamp design (Section 4.1), we also introduce a format reward enforcing outputs in the form "second{start}-second{end}":

$$R_{\text{format}} = \begin{cases} 0, & \text{if output format is wrong,} \\ 1, & \text{if output format is correct.} \end{cases} \tag{4}$$

For subtasks requiring the model to output a video or audio caption (T2A, V2A, T2V, A2V), we use Meteor [3] as the reward, which is a commonly used metric for evaluating video caption quality [1, 63]:

$$R_{\text{Meteor}} = \text{Meteor}(C_{\text{pred}}, C_{\text{gt}}), \tag{5}$$

where C_{pred} and C_{gt} is the predicted and ground-truth caption, respectively.

This progressive refinement strategy effectively mitigates the discretization and exposure-bias issues of SFT by incorporating reward-based temporal supervision. As a result, the model achieves more accurate boundary localization and stronger audiovisual alignment, yielding substantially improved multimodal temporal grounding performance.

5 ChronusAV: Multimodal Dataset for Audiovisual Temporal Grounding

To support training and comprehensive evaluation of the audiovisual temporal grounding task, we construct ChronusAV. ChronusAV is a large-scale dataset tailored for the audiovisual temporal grounding task, containing separate, detailed captions for audio (including sound, speech, and music) and visual information with precise temporal boundaries.

5.1 Construction of ChronusAV

To construct a temporally-accurate, modality-complete, and cross-modal-aligned dataset, we design a systematic dataset construction pipeline:

- **Video collection.** To construct a real-world, open-domain dataset, we select English videos from the large-scale, high-resolution Panda-70M dataset [8]. To cover videos of varying durations and reflect real-world temporal complexity, we select untrimmed long videos with audio tracks and restrict durations to 60–600 seconds, ensuring sufficient context for multimodal analysis.
- **Video segmentation.** To obtain precise event timestamps, we segment long, untrimmed videos into meaningful and coherent event units. Following Panda-70M, we first split basic visual scenes and then merge semantically similar ones, ensuring that each final segment represents a complete and coherent event. The corresponding audio tracks are segmented using the same timestamps to maintain audio–visual alignment. Finally, we retain only those audio–video pairs with 5 to 30 segments, ensuring that each sample contains rich temporal dynamics without being excessively fragmented.
- **Modality-specific annotation.** For each segment, we produce fine-grained, modality-aware captions by separately describing the visual and auditory streams. We use Gemini-2.5-Flash and Gemini-2.5-Pro [9] to annotate video and audio, respectively.
- **Human verification.** To ensure quality and modality separation of LLM-generated captions, we conduct a human study on 1,000 randomly sampled segments. Annotators rate accuracy and cross-modal leakage while viewing the corresponding video and audio. Results show that video captions are mostly accurate in 96.1% of cases; audio captions in 93.5%. And modality disentanglement is strong: 99.3% of video captions and 97.5% of audio captions show no or only minor cross-modal leakage. This high level of quality

Table 1: Comparison of ChronusAV to existing related datasets. ChronusAV covers large-scale open-domain long videos and, among the listed benchmarks, is the only dataset that simultaneously provides precise timestamps, multimodal annotations across vision, sound/music, and speech, as well as separate captions for audio and video. A&V: audio and video.

Datasets	Videos	Duration	Annotations	Domain	Timestamps	Vision	Sound&Music	Speech	A&V separate captions
Charades-STA [45]	10K	30s	16K	daily activity	✓	✓	✗	✗	-
ActivityNet Caps [26]	20K	180s	72K	daily activity	✓	✓	✗	✗	-
VALOR [34]	1.18M	10s	1.18M	open	✗	✓	✓	✗	✗
VAST [7]	27M	30s	27M	open	✗	✓	✓	✓	✗
AVEL [49]	4K	10s	4K	open	✓	✓	✓	✗	✗
UnAV-100 [15]	30K	42s	84K	open	✓	✓	✓	✗	✗
Shot2Story [19]	43K	17s	181K	open	✓	✓	✗	✓	✓
VUE-TR [48]	428	907s	1598	open	✓	✓	✗	✓	✓
LongVALE [16]	8.4K	235s	108K	open	✓	✓	✓	✓	✗
ChronusAV	47K	226s	677K	open	✓	✓	✓	✓	✓

assurance validates the reliability of our automatically generated annotations for large-scale audiovisual temporal grounding training and evaluation.

