CrossVid: A Comprehensive Benchmark for Evaluating Cross-Video Reasoning in Multimodal Large Language Models

Jingyao Li[1*], Jingyun Wang[1*], Molin Tan[1*], Haochen Wang[1], Cilin Yan[1], Likun Shi[1], Jiayin Cai[1†], Xiaolong Jiang[1], Yao Hu[1]

[1]Xiaohongshu Inc., China

{lijingyao, tanmolin, shilikun, laige}@xiaohongshu.com, {clyanhh, yaoohu}@gmail.com, 1411249598@qq.com, h.wang3@uva.nl, caijy18@tsinghua.org.cn

Abstract

Cross-Video Reasoning (CVR) presents a significant challenge in video understanding, which requires simultaneous understanding of multiple videos to aggregate and compare information across groups of videos. Most existing video understanding benchmarks focus on single-video analysis, failing to assess the ability of multimodal large language models (MLLMs) to simultaneously reason over various videos. Recent benchmarks evaluate MLLMs' capabilities on multi-view videos that capture different perspectives of the same scene. However, their limited tasks hinder a thorough assessment of MLLMs in diverse real-world CVR scenarios. To this end, we introduce **CrossVid**, the first benchmark designed to comprehensively evaluate MLLMs' spatial-temporal reasoning ability in cross-video contexts. Firstly, CrossVid encompasses a wide spectrum of hierarchical tasks, comprising four high-level dimensions and ten specific tasks, thereby closely reflecting the complex and varied nature of real-world video understanding. Secondly, CrossVid provides 5,331 videos, along with 9,015 challenging question-answering pairs, spanning single-choice, multiple-choice, and open-ended question formats. Through extensive experiments on various open-source and closed-source MLLMs, we observe that Gemini-2.5-Pro performs best on CrossVid, achieving an average accuracy of 50.4%. Notably, our in-depth case study demonstrates that most current MLLMs struggle with CVR tasks, primarily due to their inability to integrate or compare evidence distributed across multiple videos for reasoning. These insights highlight the potential of CrossVid to guide future advancements in enhancing MLLMs' CVR capabilities.

Datasets — https://github.com/chuntianli666/CrossVid

Introduction

With the rapid development of multimodal large language models (MLLMs) (Bai et al. 2025; Hurst et al. 2024; Comanici et al. 2025), video reasoning (Lin et al. 2023; Chen et al. 2024; Shu et al. 2025; Zhang, Li, and Bing 2023) emerges as an important testbed for evaluating the reasoning capabilities of MLLMs. However, most existing benchmarks (Fu et al. 2025; Xiao et al. 2021; Yu et al. 2019) for video reasoning primarily focus on single-video analysis, severely restricting the assessment of MLLMs' reasoning capabilities on more complex tasks that span multiple videos.

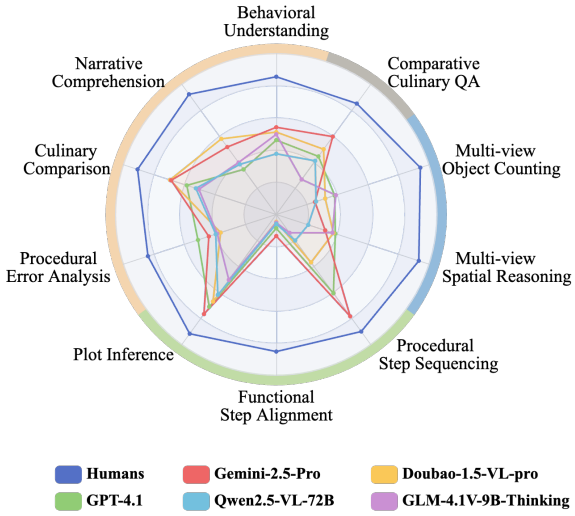


Figure 1: Performance of MLLMs on CrossVid.

Cross-Video Reasoning (CVR) is a challenging yet essential task within the domain of video reasoning. Aiming to aggregate and compare information across videos, CVR requires to simultaneously understand multiple videos. A recent research, All-Angles Bench (Yeh et al. 2025), evaluates MLLMs' performance on groups of multi-view videos, where each video captures a different perspective of the same scene. However, the task of All-Angles Bench is limited to multi-view videos showing the same scene, hindering a thorough assessment of MLLMs' CVR abilities across the diverse and complex scenarios in the real world.

To this end, we propose **CrossVid**, the first video reasoning benchmark designed to advance from previous single-query, single-video paradigms to single-query, multi-video understanding, and to comprehensively evaluate MLLMs' spatial-temporal reasoning ability for CVR. CrossVid features a wide range of hierarchical tasks, reflecting the diversity of real-world video understanding scenarios. It consists of 4 high-level dimensions, including comparative anal-

*These authors contributed equally.

†Corresponding author.

