LVOmniBench: PIONEERING LONG AUDIO-VIDEO UNDERSTANDING EVALUATION FOR OMNIMODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advancements in omnimodal large language models (OmniLLMs) have significantly improved the comprehension of audio and video inputs. However, current evaluations primarily focus on short audio and video clips ranging from 10 seconds to 5 minutes, failing to reflect the demands of real-world applications, where videos typically run for tens of minutes. To address this critical gap, we introduce **LVOmniBench**, a new benchmark designed specifically for the cross-modal comprehension of *long-form* audio and video. This dataset comprises high-quality videos sourced from open platforms that feature rich audio-visual dynamics. Through rigorous manual selection and annotation, LVOmniBench comprises 275 videos, ranging in duration from 10 to 90 minutes, and 1,014 question-answer (QA) pairs. LVOmniBench aims to rigorously evaluate the capabilities of OmniLLMs across domains, including long-term memory, temporal localization, fine-grained understanding, and multimodal perception. Our extensive evaluation reveals that current OmniLLMs encounter significant challenges when processing extended audio-visual inputs. Open-source models generally achieve accuracies below 40%, whereas Gemini 3.8-Flash reaches a peak accuracy of approximately 69%. We anticipate that this dataset, along with our empirical findings, will stimulate further research and the development of advanced models capable of resolving complex omnimodal understanding problems within long inputs.

1 INTRODUCTION

The rapid development of omnimodal large language models (OmniLLMs) has highlighted their significant perceptual and cognitive capabilities in integrating vision, audio, and text (Xu et al., 2025a;b; Sun et al., 2024; Tang et al., 2025; Cheng et al., 2024; Fu et al., 2025b; Ye et al., 2025; Li et al., 2024b; Chen et al., 2025c; Yang et al., 2025a; Tong et al., 2025; Team et al., 2025c; AI et al., 2025; Yao et al., 2024; Fu et al., 2024). These advancements demonstrate the substantial potential of OmniLLMs as foundation models that can simultaneously comprehend real-world audio and video inputs. However, real-world videos are not merely combinations of isolated modalities; they are intrinsically long-form, often spanning tens of minutes and featuring highly intertwined audio-visual streams. This extended temporal dimension amplifies the complexity of multimodal interactions, posing significant challenges for fine-grained understanding, cross-modal alignment, and reasoning. Consequently, while current OmniLLMs perform well on isolated tasks, they struggle with highly complex scenarios. Furthermore, although numerous benchmarks exist for omnimodal understanding (Chen et al., 2025a; Li et al., 2024c; Zhou et al., 2025b; Hong et al., 2025; Li et al., 2025b; Sung-Bin et al., 2024; Chao et al., 2025; Cao et al., 2025; Wang et al., 2025a; Zhang et al., 2025d; Ma et al., 2025), most are limited to reasoning over static image-audio pairs (Li et al., 2024c; Gong et al., 2024) or short-form video clips (Zhou et al., 2025b; Hong et al., 2025; Li et al., 2025b; Sung-Bin et al., 2024; Chao et al., 2025). This scarcity of evaluations for long-form joint audio-video content leaves a critical gap in comprehensively assessing and advancing robust OmniLLMs for real-world applications.

To this end, we introduce **LVOmniBench**, which, to the best of our knowledge, is the first comprehensive benchmark specifically designed for the rigorous evaluation of OmniLLMs in understanding long-form, integrated audio-visual content. We curated a dataset comprising 275 long videos across diverse scenarios, totaling 140 hours; the scale of this dataset significantly surpasses that of previous audio-visual benchmarks while ensuring rich spatiotemporal and acoustic dynamics within each

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

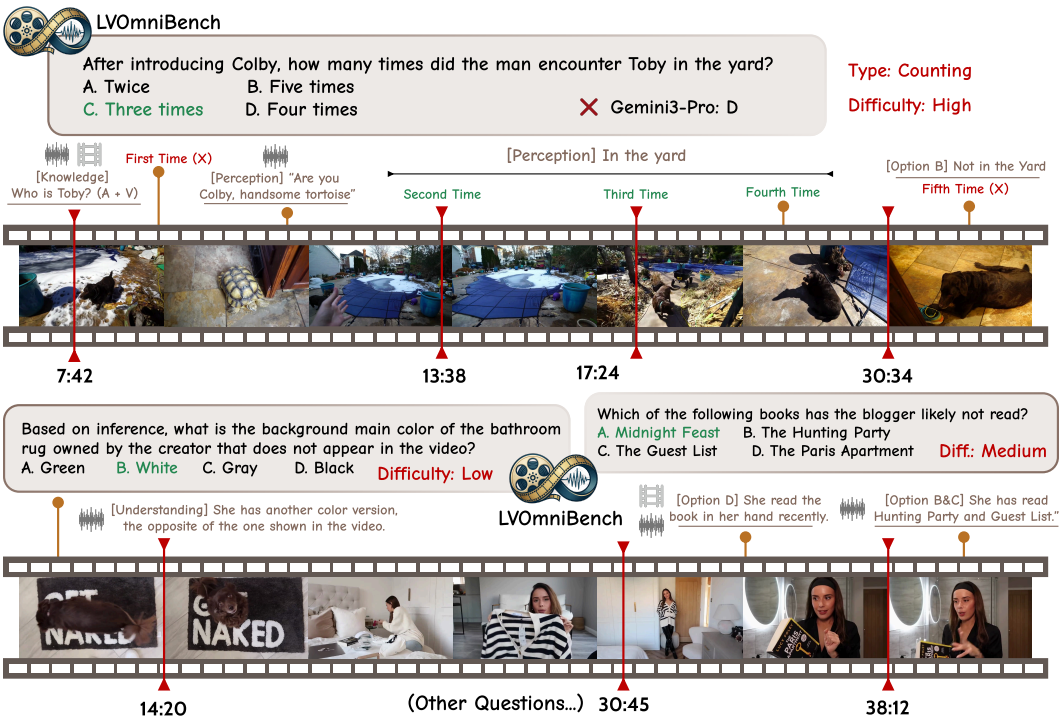

Figure 1: We introduce **LVOmniBench** to provide a rigorous assessment of the performance of OmniLLMs on *long* audio-visual inputs, comprising strictly manually curated videos and annotations. Each question is assigned a specific difficulty level to facilitate hierarchical evaluation of model performance. In the first example, the model should comprehend the entire audio-visual context. Solving this question initially requires cross-modal alignment to identify "Toby" and finally necessitates visual counting and scene recognition. Even SoTA models, such as Gemini 3 Pro, struggle to answer this question correctly. The second example presents questions at two additional difficulty tiers, demonstrating that a correct answer requires a combination of audio and visual.

video. In addition, to ensure broad generalization, the dataset encompasses a wide range of categories. We manually constructed 1,014 high-quality multiple-choice questions, which are explicitly designed to require joint reasoning across the audio and visual modalities, thereby facilitating a more comprehensive evaluation of OmniLLMs. Fig. 1 illustrates three representative question paradigms within the benchmark, which are categorized across varying levels of difficulty. As demonstrated, each task strictly necessitates cross-modal audio-visual reasoning. Furthermore, these examples highlight the inherent challenges of OmniLLMs when processing extended audio-visual inputs.

Based on extensive experiments and the rigorously constructed benchmark, the principal contributions and findings are summarized as follows:

- We introduce LVOmniBench, a new benchmark specifically designed to evaluate OmniLLMs for cross-modal comprehension of long-form audio-visual content, constructed through strictly manual video curation and annotation.
- We curated a diverse collection of long videos, with durations ranging from 10 to 90 minutes and an average duration of 2,069s. This duration represents a greater than sixfold increase in temporal scale compared to that of existing benchmarks for audio-visual understanding (as illustrated in Tab. 1).
- Within LVOmniBench, each question is classified into multiple types, including perception, understanding, inference, and complex logical reasoning, and is assigned one of three levels of difficulty, thereby allowing for a hierarchical evaluation of the performance of models.
- Experimental results demonstrate that the processing of long audio-visual sequences remains a significant challenge for OmniLLMs. Even the SoTA model, Gemini 3.8-Flash, achieves a peak accuracy of only 69%, whereas open-source counterparts struggle to surpass 40%, yielding a performance that is often marginally better than random chance.

Table 1: Comparison of different benchmarks. The modalities V, I, and A denote video, image, and audio, respectively. The open-domain metric indicates whether the video sources span a broad range of real-world scenarios. The annotation method (Anno.) specifies whether the dataset was curated manually (M) or automatically (A). We also report the video duration and the average temporal duration for each dataset, both measured in seconds. Multi-dimension (Multi-D) indicates whether the benchmark questions are hierarchically stratified by varying levels of cognitive difficulty, and multi-type indicates whether the problem involves multiple aspects.

