LongInsightBench: A Comprehensive Benchmark for Evaluating Omni-Modal Models on Human-Centric Long-Video Understanding

Anonymous ACL submission

Abstract

We introduce **LongInsightBench**, the first benchmark designed to assess models' ability to understand long videos, with a focus on human language, viewpoints, actions, and other contextual elements, while integrating **visual, audio, and text** modalities. Our benchmark excels in three key areas: **a) Long-Duration, Human-Centric Videos:** We carefully selected approximately 1,000 videos from open-source datasets FineVideo based on duration limit and multi-modal information density, focusing on content like lectures, interviews, and vlogs, which contain rich human-centric semantic and contextual attributes. **b) Diverse and Challenging Task Scenarios:** We have designed six challenging task scenarios, including both Intra-Event and Inter-Event Tasks. **c) Rigorous and Comprehensive Quality Assurance Pipelines:** We have developed a three-step, semi-automated data quality assurance pipeline to ensure the difficulty and validity of the synthesized questions and answer options. Based on LongInsightBench, we designed a series of experiments. which shows that Omni-modal models(OLMs) still face challenge in tasks requiring precise temporal localization (T-Loc) and long-range causal inference (CE-Caus). Surprisingly, extended experiments reveal the information loss in modal fusion of OLMs, which we called the ***Fusion Deficit Paradox***. [1]

1 Introduction

The rapid progress of large pre-trained models has advanced multimodal understanding to the forefront of artificial intelligence research. While Vision-Language Models (VLMs) (Bai et al., 2025, 2023; Radford et al., 2021) and Audio-Language Models (ALMs) (Huang et al., 2023; Radford et al., 2022) excel at short clips and speech, recent **Omni-modal Models (OLMs)** aim for unified perception

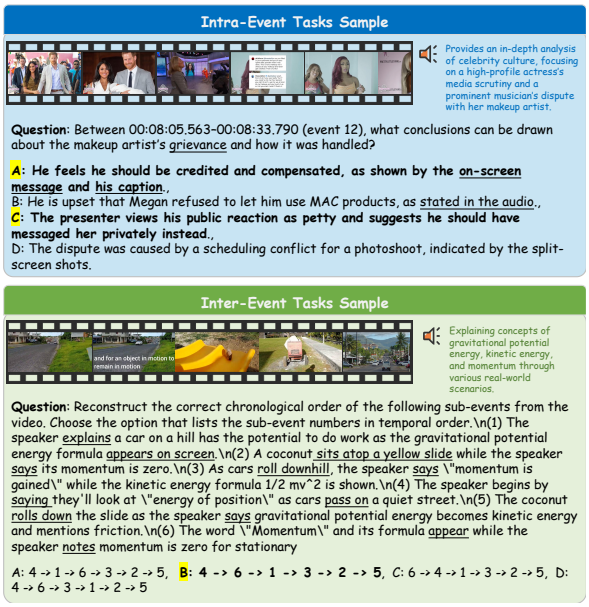

Figure 1: **Task Samples in LongInsightBench.** The upper one comes from IE-Rea(Intra-Event Reasoning) subcategory and the lower one comes from T-Recon(Timeline Reconstruction) subcategory.

across all modalities. However, evaluation remains limited: current benchmarks fail to assess a model's ability to comprehend complex, continuous, and multimodal information in **long videos**.

Existing datasets such as MSR-VTT (Xu et al., 2016), ActivityNet (Heilbron et al., 2015), and Ego4D (Grauman et al., 2021) focus on short-term tasks like action recognition or clip-based QA, requiring only local perception and reasoning within limited time windows. Datasets like HourVideo(Chandrasegaran et al., 2024) and Video-MME(Fu et al., 2025) take long video understanding into consideration, but none of them paid attention to audio information. In contrast, real-world long-form content (such as lectures, interviews and vlogs) often spans tens of minutes and carries dense cross-modal information. Understanding such content demands **long-range temporal dependency modeling**, **precise cross-modal**

[1] Our dataset and code is available at https://anonymous.4open.science/r/LongInsightBench-910F/.

alignment and fusion (especially between spoken language and visual context), and deep comprehension of **subtle contextual elements** (Wang et al., 2025a).

Compared to general video understanding, **human-centric video understanding** imposes greater challenges on models, as these tasks require not only the recognition of human actions and behaviors but also more sophisticated reasoning abilities in cross-modal and long-temporal context. Yet, current benchmarks remain limited in three aspects: (1) **Duration**, lack of evaluation for maintaining attention and contextual coherence beyond a few minutes (Wei, 2024); (2) **Modality**, neglect the critical role of the audio modality and underuse of rich linguistic information presented in long videos (Cai et al., 2025); (3) **Reasoning depth**, test surface matching only rather than distinguish deep, cross-event reasoning (Feng et al., 2025).

To address these gaps, we introduce **LongInsightBench**, the first benchmark explicitly designed for human-centric long-video omni-modal understanding. It emphasizes cues centered with human such as viewpoint, intent and actions, while integrating visual, audio and textual modalities for holistic multimodal understanding and reasoning.

LongInsightBench comprises a curated set of around 1,000 high-density long videos selected from FineVideo (Farré et al., 2024) and 6 challenging task types that span the spectrum of reasoning complexity. A semi-automated quality control pipeline ensures that each question requires genuine multi-modal reasoning rather than simple retrieval or single-modality clues.

We further perform systematic evaluations of state-of-the-art OLMs, complemented by additional experiments with VLMs, ALMs, and LLMs, as well as ablation studies examining the impact of video frame sampling. The results reveal a clear performance hierarchy: large proprietary models such as Gemini2.5 excel in global comprehension but still struggle with temporal localization and long-range causal inference. Our analysis also identifies a persistent *fusion deficit paradox*, where multimodal integration leads to information loss and bias. Ablation findings show that denser frame sampling generally enhances accuracy, though the improvement varies across models.

In summary, we have three main contributions:

1. We introduce **LongInsightBench**, the first comprehensive benchmark for human-centric long-video omni-modal understanding, featuring about 1,000 carefully selected long videos.
2. We have conducted extensive **evaluations** on LongInsightBench, which establish a clear performance hierarchy among OLMs and highlights challenges in temporal localization and long-range causal inference.
3. We reveal a consistent phenomenon, which we term the **fusion deficit paradox**, shedding light on the limitations of existing multi-modal fusion mechanisms in current OLMs.

2 Related Works

2.1 Multimodal Large Language Models

Recent multimodal large language models (MLLMs) integrate text, vision, and audio for unified reasoning. Early efforts focused on vision–language alignment (Radford et al., 2021; Jia et al., 2021), later extending to video understanding (Zhang et al., 2023; Maaz et al., 2024) and audio perception (Girdhar et al., 2023; Bai et al., 2023). Modern omni-modal models employ modular encoders for each modality (Wang et al., 2025b; Team et al., 2025), achieving unified inference but still struggling with temporal alignment and deep cross-modal fusion. Recent advances improve zero-/few-shot performance (Team and Google, 2025) and enable fine-grained visual reasoning (Wu et al., 2025), motivating further research on multimodal planning and reasoning (Huang et al., 2025; Yang et al., 2025b).

2.2 Video Understanding Datasets and Benchmarks

The evolution of multimodal benchmarks has driven progress from visual-only QA to comprehensive audiovisual reasoning. Early datasets such as MSRVTT-QA (Xu et al., 2017), and ActivityNet-QA (Yu et al., 2019) focus on text–video understanding, while recent ones including MovieChat (Song et al., 2024), Video-MME (Fu et al., 2025), and MMBench-Video (Fang et al., 2024) extend evaluation to open-ended, multi-turn, and long-video reasoning. Audiovisual datasets such as AVQA (Yang et al., 2022) and Music-AVQA (Li et al., 2022) further integrate sound cues for joint perception. Despite broader domain coverage, recent resources still face challenges: WorldSense (Hong et al., 2025) requires costly manual annotation, Daily-Omni (Zhou et al., 2025) suffers from

semantic discontinuity due to fixed-duration segmentation, and IntentBench (Yang et al., 2025a) offers limited novelty by aggregating prior datasets. Persistent gaps in event segmentation, filtering quality, multi-event reasoning, and long-video understanding motivate our benchmark for robust audiovisual reasoning.