Benchmarks	#Videos	#QA pairs	Len. (s)	#Tasks	Anno.	Closed-ended	Open-ended	Multi-video	Multi-view
TVQA (Lei et al. 2018)	2,179	15,253	11	3	M	✓	✗	✗	✗
MVBench (Li et al. 2024b)	3,641	4,000	16	20	A	✓	✗	✗	✗
ActivityNet-QA (Yu et al. 2019)	5,800	58,000	180	4	M	✗	✓	✗	✗
NExT-QA (Xiao et al. 2021)	5,440	52,044	44	2	M	✓	✓	✗	✗
LongVideoBench (Wu et al. 2024)	3,763	6,678	473	17	M	✓	✗	✗	✗
MMVU (Zhao et al. 2025)	1,529	3,000	51	27	M	✓	✓	✗	✗
Video-MME (Fu et al. 2025)	900	2,700	1,017	12	M	✓	✗	✗	✗
MLVU (Zhou et al. 2024)	1,730	3,102	930	9	M+A	✓	✓	✗	✗
Ego-Exo4D (Grauman et al. 2024)	5,035	-	156	4	M	✗	✗	✓	✓
EgoExoLearn (Huang et al. 2024)	747	-	-	4	M	✓	✓	✓	✓
All-Angles Bench (Yeh et al. 2025)	90 scenes	2,132	-	6	M	✓	✗	✓	✓
CrossVid (Ours)	5,331	9,015	215	10	M+A	✓	✓	✓	✓

Table 1: Comparison of CrossVid with existing benchmarks, including **#Videos** (number of videos), **#QA pairs** (number of QA pairs), **Len.** (average length of videos in seconds), **#Tasks** (number of distinct tasks), **Anno.** (annotation pipeline, M: manual, A: automated), presence of **Open-ended** and **Closed-ended** tasks, presence of **Multi-video reasoning** and **Multi-view reasoning**.

ysis, temporal understanding, multi-view reasoning, and free-form QA, encompassing a total of 10 specific tasks. CrossVid consists of 5,331 videos and 9,015 challenging QA pairs, covering single-choice, multiple-choice, and open-ended question formats. On average, each query requires MLLMs to comprehend approximately 770 seconds of video content. To ensure precise annotation, we develop a semi-automated annotation pipeline and employ 10 expert annotators to facilitate the construction.

Extensive experiments on CrossVid are conducted on various representative closed-source (Hurst et al. 2024; Comanici et al. 2025) and open-source MLLMs (Bai et al. 2025; Zhu et al. 2025), ranging from 7B to 78B parameters and diverse architectures. As shown in Figure 1, while current MLLMs excel in single-video tasks, they still struggle with CVR. Notably, Gemini-2.5-Pro achieves the best average accuracy of 50.4%. Furthermore, we provide multiple key insights based on the experimental results, which further verify that our proposed CrossVid establishes new pathways for MLLMs' future advancements in video reasoning. Our detailed case studies and ablation experiments confirm that, despite ongoing advances, MLLMs still struggle to aggregate and compare evidence distributed across videos—a fundamental capability for real-world CVR.

In summary, our main contributions are:

- We propose **CrossVid**, the first benchmark to systematically evaluate MLLMs' CVR capability. CrossVid incorporates hierarchical tasks spanning four high-level dimensions and ten specific tasks. The dataset is constructed with a semi-automated annotation pipeline under rigorous quality controls. It contains 9,015 high-quality QA pairs and 5,331 videos, including both closed-ended and open-ended question formats.
- We conduct extensive experiments on 22 representative closed-source and open-source MLLMs. Our detailed case analysis and ablation studies offer critical insights into current limitations of MLLMs for CVR, paving the way for improvements in future development in video understanding for MLLMs.

Related Work

Video Understanding MLLMs

MLLMs have demonstrated remarkable advancements in video understanding by integrating visual encoders with large language models and fine-tuning on downstream tasks (Zhang, Li, and Bing 2023; Lin et al. 2023; Zhang et al. 2024b). Previous works mainly focus on single-video understanding, utilizing key frame selection (Tang et al. 2025; Gong et al. 2025) and token compression (Shu et al. 2025; Song et al. 2024) to accomplish tasks such as video captioning, action recognition, and long-video comprehension. Notably, models like Qwen2.5-VL (Bai et al. 2025) have shown the ability to process hour-long videos with improved temporal reasoning.

However, despite these advances, most existing open-source MLLMs remain untrained for comprehensive CVR (*i.e.*, joint inference across multiple input videos). A few recent works (Reilly et al. 2025) begin to explore cross-perspective understanding, but do not generalize to broader multi-video settings.

Video QA Benchmarks

Video Question Answering (VQA) benchmarks have mostly focused on assessing models' abilities to understand and reason over single videos. Early works such as TVQA (Lei et al. 2018) and ActivityNet-QA (Yu et al. 2019) require understanding short or long video clips via closed- or open-ended questions. To target spatial and temporal reasoning, datasets like NExT-QA (Xiao et al. 2021) and LongVideoBench (Wu et al. 2024) are introduced. Some recent efforts, such as MVBench (Li et al. 2024b) and Video-MME (Fu et al. 2025), extend the diversity of covered tasks.

More recent works start to target multi-view reasoning. For example, EgoExoLearn (Huang et al. 2024) evaluates across exocentric and egocentric views, and All-Angles Bench (Yeh et al. 2025) introduces multi-view video QAs. However, their scale and task coverage are limited, and they primarily focus on specific domains or perspectives.

