

LiViBench: An Omnimodal Benchmark for Interactive Livestream Video Understanding

Xiaodong Wang¹, Langling Huang¹, Zhirong Wu¹, Xu Zhao², Teng Xu^{2*}, Xuhong Xia², Peixi Peng^{1*}

¹School of Electronic and Computer Engineering, Peking University

²Douyin Group

{wangxiaodong21s@stu., pypeng@}pku.edu.cn

Abstract

The development of multimodal large language models (MLLMs) has advanced general video understanding. However, existing video evaluation benchmarks primarily focus on non-interactive videos, such as movies and recordings. To fill this gap, this paper proposes the first omnimodal benchmark for interactive livestream videos, LiViBench. It features a diverse set of 24 tasks, highlighting the perceptual, reasoning, and livestream-specific challenges. To efficiently construct the dataset, we design a standardized semi-automatic annotation workflow that incorporates the human-in-the-loop at multiple stages. The workflow leverages multiple MLLMs to form a multi-agent system for comprehensive video description and uses a seed-question-driven method to construct high-quality annotations. All interactive videos in the benchmark include audio, speech, and real-time comments modalities. To enhance models’ understanding of interactive videos, we design tailored two-stage instruction-tuning and propose a Video-to-Comment Retrieval (VCR) module to improve the model’s ability to utilize real-time comments. Based on these advancements, we develop LiVi-LLM-7B, an MLLM with enhanced knowledge of interactive livestreams. Experiments show that our model outperforms larger open-source models with up to 72B parameters, narrows the gap with leading proprietary models on LiViBench, and achieves enhanced performance on general video benchmarks, including VideoMME, LongVideoBench, MLVU, and VideoEval-Pro.

Code —

<https://github.com/Wang-Xiaodong1899/LiViBench>

Introduction

The rapid advancement of Multimodal Large Language Models (MLLMs) has driven significant progress in general video understanding (Comanici et al. 2025; Guo et al. 2025). Challenging benchmarks can drive rapid progress in model capabilities. Accordingly, many recent efforts have advanced the video understanding benchmark ecosystem by introducing more challenging tasks and longer videos (Fu et al. 2025; Wang et al. 2024a). However, existing video benchmarks primarily focus on non-interactive content, such as movies, recordings, and short videos, and lack coverage

of interactive videos, such as livestreams that involve rich and frequent interactions between streamers and audiences. Given the growing prevalence of livestreams in online video consumption, enhancing models’ ability to understand such interactive content has become increasingly important.

Unlike general videos, livestreams are inherently interactive, emphasizing real-time engagement between streamers and their audiences. This interaction includes, but is not limited to, gift-giving, live conversations, audiences’ real-time comments, and multi-person co-streaming. These interactive features demonstrate the unique characteristics of livestreams. Despite recent advancements in MLLMs, it remains unclear how effectively these models can comprehend these types of interactive videos.

To address the lack of interactive video content in existing video understanding benchmarks and evaluate the comprehension capabilities of MLLMs on livestream videos, we introduce LiViBench, the first omnimodal interactive video benchmark. The benchmark focuses on human-centered themes and covers 9 vertical domains of interactive livestream (e.g., chatting and singing), and the task taxonomy comprehensively includes 24 distinct tasks (e.g., multi-person interaction and behavior reasoning). These tasks span a broad range of categories, including general perception and reasoning, knowledge-based question answering, and livestream-specific tasks that highlight the interactive characteristics of livestream scenes. LiViBench comprises 3,168 livestream videos with durations ranging from 14 seconds to 33 minutes, along with 3,175 high-quality multiple-choice questions. At the same time, we introduce rich and heterogeneous data to enable a more comprehensive evaluation, covering audio, speech, and comments.

Constructing a video benchmark poses challenges to video annotation and question quality. While prior work has either lacked transparency (Fu et al. 2025; Wu et al. 2024) or relied entirely on automated annotations (Han et al. 2025), we design a standardized semi-automatic annotation workflow, and incorporate human-in-the-loop at multiple stages. To reduce the bias introduced by MLLMs during automatic annotation, we first use multiple MLLMs to build a multi-agent system that comprehensively describes the video. For each task, we construct a seed question library through automatic generation using proprietary models, followed by human revision and augmentation. Using

*Corresponding author.

a seed-question-driven strategy, models generate candidate questions for each video, which are then screened and refined by humans. Both models and humans provide answers to the selected questions. Finally, human annotators conduct thorough quality control on the resulting multi-choice QA set to ensure clarity, correctness, and relevance.

Based on the constructed LiViBench, we perform a comprehensive evaluation of both state-of-the-art proprietary models and open-source models. Preliminary experiments reveal that proprietary models (e.g., GPT-4o and Gemini-2.5-Pro) exhibit notable limitations in understanding interactive videos, and large-scale open-source models also demonstrate constrained performance. These limitations are likely due to the lack of instruction-tuning datasets specifically designed for interactive videos. To address this, we construct an instruction-tuning dataset comprising 37,953 machine-annotated and 11,180 manually annotated samples for interactive videos. We further propose a tailored two-stage instruction tuning strategy to fully leverage the training data.