– **Test set construction.** Ultimately, we annotate 677K segments from 47K untrimmed videos. Each annotation is a {timestamp, video caption, audio caption} triplet, supporting all subtasks of audiovisual temporal grounding. We hold out 2,000 videos with corresponding audio to construct the test set and use the remaining for training. For the test set, we randomly select one segment per video and generate six QA pairs covering the six subtasks, as shown in Figure 1, yielding 12K QA pairs. All test QAs are human-verified and refined to ensure accurate, unique answers.

As a result, we get the ChronusAV, a temporally-accurate, modality-complete, and cross-modal-aligned dataset for audiovisual temporal grounding. More details of the dataset construction are elaborated in Appendix.

5.2 Analysis of ChronusAV

We provide an overview of the dataset and highlight the key characteristics that distinguish ChronusAV from existing temporal grounding datasets, which are also shown in Table 1.

– **Large-scale and long-duration videos.** ChronusAV contains 47K multi-shot videos (2,922 hours total, 60–600s range, averaging 226s). It is substantially longer than most prior benchmarks (e.g., ActivityNet Caps 180s, UnAV-100 42s), enabling robust modeling of long-term event dependencies.
– **Diverse and open-domain coverage.** ChronusAV spans 15 real-world domains, covering a wide spectrum of scenarios. This domain diversity ensures that ChronusAV aligns with the distributions of natural audiovisual content and supports strong generalization across contexts.
– **Fine-grained temporal annotations.** Unlike boundary-free datasets (e.g., VALOR, VAST), ChronusAV provides precise timestamps to facilitate temporal grounding. In the test set, QA-related segments peak at 3–6 seconds, demanding high capabilities in capturing subtle cross-modal temporal cues.

- **Modality completeness.** It simultaneously provides visual content and comprehensive audio cues (speech, music, and sound), overcoming the audio-incompleteness common in existing datasets (e.g., Charades-STA and ActivityNet Caps provide no audio annotation; VALOR and AVEL lack speech; Shot2Story and VUE-TR lack music and sound).
- **Separate audio and video captions.** ChronusAV provides modality-separated audio and video captions, enabling detailed and disentangled multimodal understanding. Most existing datasets (including LongVALE) do not offer this feature. Although Shot2Story and VUE-TR provide separate captions, their annotations lack sound-event descriptions, limiting their ability to capture real-world auditory complexity.

These features make ChronusAV a solid foundation for both training and evaluation. More statistics and cases can be found in Appendix.

6 Experiments

6.1 Implementation Details

We adopt the multimodal LLM Ola [36] as our backbone, which is built upon Qwen-2.5-7B [57]. For visual processing, we use OryxViT [35] as the vision encoder and uniformly sample 64 frames from each input video. On the audio side, we employ Whisper-Large-V3 [41] as the speech encoder and BEATs [5] as the sound and music encoder. To achieve a comprehensive representation of the audio content, the embedding features from Whisper-Large-V3 and BEATs are concatenated across the channel dimension. The resulting audio features are then downsampled by a factor of 10 to reduce token length while preserving semantic information. Finally, both visual and audio features are projected through two-layer multi-layer perceptron adapters, which transform them into unified tokens for the LLM decoder.

6.2 Training Settings

We initialize ChronusOmni with the pretrained Ola model and only finetune the LLM backbone, keeping all encoders and adapters frozen throughout training.

- **SFT stage.** In this stage, we train on a mixture of How2 [43], AVSD [2], and ChronusAV datasets. We randomly sample 10K clips of How2-300h and train on the AVSR task. For AVSD, we split each dialogue into turn-level QA in every dialogue, randomly sample 30K QA pairs, and train on the AVQA task. For ChronusAV, we randomly sample 30K videos from the training set and train on the dense video and audio caption task. We train on all 70k samples for one epoch.
- **GRPO stage.** We construct the GRPO data using the ChronusAV training set. For each untrimmed video, we randomly select one segment and form a QA pair corresponding to one of the six subtasks using the segment's timestamp and captions. In this stage, we ultimately use 4,000 pairs of QA data to train for 1,000 steps.

6.3 Evaluations on Audiovisual Temporal Grounding

Evaluation on ChronusAV. We comprehensively assess ChronusOmni's audiovisual temporal grounding performance using the ChronusAV benchmark, and compare it with previous audiovisual LLMs. For V2T and A2T subtasks, we follow the evaluation metrics for the moment retrieval task used in previous works [17, 42], reporting Recall@1 at IoU thresholds of 0.5, 0.7. For other subtasks, we use BLEU-4 [39], ROUGE-L [33], and METEOR [3] for caption quality evaluation.

