MT-Video-Bench: A Holistic Video Understanding Benchmark for Evaluating Multimodal LLMs in Multi-Turn Dialogues

Yaning Pan[1], **Zekun Wang**[2], **Qianqian Xie**[3], **Yongqian Wen**[3], **Yuanxing Zhang**[2], **Guohui Zhang**[4], **Haoxuan Hu**[3], **Zhiyu Pan**[3], **Yibing Huang**[3], **Zhidong Gan**[3], **Yonghong Lin**[3], **An Ping**[3], **Tianhao Peng**[3], **Jiaheng Liu**[3,†]

[1]Fudan University, [2]Kuaishou Technology, [3]Nanjing University, [4]University of Science and Technology of China

ynpan24@m.fudan.edu.cn liujiaheng@nju.edu.cn

Abstract

The recent development of Multimodal Large Language Models (MLLMs) has significantly advanced AI's ability to understand visual modalities. However, existing evaluation benchmarks remain limited to single-turn question answering, overlooking the complexity of multi-turn dialogues in real-world scenarios. To bridge this gap, we introduce **MT-Video-Bench**[a], a holistic video understanding benchmark for evaluating MLLMs in multi-turn dialogues. Specifically, our MT-Video-Bench mainly assesses six core competencies that focus on perceptivity and interactivity, encompassing 987 meticulously curated multi-turn dialogues from diverse domains. These capabilities are rigorously aligned with real-world applications, such as interactive sports analysis and multi-turn video-based intelligent tutoring. With MT-Video-Bench, we extensively evaluate various state-of-the-art open-source and closed-source MLLMs, revealing their significant performance discrepancies and limitations in handling multi-turn video dialogues. The benchmark will be publicly available to foster future research.

[a] https://github.com/NJU-LINK/MT-Video-Bench

1 Introduction

The rapid progress of Multimodal Large Language Models (MLLMs) has markedly advanced AI's capacity to perceive and reason over visual modalities, especially when integrated with natural language. Recent systems such as Qwen2.5-VL (Bai et al., 2025), InternVL3.5 (Wang et al., 2025a), and Gemini 2.5 (Team, 2025) demonstrate impressive performance in single-turn video question answering and long-form video comprehension (Zhang et al., 2023; Rawal et al., 2024; Sun et al., 2022; Wang et al., 2024a; Chandrasegaran et al., 2024). Yet, real-world human–AI interaction is rarely confined to single-turn queries. Instead, it typically unfolds as multi-turn dialogues, where users iteratively refine their questions, shift topics, and expect contextually coherent responses grounded in video content. This interactive setting poses unique challenges: models must not only recall and integrate prior dialogue history but also adapt to conversational dynamics, such as handling topic shifting or gracefully refusing unanswerable queries.

Despite these demands, existing video understanding benchmarks (Fu et al., 2025; Wang et al., 2024b; Zhou et al., 2025; Ma et al., 2025) predominantly focus on single-turn evaluation, emphasizing factual perception of video content—such as recognizing objects, actions, or temporal relations—while neglecting dialogue-level reasoning. A few recent efforts explore long-context or multi-shot video benchmarks, yet they fall short of capturing the interplay between perceptivity (faithfully interpreting multimodal input) and interactivity (sustaining natural, user-aware conversations). Consequently, **the community lacks a rigorous and holistic framework to measure how well MLLMs can operate in realistic multi-turn, video-grounded dialogues.**

To fill this gap, as shown in Figure 2, we introduce **MT-Video-Bench**, a holistic benchmark for evaluating MLLMs in multi-turn video dialogue. MT-Video-Bench systematically targets six core capabilities spanning perceptivity (object reference, memory recall, and content summary) and interactivity (answer refusal, topic shifting, and proactive interaction). The benchmark comprises 987 carefully curated dialogues across 135 videos, covering diverse domains such as sports, education, and daily activities.

† Corresponding Author.

Figure 1: Illustration of multi-turn dialogues under single-scene and cross-scene settings. The evaluated questions corresponding to tasks are marked with underlining, and the scenes involved in the entire multi-turn dialogues are marked with blue dotted boxes.