Benchmark	Modality	Open Domain	Video Duration	Avg. Length	Anno.	Multi-Type	Multi-D	Answer Type
AVQA (Yang et al., 2022)	V+A	✗	10	10	M	✗	✗	MC
Music-AVQA (Li et al., 2022)	V+A	✗	60	60	M	✗	✗	CLS
AVTrustBench (Chowdhury et al., 2025)	V+A	✓	10\60	-	A&M	✗	✗	MC
AV-Odyssey (Gong et al., 2024)	I+A	✓	-	-	M	✓	✗	MC
AVHBench (Sung-Bin et al., 2024)	V+A	✓	10	10	A&M	✓	✗	CLS
OmniBench (Li et al., 2024c)	I+A	✓	-	-	M	✓	✗	MC
LongVALE (Geng et al., 2025)	A+V	✓	30–600	235	A&M	✗	✗	OE
AVUT (Yang et al., 2025b)	V+A	✓	-	67.8	A&M	✓	✗	MC
Daily-Omni (Zhou et al., 2025b)	V+A	✓	30–60	42.8	A&M	✓	✗	MC
WorldSense (Hong et al., 2025)	V+A	✓	15–656	141.1	M	✓	✗	MC
OmniVideoBench (Li et al., 2025b)	V+A	✓	4–1955	384.2	M	✓	✗	MC
JointAVBench (Chao et al., 2025)	V+A	✓	-	97.2	A&M	✓	✗	MC
LVOmniBench (Ours)	**V+A**	✓	**613–5482**	**2069.7 (↑ 6×)**	**M**	✓	✓	**MC**

These findings underscore the necessity for further improvements in the processing of extended audio-visual inputs and cross-modal alignment.

2 Related Work

2.1 Omnimodal Large Language Models

Research on multimodal large language models (MLLMs) (Chen et al., 2023; Liu et al., 2023; Li et al., 2025a; Bai et al., 2025b; Wang et al., 2024a; Awadalla et al., 2023; Team et al., 2025b; Ding et al., 2025a; Bai et al., 2025a; Chen et al., 2024b; Team et al., 2025a; Ge et al., 2025; Lin et al., 2024; Zhang et al., 2025a; Feng et al., 2025a) is transitioning from isolated single-modality perception toward omnimodal architectures capable of jointly processing text, images, video, and audio, thereby facilitating practical audio-visual comprehension in real-world scenarios. This trajectory is exemplified by recent advanced models designed to seamlessly process continuous video and audio streams to generate text and speech outputs (Xu et al., 2025a;b; Sun et al., 2024; Tang et al., 2025; Cheng et al., 2024; Fu et al., 2025b; Ye et al., 2025; Li et al., 2024b; Chen et al., 2025c; Yang et al., 2025a; Tong et al., 2025; Team et al., 2025c; AI et al., 2025; Liu et al., 2025c; Shu et al., 2025; Sun et al., 2025; Yao et al., 2024). Furthermore, the Gemini series serves as a strong baseline, distinguished by robust omnimodal understanding capabilities (Team et al., 2023; 2024; Comanici et al., 2025). Despite the proliferation of SoTA models, the evaluation of audio-visual comprehension predominantly focuses on short video clips or static images. Our experiments reveal that current models struggle with tasks requiring long-range temporal reasoning, highlighting an inadequacy in processing extended audio-visual inputs. Thus, there is an urgent need to develop benchmarks specifically tailored for the comprehension of long audio-visual content.

2.2 Multimodal Benchmarks

The rapid advancement of MLLMs has been significantly propelled by evaluation benchmarks. At present, evaluation benchmarks targeting unimodal comprehension, encompassing isolated image (Li et al., 2023; Hu et al., 2025; Antol et al., 2015; Hudson & Manning, 2019; Liu et al., 2024a; He et al., 2020; Marino et al., 2019; Lu et al., 2023; Fu et al., 2023; Yue et al., 2024; Feng et al., 2025c;b), video (Fu et al., 2025a; Wang et al., 2025a; Li et al., 2024a; Mangalam et al., 2023; Xiao et al., 2021; Yu et al., 2019; Wu et al., 2024; Zhou et al., 2025a; Maaz et al., 2023; Yuan et al., 2025; Song et al., 2024; Liu et al., 2024b), and audio understanding (Panayotov et al., 2015; Chen et al., 2020), are relatively mature. However, following the emergence of OmniLLMs, evaluating joint audio-visual reasoning poses a significant challenge. Most existing benchmarks are constrained to domain-specific evaluations or the processing of static images (Yang et al., 2022; Li et al., 2024c; Yuan et al., 2025; Li et al., 2022; Geng et al., 2025). Although recent advanced benchmarks, includ-

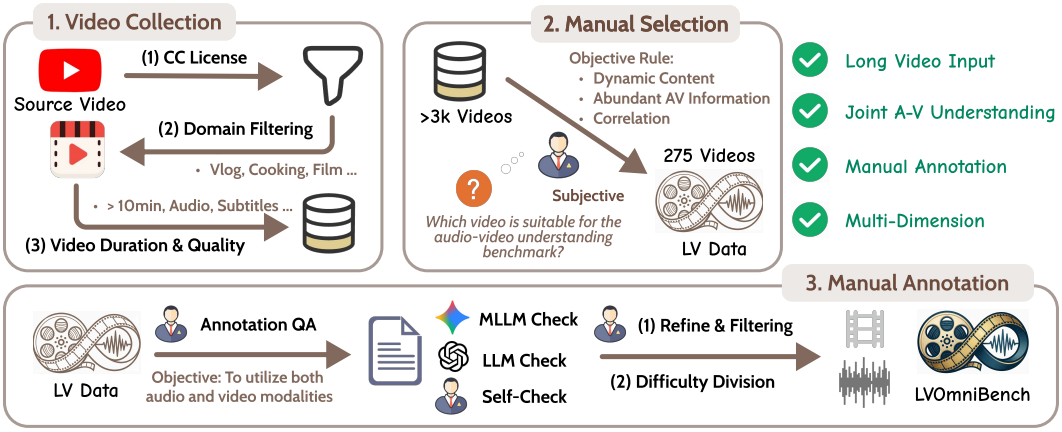

Figure 2: The construction of LVOmniBench follows a rigorous pipeline encompassing video collection, filtering, and question annotation. To guarantee both high-fidelity data quality and sufficient challenge for OmniLLMs, the videos and questions were manually curated and annotated, and all final quality-control decisions were made by human annotators.

ing WorldSense (Hong et al., 2025) and Daily-Omni (Zhou et al., 2025b), have been proposed, they predominantly focus on short video clips (Zhou et al., 2025b; Chao et al., 2025; Nguyen et al., 2025; Yang et al., 2025b; Sung-Bin et al., 2024). While OmniVideoBench (Li et al., 2025b) includes a subset of videos lasting 10 to 30 minutes, the vast majority of the dataset consists of videos lasting only a few minutes. This brief duration is misaligned with the lengths of videos typically encountered in real-world scenarios. Consequently, we introduce a new benchmark dedicated to the omnimodal understanding of long-form audio-visual inputs. The average duration of videos in LVOmniBench exceeds thirty minutes, which is six to twenty times longer than the durations found in previous benchmarks. Furthermore, the automated generation of questions using LLMs faces challenges in capturing complex, real-world reasoning requirements and remains prone to hallucination (Zhou et al., 2025b; Chao et al., 2025; Cao et al., 2025). To ensure the highest-quality evaluation, all videos and questions in LVOmniBench were manually selected and annotated by human experts.

3 LVOmniBench

3.1 Video Collection

To ensure dynamic audio-visual content and broad thematic coverage, we sourced videos from YouTube to establish a diverse corpus, as shown in Fig. 2. To ensure long-term accessibility and strictly comply with copyright regulations, all videos in LVOmniBench adhere to *Creative Commons* licenses. This guarantees that the benchmark remains open-source for the research community. To verify that the videos are well-suited for complex audio-visual reasoning tasks, we systematically categorized them into five broad domains: *entertainment, lifestyle, DIY & cooking, record, film & TV.* This process involved rigorous keyword-based screening and collection across 21 fine-grained subcategories (see Fig. 10 in the Appendix).

Subsequently, we applied strict length and quality controls, ultimately amassing an initial pool of more than 3,000 raw videos. Each video ranges from 10 to 90 minutes in duration and features a synchronized audio track. Unlike video-only benchmarks, not every video satisfies the stringent prerequisites for cross-modal audio-visual reasoning. Consequently, we meticulously filtered the initial pool to identify dynamic and informative content, curating a final set of 275 high-quality, long videos suitable for question annotation. As shown in Fig. 10, most video durations fall between 20 and 45 minutes, aligning with the typical length distribution of videos in real-world scenarios.

3.2 Question-Answer Annotation

To systematically evaluate long audio-visual comprehension, we first establish a comprehensive taxonomy of question types tailored to the capabilities of OmniLLMs (*Perception, Understanding, Inference, and Logical*). These categories aim to assess the proficiency of models in temporal fea-

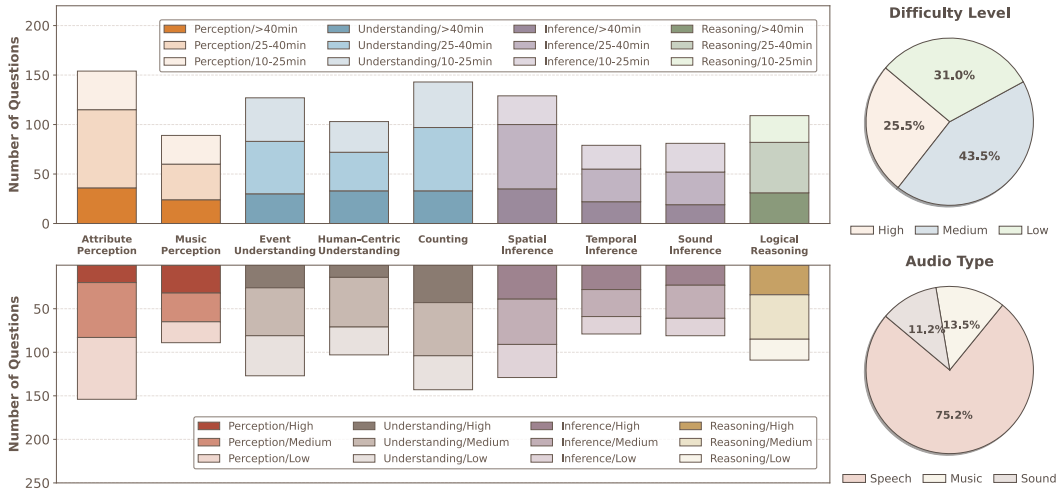

Figure 3: Distribution of questions. **Left:** LVOmniBench comprises nine question subcategories, with each demonstrating a balanced distribution across difficulty levels and video durations. **Right:** This panel illustrates the distribution of question difficulty and the corresponding audio types required to answer the questions.

ture alignment, fine-grained understanding, and complementary reasoning across multimodal inputs, including complex scenarios requiring the simultaneous application of multiple cognitive skills.