Table 2: Comparison with existing video-audio LLMs for multimodal temporal grounding task on ChronusAV benchmark. The metrics B, R, and M denote BLEU-4, ROUGE-L, and METEOR, respectively. Best results are in bold, and second-best are underlined. The last row indicates the relative performance gain of our model compared to the second-best results. For Qwen3-Omni we use Qwen3-Omni-30B-A3B, while other models are all of size 7B. "FT" denotes that the baseline is further fine-tuned on our training data.

Model	V2T		T2V			A2T		T2A			V2A			A2V		
	R@0.5	R@0.7	B	R	M	R@0.5	R@0.7	B	R	M	B	R	M	B	R	M
VideoLLaMA [61]	2.00	0.80	0.10	1.50	0.93	1.70	0.80	0.05	1.13	0.63	0.08	1.11	0.81	0.07	1.22	0.93
Ola [36]	5.80	2.65	0.26	1.66	0.99	5.50	2.55	0.18	0.78	0.31	0.15	1.15	0.53	0.18	1.60	1.08
AVicuna [47]	10.75	4.70	0.04	1.03	0.32	8.20	3.45	0.01	0.51	0.13	0.02	0.54	0.23	0.11	1.45	0.65
LongVALE-LLM [16]	9.50	3.65	0.35	2.03	0.99	4.25	1.25	0.15	1.31	0.50	0.10	1.26	0.64	0.21	1.88	1.02
Qwen2.5-Omni [55]	7.05	3.00	0.35	1.89	1.02	10.10	3.70	0.67	1.01	0.64	0.54	0.91	0.61	0.19	1.61	0.97
ARC-Hunyuan-Video [14]	36.10	23.15	0.24	1.54	1.11	36.85	24.25	0.18	1.17	0.65	0.12	0.94	0.70	0.11	1.12	0.91
Qwen3-Omni [56]	37.85	21.80	0.37	2.13	1.62	46.70	33.10	0.94	2.26	1.20	0.35	1.58	1.25	0.22	1.72	1.49
Qwen2.5-Omni (FT) [55]	53.55	34.20	1.02	3.30	1.78	49.35	35.30	5.37	5.29	3.90	3.43	4.09	2.91	0.85	2.99	1.87
ARC-Hunyuan-Video (FT) [14]	55.25	38.05	0.75	2.75	1.60	77.85	56.55	0.72	1.71	1.69	0.85	1.84	1.11	0.42	2.19	1.48
ChronusOmni	**63.15**	**45.95**	**1.16**	**3.37**	**2.12**	**90.50**	**79.85**	**6.78**	**6.86**	**4.50**	**3.61**	**4.90**	**3.27**	**1.01**	**3.17**	**2.12**
Δ to the second best	+14.3%	+20.2%	+13.7%	+2.1%	+19.1%	+16.2%	+43.5%	+26.3%	+29.7%	+15.4%	+5.2%	+19.8%	+12.4%	+18.8%	+6.0%	+13.4%

As shown in Table 2, ChronusOmni achieves state-of-the-art performance across all metrics in the six subtasks. The experimental results reveal that existing open-source large models generally struggle with complex temporal grounding tasks; even the 30B parameter Qwen3-Omni model exhibits relatively limited performance, scoring only 46.70% on the R@0.5 metric for the A2T task. After further fine-tuning Qwen2.5-Omni and ARC-Hunyuan-Video on our specific training data, the performance of these baselines experienced a substantial leap, which strongly validates the high quality and effectiveness of our dataset. However, even when compared to the fine-tuned baselines, ChronusOmni demonstrates an overwhelming advantage. Compared to the second-best results, ChronusOmni achieves an average gain of 17.3%. These results validate ChronusOmni's comprehensive and exceptional capabilities in fine-grained temporal understanding and precise cross-modal alignment.

Evaluation on LongVALE. LongVALE [16] comprises three omni temporal tasks: Omni-TVG (predict the time segment corresponding to an omni event caption), Omni-DVC (predict timestamps and captions of all omni events), and Omni-SC

(generate the omni caption for a specified time segment). We evaluate ChronusOmni on the LongVALE test set in a zero-shot setting.

Table 3: Comparison on LongVALE benchmark. "*" indicates that the model has been trained on the LongVALE training set, while all other models report zero-shot results. Best results are in bold, and second-best are underlined. The last row indicates the relative performance gain of our model compared to the second-best results.