2.3 Multimodal Intent Understanding and Emotion Recognition

Current AI research in Intent Understanding and Emotional Intelligence is comprehensive, covering practical applications like the systematic multi-modal assessments like HumanSense (Qin et al., 2025) and the psychology-based EmoBench (Sabour et al., 2024). Despite this wide coverage across modalities (text, audio, visual) and tasks (from perception to application), these works predominantly rely on static images, short dialogues, or brief video clips. This focus limits their ability to evaluate the dynamic, long-range evolution of emotional and social contexts, which require tracking subtle, non-contiguous events over extended periods.

3 LongInsightBench Dataset Construction

To ensure LongInsightBench serves as a challenging and reliable benchmark for evaluating omni-modal understanding in long videos, we established a rigorous, multi-stage construction pipeline(See Figure 2). This process involved careful video selection, stringent filtering based on multi-modal density, automated captioning, structured question generation across diverse tasks, and a comprehensive quality assurance protocol.

3.1 Video Selection and Filtering

3.1.1 Video Source and Categories

The video corpus for LongInsightBench is sourced from the publicly available FineVideo dataset (Farré et al., 2024), with a strategic focus on content categories that naturally exhibit high linguistic complexity and diverse multi-modal interactions rather than mere object recognition, which is crucial for testing deep comprehension.

3.1.2 Filtering for Temporal Scope and Information Richness

To ensure the selected videos present a significant temporal challenge and require deep, long-range reasoning, we applied two strict filtering criteria beyond the initial category selection.

Duration Constraint. All selected videos must be longer than 7 minutes. This threshold ensures that models are challenged in maintaining contextual awareness and managing long-term temporal dependencies.

Criteria for Content Richness. We established criteria for content richness across both visual and audio-textual modalities. For **visual dynamism**, we utilized `pyscenedetect` to identify scene cuts. Only videos exhibiting at least three distinct scene changes were retained, ensuring visual diversity and dynamic content. Regarding **linguistic complexity**, we transcribed the audio using WhisperX (with Whisper-medium model and Wav2Vec2.0-based alignment module) and removed non-English content. Furthermore, to verify the complexity of the argumentative structure, we employed GPT-4o for paragraph-level semantic segmentation, retaining only videos that demonstrated at least 4 distinct topic shifts throughout their duration. The details of semantic segmentation is listed in Appendix B.1.

3.2 Automated Annotation of Captions

To facilitate precise question generation and subsequent automated evaluation, we performed detailed multi-modal annotation on the filtered video corpus.

Based on the segmentation result introduced in Section 3.1.2, we generated specialized multi-modal captions for each clip. The visual captions are generated using Ovis2.5-9B, focusing on describing the actions, entities, and visual context within the clip, while the audio captions are annotated using Gemini2.0-Flash, focusing on summarizing the spoken content, identifying speaker sentiment, and describing music or background sounds.

3.3 Task Scenarios and Question Generation

LongInsightBench is designed to evaluate multi-modal reasoning abilities across two complementary dimensions: fine-grained perception within short temporal spans and holistic understanding over extended durations. To this end, we define six task scenarios, grouped into two categories: **Intra-event tasks**, targeting localized reasoning, and **Inter-event tasks**, targeting long-range reasoning across events.

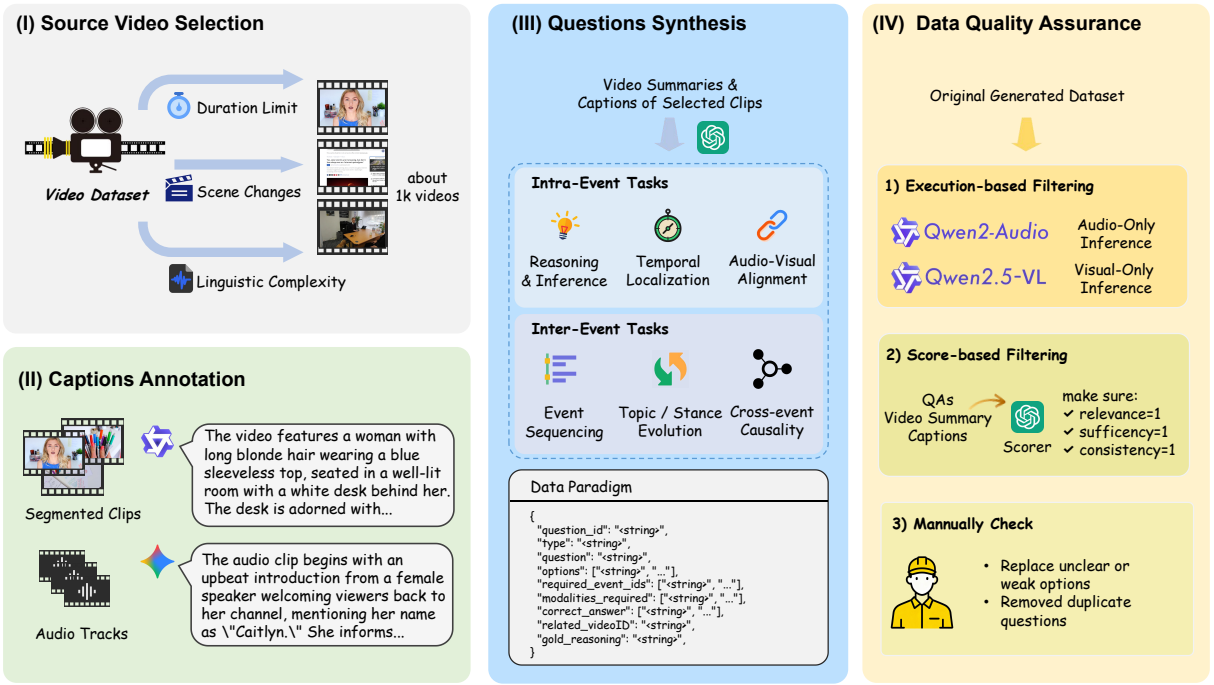

Figure 2: **Overview of the LongInsightBench construction workflow.** The pipeline begins with **video selection** from FineVideo, applying filters on duration, scene shifts, and content richness. Next, **automated annotation** integrates visual and audio descriptions via MLLMs. These annotations support **task scenario design and question generation**, spanning intra-event and inter-event reasoning tasks. Finally, a **quality assurance process** combines automatic filtering, scoring, and manual validation to ensure a high-quality QA set.

While the questions adopt a multiple-choice format, the number of valid answers is dynamically determined by the LLM, in order to increase reasoning diversity and complexity.

Intra-event Tasks. These tasks evaluate a model's ability to perceive and reason within a short temporal window for about one minutes, requiring multimodal alignment and local inference:

- **Reasoning/Inference:** Local causal or inferential reasoning based on immediately preceding or simultaneous multimodal cues.
- **Temporal Localization:** Identifying the precise timing of a specific event or action, using visual and auditory evidence.
- **Audio-Visual Alignment:** Matching spoken or textual content with corresponding visual cues to ensure cross-modal consistency.

Inter-event Tasks. These tasks assess long-range understanding, evaluating a model's ability to retain, integrate, and reason over information distributed across the video timeline, which often require combining multiple non-contiguous events.

- **Timeline Reconstruction:** Sequencing events across the video by linking distant audio–visual anchors.
- **Topic/Stance Evolution:** Tracking how specific themes, viewpoints, or arguments develop, shift, or evolve throughout the video.
- **Cross-event Causality:** Reasoning over long temporal gaps to uncover causal relationships connecting earlier triggers and later outcomes.

We employ GPT-4o to generate QA pairs for each task type using tailored prompt templates, as detailed in Appendix B.4. Additionally, we store the reference chain-of-thought (CoT) reasoning generated by GPT-4o for each question, enabling direct comparison with model-generated CoTs and facilitating a more detailed analysis of model performance. We also provide several failure case studies for selected models, as detailed in Appendix B.6.