In the livestream video domain, real-time comment is a unique and important modality. The sheer volume of these comments poses significant challenges for both the input context length and the information extraction abilities of MLLMs. To evaluate their impact on video understanding, we incorporate the real-time comments into each task. To better leverage these comments, we propose a Video-to-Comment Retrieval (VCR) module that retrieves relevant comments using video features. Together with the tailored instruction tuning, we develop LiVi-LLM-7B, a video understanding model enriched with enhanced knowledge of interactive livestreams. Experimental results demonstrate that our model outperforms larger open-source models with up to 72B parameters and narrows the gap to the best proprietary models on LiViBench. It also shows strong generalization across general video benchmarks of varying lengths, including Video-MME, LongVideoBench, MLVU, and VideoEval-Pro. The contributions are as follows:

- We propose the first omnimodal benchmark specifically designed for interactive livestream videos, LiViBench, with audio, speech, and comment modalities. The comprehensive evaluation shows that some proprietary models (e.g., GPT-4o and Gemini) have limited performance on this new benchmark.
- We propose a standardized semi-automatic annotation workflow that introduces human-in-the-loop in multiple stages. The multi-agent mechanism and seed question library are introduced to efficiently construct high-quality evaluation data and instruction-tuning data.
- We design tailored instruction tuning and a video-to-comment module to build the comprehension-enhanced model: LiVi-LLM-7B. It outperforms larger open-source models, including those with up to 72B parameters, and narrows the gap with the best proprietary models.

Related Work

Multi-Modal Large Language Models

Multimodal Large Language Models (MLLMs) (Hurst et al. 2024; Anthropic 2024; Comanici et al. 2025) have achieved

significant advancements in video understanding tasks. These models typically treat video as a sequence of images and connect the output of the visual encoder to a large language model (LLM) through a modality alignment module, enabling visual content understanding and reasoning. The architecture and training paradigms of video understanding systems continue to evolve. Early works incorporate Q-Former (Li et al. 2023) to extract informative video features while reducing the number of visual tokens (Zhang, Li, and Bing 2023; Ren et al. 2024; Wang et al. 2024b). Then, some works adopt a simpler approach by using an MLP to directly project the processed visual features into the feature space of the LLM (Zhang et al. 2024a; Liu et al. 2025c; Maaz et al. 2023). Recently, Qwen2.5-VL (Bai et al. 2025) fuses adjacent frames at the input stage and further compresses encoded multiple visual tokens into a single token, which is then connected to the language model via MLP. To enhance temporal awareness and achieve event-level localization, TimeChat (Ren et al. 2024) introduces explicit temporal textual prompts, TimeSuite (Zeng et al. 2025) incorporates the TAPE module to capture temporal structures, and Qwen2.5-VL utilizes the MRoPE to model inter-frame temporal relationships. Regarding training data, many works adopt a hybrid data training strategy. For instance, InternVL2.5 (Chen et al. 2024a) and Qwen2.5-VL are trained on a combination of single images, multi-frame image sequences, and videos. Additionally, post-training techniques are widely used to improve reasoning performance (Zhu et al. 2025a; Wang et al. 2025; Wang and Peng 2025).

Moreover, researchers are actively exploring omnimodal models capable of processing text, images, videos, and audio, e.g., (Cheng et al. 2024), (Xu et al. 2025), (Yao et al. 2024), and (Liu et al. 2025b), aiming to further expand the perceptual capabilities of omnimodal systems. Despite the strong performance of the aforementioned models in general video understanding tasks, their adaptability to live streaming scenarios remains underexplored. Related model Kwai-Keye (Team et al. 2025a) is primarily optimized for non-interactive short videos and struggles to process interactive livestream videos. Therefore, our work targets interactive livestream video understanding. We introduce an omnimodal benchmark for interactive videos and develop a model enriched with interactive knowledge.

Multi-Modal Video Benchmarks

The introduction of various benchmarks has promoted the development of MLLMs. Existing video benchmarks mainly focus on general video understanding tasks, such as short video understanding tasks (Xu et al. 2016; Yu et al. 2019; Li et al. 2024; Fang et al. 2024; Hong et al. 2025b; Liu et al. 2024; Shangguan et al. 2024), video temporal grounding tasks (Gao et al. 2017; Lei, Berg, and Bansal 2021; Rohrbach et al. 2014), video reasoning tasks (Hu et al. 2025; Zhao et al. 2025; Rasheed et al. 2025; Zhu et al. 2025b; Han et al. 2025), long video understanding tasks (Fu et al. 2025; Wang et al. 2024a; Zhou et al. 2024; Ma et al. 2025), in order to perceive the real world more comprehensively, some audio-visual benchmarks have been proposed (Li et al. 2022; Geng et al. 2025; Hong et al. 2025a).

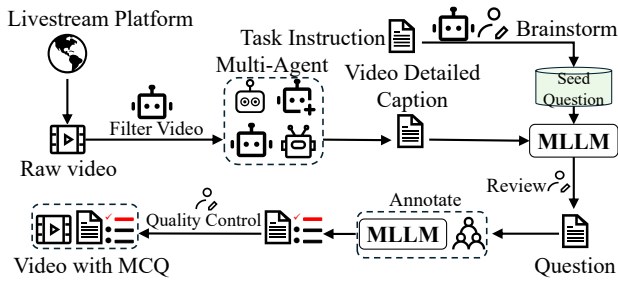


Figure 1: Dataset generation pipeline.

Some other benchmarks focus on specific domains, such as first-person videos (Mangalam, Akshulakov, and Malik 2023), cinematic language (Liu et al. 2025a), video comments (Lei et al. 2025), video quality and aesthetics (Jia et al. 2025), content moderation on short video platforms (Lu et al. 2025), and offline short videos from Kuaishou (Team et al. 2025b). Despite the emergence of benchmarks, current video understanding benchmarks mainly focus on non-interactive videos. With the rapid growth of social media platforms such as Instagram Live and TikTok Live, livestream videos have become increasingly prevalent, bringing with them complex multimodal tasks that challenge existing model capabilities. A benchmark that can evaluate the ability of MLLMs to understand such data is needed. Therefore, we propose LiViBench, the first omnimodal livestream video benchmark.