After selecting the videos and question types, we annotated the selected high-quality videos using a multiple-choice format. Specifically, each annotator generated between 1 and 20 questions per video; the exact number scaled according to the duration of the video and the density of relevant audio-visual events. Each question was formulated with four candidate options. To ensure the rigor of the benchmark, we enforced strict annotation guidelines: **(1)** Questions must necessitate joint audio-visual reasoning to prevent unimodal bias, and the correct option must be unambiguous. **(2)** Questions cannot be answered by relying solely on prior commonsense knowledge. Furthermore, the length of the four options was required to be uniform, and the distractors had to be directly derived from the video or audio. **(3)** Annotators were instructed to minimize the use of explicit timestamps in the prompts, ensuring that any temporal references provided did not offer trivial shortcuts to the solution. Consequently, this first round of annotation yielded more than 1,500 candidate QA pairs.

3.3 Question Refinement and Filtering

Following the initial annotation phase, we implemented a rigorous evaluation and filtering pipeline. First, we leveraged the Gemini model to conduct inference testing across three unimodal baselines: video-only, audio-only, and text-only. Based on the corresponding outputs, we required annotators to refine or delete QA pairs that could be answered *effectively* using a single modality, as this indicated a failure to properly integrate audio-visual cues during annotation. Additionally, we observed that certain annotators depended excessively on timestamps and explicit descriptions, inadvertently introducing unimodal bias and reducing the difficulty of temporal grounding and modality perception. Consequently, we rigorously refined or discarded such questions. Finally, after this comprehensive quality screening, we obtained 1,014 QA pairs as the final benchmark dataset. *The filtering of all annotated questions was done manually.*

Difficulty Level Annotation. To provide the community with a more informative hierarchical evaluation, we consider coarse attributes, such as video duration, question type, and audio modality, insufficient to characterize task difficulty or reliably assess model performance. Accordingly, we evaluate each QA pair along multiple dimensions, including perceptual difficulty, information granularity, temporal span, and inference complexity. We then assign each pair to one of three overall difficulty tiers: Low, Medium, or High. Harder questions require models to locate and integrate audiovisual evidence dispersed across longer temporal contexts, thereby directly assessing their capacity for complex long-video understanding. Question type characterizes model performance across tasks, whereas difficulty level captures the complexity of perceiving, locating, and reasoning over

Table 2: The performance of models on LVOmniBench. We evaluate OmniLLMs, utilizing both Video LLMs and Audio LLMs as specialized baselines. During inference, each model is evaluated using the maximum permissible number of input frames. The results demonstrate that long-form audio-visual comprehension remains a significant challenge for current SoTA models.

Model	Modality	Difficulty			Understanding	Perception	Inference	Logical	Avg.
		Low	Medium	High					
Proprietary Omnimodal LLMs									
Gemini-3.8-Flash	A + V	**81.2**	**70.9**	**52.9**	**76.1**	**65.5**	**69.6**	**69.7**	**69.5**
Gemini-3.5-Flash	A + V	76.1	70.3	42.5	71.3	60.6	65.1	67.0	65.0
Qwen3.5-Omni-Plus	A + V	63.1	52.4	42.1	55.7	56.5	48.8	46.8	53.1
Gemini-3.0-Pro	A + V	79.3	68.1	45.0	73.3	60.1	65.4	67.5	65.8
Gemini-3.0-Flash	A + V	76.6	63.0	31.0	67.7	54.7	60.6	51.8	59.0
Gemini-2.0-Flash	A + V	57.0	48.9	29.8	41.4	38.5	49.1	42.1	42.9
Gemini-3.0-Flash	V	55.6	49.3	30.6	51.4	42.9	42.3	48.5	46.2
Open-source Omnimodal LLMs									
Realtime-Venus-Omni-9B	A + V	**45.5**	37.9	32.4	**40.9**	**40.8**	39.4	27.5	38.9
Ming-Flash-Omni-2.0-100B	A + V	41.3	32.9	29.3	30.0	36.5	33.9	**39.1**	34.6
MiniCPM-o 4.5	A + V	43.4	34.1	25.1	35.7	31.9	39.1	32.7	34.8
Qwen3-Omni-30B	A + V	44.6	**38.6**	**34.8**	40.4	38.3	**42.2**	33.9	**39.5**
video-SALMONN 2+ 7B	A + V	40.9	30.2	26.7	30.0	36.3	30.3	31.8	32.7
Qwen2.5-Omni-7B	A + V	37.7	29.9	28.3	29.1	34.9	31.8	28.2	32.0
VideoLLaMA2-7B	A + V	27.0	26.8	28.2	23.9	28.2	29.4	24.6	27.2
Video LLMs (Visual)									
Qwen3-VL-30B	V	42.9	35.2	30.1	37.4	39.9	32.5	30.9	36.3
Qwen3-VL-8B	V	37.1	36.5	32.1	32.2	37.1	36.7	34.6	35.6
Audio LLMs (Audio)									
Qwen2-Audio	A	27.0	25.2	21.2	25.2	22.0	26.3	29.1	24.7

the required evidence. This distinction is intended to reveal model limitations that category-level averages may obscure.

3.4 Dataset Statistics and Comparison

As detailed in Tab. 1 and Tab. 4 in the Appendix, the proposed LVOmniBench contains 275 videos spanning five categories and 21 subclasses, with an average duration of 34 minutes and 29 seconds, which is 6-20 times longer than all previous benchmarks (Li et al., 2024c; Zhou et al., 2025b; Hong et al., 2025; Li et al., 2025b; Sung-Bin et al., 2024; Chao et al., 2025; Cao et al., 2025; Wang et al., 2025a). The dataset comprises 1,014 multiple-choice questions across nine categories, with an average question length of 16.4 words. As illustrated in Fig. 3, the distribution of requisite audio types for answering the questions, specifically speech, music, and sound, is 763:137:114, respectively, which reflects the prevalence of speech-driven interactions in real-world scenarios. Regarding the difficulty gradient, the distribution of low, medium, and high questions is 314:441:259, with all three tiers represented across all question categories. Overall, to the best of our knowledge, LVOmniBench is the first benchmark to comprehensively evaluate the comprehension of OmniLLMs in long-form audio-visual scenarios. We aim to catalyze future advancements in processing extended context lengths and joint audio-visual inputs. To this end, our benchmark establishes a robust foundation for the comprehensive assessment and in-depth analysis of OmniLLM capabilities on prolonged multimodal sequences.

4 Experimental Results

4.1 Models

To provide a comprehensive evaluation of OmniLLM performance on long-form audio-visual inputs, we benchmark several leading open-source models: Realtime-Venus-Omni-9B (Zhao et al., 2026), Ming-Flash-Omni-2.0-100B (AI et al., 2025), MiniCPM-o 4.5 (Yao et al., 2024), Qwen3-Omni-30B (Xu et al., 2025b), video-SALMONN 2+ 7B (Tang et al., 2025), Qwen2.5-Omni-7B (Xu et al., 2025a), and VideoLLaMA2-7B (Cheng et al., 2024). Furthermore, we evaluate the performance of the Video LLMs Qwen3-VL-8B and Qwen3-VL-30B (Bai et al., 2025a), alongside the Audio LLMs Qwen2-Audio (Chu et al., 2024). Finally, we incorporate Gemini (Team et al., 2023; 2024;

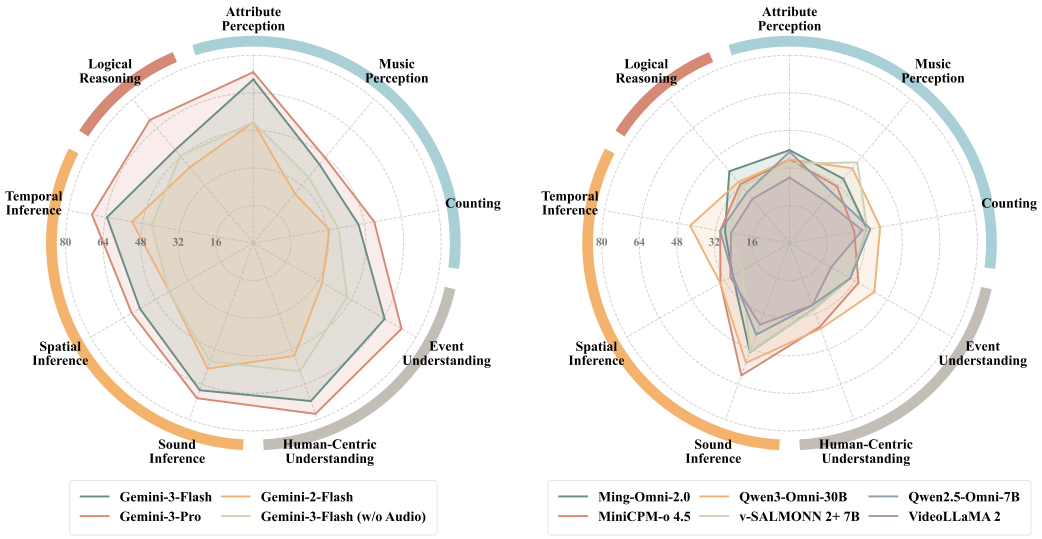

Figure 4: Comparison between proprietary models and open-source models on LVOmniBench across different tasks. The blue, red, orange, and gray in the outer circle stand for perception, logical, inference, and understanding tasks. As can be observed, proprietary models demonstrate a substantial performance advantage over their open-source counterparts. Furthermore, across all models, significant vulnerabilities remain in Counting and Music Perception.