Moreover, unlike prior datasets, MT-Video-Bench emphasizes cross-scene reasoning, long-range dependencies, and interactive adaptability, thereby aligning closely with real-world application demands.

Based on our MT-Video-Bench, we provide a detailed evaluation of both open-source and closed-source models, highlighting the current limitations and performance discrepancies in different abilities. Specifically, several insightful findings are as follows:

- The perceptual and interactive capabilities of MLLMs in multi-turn dialogues still have significant room for improvement. On MT-Video-Bench, even the strongest closed-source model Gemini 2.5 Pro achieves only 68.45% overall accuracy, while most open-sourced MLLMs exhibit accuracies below 50%, except for the Qwen2.5-VL and InternVL3.5 series.
- Performance is imbalanced across different tasks and scene types. MLLMs generally perform better on perceptual subtasks (e.g., Object Reference) than on interactive ones (e.g., Proactive Interaction), with a substantial gap between closed- and open-source models. Moreover, all models tend to perform worse in cross-scene settings compared to single-scene tasks.
- Model scaling is beneficial but not sufficient. Larger models consistently outperform smaller counterparts across most subtasks, yet scaling alone does not ensure consistent improvements. For example, in the InternVL 3.5 series, enabling the Thinking mode allows smaller models to achieve performance comparable to that of larger models, which demonstrates the significant benefit of the reasoning process in enhancing model performance.

To summarize, the contributions of this paper are as follows: We identify the critical gap in evaluating multi-turn video-grounded dialogues and propose the MT-Video-Bench, the first holistic benchmark that operationalizes this evaluation via six well-defined capabilities across 987 dialogues and 5,805 QA pairs. Then, based on extensive experiments on MT-Video-Bench, we underscore the challenges and potential directions for improvement of handling and reasoning over multi-turn dialogues, offering a roadmap for future research and development.

2 Related Work

Multimodal LLMs. MLLMs have become a central research focus in advancing general-purpose intelligence. By jointly modeling textual and visual modalities, these models are able to capture cross-modal dependencies and enhance semantic reasoning (Zhu et al., 2023; Ma et al., 2024; Zhang et al., 2024a; Wang et al., 2025b; 2024c). Recent advances have further extended MLLMs to the video domain, enabling

Table 1: Comparison with other benchmarks. **Avg. Q/V**: the average number of QA pairs per video. **Long**: whether the average video length is greater than 10 minutes. **Cross-Scene**: whether the dialogue covers more than 4 scenes.

Benchmark	#QAs	Avg. Q/V	Long	Dialogue	#Turns	Cross-Scene	Annotation
MVBench (Li et al., 2024a)	4,000	1.00	✗	✗	1.00	-	Auto
LongVideoBench (Wu et al., 2024a)	6,678	1.77	✗	✗	1.00	-	Manual
Video-MME (Fu et al., 2025)	2,700	3.00	✓	✗	1.00	-	Manual
LVBENCH (Wang et al., 2024b)	1,549	15.04	✓	✗	1.00	-	Manual
MLVU (Zhou et al., 2025)	3,102	1.79	✓	✗	1.00	-	Manual
Video-MMLU (Song et al., 2025)	15,746	14.78	✗	✗	1.00	-	Auto&Manual
ScaleLong (Ma et al., 2025)	1,747	6.49	✓	✗	1.00	-	Manual
SVBench (Yang et al., 2025)	7,374	36.87	✗	✓	4.29	✗	Auto&Manual
MT-Video-Bench (Ours)	5,805	43.00	✓	✓	5.88	✓	Auto&Manual

video understanding, which subsequently supports dialogue(Li et al., 2023; Cheng et al., 2024; Maaz et al., 2023). For example, Qwen2.5-VL (Bai et al., 2025) employs a dynamic-resolution Vision Transformer with MRoPE for spatiotemporal alignment, and connects an MLP merger to the Qwen2.5 LLM decoder. InternVL3.5 (Wang et al., 2025a) integrates InternViT as the vision encoder with a ViT-MLP-LLM paradigm, and further adopts Visual Resolution Router (ViR) with Visual Consistency Learning (ViCO) for cross-modal alignment.