Model	Omni-TVG				Omni-DVC			Omni-SC			
	R@0.3	R@0.5	R@0.7	mIoU	SODA_c	CIDEr	METEOR	BLEU-4	ROUGE-L	CIDEr	METEOR
LongVALE-LLM* [16]	15.7	8.6	3.9	11.0	2.8	7.9	4.7	5.6	22.4	20.3	10.9
VideoChat [29]	2.2	0.9	0.4	3.0	0.7	0.2	0.9	0.5	9.6	0.0	8.2
VideoChatGPT [37]	4.9	2.0	0.9	5.0	0.7	0.1	0.9	0.4	14.0	0.9	5.9
VideoLLaMA [61]	2.5	1.1	0.3	1.9	0.6	0.6	0.9	0.9	11.5	0.1	8.9
PandaGPT [46]	2.5	1.0	0.3	2.2	0.5	0.0	0.6	0.6	14.9	0.3	8.9
NExT-GPT [53]	4.3	1.9	0.7	4.0	0.2	0.1	0.3	0.4	10.2	0.0	8.1
TimeChat [42]	5.8	2.6	1.1	5.2	1.6	0.1	1.4	1.2	16.1	1.6	10.0
VTimeLLM [23]	7.5	3.4	1.3	6.4	2.4	0.2	2.0	1.0	14.5	1.6	5.5
TriSense [32]	14.8	9.3	4.7	11.2	-	-	-	4.8	21.9	18.8	10.4
ARC-Hunyuan-Video [14]	38.1	24.6	13.4	26.9	3.0	3.3	2.5	2.5	18.4	11.7	6.0
Qwen3-Omni [56]	35.1	26.3	15.8	26.1	2.7	3.5	2.7	2.6	17.5	8.3	8.0
ChronusOmni	**49.7**	**32.5**	**17.6**	**34.5**	**3.7**	**5.6**	**5.2**	**5.5**	**22.0**	**20.3**	**11.7**
Δ to the second best	+41.6%	+23.6%	+11.4%	+28.3%	+23.3%	+60.0%	+92.6%	+14.6%	+0.5%	+8.0%	+12.5%

Table 3 shows that ChronusOmni establishes a new state-of-the-art performance on LongVALE benchmark, demonstrating overwhelming superiority over other multimodal large language models across all three subtasks under a zero-shot setting. In the Omni-TVG task, ChronusOmni establishes a new standard with an mIoU of 34.5%, substantially surpassing strong audiovisual models like TriSense [32], ARC-Hunyuan-Video [14], and Qwen3-Omni [56] by massive margins. In temporal captioning tasks (Omni-DVC and Omni-SC), ChronusOmni also consistently yields the highest scores in all metrics among all zero-shot models. Compared to the second-best zero-shot results, ChronusOmni achieves an average gain of 28.8%. Most notably, while operating entirely in a zero-shot setting, ChronusOmni remains highly competitive with, and surpasses in most metrics, LongVALE-LLM, despite the latter being explicitly trained on the LongVALE training set. These results highlight our model's robust capability in audiovisual temporal grounding and understanding. Furthermore, despite being trained solely on ChronusAV for temporal supervision, our model still achieves strong cross-dataset generalization on LongVALE, which implicitly validates the high quality of our ChronusAV dataset.

6.4 Evaluations on Visual-only Temporal Grounding

Charades-STA [45] and ActivityNet [26] are widely used visual-only temporal grounding datasets. Because some of our training data (e.g., AVSD) contains Charades-STA videos, a fair zero-shot evaluation on Charades-STA is infeasible; thus, we report one-epoch GRPO fine-tuned results for ChronusOmni and compare with baselines trained on Charades-STA. For ActivityNet, we evaluate

zero-shot results on the test set. To enable fair comparison with prior VLMs, we use video-only inputs (no audio) for both datasets during training and testing.

As shown in Table 4, ChronusOmni achieves state-of-the-art performance on Charades-STA, outperforming the second-best models by +2.2 (vs. Time-R1 at R@0.3), +2.8 (vs. Time-R1 at R@0.5), and +4.0 (vs. VideoChat-R1 at R@0.7). On ActivityNet, ChronusOmni ranks second at R@0.3 and R@0.5, and achieves the best R@0.7 (22.1), surpassing Time-R1 by +0.7, indicating stronger high-precision temporal localization under zero-shot transfer. Overall, ChronusOmni consistently outperforms or matches strong baselines on visual-only temporal grounding benchmarks, demonstrating modality-robust and generalizable temporal understanding capabilities.