Comanici et al., 2025) and Qwen3.5-Omni-Plus (Team, 2026) as proprietary baselines, leveraging their SoTA omnimodal comprehension capabilities.

4.2 Quantitative Results

Performance of Proprietary Models. As shown in Tab. 2, the Gemini series, currently the leading proprietary architecture in the field of omnimodal comprehension, achieves the highest overall accuracy on LVOmniBench. This superior performance is attributable to the capacity of these models to process ultra-long video contexts, alongside robust audio comprehension and precise temporal alignment. Analyzing performance across difficulty tiers, the overall distribution of accuracy closely aligns with our annotated difficulty gradients; notably, Gemini 3.0 Pro maintains an accuracy of 45% even on high-difficulty questions. This consistency validates the effectiveness of our benchmark for rigorous hierarchical evaluation. Furthermore, as shown in Fig. 4, a granular analysis categorized by question type reveals that tasks requiring counting and music perception remain exceptionally challenging. Finally, as depicted in Fig. 5, the capacity of models to perceive non-verbal and abstract sounds within the music category is significantly lower than the performance of these models on the other two audio types. These findings highlight that robust cross-modal alignment for non-linguistic audio classes remains an urgent challenge. Additional error analyses are provided in the Appendix A.

Performance of Open-Source Models. Qwen3-Omni-30B achieves the highest accuracy of 39.5%, while most other open-source models fall below the 35% threshold. As detailed in Tab. 2, this starkly illustrates that current open-source architectures struggle to process and comprehend long-form audio-visual inputs effectively. On high-difficulty tasks, the performance of these models degrades to near-random chance. Furthermore, Fig. 4 highlights significant performance variance across distinct task categories. For instance, Qwen3-Omni excels in temporal inference, Ming-Flash-Omni-2.0 demonstrates superior logical reasoning capabilities, and MiniCPM-o 4.5 exhibits relatively better audio perception. Additionally, we observe that accuracy scores in the range of 30% to 35% largely reflect models relying on superficial visual priors rather than genuinely achieving cross-modal comprehension. The severe computational overhead introduced by the processing of extended temporal contexts remains a primary bottleneck limiting the capabilities of open-source models. Thus, OmniLLMs still require substantial optimization to bridge this performance gap.

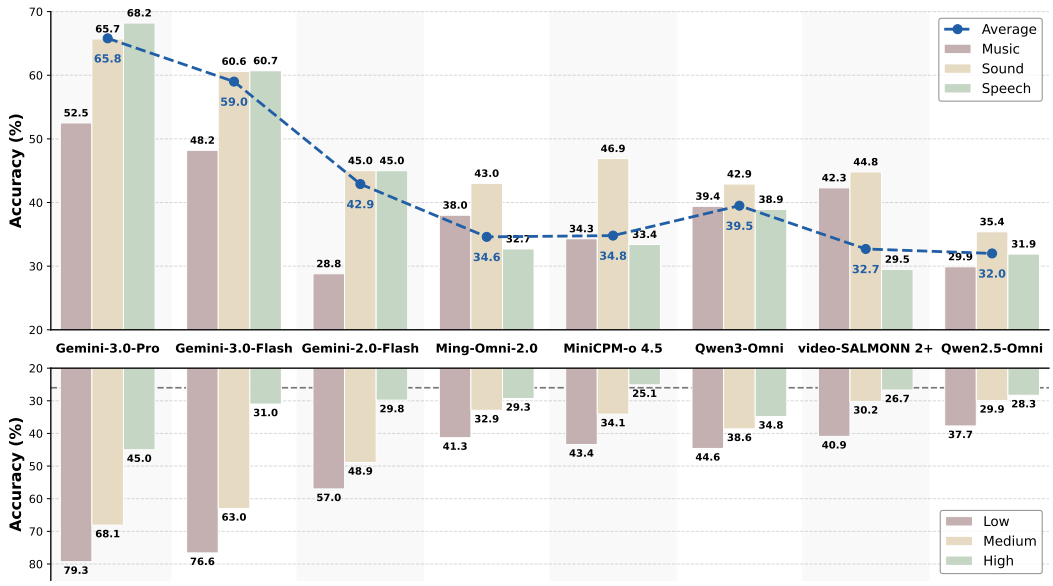

Figure 5: Results across audio type and difficulty levels. The findings demonstrate that proprietary models consistently outperform open-source models across all evaluated dimensions. OmniLLMs generally face greater challenges in perceiving musical audio; specifically, at high difficulty levels, the performance of open-source models declines to a level approaching random chance.

Effectiveness of Omnimodal Understanding. The performance of the unimodal baselines, specifically Qwen3-Omni (audio-only) and Qwen2-Audio, underscores the necessity of joint visual and audio comprehension within LVOmniBench, as illustrated in Tab. 2 and Tab. 3(a). In addition, the video-only performance of models on LVOmniBench was evaluated; for instance, Qwen3-VL, despite its specialized optimization for long-video understanding (Bai et al., 2025a; Fu et al., 2025a; Wang et al., 2025a; Wu et al., 2024), achieves only 36% accuracy. This performance represents a substantial decline relative to its results on prior benchmarks that focused exclusively on long-video understanding. Furthermore, as detailed in Tabs. 2 and 3, restricting Gemini 3 Flash to visual-only reasoning results in a significant accuracy reduction of up to 13%; this ablation study clearly validates the essential role of audio cues in the proposed benchmark.

4.3 Key Takeaways

Decisive Lead of Proprietary Models. As indicated in Tab. 2 and Figs. 4 and 5, proprietary models consistently outperform open-source baselines across all facets of long-form audio-visual comprehension, spanning both individual tasks and various audio categories. This divergence highlights a critical distinction from video-only comprehension. While advanced methodologies and frameworks have rendered open-source models highly competitive in purely long-video tasks, the inherent *complexity of joint audio-visual processing* and the rigorous demands of *fine-grained cross-modal alignment* reveal a substantial performance gap. Consequently, it is expected that LVOmniBench will catalyze future research dedicated to addressing these cross-modal challenges.

Enhancing Cross-Modal Capabilities of OmniLLMs. Tab. 3 provides a detailed performance breakdown across a suite of OmniLLMs. It is observed that most models struggle to effectively utilize or comprehend the audio modality, raising a critical question:

Can the open-source models genuinely use the audio in long videos?

The analysis suggests **no**. To rigorously investigate this hypothesis, the audio tracks were transcribed using Gemini 2.5 to obtain automatic speech recognition (ASR) transcripts. By providing this ASR text to the model as an explicit supplementary cue alongside the visual stream, significant improvements were observed for Qwen3-Omni and Ming-Flash-Omni-2.0. These results indicate that, for long audio-video inputs, open-source models typically truncate audio inputs by default due to context-length limitations, preventing them from effectively leveraging information contained in longer audio segments and thereby leading to the observed degradation in accuracy. Overall, achiev-

Table 3: The impact of audio and video information. **(a)** The evaluation of performance across various capabilities. The abbreviations are defined as follows: **AP**: Attribute Perception, **EU**: Event Understanding, **HU**: Human-Centric Understanding, **LR**: Logical Reasoning, **MP**: Music Perception, **SI**: Sound Inference, **SpI**: Spatial Inference, and **TI**: Temporal Inference. **(b)** Experiments are conducted across three input settings: video-only, video with ASR, and video with original audio.