Video Benchmarks. Significant developments have also been made in video understanding benchmarks(Wang et al., 2023; Wu et al., 2024b; Xiao et al., 2021; Li et al., 2025). For example, MVBench (Li et al., 2024a) focuses on concise video QA tasks to evaluate multimodal understanding abilities, while MLVU (Zhou et al., 2025) and LVBENCH (Wang et al., 2024b) provide a comprehensive analysis for MLLMs' long-video understanding performance. MMBench-Video (Fang et al., 2024) is a long-form, multi-shot benchmark that evaluates fine-grained abilities of MLLMs, including temporal reasoning, perception, and general reasoning in video understanding. SVBench (Yang et al., 2025) is a benchmark for temporal multi-turn dialogues in streaming videos, designed to assess the capabilities of streaming video understanding of MLLMs. However, prior benchmarks primarily focus on evaluating the video understanding capabilities of models, overlooking the multi-turn dialogue capabilities, which require not only the ability to recall contextual information but also to engage in coherent, interactive communication with users across multiple turns.

3 MT-Video-Bench

3.1 Overview

MT-Video-Bench is designed to comprehensively evaluate the "Perceptivity" and "Interactivity" of MLLMs in multi-turn video-grounded dialogues. Different from conventional video understanding benchmarks that primarily focus on single-turn question answering, MT-Video-Bench is specifically designed to mimic real-world interactive scenarios, emphasizing contextual coherence, cross-scene video comprehension, and adaptive interactivity.

MT-Video-Bench systematically evaluates six core capabilities of MLLMs through 987 meticulously curated multi-turn dialogues with 5,805 QA pairs. Each conversation requires not only accurate video perception but also contextual reasoning within or across video scenes, with representative examples shown in Figure 1.

A comprehensive comparison between our MT-Video-Bench and other related benchmarks is provided in Table 1. MT-Video-Bench presents the following critical values: (1) supports multi-turn dialogues that evaluate contextual coherence and long-range dependency, (2) supports cross-scene reasoning that requires integrating information across different video clips, and (3) provides a fine-grained assessment of perceptivity and interactivity through six tasks.

3.2 Evaluation Tasks

Perceptivity assesses the model's foundational ability to perceive and integrate information from both the visual video content and the multi-turn conversational context. This capability is essential for accurately understanding user queries and generating contextually grounded responses throughout the dialogue. It

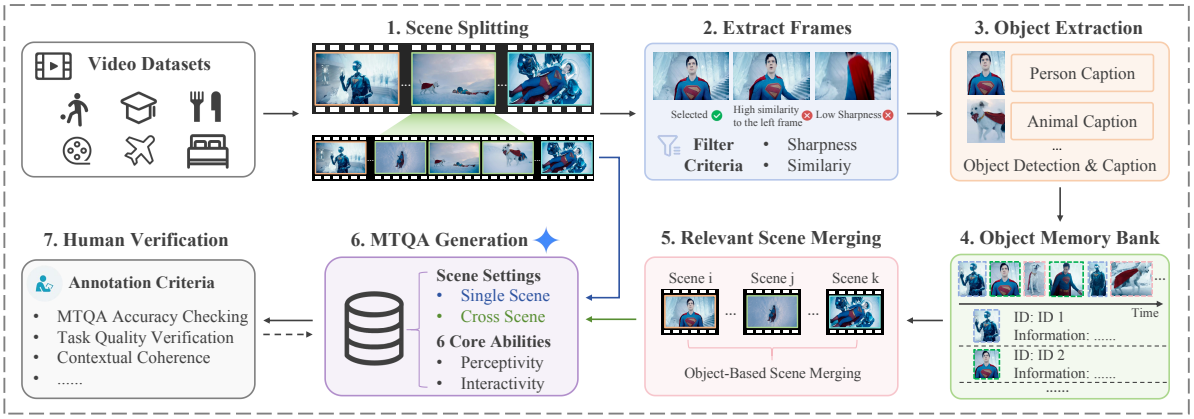


Figure 2: An overview of the semi-automatic data construction process of MT-Video-Bench.