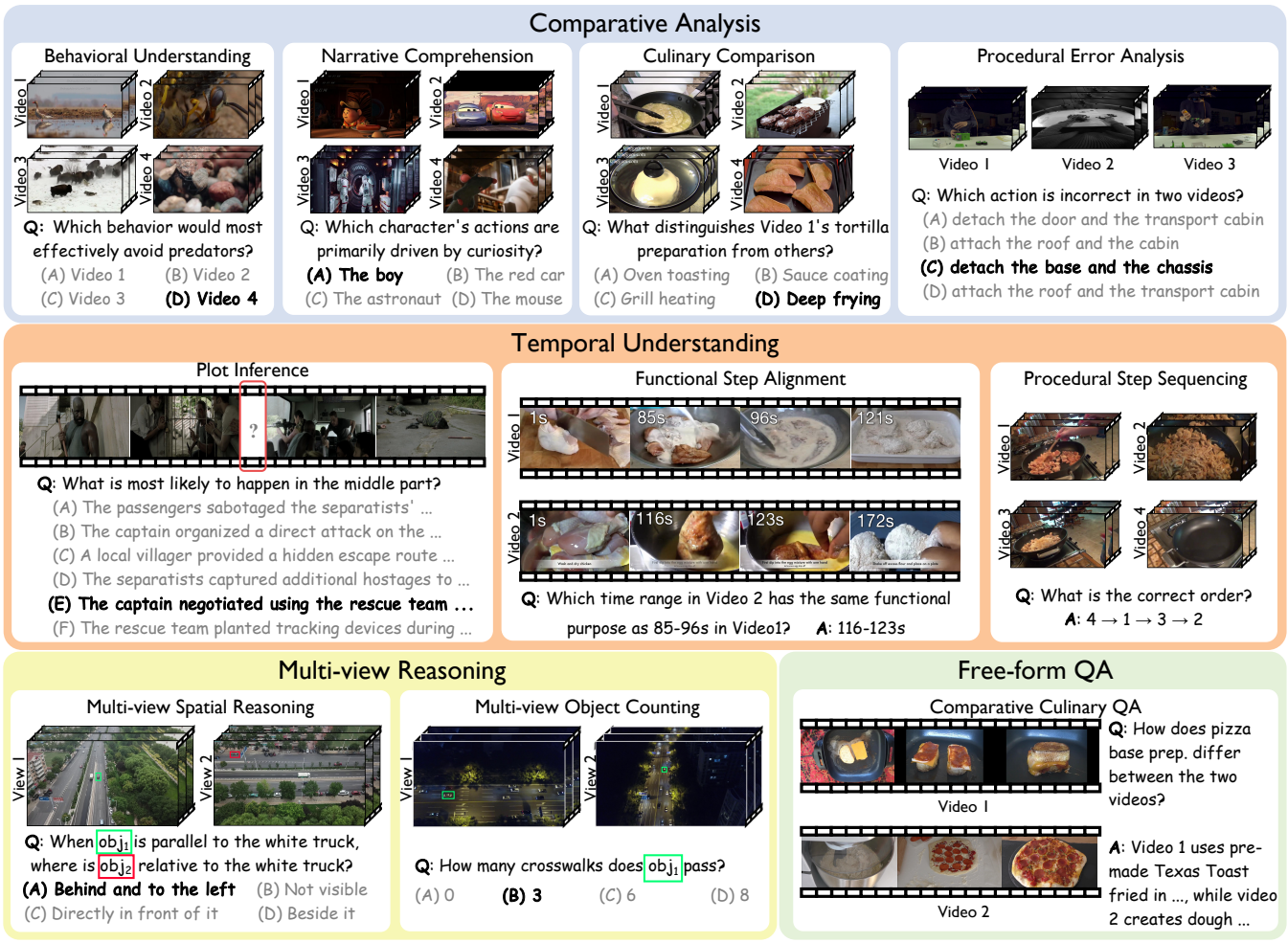


Figure 2: Overview of CrossVid. It evaluates MLLMs' CVR capability on 4 dimensions: comparative analysis, temporal understanding, multi-view reasoning, and free-form QA. It contains 10 distinct tasks: behavioral understanding (**BU**), narrative comprehension (**NC**), culinary comparison (**CC**), procedural error analysis (**PEA**), plot inference (**PI**), functional step alignment (**FSA**), procedural step sequencing (**PSS**), multi-view spatial reasoning (**MSR**), multi-view object counting (**MOC**) and comparative culinary QA (**CCQA**).

To our knowledge, there is currently no large-scale benchmark for general CVR. CrossVid is the first to comprehensively benchmark MLLMs on a diverse suite of CVR tasks, providing an important resource to spur future advances.

CrossVid Benchmark

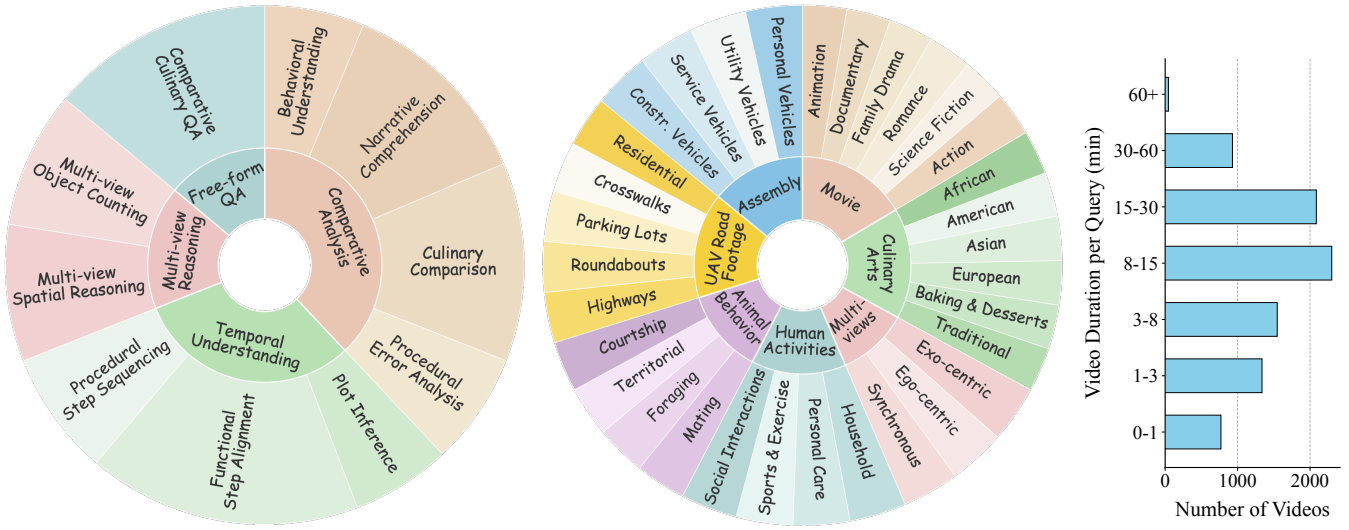
In this section, we first provide an overview of our CrossVid and its data curation process. We then present our semi-automated annotation pipeline used for its construction.