Model	Modality	AP	Counting	EU	HU	LR	MP	SI	SpI	TI	Avg.
Qwen3-Omni-30B	Audio	27.9	23.1	24.4	26.2	28.2	36.1	35.8	26.4	31.7	28.1
	+Video	35.7	39.2	41.7	38.8	33.9	41.6	54.3	34.1	43.0	$39.5_{+11.4}$
Gemini-3.0-Flash	A + V	69.7	45.5	64.6	71.6	51.8	43.8	66.7	55.8	63.8	59.0

(a) Impact of Video Information

Model	Speech			Sound			Music			Overall		
	Video	+ASR	+Audio	Video	+ASR	+Audio	Video	+ASR	+Audio	Video	+ASR	+Audio
Gemini-3-Flash	46.5	$54.2_{+7.7}$	$60.7_{+14.2}$	53.5	$48.8_{-4.7}$	$60.6_{+7.1}$	38.8	$44.0_{+5.2}$	$48.2_{+9.4}$	46.2	$52.0_{+5.8}$	$59.0_{+12.8}$
Ming-Omni-2.0-100B	32.6	$40.8_{+8.2}$	$32.7_{+0.1}$	35.1	$44.8_{+9.7}$	$43.0_{+7.9}$	37.9	$36.0_{-1.9}$	$38.0_{+0.1}$	33.6	$40.5_{+6.9}$	$34.6_{+1.0}$
MiniCPM-o 4.5	32.8	$33.7_{+0.9}$	$33.4_{+0.6}$	35.4	$28.6_{-6.8}$	$46.9_{+11.5}$	32.8	$27.1_{-5.7}$	$34.3_{+1.5}$	33.1	$32.3_{-0.8}$	$34.8_{+1.7}$
Qwen3-Omni-30B	34.4	$42.3_{+7.9}$	$38.9_{+4.5}$	34.2	$50.1_{+15.9}$	$42.9_{+8.7}$	40.1	$36.0_{-4.1}$	$39.4_{-0.7}$	35.1	$42.2_{+7.1}$	$39.5_{+4.4}$
VideoLLaMA2-7B	27.7	–	$27.1_{-0.6}$	37.7	–	$31.6_{-6.1}$	24.1	–	$24.1_{+0.0}$	28.4	–	$27.2_{-1.2}$

(b) Impact of Audio Information

ing cross-modal alignment between long audio and video within OmniLLMs remains a formidable challenge, representing a critical bottleneck that warrants further exploration.

Impact of Audio Input. In contrast, Gemini 3.0 Flash demonstrates robust multimodal processing; supplementing the visual stream with ASR text does not yield improvements in accuracy over the original raw audio. This suggests that raw audio encapsulates critical non-verbal signals, such as emotion, intonation, and musicality, that extend beyond mere textual semantics. These findings underscore the need to preserve comprehensive acoustic cues for better omnimodal understanding in real-world video scenarios. This discrepancy not only highlights the superior performance of proprietary models but also validates the *effectiveness* of our benchmark design.

Long Audio-Video Understanding Capabilities. We observe that as the duration of audio and video inputs increases, model performance degrades significantly compared to that on short-video benchmarks (Zhou et al., 2025b; Hong et al., 2025). In the context of long-form video understanding, numerous methodologies have been proposed to mitigate the information bottleneck posed by processing comprehensive video content with a limited number of input frames. These approaches encompass training-free token compression strategies (Tao et al., 2025a; Shao et al., 2025a; Liu et al., 2025a; Shao et al., 2025b; Shen et al., 2025a), dynamic input resolutions (Bai et al., 2025b;a), streaming understanding (Di et al., 2025; Xu et al., 2025c; Yao et al., 2025; Chen et al., 2025b), and specialized model architectures (Shu et al., 2024; Liu et al., 2025a; Shen et al., 2025b) and agents (Zhang et al., 2025c; 2024; Fan et al., 2024; Wang et al., 2024b; 2025c; Ding et al., 2025b) tailored for long videos. However, within the realm of joint audio-visual comprehension, this bottleneck becomes particularly pronounced (Tao et al., 2025b; Ding et al., 2026). As an extended temporal sequence, audio inherently possesses a high information density and exhibits strict sequential dependencies. Consequently, we believe that the primary limitation in long-form audio-visual comprehension currently stems from the challenges of modeling long audio sequences. Therefore, enhancing the capabilities for long-form audio-visual modeling is imperative for the next generation of OmniLLMs to effectively comprehend extended-duration real-world video inputs.

5 Conclusion

In this paper, we introduce LVOmniBench, a pioneering benchmark designed for long audio-video understanding. The benchmark features diverse long-form videos with dynamic multimodal content and manually annotated QA pairs, aiming to comprehensively evaluate the capabilities of OmniLLMs in complex joint audio-visual understanding. Extensive evaluations reveal a substantial performance disparity between proprietary and open-source models, underscoring the need for advanced architectures capable of jointly processing extended audio and video sequences. We hope that LVOmniBench will catalyze future research and drive the development of more robust and comprehensive OmniLLMs.

AI USE STATEMENT

In this work, we used generative AI tools to support dataset quality control, evaluation code, and grammar review. Specifically, a Gemini model was used to run video-only, audio-only, and text-only inference baselines (Sec. 3.3). All final filtering decisions were made *manually*, and *LLMs were not used to generate the benchmark questions, answer choices, or ground-truth labels.* We additionally used LLMs to inspect the evaluation code for possible implementation errors and to check the manuscript for grammar, clarity, and readability. All code-related suggestions were reviewed by the authors and accepted only after manual inspection and testing. We did not use generative AI tools to develop theoretical models, formulate mathematical claims or proofs, design the research methodology, or make the final interpretation of the results; the remaining required-disclosure tasks do not apply to this work. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, et al. Ming-omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*, 2025.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *CVPR*, 2015.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Yuqin Cao, Xiongkuo Min, Yixuan Gao, Wei Sun, Zicheng Zhang, Jinliang Han, and Guangtao Zhai. Xgc-avis: Towards audio-visual content understanding with a multi-agent collaborative system. *arXiv preprint arXiv:2509.23251*, 2025.

Jianghan Chao, Jianzhang Gao, Wenhui Tan, Yuchong Sun, Ruihua Song, and Liyun Ru. Jointavbench: A benchmark for joint audio-visual reasoning evaluation. *arXiv preprint arXiv:2512.12772*, 2025.

Chen Chen, ZeYang Hu, Fengjiao Chen, Liya Ma, Jiaxing Liu, Xiaoyu Li, Ziwen Wang, Xuezhi Cao, and Xunliang Cai. Uno-bench: A unified benchmark for exploring the compositional law between uni-modal and omni-modal in omni models. *arXiv preprint arXiv:2510.18915*, 2025a.

Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, 2023.

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

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *ECCV*, 2024a.

Xueyi Chen, Keda Tao, Kele Shao, and Huan Wang. Streamingtom: Streaming token compression for efficient video understanding. *arXiv preprint arXiv:2510.18269*, 2025b.

Yijing Chen, Yihan Wu, Kaisi Guan, Yuchen Ren, Yuyue Wang, Ruihua Song, and Liyun Ru. Chronusomni: Improving time awareness of omni large language models. *arXiv preprint arXiv:2512.09841*, 2025c.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024b.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Yaoting Wang, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Avtrustbench: Assessing and enhancing reliability and robustness in audio-visual llms. In *CVPR*, 2025.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, 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.

Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, Hao Jiang, et al. Streaming video question-answering with in-context video kv-cache retrieval. In *ICLR*, 2025.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025a.

Yang Ding, Yizhen Zhang, Xin Lai, Ruihang Chu, and Yujiu Yang. Videozoomer: Reinforcement-learned temporal focusing for long video reasoning. *arXiv preprint arXiv:2512.22315*, 2025b.

Yue Ding, Yiyan Ji, Jungang Li, Xuyang Liu, Xinlong Chen, Junfei Wu, Bozhou Li, Bohan Zeng, Yang Shi, Yushuo Guan, et al. Omnisift: Modality-asymmetric token compression for efficient omni-modal large language models. *arXiv preprint arXiv:2602.04804*, 2026.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *ECCV*, 2024.

Sicheng Feng, Gongfan Fang, Xinyin Ma, and Xinchao Wang. Efficient reasoning models: A survey. *arXiv preprint arXiv:2504.10903*, 2025a.

Sicheng Feng, Kaiwen Tuo, Song Wang, Lingdong Kong, Jianke Zhu, and Huan Wang. Rewardmap: Tackling sparse rewards in fine-grained visual reasoning via multi-stage reinforcement learning. *arXiv preprint arXiv:2510.02240*, 2025b.

Sicheng Feng, Song Wang, Shuyi Ouyang, Lingdong Kong, Zikai Song, Jianke Zhu, Huan Wang, and Xinchao Wang. Can mllms guide me home? a benchmark study on fine-grained visual reasoning from transit maps. *arXiv preprint arXiv:2505.18675*, 2025c.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Yuhang Dai, Meng Zhao, Yi-Fan Zhang, Shaoqi Dong, Yangze Li, Xiong Wang, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *CVPR*, 2025a.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025b.

Yuying Ge, Yixiao Ge, Chen Li, Teng Wang, Junfu Pu, Yizhuo Li, Lu Qiu, Jin Ma, Lisheng Duan, Xinyu Zuo, et al. Arc-hunyuan-video-7b: Structured video comprehension of real-world shorts. *arXiv preprint arXiv:2507.20939*, 2025.

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *CVPR*, 2025.

Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, et al. Av-odyssey bench: Can your multimodal llms really understand audio-visual information? *arXiv preprint arXiv:2412.02611*, 2024.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*, 2025.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *CVPR*, 2024.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

Xin Jin, Siyuan Li, Siyong Jian, Kai Yu, and Huan Wang. Mergemix: A unified augmentation paradigm for visual and multi-modal understanding. *arXiv preprint arXiv:2510.23479*, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *TMLR*, 2025a.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Jiafu Tang, Zhenghao Song, Dingling Zhang, et al. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. *arXiv preprint arXiv:2510.10689*, 2025b.

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

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, 2024a.

Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, Song Chen, Xu Li, Da Pan, Shusen Zhang, Xin Wu, Zheng Liang, Jun Liu, Tao Zhang, Keer Lu, Yaqi Zhao, Yanjun Shen, Fan Yang, Kaicheng Yu, Tao Lin, Jianhua Xu, Zenan Zhou, and Weipeng Chen. Baichuan-omni technical report. *arXiv preprint arXiv:2410.08565*, 2024b.