includes:

- **Object Reference (OR)** evaluates the model's ability to resolve references and pronouns in the user's input, ensuring that entities mentioned implicitly are correctly mapped to the appropriate objects, characters, or concepts.
- **Memory Recall (MR)** measures the model's capacity to retrieve, retain, and integrate relevant information from prior conversational turns or long-term history, enabling coherent reasoning and continuity across interactions.
- **Content Summary (CS)** assesses the model's effectiveness in condensing conversational and video content into succinct yet comprehensive summaries, while preserving essential details, coherent structure, and semantic fidelity.

Interactivity evaluates the model's capacity to conduct coherent, adaptive, and user-aware dialogues based on the video content. It focuses on appropriately refusing unanswerable questions, smoothly adapting to topic changes, and proactively maintaining engagement. It includes:

- **Answer Refusal (AR)** tests the ability to recognize unanswerable queries based on available evidence and explicitly decline or indicate insufficiency without hallucination.
- **Topic Shifting (TS)** evaluates how effectively the model can track and adapt to user-initiated changes in conversational focus or subject matter, while maintaining coherence, fluency, and relevance throughout the dialogue.
- **Proactive Interaction (PI)** probes the model's capacity to sustain or restore engagement through clarifications, elaborations, or novel insights when signs of disinterest or disengagement are detected, thereby fostering renewed interest and continuation of the dialogue.

3.3 Data Collection

As shown in Figure 2, the data collection process for MT-Video-Bench involves both automated construction and human verification. We first acquire videos from online platforms and split them into single-scene segments. Next, we retrieve and merge relevant scenes by extracting frames, performing object detection, and constructing an object memory bank. Multi-turn dialogues are then generated automatically for diverse evaluation tasks. Finally, human annotators are involved to ensure the accuracy and quality of the generated dialogues.

Video Collection and Single-Scene Splitting. The data collection process begins with the manual acquisition of 135 videos from various online platforms, such as YouTube, within the past year. Subsequently, we employ PySceneDetect[1] to divide the videos into shorter clips. Recognizing that these clips are often too brief to represent complete scenes, we then use the Gemini 2.5 Flash model (Team, 2025) to generate descriptive captions for each clip. Finally, the caption-based clip merging method is iteratively applied twice to combine related clips into a coherent, single-scene video, ensuring a seamless and contextually accurate representation of the scene. These refined single-scene videos serve as the core visual content for the subsequent task of generating single-scene, multi-turn dialogues.

[1]https://github.com/Breakthrough/PySceneDetect

Cross-Scene Video Merging. The generation of cross-scene, multi-turn dialogues necessitates the retrieval and merging of relevant scenes from disparate video segments, which serves as a critical step in creating coherent interactions that span across multiple visual contexts. Firstly, frames are extracted from the video at 2 FPS and then filtered based on two criteria: sharpness and similarity to the previous selected frame. The sharpness of each frame is evaluated by the Laplace Operator to ensure that only clear, visually significant frames are retained, improving the overall quality of the selected frames. To avoid redundancy, frames with high similarity to the preceding selected frame are discarded. Specifically, a histogram-based image similarity calculation method is used to compare consecutive frames, excluding those with a similarity score above 0.9. This approach ensures that the selected frames are distinct and capture key moments in the video.

Following frame selection, object detection is performed using YOLOv11 (Khanam and Hussain, 2024), and each detected object is then annotated with a caption generated by the Gemini 2.5 Flash (Team, 2025), providing detailed descriptions for each object. As the video progresses, a dynamic object memory bank is maintained, continuously expanded based on object captions and visual similarities. This memory bank associates unique object IDs with their corresponding attributes, enabling the identification of the same objects across frames. To merge relevant scenes, a retrieval step across scenes is performed to select video segments that share common objects or themes, which are then merged to ensure continuity both thematically and contextually.