Method

Video Curation

To evaluate the capability of MLLMs in understanding interactive livestream videos, we curate a diverse and comprehensive dataset from publicly accessible livestream videos. This dataset comprises synchronized multi-modal data, including video, audio, speech, and user comments, which provides rich contextual information. To facilitate human-centered video comprehension, reasoning, and knowledge-based question answering, we focus on livestream categories primarily related to entertainment, including genres like singing, dancing, and chatting. The distribution of these vertical domains is shown in the lower left part of Fig. 2.

Video data filtering To ensure video diversity, we filter out static and simple videos with minimal frame changes. We employ a proprietary model, Seed1.5-VL, to score each video’s spatiotemporal complexity on a scale from 1 to 10, and filter out those scoring below 3. We also exclude videos that focus on web games or e-commerce. After filtering over 30,000 videos, we retain 5,245 videos with durations ranging from 20 seconds to 60 minutes, which form the basis of our benchmark.

Multi-Agent Seed-guided Video QA Generation Pipeline

Constructing a video benchmark requires detailed annotation of the video. At the same time, designing and answering questions based on videos demands careful handling of temporal information, which makes the annotation process

more difficult and costly. Previous works (Fu et al. 2025; Wu et al. 2024) rely entirely on manual annotation to design questions and answers, involving high labor costs. Recent fully automated efforts, such as (Han et al. 2025), rely solely on a single model for video captioning and then use GPT-4o to generate questions and answers. However, this method introduces too many biases from the captioning model, and the result questions and options tend to be lengthy, influenced by the preferences of the language model.

To address these issues, we propose a novel video QA generation pipeline as shown in Fig. 1, characterized by multi-agent annotation, seed-based question-posing, and human-in-the-loop at multiple stages. To mitigate the impact of model bias in video descriptions, we build a multi-agent system composed of several large multimodal models with large parameter sizes, including LLaVA-Video, Qwen2.5-VL, Intern3VL, and Seed1.5-VL. Based on their unique characteristics and capabilities, we develop specialized instructions for each agent, ensuring they generate only the content necessary for their specific tasks. Through the collaboration of different experts, the resulting detailed video description is no longer limited by the capabilities of a single model, but contains richer and more comprehensive content.

We design 24 tasks grouped into 5 categories: 4 coarse-grained perception tasks, 6 fine-grained perception tasks, 3 knowledge-based reasoning tasks, 4 general reasoning tasks, and 7 livestream-specific tasks. To ensure controllability and high-quality QA generation, we introduce a seed question-driven framework that generates questions based on detailed video descriptions. Specifically, we first define task-specific instructions and employ a proprietary model to extract and summarize question patterns from previous work, generating candidate seed questions for each task. Human annotators then review the questions to remove unreasonable or overly simple questions, revising them as needed to form a curated seed question library. Leveraging this library and detailed video descriptions from the multi-agent pipeline, the proprietary model generates candidate questions tailored to specific tasks. Finally, annotators review both the videos and the generated questions, filtering or refining those that are ambiguous, overly simple, or irrelevant. For the video question set of all tasks after quality control, we employ a proprietary model and human annotators to generate answers and propose alternative distractors, forming the initial multiple-choice questions. Subsequently, for each question, annotators are required to carefully review the corresponding video content to verify the correctness of the answer and remove or refine any misleading or inappropriate distractors.

Data Analysis

This section presents a detailed data analysis. The dataset includes 3,168 videos, each with audio, comments, and ASR information. As shown in Fig. 2, the videos mainly belong to entertainment-related livestream categories. Their durations range from 14 seconds to 33 minutes and are grouped into four categories: very short, short, medium, and long. The duration and task distribution are also shown in Fig. 2.

Due to the real-time interactive nature of livestream videos, they are often accompanied by numerous real-time

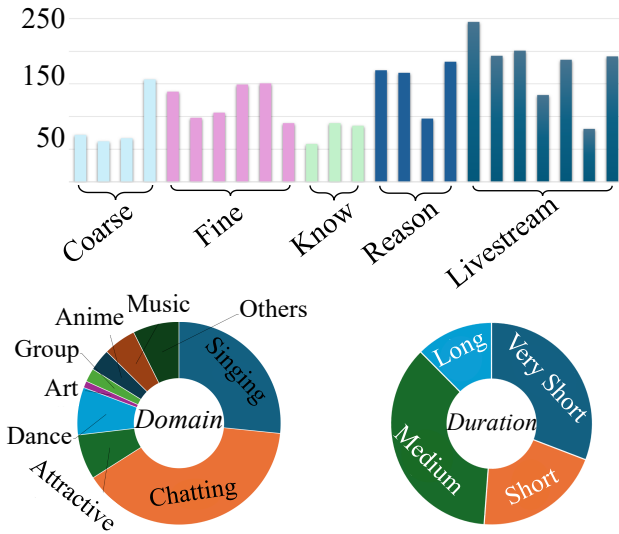


Figure 2: The statistical analysis of our LiViBench.

comments and automatic speech recognition (ASR) transcripts. We analyze the distribution of ASR and comment counts within the videos, as shown in Fig. 3. The dataset contains approximately 1.45 million comments, with an average length of 12.15 Chinese characters. To illustrate the characteristics of livestream videos, we generate word clouds based on the question and option sets, shown in Fig. 4. In the question set, terms like “performance,” “action,” and “interaction” appear frequently, while words such as “anchor” and “audience” dominate the option set. This distribution reflects the benchmark’s focus on livestream-specific features, particularly those related to performance dynamics and interactive behavior.