Yizhi Li, Yinghao Ma, Ge Zhang, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, et al. Omnibench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272*, 2024c.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *EMNLP*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Xiangrui Liu, Yan Shu, Zheng Liu, Ao Li, Yang Tian, and Bo Zhao. Video-xl-pro: Reconstructive token compression for extremely long video understanding. *arXiv preprint arXiv:2503.18478*, 2025a.

Xuyang Liu, Ziming Wang, Junjie Chen, Yuhang Han, Yingyao Wang, Jiale Yuan, Jun Song, Siteng Huang, and Honggang Chen. Global compression commander: Plug-and-play inference acceleration for high-resolution large vision-language models. *arXiv preprint arXiv:2501.05179*, 2025b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2024a.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of ACL*, 2024b.

Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment. *arXiv preprint arXiv:2502.04328*, 2025c.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Jie Ma, Zhitao Gao, Qi Chai, Jun Liu, Pinghui Wang, Jing Tao, and Zhou Su. Fortisavqa and maven: a benchmark dataset and debiasing framework for robust multimodal reasoning. *arXiv preprint arXiv:2504.00487*, 2025.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL*, 2023.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*, 2023.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019.

Le Thien Phuc Nguyen, Zhuoran Yu, Samuel Low Yu Hang, Subin An, Jeongik Lee, Yohan Ban, SeungEun Chung, Thanh-Huy Nguyen, JuWan Maeng, Soochahn Lee, et al. See, hear, and understand: Benchmarking audiovisual human speech understanding in multimodal large language models. *arXiv preprint arXiv:2512.02231*, 2025.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *ICASSP*, 2015.

Mengxue Qu, Xiaodong Chen, Wu Liu, Alicia Li, and Yao Zhao. Chatvtg: Video temporal grounding via chat with video dialogue large language models. In *CVPR*, 2024.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *ICCV*, 2025.

Kele Shao, Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Holitom: Holistic token merging for fast video large language models. *arXiv preprint arXiv:2505.21334*, 2025a.

Kele Shao, Keda Tao, Kejia Zhang, Sicheng Feng, Mu Cai, Yuzhang Shang, Haoxuan You, Can Qin, Yang Sui, and Huan Wang. When tokens talk too much: A survey of multimodal long-context token compression across images, videos, and audios. *arXiv preprint arXiv:2507.20198*, 2025b.

Kele Shao, Keda Tao, Kejia Zhang, Sicheng Feng, Mu Cai, Yuzhang Shang, Haoxuan You, Can Qin, Yang Sui, and Huan Wang. When tokens talk too much: A survey of multimodal long-context token compression across images, videos, and audios. *arXiv preprint arXiv:2507.20198*, 2025c.

Leqi Shen, Guoqiang Gong, Tao He, Yifeng Zhang, Pengzhang Liu, Sicheng Zhao, and Guiguang Ding. Fastvid: Dynamic density pruning for fast video large language models. *arXiv preprint arXiv:2503.11187*, 2025a.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. In *ICML*, 2025b.

Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. Audio-visual llm for video understanding. In *CVPR*, 2025.

Yan Shu, Peitian Zhang, Zheng Liu, Minghao Qin, Junjie Zhou, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv:2409.14485*, 2024.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *CVPR*, 2024.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-salmonn: Speech-enhanced audio-visual large language models. *arXiv preprint arXiv:2406.15704*, 2024.

Wei Sun, Linhan Cao, Yuqin Cao, Weixia Zhang, Wen Wen, Kaiwei Zhang, Zijian Chen, Fangfang Lu, Xiongkuo Min, and Guangtao Zhai. Engagement prediction of short videos with large multimodal models. In *CVPR*, 2025.

Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. *arXiv preprint arXiv:2410.18325*, 2024.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn 2: Captioning-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025.

Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Dycoke: Dynamic compression of tokens for fast video large language models. In *CVPR*, 2025a.

Keda Tao, Kele Shao, Bohan Yu, Weiqiang Wang, Huan Wang, et al. Omnizip: Audio-guided dynamic token compression for fast omnimodal large language models. *arXiv preprint arXiv:2511.14582*, 2025b.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025a.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025b.

Meituan LongCat Team, Bairui Wang, Bin Xiao, Bo Zhang, Bolin Rong, Borun Chen, Chang Wan, Chao Zhang, Chen Huang, Chen Chen, et al. Longcat-flash-omni technical report. *arXiv preprint arXiv:2511.00279*, 2025c.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Wenwen Tong, Hewei Guo, Dongchuan Ran, Jiangnan Chen, Jiefan Lu, Kaibin Wang, Keqiang Li, Xiaoxu Zhu, Jiakui Li, Kehan Li, et al. Interactiveomni: A unified omni-modal model for audio-visual multi-turn dialogue. *arXiv preprint arXiv:2510.13747*, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *CVPR*, 2025a.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *ECCV*, 2024b.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. *arXiv preprint arXiv:2503.13377*, 2025b.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *CVPR*, 2025c.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *NeurIPS*, 2024.

Jianlong Wu, Wei Liu, Ye Liu, Meng Liu, Liqiang Nie, Zhouchen Lin, and Chang Wen Chen. A survey on video temporal grounding with multimodal large language model. *TPAMI*, 2025.

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

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. Streamingvlm: Real-time understanding for infinite video streams. *arXiv preprint arXiv:2510.09608*, 2025c.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *ACM MM*, 2022.

Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. Humanomniv2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*, 2025a.

Yudong Yang, Jimin Zhuang, Guangzhi Sun, Changli Tang, Yixuan Li, Peihan Li, Yifan Jiang, Wei Li, Zejun Ma, and Chao Zhang. Audio-centric video understanding benchmark without text shortcut. In *EMNLP*, 2025b.

Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, Yuanxin Liu, Kun Ouyang, Lean Wang, Shicheng Li, Sida Li, Lingpeng Kong, Qi Liu, Yuanxing Zhang, and Xu Sun. Timechat-online: 80*arXiv preprint arXiv:2504.17343*, 2025.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *arXiv preprint arXiv:2510.15870*, 2025.

Yufei Yin, Qianke Meng, Minghao Chen, Jiajun Ding, Zhenwei Shao, and Zhou Yu. Videoarm: Agentic reasoning over hierarchical memory for long-form video understanding. *arXiv preprint arXiv:2512.12360*, 2025.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, 2019.

Huaying Yuan, Zheng Liu, Junjie Zhou, Ji-Rong Wen, and Zhicheng Dou. Videodeepresearch: Long video understanding with agentic tool using. *arXiv preprint arXiv:2506.10821*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, et al. Timesuite: Improving mllms for long video understanding via grounded tuning. *arXiv preprint arXiv:2410.19702*, 2024.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, Ying Shan, and Limin Wang. Timelens: Rethinking video temporal grounding with multimodal llms. *arXiv preprint arXiv:2512.14698*, 2025b.

Lu Zhang, Tiancheng Zhao, Heting Ying, Yibo Ma, and Kyusong Lee. Omagent: A multi-modal agent framework for complex video understanding with task divide-and-conquer. *arXiv preprint arXiv:2406.16620*, 2024.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. *arXiv preprint arXiv:2505.18079*, 2025c.

Yiman Zhang, Ziheng Luo, Qiangyu Yan, Wei He, Borui Jiang, Xinghao Chen, and Kai Han. Omnieval: A benchmark for evaluating omni-modal models with visual, auditory, and textual inputs. *arXiv preprint arXiv:2506.20960*, 2025d.

Ruixiang Zhao, Hualei Wang, Renhe Sun, Enzhi Zhou, Jincenzi Wu, Xujie Song, Kexin Shi, Zihang Liu, Pengcheng Zhu, Jiayi Zhou, et al. Realtime-venus: A full-duplex interaction system with asynchronous delegation. *arXiv preprint arXiv:2609.13814*, 2026.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *CVPR*, 2025a.

Ziwei Zhou, Rui Wang, and Zuxuan Wu. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025b.

SUPPLEMENTARY MATERIAL

A ERROR ANALYSIS

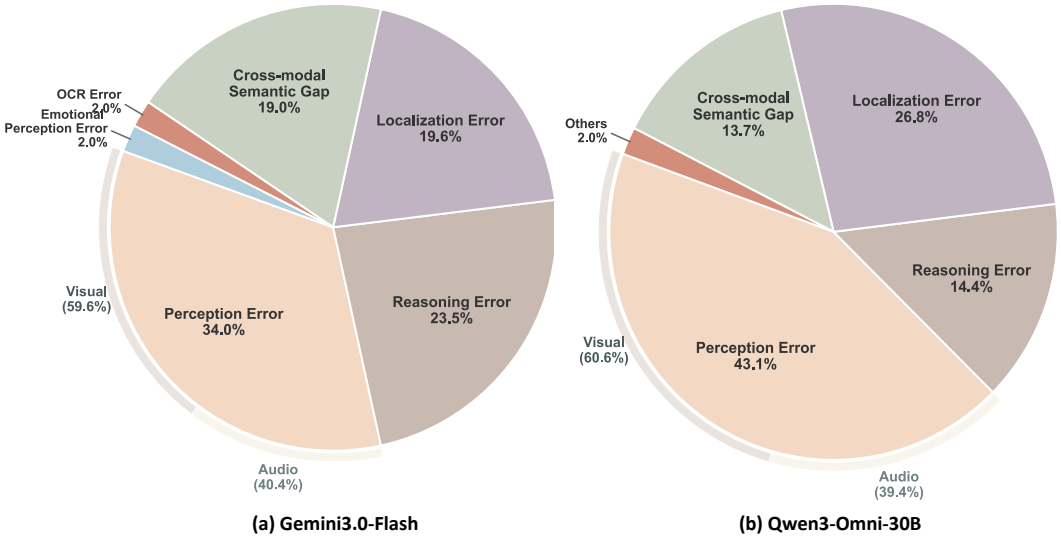

Figure 6: Error distributions for 153 annotated errors made by (a) Gemini 3 Flash and (b) Qwen3-Omni-30B on LVOmniBench. Compared with Gemini 3 Flash, the open-source model exhibits a higher proportion of errors in fundamental capabilities.