Overview

CrossVid is the first large-scale benchmark to systematically evaluate MLLMs' capabilities for CVR. This benchmark specifically assesses models' ability to integrate, compare, and reason over information from a group of related videos.

Video curation CrossVid consists of 5,331 video clips curated from six diverse, publicly available datasets: Animal Kingdom (Ng et al. 2022), MovieChat-1K (Song et al. 2024), YouCook2 (Zhou, Xu, and Corso 2018), VisDrone (Liu et al. 2023), Charades (Sigurdsson et al. 2016), and Assembly101 (Sener et al. 2022). With the various sources, CrossVid covers a wide range of video lengths and various degrees of visual complexity. During video selection, we emphasize scenario diversity, action complexity, and inter-video correlation, ensuring that the resulting multi-video groups are both challenging and appropriate for CVR tasks.

Hierarchical tasks Based on the curated video clips, we further propose 9,015 high-quality QA pairs. Each piece in CrossVid is composed of a group of semantically related videos, a task-specific query targeting CVR, and a carefully verified reference answer. As shown in Figure 2, all queries in CrossVid form hierarchical tasks with 4 high-level dimensions, including comparative analysis, temporal understanding, multi-view reasoning, and free-form QA. These 4 high-



(a) Types of dimensions and tasks.

(b) Categories and genres of videos.

(c) Video duration.

Figure 3: Statistical analysis of our CrossVid dataset. It consists of 4 high-level dimensions and 10 specific tasks, covering a wide range of video durations and video sources of 7 primary categories and 32 genres.

level dimensions are further divided into 10 distinct tasks, which are shown in Figure 3a.

Besides, CrossVid encompasses 32 various genres in Figure 3b, fully capturing representative CVR scenarios encountered in the real world. Furthermore, CrossVid features a hierarchical distribution of video duration per query, ranging from 1 minute to over 1 hour, which is presented in Figure 3c. More statistical analysis and details of the tasks in CrossVid can be found in the Appendix.

Comparison with previous benchmarks Table 1 summarizes the characteristics of existing VQA benchmarks. Compared with previous single-video understanding benchmarks, CrossVid innovatively introduces multi-video input and cross-video understanding. Compared with current multi-view benchmarks, CrossVid significantly extends the types of tasks, the formats of questions, and the application scenarios. Therefore, CrossVid is the first benchmark comprehensively evaluating MLLMs' CVR capability with a wide coverage of tasks and question formats.

Data Annotation

We design a semi-automated multi-stage pipeline to construct CrossVid. The overall process is shown in Figure 4.

Frame Caption We first densely extract frames from the source video and leverage Qwen2.5-VL-72B (Bai et al. 2025) to generate concise captions for each extracted frame. In order to enrich the contextual information for the caption, we also incorporate metadata from the original datasets (*e.g.*, plot summaries, scene descriptions, and action labels) during the generation process.

QA Generation Firstly, we manually assign the most suitable videos to the predefined tasks. For example, cooking videos inherently comprise multiple sequential steps, making them appropriate for temporal understanding tasks. More details about the assignment process can be found in the Appendix. Subsequently, for each task, videos are clustered into different sets based on their labels in the original datasets. Videos sharing the same labels (*e.g.* , the same film genre in MovieChat-1K or the same recipe in YouCook2) are grouped together. We then randomly sample from the same set the number of videos required for each question, and provide their frame-level captions to DeepSeek-R1 (Guo et al. 2025b) for automatic QA generation. We strictly retrieve videos from the same set to ensure strong semantic relevance and comparability across videos. For each task, we design a customized prompt comprising three key components: 1) The prompt explicitly instructs DeepSeek-R1 to analyze the relationships among all given videos. 2) The prompt guides DeepSeek-R1 to generate QA pairs that are closely aligned with the specific requirements of the task (*e.g.*, behavioral understanding tasks may prioritize comparison of action patterns and aims). 3) The prompt requires DeepSeek-R1 to provide a detailed explanation to support the correctness of its answer. Such prompts reduce DeepSeek-R1's hallucinations during the QA generation process and enhance the reliability of the generated QA pairs. Furthermore, this ensures that the generated QA pairs are challenging and require performing integrative reasoning across multiple videos.

Data Filtration To ensure data quality, we conduct a rigorous manual review with ten expert annotators. During this coarse filtering stage, we sequentially eliminate unsuitable QA pairs through three steps. First, we filter out questions unrelated to video understanding. Then, we exclude questions referencing only specific queried videos (*e.g.*, "In video three, what color is the car?"). Finally, we discard subjective or overly complex questions, such as those requiring philosophical reasoning or domain-specific expertise.