Table 4: Comparison with 7B VLMs capable of temporal grounding on Charades-STA and ActivityNet. Due to the presence of Charades-STA videos in our pre-training data, we report one-epoch GRPO fine-tuned results for it and compare with baselines trained on Charades-STA, while evaluating ActivityNet in a zero-shot setting.

Model	Charades-STA			ActivityNet		
	R@0.3	R@0.5	R@0.7	R@0.3	R@0.5	R@0.7
TimeChat [42]	-	46.7	23.7	36.2	20.2	9.5
Hawkeye [51]	72.5	58.3	28.8	49.1	29.3	10.7
VTimeLLM [23]	-	-	-	44.0	27.8	14.3
VTG-LLM [18]	-	57.8	33.9	-	-	-
TRACE [17]	-	61.7	41.4	-	-	-
TimeSuite [60]	79.4	67.1	43.0	-	-	-
VideoChat-R1 [31]	-	71.7	<u>50.2</u>	-	33.4	17.7
Time-R1 [50]	<u>82.9</u>	<u>72.2</u>	50.1	**58.6**	**39.0**	21.4
ChronusOmni	**85.1**	**75.0**	**54.2**	<u>56.9</u>	<u>38.2</u>	**22.1**

6.5 Evaluations on General Video and Audio Understanding

To evaluate whether our temporal modeling affects general audiovisual understanding, we conduct evaluations on six widely used benchmarks: Video-MME [11], Librispeech [38], VisSpeech [12], MUSIC-AVQA [28], WorldSense [22], and Daily-Omni [64]. Video-MME is vision-only, Librispeech is audio-only, while the other four benchmarks utilize both visual and audio modalities.

Table 5: Evaluation on general video and audio understanding benchmarks.

Model	**Video-MME** (ACC ↑)	**LibriSpeech** (WER ↓)	**VisSpeech** (WER ↓)	**MUSIC-AVQA** (ACC ↑)	**WorldSense** (ACC ↑)	**Daily-Omni** (ACC ↑)
Base (Ola)	**62.7**	**3.2**	12.3	62.2	46.7	60.1
ChronusOmni	62.4	3.5	**9.1**	**66.0**	**47.8**	**72.3**

Table 5 compares ChronusOmni against its base model across these six video and audio understanding benchmarks. ChronusOmni shows only a minor degradation on video QA (Video-MME) and speech recognition (LibriSpeech) tasks, while yielding substantial improvements on visual speech recognition (VisSpeech:

9.1 vs. 12.3 WER, a 26% reduction) and audiovisual QA (MUSIC-AVQA: +3.8, WorldSense: +1.1, Daily-Omni: +12.2). It indicates that our temporal interleaved tokenization and temporal optimization do not impair general modality-specific abilities. Instead, they enhance the model's ability to perform joint audiovisual reasoning, demonstrating stronger multimodal integration.

6.6 Ablation Study

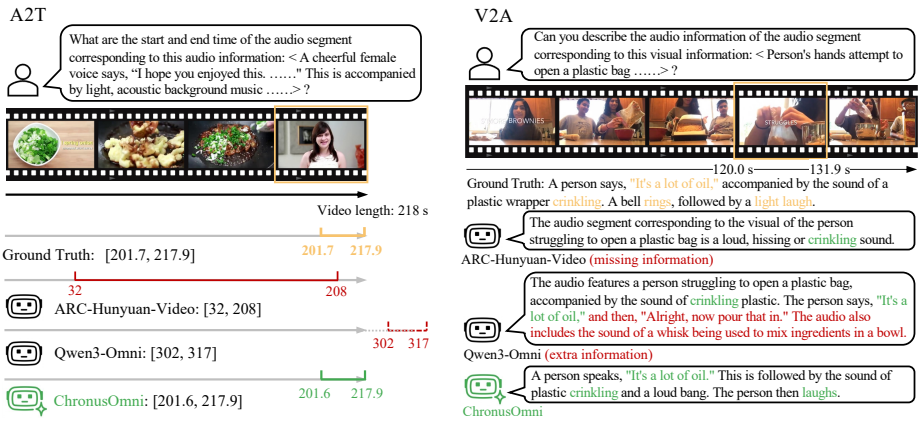
Table 6: Ablation study on our temporal interleaved tokenization method and training strategy. TIT denotes our proposed temporal interleaved tokenization.