LiVi-LLM

To enhance the ability of open-source MLLMs to understand interactive livestream videos, we construct an instruction-tuning dataset enriched with interactive knowledge. Based on a carefully designed two-stage instruction tuning strategy and the efficient use of comments through a Video-to-Comment Retrieval (VCR) module in Fig. 5, we develop LiVi-LLM-7B, a high-performance and efficient omnimodal model for livestream video understanding. The following sections detail the model’s training and inference processes.

Instruction Tuning

The left part of Fig. 5 illustrates the architecture and training process of the model. Given a livestream video, we extract the video frames and audio from it, respectively. For continuous video frames, we use Qwen2.5-VL’s visual encoder for spatiotemporal encoding and convert them into video tokens. For audio streams, we use Qwen2-Audio to encode them into audio tokens. To fuse the video and audio representations, a transformer decoder is used to aggregate the features. The fused tokens are then input into the large language model along with the query’s text tokens. The parameters of the model are initialized from Qwen-2.5-Omni.

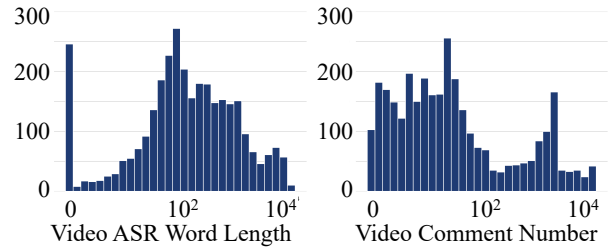


Figure 3: The ASR and comment distribution.



Figure 4: The word clouds of our LiViBench.

The data generation pipeline proposed in Sec. can be directly extended to the construction of instruction fine-tuning data in the livestream field. We collect over 100,000 livestream videos, each ranging from 1 to 5 minutes in length. To ensure content richness, we use a multimodal large language model to analyze the number of distinct scenes in each video and retain only those containing more than one scene. For each selected video, agents consisting of one or more models generate 1 to 3 questions along with corresponding answers. To improve the accuracy of instruction tuning data while balancing annotation costs, we sample a portion of the data for human review and annotation. These higher-quality samples are used for fine-grained tuning of the model. In total, we obtain 37,953 synthetic samples without human review and 11,180 manually refined samples.

During instruction tuning, we adopt a carefully designed two-stage training strategy. In the first stage, the model is fine-tuned on synthetic data without manual review to align the model to the interactive video domain. To balance domain specificity with generalization, we also incorporate general video data (Zhang et al. 2024b), helping the model retain general video understanding capabilities. In the second stage, we perform fine-grained tuning using manually annotated data to further enhance the model’s accuracy and robustness on video understanding tasks.

Inference with Video-to-Comment Retrieval

Instruction tuning on the specific interactive video domain can enhance the understanding ability of MLLMs for livestream videos. But the core difference between livestream videos and general videos is real-time interaction, such as real-time user comments during live streaming. However, the massive amount of real-time comments poses a huge challenge to the model’s context and information extraction capabilities. To this end, this paper proposes a Video-to-Comment Retrieval (VCR) module that aims to obtain more relevant comments from massive comments.

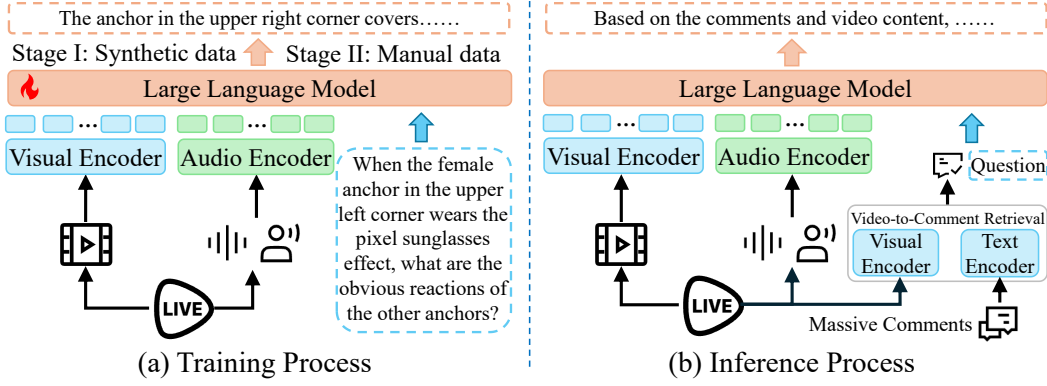


Figure 5: LiVi-LLM architecture. (a) Training process: In the first stage, the model is aligned to the interactive video domain using synthetic data; in the second stage, it undergoes fine-grained tuning with manual data. (b) Inference process: The model integrates a video-to-comment retrieval module to fully use omni-modality to enhance the comprehensive understanding.

For efficient retrieval, we uniformly sample video frames from a video, and use Chinese-CLIP (Yang et al. 2022) to obtain embeddings for each frame. We use a text encoder to encode all comments into text embeddings. By calculating the similarity between frame embeddings and text embeddings, we can obtain the top-k relevant comments corresponding to each frame. All retrieved relevant comments are sorted in chronological order, and together with the question as text context, they are input into the fine-tuned model.

Experiment

Settings

The proposed benchmark includes a total of 3175 Video QA. We evaluate 24 models and use the most suitable number of frames for each model for reasoning while ensuring the context does not overflow. The benchmark comprises 24 subtasks, categorized into five evaluation groups: 4 coarse-grained general perception tasks (Coarse), 6 fine-grained general perception tasks (Fine), 3 knowledge-based question answering tasks (Know), 4 general reasoning tasks (Reason), and 7 livestream-specific tasks (Livestream). For evaluation, we report the average score of all models within each category and the overall score of the entire benchmark. More details refer to the supplementary material.