QA Refinement The retained QA pairs then undergo refinement consisting of three steps: 1) Annotators revise the questions to eliminate ambiguities. 2) Each annotator then

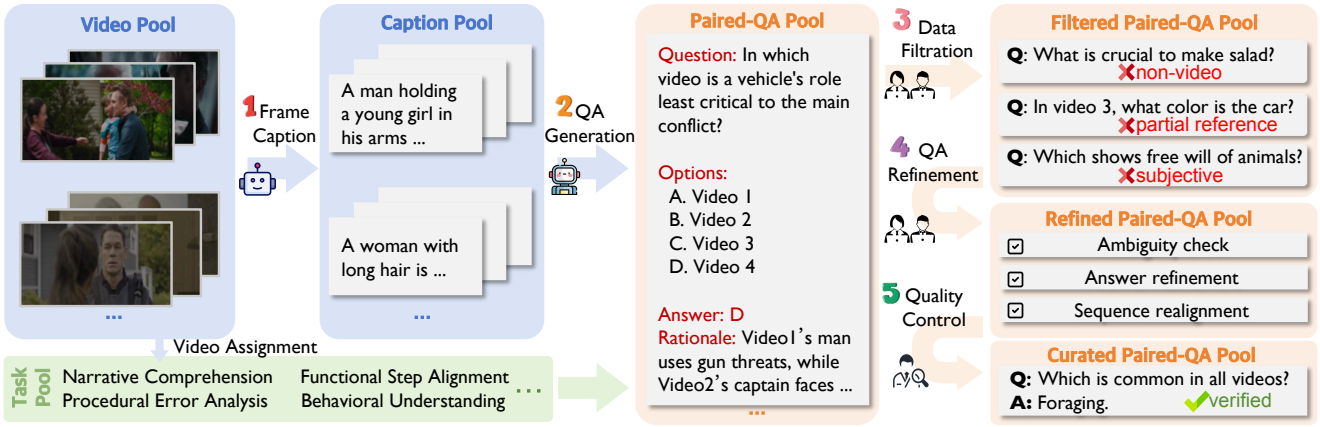


Figure 4: Illustration of the CrossVid annotation pipeline. The process consists of the following main stages: (1) Frames are extracted from videos and captioned by Qwen2.5-VL-72B; (2) Deepseek-R1 generates QA pairs using task-specific prompts; (3) The QA pairs undergo rigorous human quality review, including data filtering, refinement, and quality control.

answers questions without consulting DeepSeek R1's outputs or explanations. 3) Based on the annotators' responses, task-specific refinements are further conducted. Specifically, for single- and multi-choice questions, both ground truths and other false options are refined to ensure unique correctness. For the functional step sequencing task in temporal understanding, we address potential shortcut learning (reliance on camera angle continuity) by temporal realignment, *i.e.*, each preceding clip is advanced by 1-5 seconds while equivalent offsets delay subsequent clips. Such a strategy creates intentional discontinuity in visual features across clip boundaries, forcing models to infer temporal relationships through semantic content rather than low-level consistency. For open-ended questions, annotators check whether the scoring points align with the generated standard answer and cover all key information related to the question.

Quality Control At the quality control step, an independent group of experts further assesses the refined paired QA pool and forms the curated pool. These processes are conducted via our designed interface, shown in the Appendix.

Through this semi-automated pipeline, a large number of high-quality QA pairs are generated. More importantly, our curation process ensures that each QA pair is built on meaningful inter-video relationships and requires integrative CVR. This meets the objectives of our CrossVid.

Experiments and Analysis

We conduct a comprehensive evaluation of existing MLLMs on CrossVid. In this section, we first describe experimental settings, followed by a detailed analysis of model performance. We then present ablation studies and key findings.

Experimental Settings

We evaluate 22 MLLMs on CrossVid, including both closed-source models (*e.g.*, GPT-4.1 and Gemini-2.5-Pro) and a wide range of open-source models (*e.g.*, Qwen2.5-VL, InternVL3) with parameter sizes spanning from 7B to 78B. Other architectures, like Mixture of Experts (MoE), are also included. For video pre-processing, we evenly distribute the total number of frames across all input videos and uniformly sample frames within each video. For each QA pair, the frames of all videos and the question prompt are simultaneously input into the MLLM in one turn. We adopt a zero-shot strategy and require the MLLMs to give their answers directly. Inference on open-source models is conducted according to their official implementations, whereas closed-source models are accessed via their official APIs. We report accuracy as our evaluation metric. More details of the implementation and evaluation can be found in the Appendix.

Main Results

We present the performance of MLLMs on CrossVid, including the accuracy for each task, the average accuracy of each dimension, and the overall average on all tasks, in Table 2. Based on these, we highlight three main observations:

1) CVR is challenging for existing MLLMs. The average performance of all MLLMs is notably lower than the human performance of 89.2%. Even the best-performing MLLM, Gemini-2.5-Pro, only achieves an overall average accuracy of 50.4%. Notably, MLLMs perform worse than humans on multi-view reasoning, a task focusing on spatial reasoning. Specifically, the leading MLLM, InternVL3-8B, only achieves an average accuracy of 40.7%, while humans reach 93.7%. The gap between MLLM and humans becomes even larger in temporal understanding. For example, on the step alignment task, Gemini-2.5-Pro with the highest accuracy achieves only 13.4% versus 85.2% for humans. These reveal critical limitations in existing MLLMs' capability for both temporal and spatial understanding in CVR.