Model	V2T		T2V			A2T		T2A			V2A			A2V		
	R@0.5	R@0.7	B	R	M	R@0.5	R@0.7	B	R	M	B	R	M	B	R	M
w/o TIT	16.05	6.80	0.64	2.51	1.67	26.05	13.00	2.43	3.29	2.99	2.56	3.74	2.75	0.67	2.58	1.79
w/o SFT	59.40	37.20	0.53	2.39	1.81	84.45	69.25	2.45	3.29	1.89	2.12	3.15	1.96	0.49	2.34	1.84
w/o GRPO	30.00	15.45	0.46	2.53	1.07	34.55	19.00	0.04	1.48	0.35	0.54	1.40	0.73	0.12	1.81	0.66
ChronusOmni	**63.15**	**45.95**	**1.16**	**3.37**	**2.12**	**90.50**	**79.85**	**6.78**	**6.86**	**4.50**	**3.61**	**4.90**	**3.27**	**1.01**	**3.17**	**2.12**

To evaluate the effectiveness of our proposed components, we conduct a comprehensive ablation study (Table 6) on the temporal interleaved tokenization (TIT) method and our training strategies (SFT and GRPO) across six diverse cross-modal tasks. Removing the TIT module leads to severe performance degradation across all subtasks, with R@0.5 scores in V2T and A2T plummeting by 47.10 and 64.45 absolute points respectively, demonstrating that interleaved tokenization is crucial for preserving fine-grained temporal perception ability. Furthermore, the w/o SFT variant exhibits a noticeable decline particularly in temporal captioning tasks, despite retaining basic cross-modal grounding capabilities, and the absence of the GRPO stage (w/o GRPO) causes a severe performance degradation across all subtasks. Our full model, ChronusOmni, consistently achieves the best performance across all metrics, validating the necessity of each module.

6.7 Qualitative Results

We provide qualitative results of ChronusOmni and the other two audiovisual LLMs in two subtasks of ChronusAV benchmark in Figure 4. As shown in Figure 4(a), ChronusOmni can locate more precise time boundaries. Despite the video being only 218 seconds long, Qwen3-Omni experienced hallucinations, situating events at over 300 seconds. Our explicit text timestamp design effectively prevents such hallucinations from occurring. As Figure 4(b) shows, given a specific visual segment query, Qwen3-Omni generates a description containing audio events occurring after the queried time period. Conversely, ARC-Hunyuan-Video exhibits incomplete temporal coverage; while it avoids generating out-of-bounds content, it only captures the initial salient sound (crinkling) and misses



(a) Qualitative results on A2T subtask. (b) Qualitative results on V2A subtask.

Fig. 4: Qualitative results. The sample is from the ChronusAV benchmark.

the subsequent speech and background sounds within the queried time period. Both baseline models demonstrate imprecise fine-grained temporal localization, whereas ChronusOmni accurately grounds the complete audio sequence corresponding strictly to the queried visual timeframe. These demonstrate ChronusOmni's stronger temporal awareness.

7 Conclusion

In this paper, we formally define the audiovisual temporal grounding task and present ChronusOmni, a multimodal LLM tailored for this task. Our approach builds a temporally synchronized audiovisual representation and uses a progressive refinement training strategy to strengthen fine-grained temporal understanding. We also release ChronusAV, a comprehensive and standardized dataset for training and evaluation of audiovisual temporal grounding. Experimental results demonstrate that ChronusOmni achieves state-of-the-art performance on audiovisual temporal grounding datasets, and delivers top-tier performance on visual-only temporal grounding datasets. These achievements highlight the strong temporal awareness of our model across modalities, while preserving its robust capabilities in general video and audio understanding. Future work will focus on deploying our model in real-world interactive scenarios and scaling it to hour-long videos.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62276268) and Baichuan Inc.