Main Results on LiViBench

The comprehensive evaluation results for LiViBench are presented in Tab. 1. The results indicate that general perception tasks (e.g., Coarse and Fine) consistently outperform other task types in terms of accuracy, for both proprietary and open-source models. The livestream-specific tasks are the most challenging for all models. Among proprietary models, Gemini-2.5-Pro and GPT-4o show limited performance on this benchmark. In contrast, Doubao-Seed-1.6 and Seed1.5-VL show promising results, with Seed1.5-VL attaining the highest overall score of 66.2%. Doubao-Seed-1.6 performs best on Coarse and Fine tasks, while Seed1.5-VL leads in tasks including knowledge, reasoning, and livestream-specific, which indicates that the model

has superior expert knowledge and reasoning ability in the livestream video domain. Among all open-source models, our model LiVi-LLM-7B significantly outperforms other large-scale open-source models with up to 72B parameters, including Qwen2.5-VL-72B. It also surpasses proprietary models such as GPT-4o and Gemini-2.5-Pro. Notably, LiVi-LLM achieves the best overall accuracy of 64.4%, matching the top-performing InternVL3-78B model. To showcase the unique features of the benchmark, we provide qualitative examples in Fig. 6. Compared to general proprietary models, our model demonstrates enhanced understanding and reasoning abilities in interactive video scenarios.

Results on General Video Benchmarks

To demonstrate the comprehensive capabilities of our model, we conduct experiments on various general video benchmarks, including Video-MME (Fu et al. 2025), MLVU (Zhou et al. 2024), LongVideoBench (LongVB) (Wu et al. 2024), and VideoEval-Pro (Ma et al. 2025). In Tab. 3, compared with the state-of-the-art models of similar parameter size, our model demonstrates promising results across all benchmarks, including the best scores on all tasks of Video-MME and VideoEval-Pro. These results indicate that our model not only excels in interactive live video understanding but also exhibits strong generalization capabilities.

Analysis

Audio impact analysis The results of the impact of audio for omnimodal models are in Tab. 2, including MiniCPM-o-26, Qwen2.5-Omni and LiVi-LLM-7B. MiniCPM-o-26 shows a performance drop across the first four categories, likely due to its limited ability to process audio. In comparison, both Qwen2.5-Omni and our model show notable improvement in most categories. We observe that all 3 models showed significant improvements in the livestream-specific category. This suggests that in the interactive video domain, audio plays a crucial role in improving video understanding.

Speech impact analysis The results of the impact of speech modality on model evaluation are also shown in



Figure 6: Qualitative examples of multi-choice questions from our LiViBench. The correct options are marked in green.

Model	Overall	Coarse	Fine	Know	Reason	Livestream
Proprietary MLLMs						
Gemini 2.5 Flash (Comanici et al. 2025)	53.0	63.6	62.9	56.4	51.5	43.9
Gemini 2.5 Pro (Comanici et al. 2025)	56.1	65.0	68.4	58.1	51.3	48.2
GPT-4o (Hurst et al. 2024)	56.3	67.0	66.5	57.6	55.2	47.4
Seed1.5-VL (Guo et al. 2025)	66.2	70.9	71.4	68.8	70.7	59.1
Doubao-Seed-1.6(ByteDance 2025)	64.9	72.9	73.2	60.2	68.4	56.8
Open-Source MLLMs						
LLaVA-Video-72B (Zhang et al. 2024b)	60.0	65.3	70.2	63.6	63.8	49.8
Qwen2.5-VL-32B (Bai et al. 2025)	59.4	73.1	69.1	57.6	61.3	49.1
Qwen2.5-VL-72B (Bai et al. 2025)	62.3	73.4	72.4	61.9	64.6	52.0
InternVL3-14B (Zhu et al. 2025a)	62.7	71.7	68.9	65.3	67.0	53.7
InternVL3-38B (Zhu et al. 2025a)	64.1	70.9	72.6	66.6	68.3	54.5
InternVL3-78B (Zhu et al. 2025a)	64.4	72.0	69.8	65.8	69.3	56.3
LLaVA-NeXT-Video (Zhang et al. 2024a)	37.5	40.7	40.5	44.8	32.3	36.1
InternVL2-8B (Chen et al. 2024b)	49.8	59.7	59.2	53.8	53.6	38.7
MiniCPM-v-26 (Yao et al. 2024)	50.5	58.9	59.9	58.5	54.2	39.1
LLaVA-Video-7B (Zhang et al. 2024b)	52.6	60.0	59.0	55.1	58.3	43.5
NVILA-8B-Video (Liu et al. 2025c)	53.3	61.1	56.2	58.1	56.8	46.6
Video-LLaMA3-8B (Zhang et al. 2025)	54.1	60.0	62.1	61.9	58.8	43.7
Keye-VL-8B-Preview (Team et al. 2025a)	55.2	68.7	65.4	51.7	52.6	47.2
MiniCPM-o-26 [†] (Yao et al. 2024)	56.0	65.3	62.1	61.9	59.9	46.5
InternVL2.5-8B (Chen et al. 2024a)	56.6	68.1	62.9	58.5	59.9	47.5
Qwen2.5-VL-7B (Bai et al. 2025)	58.3	65.5	69.2	57.6	59.7	49.1
InternVL3-8B (Zhu et al. 2025a)	59.8	68.4	66.9	58.5	63.0	51.7
InternVL3-9B (Zhu et al. 2025a)	60.0	67.3	67.8	59.4	63.0	51.7
Qwen2.5-Omni-7B [†] (Xu et al. 2025)	60.3	68.1	68.5	59.4	60.7	53.1
LiVi-LLM-7B [†] (Ours)	64.4	70.1	68.7	62.8	63.6	60.9

Table 1: LiViBench evaluation results across all categories. [†] indicates omnimodal models.