3.4 Rigorous Quality Assurance Pipeline

To ensure that the generated questions truly required multi-modal, long-range reasoning. We implemented a three-step, semi-automated filtering pipeline.

Step 1: Execution-Based Filtering. We used state-of-the-art single-modality models as solvers to identify and remove questions not requiring multimodal fusion. Questions solvable by **Qwen2.5-VL-7B-Instruct** (vision-text only) or **Qwen2-Audio-7B-Instruct** (audio-text only) were dis-

Benchmarks	Mod.	#Vids	Dur.(s)	#QA	Anno.	Multi	Open	A-V Corr.	Emo&Insight
MSRVTT-QA	V	2,990	15.2	72,821	A	✗	✓	✗	✗
ActivityNet-QA	V	800	111.4	8,000	M	✗	✗	✗	✗
MVBench	V	3,641	16.0	4,000	A	✗	✓	✗	✗
MovieChat	V	130	500.0	1,950	M	✗	✓	✗	✗
Video-Bench	V	5,917	56.0	17,036	A&M	✗	✓	✗	✗
EgoSchema	V	5,063	180.0	5,063	A&M	✓	✓	✗	✗
Video-MME	V	900	1017.9	2,700	M	✗	✓	✓	✗
MMBench-Video	V	609	165.4	1,998	M	✓	✓	✗	✗
AVQA	A+V	57,000	10	57,335	M	✗	✓	✓	✗
Music-AVQA	A+V	9,288	60	45,867	M	✗	✓	✓	✗
OmniBench	A+I	-	-	1,142	M	✓	✓	✗	✗
AV-Odyssey	A+I	-	-	4,555	M	✓	✓	✗	✗
LongVALE	A+V	8,400	235	-	A&M	✓	✓	✓	✗
WorldSense	A+V	1,662	141.1	3,172	M	✓	✓	✓	✗
Daily-Omni	A+V	684	30–60	1,197	A&M	✓	✓	✓	✗
LongInsightBench	A+V	1,001	539.1	4,781	A&M	✓	✓	✓	✓

Table 1: **Statistics of representative video QA benchmarks.** Mod. denotes modality. Dur.(s) is mean video duration in seconds. Anno. indicates automatic (A) or manual (M) annotations. Multi shows whether the dataset includes multiple question types. Open signifies coverage of diverse domains. A-V Corr. specifies if multimodal integration is required. Emo&Insight highlights whether the benchmark focuses on recognizing human intentions, emotions, and other human-centric elements.

carded.

Step 2: Score-Based Filtering. Each remaining QA pair was automatically reviewed by **GPT-4o** along three scoring dimensions: (1) **Relevance** — whether the question truly depends on the video content rather than general knowledge or common sense; (2) **Sufficiency** — whether the provided multi-modal inputs contain enough evidence for a correct answer; and (3) **Consistency** — whether the answer is factually correct and logically aligned with the ground truth from the video segment. Each criterion was rated between 0 and 1, and only QA pairs achieving high scores across all three dimensions were retained. Details of the scoring prompt and criteria are provided in Appendix B.5.

Step 3: Manual Inspection. A random sample of the filtered QA pairs was checked by human annotators. During this review, the annotators replaced unclear or weak answer options, removed questions that were too similar to each other, and adjusted the difficulty level where needed. See details in Appendix B.7.

3.5 LongInsightBench Statistics

As summarized in Table 1, our proposed benchmark consists of 1001 videos and 4781 high-quality QA pairs after a rigorous three-step filtering pipeline. The average video duration is 539 seconds, which is substantially longer than existing **audio-visual understanding benchmarks**. The approximately 1,000 selected videos are divided into three main categories: the lecture category, which includes 517 videos across 8 subcategories; the interview category, comprising 258 videos across 4 subcategories; and the Vlogs/Film Trailers category, with 230 videos spanning 4 subcategories. Detailed descriptions of the video categories are provided in Table 6. The distribution of the collected videos across all subcategories is visualized in Figure 4. The task-type distribution in Table 2 demonstrates that the benchmark covers a wide spectrum of reasoning scenarios. From localized Intra-event Reasoning and Audio-Visual Alignment to global-level Cross-event Causality

Task Type	Initial	Filtered
Intra-event Reasoning	2002	855
Temporal Localization	2002	857
Audio-Visual Alignment	2002	1307
Timeline Reconstruction	2002	506
Topic/Stance Evolution	626	165
Cross-event Causality	2002	1091
Total	10636	4781

Table 2: **Final dataset statistics** after multi-stage filtering.

Model	Intra-event			Inter-event			Overall
	IE-Rea	T-Loc	AV-Align	T-Recon	Topic Evo&Sum	CE-Caus	
Qwen2.5-Omni-7B	61.75	20.77	49.66	52.57	**89.70**	**49.95**	48.40
VideoLLama3	59.30	9.92	56.69	45.85	63.03	19.52	39.36
VideoLLama2	38.36	13.89	37.03	47.04	46.67	9.35	28.19
Ola-7B	**70.18**	**27.19**	**63.20**	**63.44**	63.03	39.14	**52.52**
Unified-IO-2 L	27.25	**2.10**	**24.10**	25.49	36.36	8.80	**17.80**
Unified-IO-2 XL	**18.25**	6.07	39.40	**22.92**	**29.09**	2.47	19.12
Unified-IO-2 XXL	36.84	17.85	54.55	32.81	50.30	**1.10**	30.16
Gemini2.5-Flash	**89.01**	**38.86**	**76.05**	**86.76**	**84.24**	**41.25**	**65.17**

Table 3: **Average accuracy (%) of various OLMs across different tasks.** Abbreviations: IE-Rea (Intra-event Reasoning), T-Loc (Multimedia Temporal Localization), AV-Align (Audio-Visual Alignment), T-Recon (Timeline Reconstruction), Topic Evo&Sum (Topic/Stance Evolution Summarization), CE-Caus (Cross-event Causality).

and Timeline Reconstruction, the benchmark requires models to handle both fine-grained grounding and complex long-horizon dependencies. Such diversity makes the evaluation more comprehensive.

4 Experiments and Analysis

4.1 Settings

We evaluate four categories of multimodal large language models (MLLMs): (1) **OLMs**, including open-source contenders like VideoLLaMA 2 (Cheng et al., 2024), VideoLLaMA 3 (Zhang et al., 2025), Unified-IO-2 (Lu et al., 2023), Qwen2.5-Omni (Xu et al., 2025), Ola (Liu et al., 2025), and the proprietary model Gemini2.5-Flash (Comanici et al., 2025); (2) **VLMs**, represented by the open-source Ovis2.5 (Lu et al., 2025) and proprietary GPT-4o (OpenAI et al., 2024), Gemini2.5-Flash; (3) **ALMs**, such as Gemini2.5-Flash; and (4) **LLMs**, represented by GPT-4o. All evaluations follow each model's official inference pipeline and pre-processing configurations. To ensure consistency across model types, we set the number of video frames to 64 for open-source models, while proprietary models are evaluated through official APIs using default multimodal input settings.

Our experiments consist of three parts. First, we evaluate a wide range of OLMs on the full benchmark to compare their abilities in perceiving multimodal linguistic cues in long videos. Second, we replaced one or two modalities of the input to OLMs with text descriptions to examine whether OLMs can effectively fuse and utilize multi-modal information. Third, we conduct ablation studies varying the number of sampled video frames to assess open-source OLMs under different frame sampling rates. Performance in all experiments is measured by accuracy, with a prediction considered correct only if all options are exactly selected, ensuring a strict assessment of the model's multimodal understanding.

To estimate the total computational budget, we consider the local model deployed on our own servers, utilizing 16 NVIDIA A800-SXM4-40GB GPUs, which require approximately 250 GPU Hours for the entire experiment, with no additional costs. The server is equipped with CUDA version 13.0 and NVIDIA driver version 580.65.06, supporting the necessary computational load. For GPT-4o and Gemini2.5-Flash models, cloud services are used, resulting in a total estimated cost of $200 for the API calling.