In this section, we aim to provide an in-depth analysis of model predictions and failure modes to inform future research directions. Generally, OmniLLMs exhibit notable limitations in interpreting abstract audio representations, such as music, and struggle with tasks requiring precise counting or spatial reasoning. We manually analyzed 153 incorrect predictions from Gemini 3 Flash, categorizing the root causes to provide a comprehensive statistical breakdown and discussion, as illustrated in Fig. 6 (a).

Perception Errors (34%). Perception errors are broadly categorized into audio and visual modalities. This categorization assesses the ability of models to comprehend fundamental audio and video information. Concerning audio perception failures, we observe a prevailing modality bias: models disproportionately rely on visual cues while *neglecting corresponding audio signals*, which ultimately results in incorrect predictions. Furthermore, accurately interpreting fine-grained acoustic properties, such as sound intensity, timbre, and pitch, poses a significant challenge for current models. Conversely, visual perception errors predominantly stem from a lack of spatial reasoning and detailed comprehension, as models frequently struggle to identify object counts, spatial directions, and visual attributes.

Localization Errors (19.6%). Given that long-form video and audio inputs span tens of minutes, precise temporal grounding is paramount. While numerous prior studies have focused extensively on video-only temporal grounding (Wang et al., 2025b; Wu et al., 2025; Zhang et al., 2025b; Zeng et al., 2024; Qu et al., 2024; Huang et al., 2024), the joint localization of audio and video modalities remains largely underexplored. By highlighting this critical gap, we aim to inspire future research toward unified omnimodal temporal localization.

Cross-Modal Semantic Gap (19%). Models frequently process the two modalities in isolation, which makes it highly challenging to fuse and align their complementary information in a manner consistent with human cognition. Although models can often independently comprehend isolated audio and video streams, they struggle with cross-modal correspondence. Consequently, they exhibit suboptimal performance in tasks that require fine-grained alignment, such as human-centered understanding and attribute perception.

Reasoning Errors (23.5%). Even when a model successfully perceives both modalities and their cross-modal alignment, it frequently struggles with logical deduction and mathematical reasoning.

Furthermore, deciphering complex temporal dynamics, such as action sequencing and event causality, alongside spatial relationships extracted from audio-visual cues, remains a significant challenge. Consequently, moving beyond fundamental perception and enhancing reasoning capabilities is critical to addressing these limitations.

Other Errors and Summary. The remaining error categories encompass OCR (2%) and emotional perception errors (2%), stemming from the limited capacity of models to extract embedded textual data and to decipher implicit affective cues within audio streams.

Open-Source Models vs. Proprietary Models. For a more comprehensive assessment, we analyze the errors produced by Qwen3-Omni-30B under the same setting, as shown in Fig. 6 (b). The results indicate that the evaluated open-source model makes errors more frequently in fundamental capabilities, particularly perception and localization. Because higher-level logical and cross-modal reasoning depends on these capabilities, fewer errors are attributed to these subsequent reasoning processes. This error distribution further demonstrates the performance gap between the evaluated open-source model and closed-source models and provides clearer guidance for future improvements in model capabilities.

Overall, OmniLLMs encounter significantly greater bottlenecks in long-form audio-visual comprehension compared to those observed in unimodal or short-video benchmarks. Through this exhaustive analysis, we aim to inspire future architectural and algorithmic enhancements for OmniLLMs.

B Error Analysis Cases

In this section, we present representative cases across various error types to illustrate the challenges models face in contemporary audio-visual understanding. Specifically, Fig. 7 depicts perception errors across both audio and visual modalities, while Fig. 8 illustrates localization errors. Furthermore, Fig. 9 demonstrates the cross-modal semantic gap and highlights reasoning errors. Overall, enhancing the capabilities of OmniLLMs necessitates improvements in long-form audio-visual comprehension, ranging from basic perception and localization to advanced reasoning and cross-modal understanding. Consequently, substantial room for improvement remains, and LVOmniBench aims to provide valuable insights and guidance for future research.

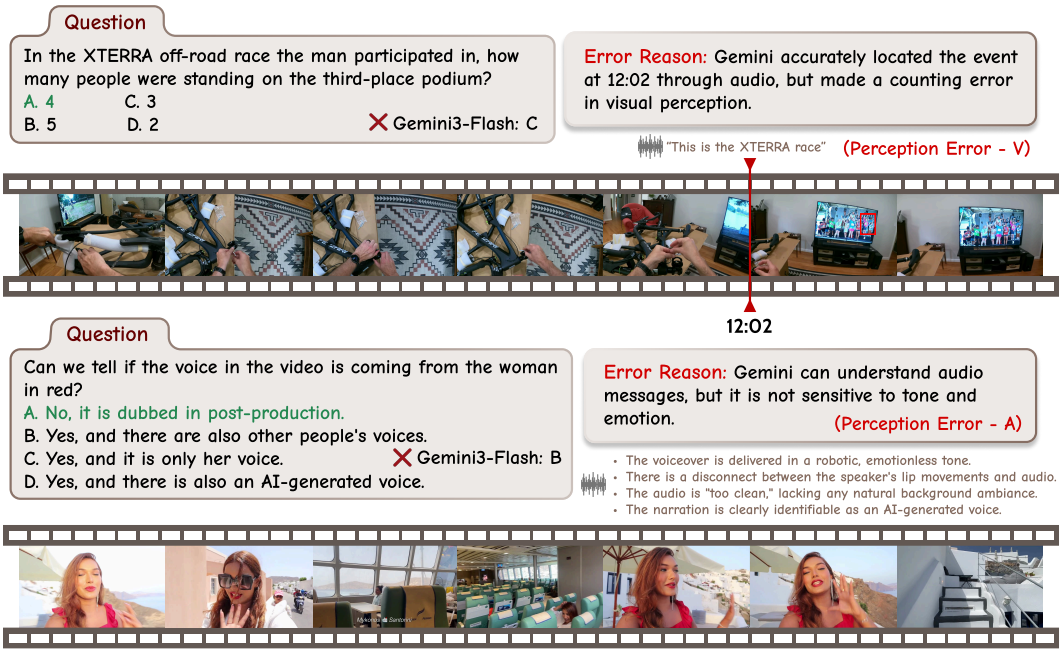

Figure 7: Cases of perception errors. We present examples of failures that arise from the misperception of underlying visual or audio information by models.

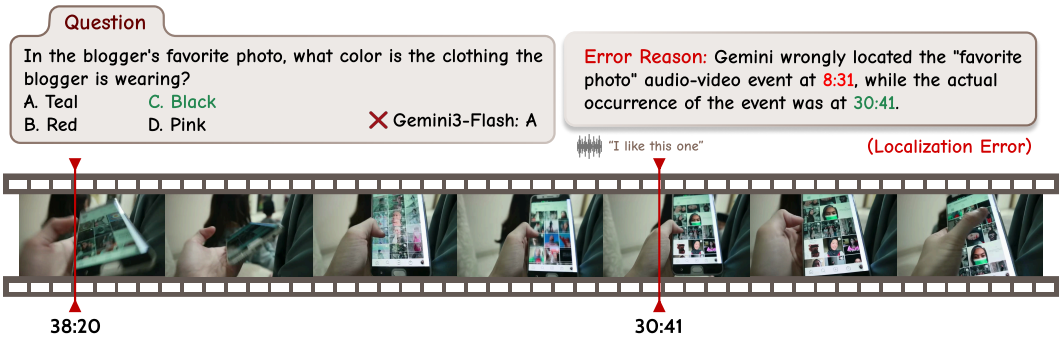

Figure 8: Cases of localization errors. It is observed that models are unable to accurately locate the temporal occurrences of events within long-form audio and video inputs, leading to incorrect predictions.