Model	Overall			Coarse			Fine			Know			Reason			Livestream		
	V	+A	+S	V	+A	+S	V	+A	+S	V	+A	+S	V	+A	+S	V	+A	+S
LLaVA-Video-7B	52.6	NA	55.4↑	60.0	NA	60.6↑	59.0	NA	61.0↑	55.1	NA	58.1↑	58.3	NA	58.8↑	43.5	NA	48.4↑
MiniCPM-o-26	56.0	54.7	57.9↑	65.3	65.3	67.0↑	62.1	61.2	61.8	61.9	58.5	62.8↑	59.9	51.6	59.7	46.5	48.5↑	51.2↑
Qwen2.5-Omni-7B	57.8	60.3↑	60.2↑	66.2	68.1↑	67.5↑	67.4	68.5↑	66.9	57.2	59.4↑	59.4↑	62.6	60.7	63.1↑	47.4	53.1↑	52.8↑
Qwen2.5-VL-7B	58.3	NA	59.8↑	65.5	NA	67.0↑	69.2	NA	68.4	57.6	NA	60.2↑	59.7	NA	59.4	49.1	NA	52.7↑
InternVL3-8B	59.8	NA	61.4↑	68.4	NA	70.6↑	66.9	NA	65.3	58.5	NA	59.4↑	63.0	NA	63.3↑	51.7	NA	55.8↑
LiVi-LLM-7B	61.4	63.9↑	63.4↑	70.6	70.9↑	68.7	69.1	68.6	68.7	63.2	65.3↑	62.8	61.7	64.2↑	62.1↑	53.6	58.7↑	59.6↑

Table 2: Analysis of the impact of audio and speech. V: Video, A: Audio, S: Speech. We use ASR to represent speech modality.

Model	Video-MME				MLVU	LongVB	VideoEval-Pro
	Short	Med	Long	Overall	M-Avg	val total	MCQ
Proprietary MLLMs							
GPT-4o (Hurst et al. 2024)	80.0	70.3	65.3	71.9	64.6	66.7	59.5
Seed1.5-VL (Guo et al. 2025)	-	-	-	77.9	82.1	74.0	66.6
Doubao-Seed-1.6(ByteDance 2025)	84.1	75.3	70.1	76.5	76.0	71.3	62.4
Open-Source MLLMs							
InternVL2-8B (Chen et al. 2024b)	68.0	52.0	48.9	56.3	56.3	54.6	39.9
Qwen2.5-VL-7B (Bai et al. 2025)	75.9	66.8	54.1	65.6	65.1	61.0	46.9
InternVL2.5-8B (Chen et al. 2024a)	75.3	61.5	55.8	64.2	68.9	60.0	45.5
InternVL3-8B (Zhu et al. 2025a)	77.5	67.3	54.1	66.3	71.4	58.8	48.4
Qwen2.5-Omni-7B (Xu et al. 2025)	78.1	67.4	57.6	67.7	70.0	58.7	48.9
LiVi-LLM-7B (Ours)	79.6	69.4	56.6	68.5	70.5	59.6	50.5

Table 3: Evaluation results on general video benchmarks.

Level	InternVL3-8B			Qwen2.5-VL-7B			LLaVA-Video-7B			Qwen2.5-Omni-7B			LiVi-LLM-7B		
	V	+Raw	+VCR	V	+Raw	+VCR	V	+Raw	+VCR	V	+Raw	+VCR	V	+Raw	+VCR
[0, 20)	60.6	59.3	58.8	59.9	59.3	59.0	52.8	54.1↑	54.3↑	60.4	60.8↑	61.3↑	63.7	63.6	63.8↑
[20, 100)	56.6	60.7↑	61.6↑	53.5	55.5↑	56.3↑	47.9	51.1↑	53.0↑	60.3	58.7	57.7	64.2	65.7↑	65.6↑
[100, 1k)	62.0	61.0	61.2	62.6	56.9	58.8	56.4	57.2↑	57.7↑	62.3	62.0	64.5↑	64.5	64.5	66.1↑
[1k, ∞)	60.6	59.9	63.4↑	57.5	49.0	54.9	57.1	29.5	58.8↑	58.2	54.7	59.7↑	63.8	55.3	63.0
Overall	59.8	60.0↑	60.4↑	58.3	56.6	57.7	52.6	50.2	55.0↑	60.3	59.5	60.5↑	63.9	63.0	64.4↑

Table 4: Analysis video comments impact. Raw: using raw comments. VCR: using Video-to-Comment Retrieval module.

Stage I	Stage II	LiViBench	Video-MME		
		Overall	Short	Med	Long
Baseline	-	60.3	78.1	67.4	57.6
Ours	X	62.9	<u>80.0</u>	<u>69.2</u>	<u>57.7</u>
Ours	Ours	63.9	79.6	69.4	56.6
LV+Ours	X	62.2	80.6	69.4	57.8
LV+Ours	Ours	<u>63.1</u>	80.6	69.4	57.3

Table 5: Ablation study of the data used in different training stages. LV indicates data from LLaVA-Video-178k.

Tab. 2. We use ASR to represent the speech modality. All models perform better in most categories and obtain an overall performance improvement. Using speech in fine-grained and reasoning tasks sometimes degrades model performance, suggesting that speech can sometimes be noisy and affect the understanding of video details. Comparing audio and speech modalities, we find that audio modality is more helpful. These findings highlight the importance of effectively leveraging modality information in interactive video understanding and may inspire future work.