4.2 Main Results

The experimental results presented in Table 1 provide a comprehensive evaluation of various Omni-modal Language Models (OLMs) across the six different task scenarios defined in LongInsightBench. The analysis clearly shows the current performance gap between proprietary, large-scale models and their open-source counterparts, while also highlighting specific areas of weakness that are common across all models, particularly in long-range temporal and causal reasoning.

Performance Ceiling and Model Hierarchy The results clearly show that **Gemini2.5-Flash** is the best performer on this benchmark, with an overall accuracy of 0.6517. This closed-source model scored the highest or second-highest in five out of the six individual task categories, demonstrating strong performance and reliability across both localized and long-range tasks.

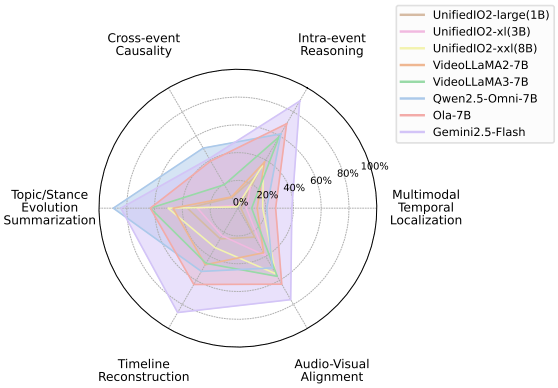

Figure 3: **Fine-grained performance across task categories.** Different OLMs' accuracies are shown over six question types, highlighting each model's strengths and weaknesses across categories.

Among the open-source models, **Ola-7B** stands out as the best competitor, achieving the second-highest overall score (0.5252) and ranking second in all three Intra-event tasks (IE-Rea, T-Loc, AV-Align) and Timeline Reconstruction (T-Recon). This suggests that Ola-7B has a solid base for multi-modal understanding and localized inference.

On the other hand, the **Unified-IO-2** models generally performed poorly, placing in the lowest performance groups. Unified-IO-2 L recorded the lowest overall accuracy (0.1780), and its variants performed poorly in several key areas, such as Intra-event Reasoning (IE-Rea), Temporal Localization (T-Loc), and Cross-event Causality (CE-Caus), suggesting that these models lack the needed multi-modal alignment and long-context processing capabilities for this benchmark.

Analysis of Task Categories Figure 3 visualizes the model performance hierarchy ,showing clear differences between tasks in difficulty, effectively testing the limits of current OLMs.

Performance on the Intra-event tasks (IE-Rea, T-Loc, AV-Align) is generally better than on the Inter-event tasks, reflecting the relative ease of localized processing. However, Temporal Localization (T-Loc) proved to be a major challenge for all models. Even the top performer, Gemini2.5-Flash, only scored 0.3886, and the scores were much lower for other models (e.g., Unified-IO-2 L scored 0.0210). This shows that accurately pinpointing event timing using multi-modal data remains a major challenge, requiring further improvements in fine-grained temporal modeling. Audio-Visual Alignment (AV-Align) saw strong performance from Gemini (0.7605), confirming its ability to align audio and visual data effectively.

The Inter-event tasks, which require long-term memory and synthesis, showed the most variation. The task Topic/Stance Evolution Summarization seemed relatively easier, with Qwen2.5-Omni-7B unexpectedly achieving the highest score (0.8970) and Gemini close behind (0.8424). This indicates that many models are good at tracking and summarizing the main theme or narrative flow across the video.

In contrast, Cross-event Causality (CE-Caus) proved to be the most difficult long-range task. This task involves identifying cause-and-effect relationships across different parts of the video, and performance was generally low. Gemini achieved the highest score at 0.4125, while several models, including Unified-IO-2 XXL, scored near chance level (0.0110). This highlights the current limitations of OLMs in maintaining and reasoning over complex causal links within long videos.

In summary, LongInsightBench effectively shows the differences in model capabilities, demonstrating the clear advantage of large proprietary models like Gemini2.5-Flash. The benchmark also shows that while models are improving in tracking narrative themes (Topic Evo&Sum), they still face significant challenges in precise temporal localization (T-Loc) and complex, long-range causal reasoning (CE-Caus).

Caption Models	Intra-event	Inter-event	Overall
VLM + audio captions			
Gemini2.5-Flash	65.35	66.29	65.70
Gemini2.0-Flash	72.28	60.11	67.78
ALM + visual captions			
Ovis2.5-9B	72.94	64.04	**69.65**
Gemini2.5-Flash	69.64	61.80	66.74
Gemini2.0-Flash	67.66	65.17	66.74
LLM + both captions			
Gemini2.5-Flash	67.00	62.36	65.28
Gemini2.0-Flash	71.62	64.04	68.81
SOTA OLM			
None	69.31	58.43	**65.17**

Table 4: **Gemini2.5-Flash Performance comparison** (accuracy, %) of VLMs(with audio captions), ALMs(with visual captions), LLMs(with both captions) and SOTA OLMs.

4.3 The Revealing of Fusion Deficit Paradox

The comparative analysis presented in Table 4 contrasts the performance of SOTA model Gemini2.5-Flash as dedicated Signal-modal Model (VLM,

Model	Input Frame Number		
	32	64	128
VideoLLama3	39.29	41.37	43.45
Ola-7B	53.64	55.30	56.96
Unified-IO-2 XXL	30.98	31.19	31.39
Qwen2.5-Omni-7B	51.35	51.98	51.98

Table 5: Overall Accuracy (%) with **Different Input Frame Numbers**.

ALM) and LLM, which replace one or two of the input modalities with textual descriptions, against true Omni-modal Model) OLM setting, revealing critical insights into the current state of multi-modal fusion.

The Paradox of Omni-modal Fusion We replaced the caption model with several other MLLMs, even with Gemini2.5-Flash itself, and observed a consistent performance increase compared to the vanilla OLM setting. When processing both raw visual and raw audio data simultaneously (OLM: 0.6517), its performance is significantly lowered by a max decreasing rate of 4.00%, 6.87% and 5.59% for VLM setting, ALM setting and LLM setting respectively. The highest overall score is achieved by the Audio-Language Model (ALM) configuration, suggesting that the model excels when the visual information is provided in a distilled, textually abstracted format.

This phenomenon strongly suggests that current OLM fusion mechanisms suffer from a information loss introduced during the process of integrating two raw modalities (pixels and waveforms), which we called *Fusion Deficit Paradox*.

Explaining the Superiority of Textual Proxies The superior performance of VLM and ALM configurations, which use textual descriptions for one modality, stems from three factors.

First, textual descriptions (captions/summaries) act as effective, pre-processed proxies, filtering noise and redundancy from raw streams. This allows the model to bypass complex, error-prone low-level feature extraction and alignment for that modality. Second, receiving one modality as text enables the model to dedicate its full capacity (e.g., attention) to robustly aligning the remaining raw modality with the pre-parsed text. This focused processing leads to a more accurate overall understanding. Third, these models are fundamentally rooted in Large Language Models (LLMs), which operate optimally with high-quality textual input. This is supported by Gemini2.5-Flash's competitive score (0.6881) using purely textual input. Ultimately, when the raw fusion mechanism is imperfect, the quality of textual representation can often outweigh the benefit of processing raw multi-modal data.

4.4 The Effect of Video Frame Sampling Rate

The ablation study presented in Table 5 examines the impact of visual information density by varying the number of sampled input video frames (32, 64, and 128) on open-source OLM performance. The results generally confirm that increased frame sampling leads to improved overall accuracy, though the degree of benefit is highly model-dependent.