2) Closed-source MLLMs substantially outperform open-source counterparts. All closed-source MLLMs obtain higher overall average accuracies than open-source MLLMs, and the advantage of closed-source MLLMs is much more prominent on several crucial tasks. Notably, for temporal understanding, closed-source MLLMs consistently outperform open-source MLLMs. The closed-source GPT-

Models	O.Avg	Comparative Analysis					Temporal Understanding				Multi-view Reasoning			Free-form
		BU	NC	CC	PEA	**C.Avg**	PI	FSA	PSS	**T.Avg**	MSR	MOC	**M.Avg**	CCQA
Human	89.2	85.6	92.3	90.7	83.9	88.1	91.6	85.2	89.9	88.9	93.2	94.2	93.7	85.2
Closed-Source Models														
GPT-4.1 (2025)	45.2	46.2	34.6	58.5	**51.2**	47.6	70.9	8.6	60.5	46.7	**38.6**	38.2	38.4	44.6
GPT-4o (2024)	36.8	38.2	34.3	50.7	49.1	43.1	57.8	9.1	39.7	35.5	15.3	39.4	27.4	34.2
Doubao-1.5-VL-pro (2025a)	44.3	51.2	**58.1**	**69.5**	36.4	53.8	66.9	4.6	36.8	36.1	37.4	32.0	34.7	50.1
Gemini-2.5-Pro (2025)	**50.4**	**54.2**	51.8	68.7	44.1	**54.7**	76.5	**13.4**	**78.2**	**56.0**	32.0	25.3	28.7	**59.8**
Open-Source Models ∼ MoE														
Kimi-VL-A3B-Thinking (2025)	28.2	29.4	33.3	36.8	34.0	33.4	40.6	3.8	9.2	17.9	28.4	36.9	32.7	29.2
ERNIE-4.5-VL-A3B (2025)	24.8	12.6	28.2	24.2	36.4	25.4	52.6	4.0	2.4	19.7	29.6	35.3	32.5	22.5
Open-Source Models <10B														
Qwen2.5-VL-7B (2025)	18.3	19.6	19.0	23.4	15.0	19.3	58.6	1.2	0.3	20.0	11.8	21.7	16.8	12.0
InternVL3-8B (2025)	25.6	15.2	22.8	24.3	42.1	26.1	56.2	3.2	1.5	20.3	34.0	**47.3**	**40.7**	9.7
LongVA-7B-DPO (2024a)	18.0	16.2	20.6	18.2	39.0	23.5	18.7	2.1	1.8	7.5	24.2	28.4	26.3	10.7
VideoLLaMA3-7B (2025)	15.3	14.7	19.5	22.2	26.6	20.8	11.6	5.1	3.5	6.7	18.7	20.8	19.8	9.8
Qwen2.5-Omni-7B (2025)	24.6	27.5	26.0	32.7	20.4	26.7	60.2	0.4	4.1	21.6	23.2	36.0	29.6	15.3
Phi-3.5-vision (2024)	21.5	18.3	22.0	21.8	41.5	25.9	46.2	1.2	4.1	17.2	28.4	26.7	27.6	4.3
MiniCPM-O 2.6 (2024)	25.6	20.3	21.8	20.1	42.6	26.2	72.1	2.9	4.1	26.4	27.1	35.7	31.4	9.0
MiMo-7B (2025)	28.3	22.3	30.6	39.2	32.8	31.2	54.6	2.8	11.6	23.0	25.8	41.3	33.6	22.0
Video-R1-7B (2025)	21.6	14.7	23.0	19.9	16.3	18.5	**77.3**	1.9	1.5	26.9	19.4	34.4	26.9	8.0
GLM-4.1V-9B-Thinking (2025)	35.1	49.8	39.9	50.6	38.6	44.7	50.2	5.1	14.1	23.1	36.7	38.9	37.8	26.9
Open-Source Models ∼30B														
Qwen2.5-VL-32B (2025)	33.7	31.4	39.5	48.6	33.7	38.3	65.7	5.1	8.7	26.5	23.7	39.6	31.7	41.2
InternVL3-38B (2025)	23.5	15.9	33.7	33.6	27.9	27.8	24.3	4.5	1.5	10.1	40.4	36.8	38.6	16.2
Open-Source Models ∼70B														
Qwen2.5-VL-72B (2025)	34.4	37.7	38.5	52.6	39.6	42.1	61.8	5.9	20.0	29.2	20.9	26.0	23.5	41.2
InternVL3-78B (2025)	25.8	27.1	29.4	33.1	42.7	33.1	37.5	4.9	4.4	15.6	22.9	33.2	28.1	23.2
LLaVA-Video-72B (2024b)	27.5	26.5	26.5	34.0	48.7	33.9	51.8	3.2	11.1	22.0	26.8	29.0	27.9	17.8
LLaVA-OV-72B (2024a)	27.5	22.1	23.3	29.2	37.0	27.9	74.1	8.8	5.0	29.3	25.9	35.0	30.5	14.6

Table 2: Performance of 22 evaluated MLLMs on CrossVid. **O.Avg**: the average accuracy of ten tasks. **C.Avg**: the average accuracy of comparative analysis; **T.Avg**: the average accuracy of temporal understanding; **M.Avg**: the average accuracy of multi-view reasoning. The top result in each task is highlighted in bold.

4o with the lowest score achieves an average accuracy of 35.5%, which is still 6.2% higher than the leading open-source model LLaVA-OV.

3) "Thinking" -enabled models demonstrate performance gains. Among closed-source models, those featuring explicit reasoning modules (*e.g.*, Gemini-2.5-Pro) consistently achieve the highest overall and per-task accuracies. In the 10B parameter group, the top two models are thinking-enabled GLM-4.1V-9B-Thinking (35.1%) and MiMo-7B (28.3%), which outperform the third best by margins of 9.5% and 2.7%, respectively. Therefore, internal "thinking" mechanism allows models to better structure multi-step reasoning processes, which contributes to enhanced performance on complex cross-video tasks.