References

1. Abdar, M., Kollati, M., Kuraparthi, S., et al.: A review of deep learning for video captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
2. AlAmri, H., Cartillier, V., Das, A., et al.: Audio visual scene-aware dialog. In: CVPR. pp. 7558–7567 (2019)
3. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
4. Boris, M., Anil, B., Anna, R., Rohrbach, M.: The surprising effectiveness of multimodal large language models for video moment retrieval. CoRR **abs/2406.18113** (2024)
5. Chen, S., Wu, Y., Wang, C., et al.: Beats: Audio pre-training with acoustic tokenizers. arXiv preprint arXiv:2212.09058 (2022)
6. Chen, S., Lan, X., Yuan, Y., Jie, Z., Ma, L.: Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability. arXiv preprint arXiv:2411.18211 (2024)
7. Chen, S., Li, H., Wang, Q., Zhao, Z., Sun, M., Zhu, X., Liu, J.: VAST: A vision-audio-subtitle-text omni-modality foundation model and dataset. In: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023)
8. Chen, T.S., Siarohin, A., Menapace, W., 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. pp. 13320–13331 (2024)
9. Comanici, G., Bieber, E., Schaekermann, M., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
10. DeepSeek-AI, Guo, D., Yang, D., Zhang, H., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. CoRR **abs/2501.12948** (2025)
11. Fu, C., Dai, Y., Luo, Y., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. CoRR **abs/2405.21075** (2024)
12. Gabeur, V., Seo, P.H., Nagrani, A., Sun, C., Alahari, K., Schmid, C.: AVATAR: unconstrained audiovisual speech recognition. In: 23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022. pp. 2818–2822 (2022)
13. Gao, J., Sun, C., Yang, Z., Nevatia, R.: TALL: temporal activity localization via language query. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 5277–5285 (2017)
14. Ge, Y., Ge, Y., Li, C., et al.: Arc-hunyuan-video-7b: Structured video comprehension of real-world shorts. CoRR **abs/2507.20939** (2025)
15. Geng, T., Wang, T., Duan, J., Cong, R., Zheng, F.: Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 22942–22951 (2023)
16. Geng, T., Zhang, J., Wang, Q., Wang, T., Duan, J., Zheng, F.: Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In: CVPR. pp. 18959–18969 (2025)

17. Guo, Y., Liu, J., Li, M., Liu, Q., Chen, X., Tang, X.: TRACE: temporal grounding video LLM via causal event modeling. In: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net (2025)
18. Guo, Y., Liu, J., Li, M., et al.: VTG-LLM: integrating timestamp knowledge into video llms for enhanced video temporal grounding. In: AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA. pp. 3302–3310 (2025)
19. Han, M., Yang, L., Chang, X., Yao, L., Wang, H.: Shot2story: A new benchmark for comprehensive understanding of multi-shot videos. In: ICLR (2025)
20. Hendricks, L.A., Wang, O., Shechtman, E., et al.: Localizing moments in video with natural language. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 5804–5813. IEEE Computer Society (2017)
21. Hendricks, L.A., Wang, O., Shechtman, E., et al.: Localizing moments in video with temporal language. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018. pp. 1380–1390 (2018)
22. Hong, J., Yan, S., Cai, J., et al.: Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. arXiv preprint arXiv:2502.04326 (2025)
23. Huang, B., Wang, X., Chen, H., Song, Z., Zhu, W.: Vtimellm: Empower llm to grasp video moments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14271–14280 (2024)
24. Huang, D., Liao, S., Radhakrishnan, S., et al.: LITA: language instructed temporal-localization assistant. In: Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXIV. Lecture Notes in Computer Science, vol. 15122, pp. 202–218 (2024)
25. Kim, M., Kim, H.B., Moon, J., Choi, J., Kim, S.T.: Do you remember? dense video captioning with cross-modal memory retrieval. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 13894–13904 (2024)
26. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: Proceedings of the IEEE international conference on computer vision. pp. 706–715 (2017)
27. Lei, J., Yu, L., Berg, T.L., Bansal, M.: Tvr: A large-scale dataset for video-subtitle moment retrieval. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16. pp. 447–463. Springer (2020)
28. Li, G., Wei, Y., Tian, Y., Xu, C., Wen, J., Hu, D.: Learning to answer questions in dynamic audio-visual scenarios. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 19086–19096 (2022)
29. Li, K., He, Y., Wang, Y., et al.: Videochat: Chat-centric video understanding. Sci. China Inf. Sci. **68**(10) (2025)
30. Li, X., Wang, Y., Yu, J., et al.: Videochat-flash: Hierarchical compression for long-context video modeling. CoRR **abs/2501.00574** (2025)
31. Li, X., Yan, Z., Meng, D., et al.: Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. arXiv preprint arXiv:2504.06958 (2025)
32. Li, Z., Zhang, X., Guo, Y., et al.: Watch and listen: Understanding audio-visual-speech moments with multimodal LLM. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)