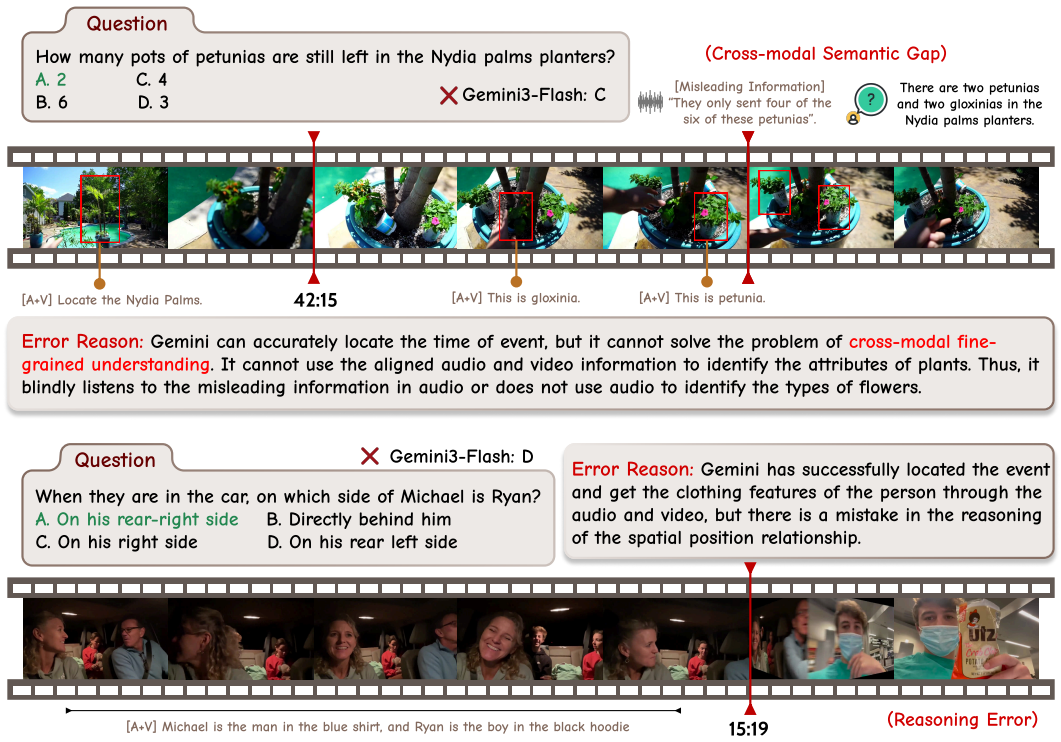

Figure 9: Cases of cross-modal semantic gaps and reasoning errors. For complex problems, models fail to fully exploit audio and video information; instead, they often over-rely on a single modality during reasoning, hindering the alignment of cross-modal features. Addressing this limitation represents a critical direction for future enhancements to OmniLLMs. Furthermore, models frequently exhibit errors in reasoning tasks, such as spatial localization.

C More Insights and Future Work

The Efficiency of Long-Context Capabilities. In the realm of long-video understanding, excessively long contextual token sequences pose a significant challenge, as they complicate model inference and incur substantial computational overhead. While numerous established strategies have effectively addressed the long-sequence bottleneck for the video modality (Chen et al., 2024a; Liu et al., 2025b; Tao et al., 2025a; Shen et al., 2025b; Shao et al., 2025c;a; Tao et al., 2025b; Yin et al., 2025; Ding et al., 2025b; Yao et al., 2025; Shang et al., 2025; Liu et al., 2025a; Ding et al., 2026; Chen et al., 2025b; Jin et al., 2025), this issue persists for the audio modality. Specifically, audio presents a more formidable compression challenge due to the strict temporal continuity of its signals. Consequently, the development of efficient processing mechanisms for joint long audio-video inputs remains a critical open problem.

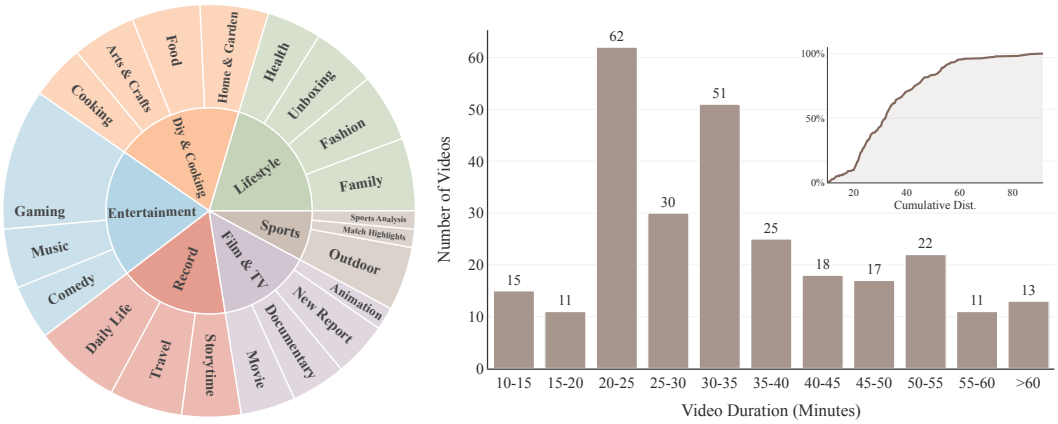

Figure 10: Distribution of videos. **Left:** Videos in LVOmniBench span five primary categories encompassing 21 fine-grained subcategories. We select each video to ensure sufficient audio-visual information and dynamic variation. **Right:** The collected videos range from 10 to 90 minutes, with most between 20 and 50 minutes.

Temporal Grounding Capabilities and Benchmarks. Our experiments demonstrate that the ability of models to perform temporal localization in long-form audio and video requires significant improvement, as a majority of errors stem from event localization failures during initial inference. In the context of audiovisual understanding, audio event localization and joint temporal alignment across the two modalities are required, in addition to purely visual scene localization (Wang et al., 2025b; Wu et al., 2025; Zhang et al., 2025b; Zeng et al., 2024; Qu et al., 2024). However, research addressing temporal localization across omnimodal remains scarce (Chen et al., 2025c). Consequently, the lack of comprehensive test benchmarks and methodological strategies constitutes a significant gap in current research.

Cross-Modal Alignment and Fine-Grained Understanding. Furthermore, we observe that the ability of current models to integrate visual and audio information requires significant improvement. In certain instances, models exhibit an over-reliance on information from a specific modality. More specifically, OmniLLMs frequently ignore information from one modality by disproportionately prioritizing the features extracted from the other. Consequently, this modality bias introduces substantial errors during cross-modal feature alignment. Additionally, the alignment training of audio and visual modalities remains highly complex, which makes the achievement of fine-grained cross-modal comprehension a persistent challenge.

Abstract Audio Perception Capabilities. Our experiments further reveal that models continue to exhibit significant errors in fine-grained perception of audio attributes, including frequency of occurrence, timbre, emotional tone, and sound pressure levels. Furthermore, such non-linguistic information cannot be transcribed into text for comprehension via ASR or alternative methods. Consequently, we emphasize the importance of future efforts dedicated to enhancing the ability of models to comprehend raw audio signals.

D More Details on Question Annotation

D.1 Taxonomy

Perception. This dimension focuses on extracting multimodal information. It evaluates the capacity to perceive fundamental acoustic and visual features, including object attributes (e.g., color, texture, shape), quantities, and musical elements. This layer is crucial for validating the ability of the model to extract fine-grained details from long-context inputs. This dimension comprises the following subcategories: *Counting, Attribute Perception, and Music Perception.*

Understanding. This category aims to evaluate the proficiency in recognizing entities, actions, and their contextual roles within complex scenes. Tasks encompass *Human-Centric Understanding* (e.g., identity tracking, emotion recognition, and behavioral modeling) and fine-grained *Event Un-*

Table 4: Key statistics of our LVOmniBench benchmark.

Video Information		Question Information	
Total Video	275	Total Questions	1014
Categories	21	Question Type	9
Audio Types	Sp+So+Mu	Difficulties (Low:Med:High)	314:441:259
Duration	613–5482	Avg Question Length	16.4
Avg. Duration	34:29	Multiple-Choice	4 (A/B/C/D)

derstanding. These evaluations rigorously test the ability of the model to synthesize complementary audio-visual cues across long-term contexts.

Inference. This dimension evaluates the ability to comprehend complex spatiotemporal dynamics and reason about sound events within long-form audio-visual inputs. Such inference requires rigorous cross-modal alignment across both temporal and spatial dimensions. Specifically, this category encompasses three distinct subtasks: *Sound Inference*, *Spatial Inference*, and *Temporal Inference*.

Logical. This evaluative dimension necessitates multi-step reasoning, causal tracking, and complex inference grounded in complementary cross-modal information. Crucially, these tasks cannot be resolved through superficial feature matching or basic audio-visual alignment. Instead, they require the model to comprehend extended contextual dependencies and construct robust logical reasoning chains across modalities.

D.2 Consistency of Difficulty Annotations

To assess the consistency of the difficulty-level annotations, we calculate an inter-annotator agreement rate of *81%*. We resolve disagreements through discussion and assign consensus labels, further supporting the reliability of the dataset.

E License

LVOmniBench is released under the CC BY-NC-SA 4.0 License.

F More Details on Experimental Settings

F.1 Evaluation Prompt

For the normal evaluation of open-source models and proprietary models, we follow the following prompt:

Prompts

Question: ⟨question⟩
Options: ⟨options_str⟩
Select the best answer from the options above. Directly provide the letter representing your choice (A/B/C/D) and nothing else. Do not include the full text of the option; do not provide any explanation.

F.2 Implementation Details

We use the official configurations for each model and try to evaluate using the maximum permissible number of frames. For Qwen2.5-Omni, Qwen3-Omni, Qwen3-VL, and video-SALMONN 2+, we set the number of input frames to 768, ensuring maximal utilization of the model's context length without exceeding architectural limits. Conversely, because of stricter context-length limitations, we restricted the number of input frames for MiniCPM-o 4.5 and VideoLLaMA2-7B to 64 and 16, respectively. We conducted all local experiments on NVIDIA H100 (80GB) and L40S (48GB) GPUs. For the Gemini series models, we set the input frame rate to 1 frame per second (FPS).