Video comment impact analysis In order to analyze the processing capabilities of different models for comments, we divide all the comments in the benchmark into 4 levels according to comment number, including [0, 20), [20, 100), [100, 1k), [1k, ∞), as shown in Tab. 4. Using raw comments degrades the performance of most models, but InternVL3 is an exception since it uses far fewer frames and thus has more context space. To mitigate the adverse effect of massive comments, we propose VCR module. This module retrieves key comments, reduces the adverse effects of massive comments, and brings better performance than video-only.

Ablation Study

Impact of training data domain We test the data domain used in different stages in Tab. 5. We test different data used in the first stage. We can see that the first stage alignment training using only our synthetic data achieved better results on LiViBench. When general data (Zhang et al. 2024b) is added, performance on LiViBench decreases but improves across all three categories of Video-MME. In the second stage, further fine-tuning using our manual data continuously improves the performance of LiViBench while maintaining good generalization. This shows that our training data is effective in improving the performance on both LiViBench and general benchmarks. However, there is still a trade-off between interactive understanding tasks and general tasks. These findings indicate that interactive video data contributes meaningfully in both training stages, with the second stage playing a key role in strengthening the model’s interactive livestream knowledge.

Conclusion

This paper introduces LiViBench, the first omnimodal benchmark for interactive livestream video understanding. To construct LiViBench, we develop a standardized semi-automatic workflow for efficient annotation and data quality. We incorporate audio, speech, and real-time comments to build an omnimodal benchmark and perform comprehensive evaluation and analysis. In addition, we carefully build LiVi-LLM-7B, a video model enhanced with interactive knowledge that outperforms open-source models with up to 72B parameters, establishing a strong baseline and a new foundation for future research in interactive video understanding.

Acknowledgments

The study was funded by the Shenzhen Science and Technology Program (KQTD20240729102051063), the National Natural Science Foundation of China under contracts No. 62422602, No. 62372010, No. 62425101, No. 62332002, No. 62372010, and No. 62206281.

References

- Anthropic. 2024. Introducing the next generation of Claude, 2024. <https://www.anthropic.com/news/claude-3-5-sonnet>.
- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- ByteDance. 2025. Introduction to Techniques Used in Seed1.6. https://seed.bytedance.com/en/seed1_6.
- Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024a. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.
- Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; et al. 2024b. How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites. *arXiv preprint arXiv:2404.16821*.
- Cheng, Z.; Leng, S.; Zhang, H.; Xin, Y.; Li, X.; Chen, G.; Zhu, Y.; Zhang, W.; Luo, Z.; Zhao, D.; et al. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.
- Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Fang, X.; Mao, K.; Duan, H.; Zhao, X.; Li, Y.; Lin, D.; and Chen, K. 2024. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *NeurIPS*, 37: 89098–89124.
- Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 2025. Videomme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the CVPR*, 24108–24118.
- Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, 5267–5275.
- Geng, T.; Zhang, J.; Wang, Q.; Wang, T.; Duan, J.; and Zheng, F. 2025. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *Proceedings of the CVPR*, 18959–18969.
- Guo, D.; Wu, F.; Zhu, F.; Leng, F.; Shi, G.; Chen, H.; Fan, H.; Wang, J.; Jiang, J.; Wang, J.; et al. 2025. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*.
- Han, S.; Huang, W.; Shi, H.; Zhuo, L.; Su, X.; Zhang, S.; Zhou, X.; Qi, X.; Liao, Y.; and Liu, S. 2025. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In *Proceedings of the CVPR*, 26181–26191.
- Hong, J.; Yan, S.; Cai, J.; Jiang, X.; Hu, Y.; and Xie, W. 2025a. Worldsense: Evaluating real-world omni-modal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*.
- Hong, W.; Cheng, Y.; Yang, Z.; Wang, W.; Wang, L.; Gu, X.; Huang, S.; Dong, Y.; and Tang, J. 2025b. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *Proceedings of the CVPR*, 8450–8460.
- Hu, K.; Wu, P.; Pu, F.; Xiao, W.; Zhang, Y.; Yue, X.; Li, B.; and Liu, Z. 2025. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*.
- Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Jia, Z.; Zhang, Z.; Zhang, Z.; Liang, Y.; Zhu, X.; Li, C.; Han, J.; Wu, H.; Wang, B.; Zhang, H.; et al. 2025. Scaling-up Perceptual Video Quality Assessment. *arXiv preprint arXiv:2505.22543*.
- Lei, J.; Berg, T. L.; and Bansal, M. 2021. Detecting moments and highlights in videos via natural language queries. *NeurIPS*, 34: 11846–11858.
- Lei, Y.; Zhang, C.; Liu, Z.; Leng, H.; Liu, S.; Gao, T.; Liu, Q.; and Wang, Y. 2025. GODBench: A Benchmark for Multimodal Large Language Models in Video Comment Art. *arXiv preprint arXiv:2505.11436*.
- Li, G.; Wei, Y.; Tian, Y.; Xu, C.; Wen, J.-R.; and Hu, D. 2022. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the CVPR*, 19108–19118.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.
- Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the CVPR*, 22195–22206.
- Liu, H.; He, J.; Jin, Y.; Zheng, D.; Dong, Y.; Zhang, F.; Huang, Z.; He, Y.; Li, Y.; Chen, W.; et al. 2025a. ShotBench: Expert-Level Cinematic Understanding in Vision-Language Models. *arXiv preprint arXiv:2506.21356*.
- Liu, Y.; Li, S.; Liu, Y.; Wang, Y.; Ren, S.; Li, L.; Chen, S.; Sun, X.; and Hou, L. 2024. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*.
- Liu, Z.; Dong, Y.; Wang, J.; Liu, Z.; Hu, W.; Lu, J.; and Rao, Y. 2025b. Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment. *arXiv e-prints*, arXiv–2502.