Ablation Study

We conduct ablations to better understand MLLMs' performance on CVR and present more insights.

Impact of frame number

The number of input frames determines the amount of visual information available to the model for reasoning. To assess its impact, we evaluate Qwen2.5-VL-72B on CrossVid using 32, 64, 128, and 256 total input frames. For settings with fewer than 256 frames, frames are uniformly sampled from the full set of 256 frames. Results are reported in Table 3.

It can be observed that increasing the number of input frames generally improves model performance. The improvements are most pronounced in tasks that require comprehensive context, such as comparative analysis and free-form QA. For Qwen2.5-VL-72B, the overall accuracy increases by 5.7% (from 33.8% to 39.5%) as frame count rises from 32 to 256, with even larger improvements (up to 15.1%) observed on the open-ended CCQA task.

With more frames, the model can access richer visual information, which is crucial for answering cross-video questions requiring precise details. For instance, when answering an open-ended question to tell the difference between cooking methods in two videos, the model can only identify superficial actions at 32 frames. When increasing the frame count to 64, the model is able to distinguish specific core techniques. At 256 frames, its analysis became granular enough to recognize secondary ingredients.

However, excessive irrelevant frames may introduce

noise, which leads to information redundancy and distracts the model with irrelevant content. For example, when solving a plot inference problem with the beginning and ending of a war film, the model correctly identifies the key events (*e.g.*, a troop convoy and a negotiation scene) with lower frame counts like 32 and 64. However, as more frames are provided, irrelevant atmospheric information (*e.g.*, generic shots of injured soldiers) distracts the model from the primary causal chain, leading to the incorrect answer based on broad military planning associations.

These findings provide guidance for future advancements on CVR. On one hand, expanding the model's context window allows it to perceive more information; On the other hand, key frame selection helps to filter out irrelevant clues and make models concentrate on core information.

#Frames	O.Avg	C.Avg	T.Avg	M.Avg	CCQA
32	33.8	37.0	33.8	35.1	18.9
64	36.9	39.8	37.4	35.9	25.9
128	39.1	45.7	**34.5**	**36.4**	32.0
256	**39.5**	**47.5**	33.9	34.9	**34.0**

Table 3: Comparison results of performance under different numbers of input frames.

Effectiveness of CoT prompts

Previous results show that internally thinking-enabled MLLMs outperform their counterparts. To assess the effectiveness of Chain-of-Thought (CoT) on non-thinking MLLMs, we design prompts to explicitly require a three-stage process: 1) comprehend the question, 2) analyze the frames for each input video, and 3) aggregate information from all videos before answering. Details of prompts can be found in the Appendix. MLLMs are required to output each step of their reasoning process. We conduct experiments on GPT-4.1 and three other open-source MLLMs in different parameter size groups. We keep the frames for each MLLM the same as the previous direct answering strategy.

Table 4 shows the performance comparison with and without CoT prompting. For most MLLMs, CoT brings performance gains on temporal understanding and multi-view reasoning tasks. This indicates that CoT facilitates more systematic reasoning across videos on both temporal and spatial understanding tasks. Notably, CoT prompting does not consistently improve accuracy on each task, whereas open-source MLLMs with more parameters exhibit larger overall performance gains. This suggests that larger MLLMs are more capable of benefiting from prompt-based optimization.

Error Analysis

To further examine the CVR capabilities of current MLLMs and better understand their limitations, we manually analyze the errors in their reasoning steps and identify four primary error types. The percentage and detailed examples for each error are presented in the Appendix.

(a) Key frame loss: Compared with previous single-video understanding, our tasks require multiple videos to be input simultaneously, which reduces the number of frames for each video. This may result in the loss of core information. As a result, MLLMs may fail to obtain the necessary information to answer the question and thus provide incorrect answers.

Method	O.Avg	C.Avg	T.Avg	M.Avg	CCQA
GPT-4.1					
w/o CoT	**45.2**	**47.6**	46.7	38.4	**44.6**
w/ CoT	44.9	46.7	**48.2**	**40.4**	36.7
MiniCPM-o 2.6					
w/o CoT	**25.6**	26.2	**26.4**	31.4	**9.0**
w/ CoT	23.7	**26.7**	18.7	**33.3**	7.2
InternVL3-38B					
w/o CoT	23.5	**27.8**	10.1	**38.6**	16.2
w/ CoT	**24.4**	26.3	**16.7**	35.2	**18.0**
Qwen2.5-VL-72B					
w/o CoT	34.4	42.1	29.2	23.5	**41.2**
w/ CoT	**39.5**	**47.5**	**33.9**	**34.9**	34.0

Table 4: Comparison of performances with and without CoT prompting.

(b) Video understanding error: In this type of error, although MLLMs capture the key information from each video, they still might fall short in cross-video understanding. Since their analysis of individual videos might be insufficient due to the requirement of simultaneously processing multiple videos, the failure to understand any single video leads to errors in multi-video understanding as a whole.

(c) Cross-video comparison error: Although MLLMs are capable of correctly understanding each individual video, they may still struggle with cross-video comparison.

For example, when the MLLM is asked to identify the hug in which film represents the crisis resolution, the MLLM successfully identifies the hugs in all videos in the group but fails to reason and compare their meanings in the context.

(d) Format error: CrossVid contains tasks requiring specific output formats, *e.g.*, a time interval with both beginning and ending timestamps in the functional step alignment task. However, some MLLMs might fail to accurately follow the specific instructions or constraints described in the prompt, resulting in the failure of answer extraction.