33. Lin, C., Och, F.J.: Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain. pp. 605–612 (2004)
34. Liu, J., Chen, S., He, X., Guo, L., Zhu, X., Wang, W., Tang, J.: VALOR: vision-audio-language omni-perception pretraining model and dataset. IEEE Trans. Pattern Anal. Mach. Intell. **47**(2), 708–724 (2025)
35. Liu, Z., Dong, Y., Liu, Z., et al.: Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. In: The Thirteenth International Conference on Learning Representations (2025)
36. Liu, Z., Dong, Y., Wang, J., et al.: Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment. CoRR (2025)
37. Maaz, M., Rasheed, H.A., Khan, S., Khan, F.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024. pp. 12585–12602 (2024)
38. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: An ASR corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015. pp. 5206–5210 (2015)
39. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA. pp. 311–318 (2002)
40. Qian, L., Li, J., Wu, Y., et al.: Momentor: Advancing video large language model with fine-grained temporal reasoning. In: Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024 (2024)
41. Radford, A., Kim, J.W., Xu, T., et al.: Robust speech recognition via large-scale weak supervision. In: International conference on machine learning. pp. 28492–28518 (2023)
42. Ren, S., Yao, L., Li, S., et al.: Timechat: A time-sensitive multimodal large language model for long video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14313–14323 (2024)
43. Sanabria, R., Caglayan, O., Palaskar, S., et al.: How2: A large-scale dataset for multimodal language understanding. CoRR **abs/1811.00347** (2018)
44. Shao, Z., Wang, P., Zhu, Q., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. CoRR **abs/2402.03300** (2024)
45. Sigurdsson, G.A., Varol, G., Wang, X., et al.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I. Lecture Notes in Computer Science, vol. 9905, pp. 510–526 (2016)
46. Su, Y., Lan, T., Li, H., et al.: Pandagpt: One model to instruction-follow them all. In: Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants! pp. 11–23 (2023)
47. Tang, Y., Shimada, D., Bi, J., et al.: Empowering llms with pseudo-untrimmed videos for audio-visual temporal understanding. In: AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA. pp. 7293–7301 (2025)

48. Team, V., Liu, C., Kuo, C., et al.: Vidi: Large multimodal models for video understanding and editing. CoRR **abs/2504.15681** (2025)
49. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II. Lecture Notes in Computer Science, vol. 11206, pp. 252–268. Springer (2018)
50. Wang, Y., Wang, Z., Xu, B., et al.: Time-r1: Post-training large vision language model for temporal video grounding. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
51. Wang, Y., Meng, X., Liang, J., Wang, Y., Liu, Q., Zhao, D.: Hawkeye: Training video-text llms for grounding text in videos. CoRR **abs/2403.10228** (2024)
52. Wu, J., Liu, W., Liu, Y., et al.: A survey on video temporal grounding with multimodal large language model. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
53. Wu, S., Fei, H., Qu, L., Ji, W., Chua, T.S.: Next-gpt: Any-to-any multimodal llm. In: Forty-first International Conference on Machine Learning (2024)
54. Xiong, B., Kalantidis, Y., Ghadiyaram, D., Grauman, K.: Less is more: Learning highlight detection from video duration. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 1258–1267. Computer Vision Foundation / IEEE (2019)
55. Xu, J., Guo, Z., He, J., et al.: Qwen2.5-omni technical report. CoRR **abs/2503.20215** (2025)
56. Xu, J., Guo, Z., Hu, H., et al.: Qwen3-omni technical report. CoRR **abs/2509.17765** (2025)
57. Yang, A., Yang, B., Zhang, B., et al.: Qwen2.5 technical report. CoRR **abs/2412.15115** (2024)
58. Yang, A., Nagrani, A., Seo, P.H., Miech, A., Pont-Tuset, J., Laptev, I., Sivic, J., Schmid, C.: 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. pp. 10714–10726 (2023)
59. Zala, A., Cho, J., Kottur, S., et al.: Hierarchical video-moment retrieval and step-captioning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 23056–23065 (2023)
60. Zeng, X., Li, K., Wang, C., et al.: Timesuite: Improving mllms for long video understanding via grounded tuning. In: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net (2025)
61. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 543–553 (2023)
62. Zhou, L., Xu, C., Corso, J.J.: Towards automatic learning of procedures from web instructional videos. In: McIlraith, S.A., Weinberger, K.Q. (eds.) Proceedings of the AAAI conference on artificial intelligence. pp. 7590–7598 (2018)
63. Zhou, X., Arnab, A., Buch, S., et al.: Streaming dense video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18243–18252 (2024)
64. Zhou, Z., Wang, R., Wu, Z.: Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. arXiv preprint arXiv:2505.17862 (2025)