- Liu, Z.; Zhu, L.; Shi, B.; Zhang, Z.; Lou, Y.; Yang, S.; Xi, H.; Cao, S.; Gu, Y.; Li, D.; et al. 2025c. Nvila: Efficient frontier visual language models. In *Proceedings of the CVPR*, 4122–4134.
- Lu, X.; Zhang, T.; Meng, C.; Wang, X.; Wang, J.; Zhang, Y.; Tang, S.; Liu, C.; Ding, H.; Jiang, K.; et al. 2025. VLM as Policy: Common-Law Content Moderation Framework for Short Video Platform. *arXiv preprint arXiv:2504.14904*.
- Ma, W.; Ren, W.; Jia, Y.; Li, Z.; Nie, P.; Zhang, G.; and Chen, W. 2025. Videoeval-pro: Robust and realistic long video understanding evaluation. *arXiv preprint arXiv:2505.14640*.
- Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.
- Mangalam, K.; Akshulakov, R.; and Malik, J. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *NeurIPS*, 36: 46212–46244.
- Rasheed, H.; Shaker, A.; Tang, A.; Maaz, M.; Yang, M.-H.; Khan, S.; and Khan, F. S. 2025. VideoMathQA: Benchmarking Mathematical Reasoning via Multimodal Understanding in Videos. *arXiv preprint arXiv:2506.05349*.
- Ren, S.; Yao, L.; Li, S.; Sun, X.; and Hou, L. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the CVPR*, 14313–14323.
- Rohrbach, A.; Rohrbach, M.; Qiu, W.; Friedrich, A.; Pinkal, M.; and Schiele, B. 2014. Coherent multi-sentence video description with variable level of detail. In *German conference on pattern recognition*, 184–195. Springer.
- Shangguan, Z.; Li, C.; Ding, Y.; Zheng, Y.; Zhao, Y.; Fitzgerald, T.; and Cohan, A. 2024. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models. *arXiv preprint arXiv:2410.23266*.
- Team, K. K.; Yang, B.; Wen, B.; Liu, C.; Chu, C.; Song, C.; Rao, C.; Yi, C.; Li, D.; Zang, D.; et al. 2025a. Kwai Key-VL Technical Report. *arXiv preprint arXiv:2507.01949*.
- Team, K. K.; Yang, B.; Wen, B.; Liu, C.; Chu, C.; Song, C.; Rao, C.; Yi, C.; Li, D.; Zang, D.; et al. 2025b. Kwai Key-VL Technical Report. *arXiv preprint arXiv:2507.01949*.
- Wang, W.; He, Z.; Hong, W.; Cheng, Y.; Zhang, X.; Qi, J.; Gu, X.; Huang, S.; Xu, B.; Dong, Y.; et al. 2024a. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*.
- Wang, X.; Huang, J.; Yuan, L.; and Peng, P. 2025. LeanPO: Lean Preference Optimization for Likelihood Alignment in Video-LLMs. *arXiv preprint arXiv:2506.05260*.
- Wang, X.; and Peng, P. 2025. Open-r1-video.
- Wang, Y.; Li, K.; Li, X.; Yu, J.; He, Y.; Chen, G.; Pei, B.; Zheng, R.; Wang, Z.; Shi, Y.; et al. 2024b. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, 396–416. Springer.
- Wu, H.; Li, D.; Chen, B.; and Li, J. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *NeurIPS*, 37: 28828–28857.
- Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; et al. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.
- Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the CVPR*, 5288–5296.
- Yang, A.; Pan, J.; Lin, J.; Men, R.; Zhang, Y.; Zhou, J.; and Zhou, C. 2022. Chinese CLIP: Contrastive Vision-Language Pretraining in Chinese. *arXiv preprint arXiv:2211.01335*.
- Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800*.
- Yu, Z.; Xu, D.; Yu, J.; Yu, T.; Zhao, Z.; Zhuang, Y.; and Tao, D. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI*, volume 33, 9127–9134.
- Zeng, X.; Li, K.; Wang, C.; Li, X.; Jiang, T.; Yan, Z.; Li, S.; Shi, Y.; Yue, Z.; Wang, Y.; Wang, Y.; Qiao, Y.; and Wang, L. 2025. TimeSuite: Improving MLLMs for Long Video Understanding via Grounded Tuning. In *The Thirteenth International Conference on Learning Representations*.
- Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; et al. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.
- Zhang, H.; Li, X.; and Bing, L. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.
- Zhang, Y.; Li, B.; Liu, h.; Lee, Y. j.; Gui, L.; Fu, D.; Feng, J.; Liu, Z.; and Li, C. 2024a. LLaVA-NeXT: A Strong Zero-shot Video Understanding Model.
- Zhang, Y.; Wu, J.; Li, W.; Li, B.; Ma, Z.; Liu, Z.; and Li, C. 2024b. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.
- Zhao, Y.; Zhang, H.; Xie, L.; Hu, T.; Gan, G.; Long, Y.; Hu, Z.; Chen, W.; Li, C.; Xu, Z.; et al. 2025. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the CVPR*, 8475–8489.
- Zhou, J.; Shu, Y.; Zhao, B.; Wu, B.; Xiao, S.; Yang, X.; Xiong, Y.; Zhang, B.; Huang, T.; and Liu, Z. 2024. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv e-prints*, arXiv–2406.
- Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025a. Internv13: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.
- Zhu, K.; Jin, Z.; Yuan, H.; Li, J.; Tu, S.; Cao, P.; Chen, Y.; Liu, K.; and Zhao, J. 2025b. MMR-V: What’s Left Unsaid? A Benchmark for Multimodal Deep Reasoning in Videos. *arXiv preprint arXiv:2506.04141*.