Models Showing Strong Context Utilization and Saturation (Ola-7B, VideoLLama3, Qwen2.5-Omni-7B, and UnifiedIO 2-XXL): Ola-7B and VideoLLama3 exhibited a clear positive correlation between frame count and accuracy, with Ola-7B improving from 0.5364 (32 frames) to 0.5696 (128 frames) and VideoLLama3 showing a similar trend (0.3929 to 0.4345). These results suggest that both models effectively integrate temporal information to enhance long-context reasoning. In contrast, Qwen2.5-Omni-7B and UnifiedIO 2-XXL showed limited improvements, with Qwen2.5-Omni-7B reaching saturation (0.5135 to 0.5198) and UnifiedIO 2-XXL showing negligible change (0.3098 to 0.3139), indicating challenges in utilizing denser visual evidence or reaching capacity limits in their visual processing mechanisms.

5 Conclusion

In conclusion, this paper introduces LongInsight-Bench, a pioneering benchmark for human-centric long-video omni-modal understanding, featuring a challenging dataset of 4,781 carefully-designed questions. This benchmark serves as a realistic testbed for next-generation OLMs, focusing on human-centric cues such as viewpoint, sentiment, and action. Our experimental results demonstrate OLMs still face challenges in tasks like temporal localization and long-range causal reasoning. Additionally, extended experiment suggests that current omni-modal fusion mechanisms may suffer from a fusion deficit. The ablation study further reveals that frame sampling improves model accuracy, though the benefits vary across models.

Limitations

The development of LongInsightBench involved significant operational costs due to the reliance on proprietary models (GPT-4o and Gemini2.0-Flash) for high-fidelity filtering, QA generation, and rigorous quality assurance. This high API calling expense restricts the pace and scale at which we can expand the dataset, despite the need for dense, multi-modal content. This financial constraint is a major limitation on scalability. Future work will focus on developing more cost-efficient, open-source or human-in-the-loop pipelines to mitigate this expense and facilitate larger-scale expansion.

Ethics Statement

We have carefully curated the video data used in LongInsightBench to ensure the exclusion of dangerous, discriminatory, or unhealthy content. Furthermore, we strictly adhere to all terms and conditions mandated by the source dataset, FineVideo. To respect the rights and privacy of the original video creators, we do not host the raw FineVideo content. Instead, we only release the synthesized data (questions, answers, and annotations) derived from the videos, maintaining responsible data usage within reasonable limits.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Yuxuan Cai, Jiangning Zhang, Zhenye Gan, Qingdong He, Xiaobin Hu, Junwei Zhu, Yabiao Wang, Chengjie Wang, Zhucun Xue, Chaoyou Fu, Xinwei He, and Xiang Bai. 2025. Humanvideo-mme: Benchmarking mllms for human-centric video understanding. *Preprint*, arXiv:2507.04909.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristobal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Fei-Fei Li. 2024. Hourvideo: 1-hour video-language understanding. In *Advances in Neural Information Processing Systems*, volume 37.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 314 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. 2024. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Preprint*, arXiv:2406.14515.

Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. 2024. Finevideo. https://huggingface.co/datasets/HuggingFaceFV/finevideo.

Bo Feng, Zhengfeng Lai, Shiyu Li, and 1 others. 2025. Vbenchcomp: Disentangling video-language model evaluation on knowledge, spatial perception, or real temporal understanding? *arXiv preprint arXiv:2505.14321*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *Preprint*, arXiv:2405.21075.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. *Preprint*, arXiv:2305.05665.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, and 1 others. 2021. Ego4d: Around the world in 3,000 hours of egocentric video. *arXiv preprint arXiv:2110.07058*.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970.

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

Chi-Pin Huang, Yueh-Hua Wu, Min-Hung Chen, Yu-Chiang Frank Wang, and Fu-En Yang. 2025. Thinkact: Vision-language-action reasoning via reinforced visual latent planning. *arXiv preprint arXiv:2507.16815*.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Zhou Zhao, and Shinji Watanabe. 2023. Audiogpt: Understanding and generating speech, music, sound, and talking head. *Preprint*, arXiv:2304.12995.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. *Preprint*, arXiv:2102.05918.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. Learning to answer questions in dynamic audio-visual scenarios. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. 2023. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. *arXiv preprint arXiv:2312.17172*.

Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, Yuxuan Han, Haijun Li, Wanying Chen, Junke Tang, Chengkun Hou, Zhixing Du, Tianli Zhou, Wenjie Zhang, Huping Ding, and 23 others. 2025. Ovis2.5 technical report. *arXiv:2508.11737*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. 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 (ACL 2024)*.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Zheng Qin, Ruobing Zheng, Yabing Wang, Tianqi Li, Yi Yuan, Jingdong Chen, and Le Wang. 2025. Humansense: From multimodal perception to empathetic context-aware responses through reasoning mllms. *arXiv preprint arXiv:2508.10576*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. 2024. Moviechat: From dense token to sparse memory for long video understanding. *Preprint*, arXiv:2307.16449.

Gemini Team and Google. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *ArXiv*, 2507.06261.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 69 others. 2025. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Jue Wang, Wentao Zhu, Pichao Wang, and 1 others. 2025a. Videorag: Retrieval-augmented generation with extreme long-context videos. *arXiv preprint arXiv:2502.01549*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025b. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*.

Fangyun Wei. 2024. A large-scale human-centric benchmark for referring expression comprehension in the lmm era.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, and 20 others. 2025. Qwen-image technical report. *Preprint*, arXiv:2508.02324.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, page 1645–1653, New York, NY, USA. Association for Computing Machinery.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. 2022. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3480–3491.

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

Senqiao Yang, Junyi Li, Xin Lai, Bei Yu, Hengshuang Zhao, and Jiaya Jia. 2025b. Visionthink: Smart and efficient vision language model via reinforcement learning. *arXiv preprint arXiv:2507.13348*.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. *Preprint*, arXiv:1906.02467.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *Preprint*, arXiv:2306.02858.

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

A Video Source Details

Table 6 shows the details of all categories of video source. Figute 4 visualizes the distribution of videos in our dataset across all subcategories.

B Implementation Details

B.1 Details of Semantic Segmentation

To prevent the model from modifying the original transcript due to hallucinations, we design a two-stage semantic segmentation procedure. Both stages are performed using GPT-4o to ensure consistency in linguistic style and reasoning. The prompts used in each stage are provided in B.3.

In the first stage, the model outlines the overall thematic structure of the transcript. Given the full transcript of the video's speech, it is prompted to estimate the number of distinct semantic topics and to generate a tentative title for each. This serves as a preparatory step for segmentation. Since the transcripts are often lengthy, identifying the expected number and focus of segments in advance helps reduce the model's cognitive load in the subsequent stage.

In the second stage, the model determines the precise semantic boundaries between segments. A second-stage prompt instructs the model to locate transition points between topics without modifying the original text. To indicate these transitions, the model outputs a few words surrounding each boundary, from which the full text segments are later extracted using regular expressions. The prompt explicitly requires that boundaries be placed between sentences. However, if a predicted boundary still falls within a sentence, the entire sentence is assigned to the following segment, while the preceding sentence serves as the closure of the previous one.

We adopt this boundary-marking strategy rather than asking the model to directly output segmented text, in order to avoid discontinuities and hallucinations. LLMs may inadvertently add, omit, or alter portions of the transcript especially when handling long passages, resulting in inconsistencies or incomplete context for subsequent audio-captioning tasks.

B.2 Prompts for visual/audio caption

Prompt in Figure 5 is used for creating visual caption by Ovis2.5-9B, and Prompt in Figure 6 is used for creating audio caption by Gemini2.5-flash.

B.3 Prompts for Semantic Segmentation

Figure 7 and 8 illustrate the prompt design for semantic segmentation. Particularly, the SEGMENT_COUNT_PROMPT shown in Figure 7 is used in the first stage of segmentation, while the

Category	Subcategories	Count	Description
Lectures	8	514	Academic talks and tutorials covering diverse topics including AI, astronomy, biology, chemistry, science explanations, software tutorials, and TED talks. High focus on spoken content and visual aids.
Interviews	4	258	Dialogues and interviews including celebrity, expert, and political interviews and sports talk shows. Rich in viewpoint tracking, sentiment changes, and nuanced discussion.
Vlogs / Film Trailers	4	229	Narrative-driven content including camping, hiking, travel vlogs and film trailers. Emphasizes temporal coherence, action-language alignment, and multimodal dynamics.
Total	–	1001	

Table 6: **Video source categories** from FineVideo used in LongInsightBench, updated with precise subcategory counts and descriptions.