Conclusion

We present CrossVid, the first comprehensive benchmark for evaluating MLLMs on CVR. CrossVid is constructed through a semi-automated pipeline and strict multi-stage quality control, resulting in 10 diverse tasks spanning 4 key reasoning dimensions. Our experimental evaluation of 22 frontier MLLMs reveals significant challenges in CVR, with the best model (Gemini-2.5-Pro) achieving only moderate performance far below human levels. Extensive ablation studies and error analyses provide deep insights into the limitations of current models. We hope CrossVid will serve as a valuable resource to drive advances in multi-video understanding and robust, generalizable visual reasoning.

References

Abdin, M.; Aneja, J.; Behl, H.; Bubeck, S.; Eldan, R.; Gunasekar, S.; Harrison, M.; Hewett, R. J.; Javaheripi, M.; Kauffmann, P.; et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

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*.

Chen, L.; Wei, X.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Tang, Z.; Yuan, L.; et al. 2024. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37: 19472–19495.

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*.

Feng, K.; Gong, K.; Li, B.; Guo, Z.; Wang, Y.; Peng, T.; Wu, J.; Zhang, X.; Wang, B.; and Yue, X. 2025. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*.

Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24108–24118.

Gong, S.; Zhuge, Y.; Zhang, L.; Yang, Z.; Zhang, P.; and Lu, H. 2025. The devil is in temporal token: High quality video reasoning segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29183–29192.

Grauman, K.; Westbury, A.; Torresani, L.; Kitani, K.; Malik, J.; Afouras, T.; Ashutosh, K.; Baiyya, V.; Bansal, S.; Boote, B.; et al. 2024. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19383–19400.

Guo, D.; Wu, F.; Zhu, F.; Leng, F.; Shi, G.; Chen, H.; Fan, H.; Wang, J.; Jiang, J.; Wang, J.; et al. 2025a. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025b. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hong, W.; Yu, W.; Gu, X.; Wang, G.; Gan, G.; Tang, H.; Cheng, J.; Qi, J.; Ji, J.; Pan, L.; et al. 2025. GLM-4.1 V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning. *arXiv preprint arXiv:2507.01006*.

Huang, Y.; Chen, G.; Xu, J.; Zhang, M.; Yang, L.; Pei, B.; Zhang, H.; Dong, L.; Wang, Y.; Wang, L.; et al. 2024. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22072–22086.

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*.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. L. 2018. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22195–22206.

Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Liu, Z.; Shang, Y.; Li, T.; Chen, G.; Wang, Y.; Hu, Q.; and Zhu, P. 2023. Robust multi-drone multi-target tracking to resolve target occlusion: A benchmark. *IEEE Transactions on Multimedia*, 25: 1462–1476.

Ng, X. L.; Ong, K. E.; Zheng, Q.; Ni, Y.; Yeo, S. Y.; and Liu, J. 2022. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19023–19034.

Reilly, D.; Govind, M. K.; Xue, L.; and Das, S. 2025. From My View to Yours: Ego-Augmented Learning in Large Vision Language Models for Understanding Exocentric Daily Living Activities. *arXiv preprint arXiv:2501.05711*.

Sener, F.; Chatterjee, D.; Shelepov, D.; He, K.; Singhania, D.; Wang, R.; and Yao, A. 2022. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21096–21106.

Shu, Y.; Liu, Z.; Zhang, P.; Qin, M.; Zhou, J.; Liang, Z.; Huang, T.; and Zhao, B. 2025. Video-xl: Extra-long vision language model for hour-scale video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 26160–26169.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Laptev, I.; Farhadi, A.; and Gupta, A. 2016. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. *ArXiv e-prints*.

Song, E.; Chai, W.; Wang, G.; Zhang, Y.; Zhou, H.; Wu, F.; Chi, H.; Guo, X.; Ye, T.; Zhang, Y.; et al. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18221–18232.

Tang, X.; Qiu, J.; Xie, L.; Tian, Y.; Jiao, J.; and Ye, Q. 2025. Adaptive keyframe sampling for long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29118–29128.

Team, B. E. 2025. ERNIE 4.5 Technical Report.

Team, K.; Du, A.; Yin, B.; Xing, B.; Qu, B.; Wang, B.; Chen, C.; Zhang, C.; Du, C.; Wei, C.; et al. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Wu, H.; Li, D.; Chen, B.; and Li, J. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37: 28828–28857.

Xiao, J.; Shang, X.; Yao, A.; and Chua, T.-S. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9777–9786.

Xiaomi, L.; Xia, B.; Shen, B.; Zhu, D.; Zhang, D.; Wang, G.; Zhang, H.; Liu, H.; Xiao, J.; Dong, J.; et al. 2025. MiMo: Unlocking the Reasoning Potential of Language Model–From Pretraining to Posttraining. *arXiv preprint arXiv:2505.07608*.

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*.

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*.

Yeh, C.-H.; Wang, C.; Tong, S.; Cheng, T.-Y.; Wang, R.; Chu, T.; Zhai, Y.; Chen, Y.; Gao, S.; and Ma, Y. 2025. Seeing from another perspective: Evaluating multi-view understanding in mllms. *arXiv preprint arXiv:2504.15280*.

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 Conference on Artificial Intelligence*, volume 33, 9127–9134.

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, P.; Zhang, K.; Li, B.; Zeng, G.; Yang, J.; Zhang, Y.; Wang, Z.; Tan, H.; Li, C.; and Liu, Z. 2024a. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*.

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 Computer Vision and Pattern Recognition Conference*, 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.

Zhou, L.; Xu, C.; and Corso, J. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.