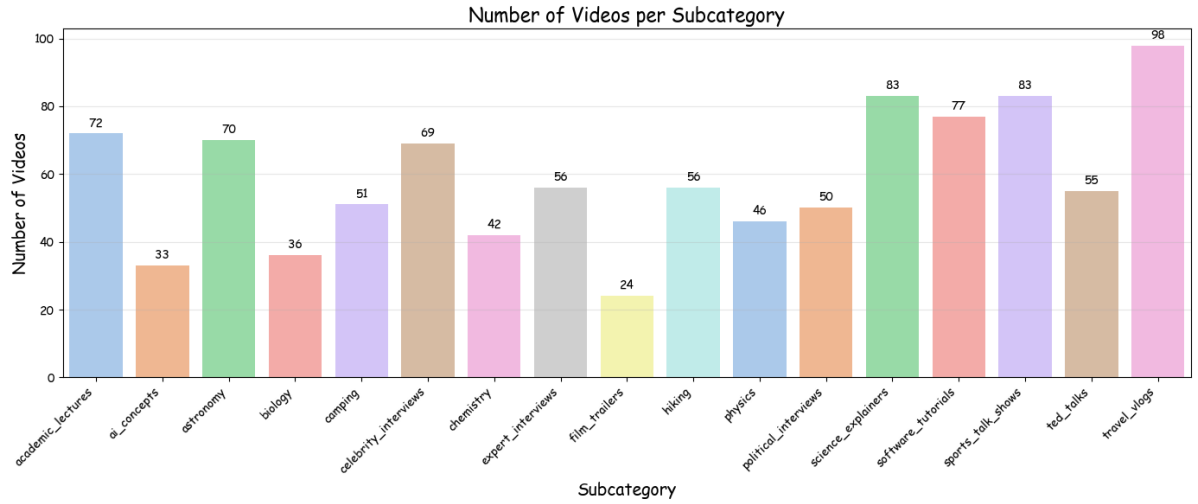

Figure 4: The **distribution** of videos in LongInsightBench across all subcategories.

BOUNDARY_DETECTION_PROMPT shown in Figure 8 is used in the second stage.

B.4 Details of QA Construction

To ensure the generated questions require deep contextual understanding—especially for Inter-event tasks—the input to the LLM included a high-level summary of the video (borrowed from the FineVideo dataset). This summary is necessary to provide the LLM with the global narrative context and thematic structure of the video, preventing the generation of questions based solely on localized, isolated events. Along with this summary, we provided randomly sampled event IDs (typically 2–3) and their corresponding visual and audio captions. The LLM was instructed to output a JSON-formatted selection question. The choice between single-choice and multi-choice was determined dynamically based on the complexity inherent in the task scenario. Prompt in Figure 9 is the system prompt used by GPT-4o in QA construction stage, while prompts in Figures 10, 11, 12, 13, 14 and 15 are the user prompts specified to generate QAs of different task types.

B.5 Prompts for Scoring QAs

The prompts take the complete visual captions, audio captions, and QA pairs as input, and return a JSON-formatted output containing dimension-wise scores for each QA pair. As illustrated in Figure 16,

VISUAL CAPTION PROMPT

You are an expert video describer.

Provide a detailed description of the given video segment as a single concise paragraph.

Focus on:
- People: their actions, gestures, clothing, and facial expressions (use distinguishing features to tell individuals apart)
- Objects and text: describe visible objects and any on-screen text (state the text in its original language, give an English translation in parentheses, and explain its contextual meaning)
- Environment: the setting, background details, and atmosphere
- Visual changes: transitions, movements, or notable differences between frames

Guidelines:
- Describe the sequence of frames as a continuous narrative, not isolated snapshots
- Emphasize how the scene evolves over time
- Avoid speculation beyond what is visually shown

Figure 5: Prompt used for providing **visual captions** for each video.

the scoring prompt is designed to guide GPT-4o in evaluating the initial QA pairs along three dimensions: *sufficiency*, *consistency*, and *relevance*.

To ensure reliable and fine-grained judgments, the prompt provides explicit scoring criteria, where each dimension is assigned a score between 0 and 1. Unlike binary filtering, this soft scoring process allows finer discrimination of borderline cases. Only QA pairs that achieve a score of 1 across all three dimensions are retained in the final dataset.

B.6 Failure Case Study

Figure 17 a case from Gemini2.5-flash in the Localization Temporal Localization task. The model's failure stems from its inability to correctly identify the multimodal elements in the video—both visually and acoustically—and match them to the events described in the question. It incorrectly asserts that the "interview title card" appears at 00:14:50 and that the "interviewer introducing his guest" occurs at 00:21:50, neither of which aligns with the actual video. These errors likely reflect misrecognition of superficially similar cues or outright hallucinations of the model. This example shows that the model's weakness in temporal localization arises from deficiencies in multimodal fusion and multi-hop reasoning, rather than from any limitation in our task design.

Figure 18 shows another failure case from Gemini2.5-flash in the Cross-event Causality task also highlights a common issue: models struggle to integrate cues over time, focusing instead on the most recent signal. In this example, the golden reasoning requires linking earlier events into a causal chain, but the model only attends to the latest cue, showing salience bias and limited multi-step integration rather than an annotation issue.

B.7 Details of Human Inspection and Filtering on QA pairs

After two rounds of automated filtering, we conducted a structured human quality control procedure with explicit scoring criteria. Specifically, we randomly sampled approximately 10% of the remaining QA pairs from each task category for manual review. Each sampled QA pair was evaluated by human annotators following a predefined quality rubric. If more than 25% of the inspected QA pairs within a task category were judged to require modification or removal, we continued to randomly sample another batch of QA pairs from the remaining pool and repeated the inspection until the proportion of low-quality items fell below this threshold, ensuring consistent quality across categories.

Each QA pair was assessed along three dimen-

AUDIO CAPTION PROMPT

You are an expert audio describer.

Provide a chronological description of the audio clip without mentioning timestamps.

For speech, include:
- Content (quote short phrases verbatim, summarize longer parts)
- Speaking tone
- Number of speakers, distinguishable by gender, speaking style, or any other perceivable audio features

For music, include:
- Genre or style
- Mood or tone
- Main instruments or notable features

For background or ambient sounds, include:
- Sound characteristics (volume, rhythm, consistency, etc.)
- Environmental cues or setting inferred from these sounds

For other sounds, include:
- Type of sound and how it contributes to the scene

Guidelines:
- Avoid guessing when uncertain
- Write the description as a single concise paragraph, highlighting transitions between different sounds

Figure 6: Prompt used for providing **audio captions** for each video.

sions: (1) semantic alignment, evaluating whether the question, answer options, and gold answer were correctly grounded in the video content across relevant modalities; (2) clarity and unambiguity, assessing whether the question was clearly phrased and admitted a single correct answer under the intended reasoning process; and (3) task appropriateness, determining whether the QA pair matched the intended task type and difficulty level without being trivial or ill-posed. Based on these criteria, annotators assigned an overall quality score on a five-point scale, interpreted as follows:

- **Score 5 (Excellent):** The QA pair is fully correct, unambiguous, and well aligned with the video content, with no redundancy or extraneous cues. Such QA pairs are retained without modification.
- **Score 4 (Good):** The QA pair is largely correct and well grounded, but may contain minor issues (e.g., slightly similar distractor options) that do not affect the correctness. These QA pairs are retained as-is or lightly edited.
- **Score 3 (Borderline):** The QA pair is correct in principle but suffers from issues such as vague phrasing, overly similar answer choices, or suboptimal difficulty calibration. These QA pairs are revised.
- **Score 2 (Poor):** The QA pair exhibits clear deficiencies, including misalignment with the video content, unclear question intent, or weak reasoning signals. These QA pairs are removed.
- **Score 1 (Invalid):** The QA pair is fundamentally flawed due to incorrect gold answers, hallucinated events, severe ambiguity, or redundancy with other QA pairs. These QA pairs are removed.

Following manual inspection, approximately 18% of the QA pairs were modified, primarily by refining ambiguous wording, adjusting answer

SEGMENT COUNT PROMPT

You are an expert in transcript chunking and topic boundary detection for long videos.

Given a piece of text transcribed from the audio of a video, your task is to:
1. Identify how many distinct semantic chunks it contains.
2. For each chunk, provide a short title (a few words) summarizing its main theme or idea.

Guidelines:
- Each chunk should correspond to a coherent theme, explanation, or dialogue unit.
- Avoid making chunks too short or too long.
- The goal of chunking is to create useful and self-contained units of text for downstream tasks such as captioning and retrieval, not to detect strict topic shifts.
- The short titles should be concise, descriptive, and capture the main semantic focus of the chunk.

Output Format (strictly follow this structure):
Chunk count: <integer>
Titles:
1. <short title for chunk 1>
2. <short title for chunk 2>
...
N. <short title for chunk N>

Now, analyze the following text:
{text}

Figure 7: **Prompt used in the first stage of semantic segmentation.** In this stage, the model is asked to identify the number of topics of the given transcript.

choices, or calibrating question difficulty. Around 13% of the QA pairs were removed due to vague or ill-defined questions, misalignment with the video content, redundancy, or overly simplistic formulations. The remaining QA pairs met the predefined quality standards and were retained in the final benchmark.

Due to limited funding, all manual inspection work was conducted by the authors of this paper without external compensation, with a total time cost of approximately 120 person-hours. Thanks for their hard working.

BOUNDARY DETECTION PROMPT

You are an expert in transcript segmentation for long videos.

Your task:
1. Identify EXACTLY {boundary_count} semantic boundaries in the transcript, based on the {topic_count} chunks and their titles.
2. Each boundary MUST be represented as:
<last few words of previous sentence>[BORDER]<first few words of next sentence>

Guidelines:
- You must output ONLY one boundary per line. No explanations, no numbering, no extra words.
- The words on the left and right of [BORDER] MUST appear **exactly as in the original transcript**, with no paraphrasing.
- The left part must be the END of a sentence. The right part must be the START of the next sentence.
- Boundaries must align with the given semantic titles.
- Do not add or skip boundaries. The number of output lines MUST equal {boundary_count}.

Output Format (strict):
<previous sentence ending>[BORDER]<next sentence beginning>
(repeated {boundary_count} times, one per line)

Now process the following transcript:

Transcript:
{text}

Topic count: {topic_count}
Chunk titles:
{titles}

Output:

Figure 8: **Prompt used in the second stage of semantic segmentation.** In this stage, the model is asked to output segment borders and several surrounding words based on the given topics and titles.

SYSTEM PROMPT

You are a multimodal question generator specializing in video understanding.
Your task is to create high-quality multiple-choice questions (MCQs) for video understanding.

You are restricted to using ONLY the provided event list (visual_caption, audio_caption, timestamps).
Do not use external knowledge, hallucinated facts, or information not present in the events.
Each generated question must strictly follow the JSON schema below.

Figure 9: **System prompt** used in QA construction stage.

INTRA EVENT REASONING USER PROMPT

video_id: {video_id}
summary: {summary}
events: {events_str}

Task requirements:
1) Generate N=2 **Intra-event Reasoning questions.** Each question MUST:
- Focus on a single event, querying **causation or conclusions within its timestamp range** (e.g., "Why X happened / How Y was achieved / What Z signifies between [start_time] and [end_time]?").
- Use **exactly 1 event_id**, referring only to its content and **timestamps**.
- Require BOTH visual and audio evidence from that event; single modality is INSUFFICIENT.
- Offer plausible options **strictly based on the specific event's details**.
- Be "single_choice_question" or "multiple_choice_question".

2) For answer options:
- Provide exactly 4 options: A, B, C, D.
- Distractors must be consistent with event details but incorrect.
- For "single_choice_question": exactly one correct option.
- For "multiple_choice_question": at least two correct options.

3) For explanations:
- Justify the correct answer(s) by synthesizing information **as if you were observing the video directly** and field the ‘"gold_reasoning"‘.
- Reasoning must detail inference steps, **explicitly referencing the single event_id and both visual + audio evidence.**

Figure 10: Prompt used in QA construction in the **Intra-event Reasoning** task.

MULTIMODAL TEMPORAL LOCALIZATION USER PROMPT

video_id: {video_id}
summary: {summary}
events: {events_str}

Task requirements:
1) Generate N=2 **Multimodal Temporal Localization questions.** Each question MUST:
- Focus on localizing a specific event, which is defined by the **simultaneous occurrence or strong correlation of a distinct visual action/cue AND associated audio information (e.g., speech content, specific sounds)**.
- Ask for the exact time segment(s).
- Use **exactly 1 event_id**. The question should provide enough detail from both visual and audio captions to uniquely identify the correct time segment(s).
- Require BOTH visual and audio evidence from that event; single modality is INSUFFICIENT.
- Offer plausible options **strictly based on the specific event's details**.
- Be "single_choice_question" or "multiple_choice_question".

2) For answer options:
- Provide exactly 4 options: A, B, C, D. Each option's value **must be a timestamp string** in "[HH:MM:SS - HH:MM:SS]" format.
- Distractor time segments must be plausible but incorrect for the queried event, ideally from other events or incorrect parts of the correct event.
- For "single_choice_question": exactly one correct time segment option.
- For "multiple_choice_question": at least two correct time segment options.

3) For explanations:
- Justify the correct answer(s) by synthesizing information **as if you were observing the video directly** and field the ‘"gold_reasoning"‘.
- Reasoning must detail inference steps, **explicitly referencing the required event_ids and both visual + audio evidence used to pinpoint the exact time segment.**

Figure 11: Prompt used in QA construction in the **Multimodal Temporal Localization** task.

AUDIO VISUAL ALIGNMENT USER PROMPT

video_id: {video_id}
summary: {summary}
events: {events_str}

Task requirements:
1) Generate N=2 **Audio-Visual Alignment questions.** Each question MUST:
- Focus on **identifying the corresponding visual characteristic/expression given an audio event**, **OR** **identifying the corresponding audio event given a visual characteristic** within a specific event.
- Use **exactly 1 event_id**. The question should target an event's [start_time] and [end_time] where the specified audio and visual elements occur concurrently.
- Require BOTH visual and audio evidence to correctly identify the aligning characteristic; single modality is INSUFFICIENT.
- Offer plausible options **strictly based on the specific event's details**.
- Be "single_choice_question" or "multiple_choice_question".

2) For answer options:
- Provide exactly 4 options: A, B, C, D. Each option's value **must be a descriptive string** that aligns with the modality being queried (i.e., visual characteristics for visual questions, or audio events for audio questions).
- Distractor options must be plausible within the event but not aligned with the queried information, or entirely incorrect.
- For "single_choice_question": exactly one correct descriptive option.
- For "multiple_choice_question": at least two correct descriptive options.

3) For explanations:
- Justify the correct answer(s) by synthesizing information **as if you were observing the video directly** and field the ‘"gold_reasoning"‘.
- Reasoning must detail inference steps, **explicitly referencing the single event_id and both visual + audio evidence used to align the audio event with its visual manifestation.**

Figure 12: Prompt used in QA construction in the **Audio-Visual Alignment** task.

TIMELINE RECONSTRUCTION USER PROMPT

video_id: {video_id}
summary: {summary}
events: {events_str}

Task requirements:
1) Generate N=2 **Timeline Reconstruction question.** The question MUST:
- Present a list of 4-10 distinct sub-events in a shuffled, non-chronological order. Each sub-event should be explicitly numbered (e.g., "(1) [Description of sub-event A]", "(2) [Description of sub-event B]").
- Each sub-event description should be **concise and focuses on a single, atomic action or observation**.
- Sub-events should be drawn from **at least 3 different** event_ids.
- Require the reconstruction of the correct chronological order of these sub-events.
- Require BOTH visual(e.g., character movements, object appearance/disappearance) and audio(e.g., specific sound effects, spoken time indicators) evidence to determine the correct sequence; single modality is INSUFFICIENT.
- Be "single_choice_question".

2) For answer options:
- Provide exactly 4 options: A, B, C, D. Each option's value **must be a sequence of the sub-event numbers**, joined by " -> ".
- Provide exactly 1 correct option which represents the correct chronological sequence of the numbered sub-events.
- Provide exactly 3 distractor options, which must be plausible but incorrect sequences.
- Ensure all sub-event numbers included in the question are used exactly once in each answer option's sequence.

3) For explanations:
- Justify the correct answer(s) by synthesizing information **as if you were observing the video directly** and field the g̈old_reasoning.̈
- Reasoning must detail inference steps, **explicitly referencing the required event_ids and both visual + audio evidence used to reconstruct the sub-events.**

Figure 13: Prompt used in QA construction in the **Timeline Reconstruction** task.

TOPIC STANCE EVOLUTION SUMMARIZATION USER PROMPT

video_id: {video_id}
summary: {summary}
events: {events_str}

Task requirements:
1) Generate N=2 **Topic/Stance Evolution Summarization question.** The question MUST:
- Focus on summarizing the **evolution or development of a key topic or a character's stance/viewpoint** across multiple relevant events.
- Involve **at least 3 different event_ids**.
- Require BOTH visual(e.g., speaker's gestures, on-screen text, changes in setting) and audio(e.g., spoken content, tone shifts, emphasis) evidence to formulate a comprehensive summary; single modality is INSUFFICIENT.
- Offer plausible options **strictly based on the video's main idea**.
- Be "single_choice_question" or "multiple_choice_question".

2) For answer options:
- Provide exactly 4 options: A, B, C, D. Each option's value **must be a concise, multi-sentence paragraph (2-4 sentences)** describing a potential progression or evolution of the topic/stance across the selected events.
- Distractor options must be plausible descriptions of an evolution, but either not aligned with the actual progression or entirely incorrect.
- For "single_choice_question": exactly one correct option.
- For "multiple_choice_question": at least two correct options.

3) For explanations:
- Justify the correct answer(s) by synthesizing information **as if you were observing the video directly** and field the gold_reasoning.
- Reasoning must detail inference steps, **explicitly referencing the required event_ids and both visual + audio evidence to support the stated progression or evolution.**

Figure 14: Prompt used in QA construction in the **Topic/Stance Evolution Summarization** task.

CROSS EVENT CAUSALITY USER PROMPT

video_id: {video_id}
summary: {summary}
events: {events_str}

Task requirements:
1) Generate N=2 **Cross-event Causality Reasoning** question. The question MUST:
- Choose a specific **"result sub-event"**, which is a localized action or state change within a larger event_id.
- Ask to identify the preceding event_id(s) and/or specific sub-event(s) within those event_id(s) that most plausibly served as the direct cause or primary contributing factor to the target result sub-event.
- The causal relationship must span **at least 3 different event_ids**.
- Require BOTH visual and audio evidence to robustly establish the causal link; single modality is INSUFFICIENT.
- Offer plausible options **strictly based on the video's main idea**.
- Be "single_choice_question" or "multiple_choice_question".

2) For the answer:
- Provide exactly 4 options: A, B, C, D. Each option should be an event_ids or a descriptive string of specific sub-events within an event_id.
- Distractor options must be plausible as preceding events/sub-events, but either not align with the actual causal chain, or are entirely incorrect.
- For "single_choice_question": exactly one correct option.
- For "multiple_choice_question": at least two correct options.

3) For explanations:
- Justify the correct answer(s) by synthesizing information **as if you were observing the video directly** and field the g̈old_reasoning..
- Reasoning must detail inference steps, **explicitly referencing the required event_ids and both visual + audio evidence to support the stated progression or evolution.**
- Clearly explain **how** the referenced cause event(s)/sub-event(s) led to the state change or outcome observed in the target result sub-event.

Figure 15: Prompt used in QA construction in the **Cross-event Causality** task.

SCORING PROMPT

You are an evaluator for long audiovisual QA.

You will receive three inputs:
- Audio caption: a textual description of audio events
- Video caption: a textual description of visual events
- QA pair: a question and its proposed answer

Your task:
Evaluate the QA from three perspectives:
1. Sufficiency — Do the captions provide enough evidence to support the answer?
2. Consistency — Is the answer consistent with the described events (no contradictions)?
3. Relevance — Are the captions relevant to the question being asked?

Scoring:
- Each dimension should be assigned a score between 0 and 1.
* 0 = completely unsupported / inconsistent / irrelevant
* 0.5 = partially supported / somewhat consistent / weakly relevant
* 1 = fully supported / consistent / highly relevant

Audio caption: {audio_caption}
Video caption: {video_caption}
QA pair: {qa}

Figure 16: Prompt used for **scoring QAs**.

Failure Case 1: Temporal Localization (Gemini-2.5-Flash)

Question:
When do we see the animated black-and-gray squares transition into the Wamda logo and interview title card while upbeat electronic music plays, followed by the interviewer introducing his guest as the COO focused on creative content generation and training?

Options:

- A: [00:00:00 – 00:00:47]
- B: [00:00:47 – 00:02:15]
- C: [00:08:09 – 00:09:34]
- D: [00:09:34 – 00:10:41]

Correct Answer: A

Gold Reasoning:
Using event_id 0, the visual evidence is the opening sequence with moving black/gray squares leading to the Wamda logo and the on-screen identification of Wayne Borg's interview. Simultaneously, the audio features an upbeat electronic music bed that transitions into the interviewer's spoken introduction of a COO focused on creative content generation and training. This combination of intro animation, music, and initial guest introduction only occurs in 00:00:00–00:00:47. Later segments (options B–D) show two men talking without the opening animation or electronic music intro.

Model Answer: D

Model Reason:
The model claims that the interview title card appears at 00:14:50 and the interviewer introduction occurs at 00:21:50, which do not align with the actual video content. Among the provided options, it selects D as the closest match based on a misinterpreted transition cue, reflecting misrecognition of superficially similar signals or hallucinated timestamps.

Figure 17: Failure Case in Temporal Localization (Gemini-2.5-Flash)

Failure Case 2: Cross-event Causality (Gemini-2.5-Flash)

Question:
In event 11, the video shows a close-up of Andre Bright's business card while he recites office phone numbers, followed by a property presentations slate. Which earlier events/sub-events most directly caused the video to culminate with displaying and announcing Andre's contact details? Select all that apply.

Options:

- A: Event 1 — The host introduces Andre as the fastest solicitor in North Wales and invites him to answer buyers'/sellers' questions.
- B: Event 2 — Andre explains the 6–8 week timeline and stresses early instruction and preparation.
- C: Event 9 — Andre reassures viewers about fixed-fee conveyancing and encourages frequent contact.
- D: Event 0 — Branding cutaway showing the "andy the estate agent" logo.

Correct Answers: A, B, C

Gold Reasoning:
The causal chain builds across the interview before culminating in event 11's on-screen contact reveal. Event 1 establishes Andre's authority and frames the interview as an expert Q&A. Event 2 reinforces this expert advisory role through substantive guidance on timelines and preparation. Event 9 explicitly encourages communication by addressing cost concerns and promoting fixed-fee conveyancing, priming a call-to-action. These steps together motivate the final display of Andre's contact details in event 11. In contrast, the logo cutaway in event 0 is general branding and does not directly contribute to the causal progression.

Model Answer: C

Model Reason:
The model focuses solely on Event 9, as it directly encourages viewers to contact Andre, but fails to integrate earlier events into a multi-step causal chain. This reflects salience bias and limited cross-event integration rather than annotation ambiguity.

Figure 18: Failure Case in Cross-event Causality (Gemini-2.5-Flash)