Multi-Turn Dialogues Generation. This process employs the Gemini 2.5 Pro (Team, 2025) to automate the generation of both single-scene and cross-scene multi-turn dialogues, based on the six evaluation tasks defined earlier. For each video, we generate multiple multi-turn dialogues, each corresponding to different scenes. To determine the most appropriate task for each scene, we prompt MLLMs to evaluate the scene's capabilities, scoring them on a scale from 1 to 6. Only those tasks that receive a score of 5 or 6 are selected for dialogue generation. For multi-turn dialogues spanning multiple scenes, we specifically adopt an object-centered approach for cross-scene question design since objects often serve as the central element around which events unfold. This approach emphasizes the continuity and relationships of objects across scenes, enabling the generation of dialogues that are both contextually consistent and thematically coherent.

3.4 Quality Control

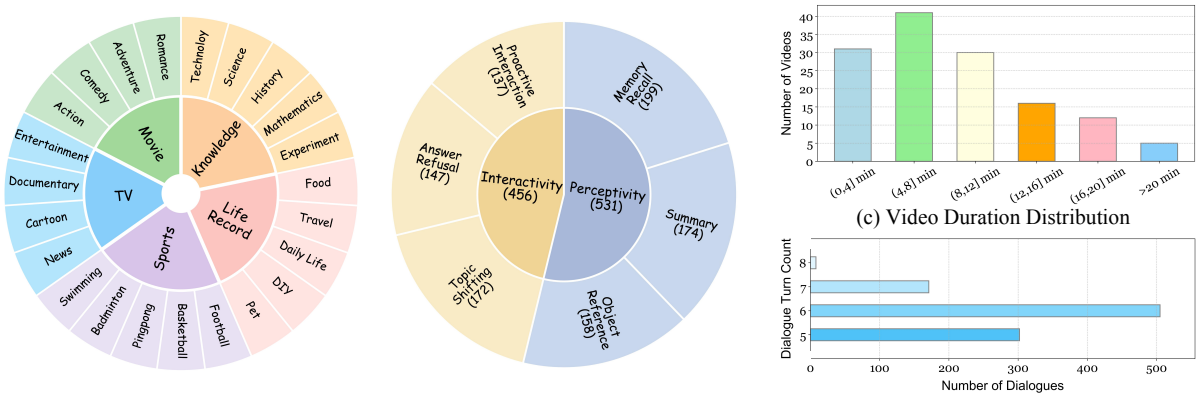
Following automated data collection, we employ the following two-stage human verification process to enhance dataset quality.

Stage 1: Eliminating information leakage. We categorize all benchmark questions into two types: (1) context-dependent, which can be answered solely based on dialogue history; and (2) video-dependent, which require direct grounding in the video content. We observe that in some generated dialogues, earlier QA pairs embedded excessive background hints, enabling models to answer subsequent questions without relying on the video. This led to an overrepresentation of context-dependent items, thereby weakening the evaluation of video understanding. To mitigate this, we systematically removed such cases to ensure that the majority of questions required genuine video-based reasoning.

Stage 2: Human verification and validation. After the first filtering, human annotators conducted a secondary review from a human perspective. We verify whether each question and answer pair was factually aligned with the video and free from ambiguities. Beyond factual correctness, we also examine whether each question is properly aligned with its intended ability dimension. For example, answer refusal questions must explicitly test whether a model can recognize "events absent from the video," while object reference questions must involve pronoun disambiguation. Any misaligned samples are discarded. Finally, we filter out overly simple questions, as they can be trivially solved by most models and fail to highlight multi-turn reasoning and video comprehension capacities.

3.5 Dataset Statistics

Figure 3 presents the statistics of MT-Video-Bench. It covers a broad range of topics across five main categories: Movie, TV, Sports, Knowledge, and Life Record, each with multiple sub-topics, ensuring a diverse and balanced data distribution. With a total of 987 multi-turn dialogues, the data distribution across the six primary tasks in MT-Video-Bench is relatively balanced, as shown in Figure 3 (b). Furthermore, our dataset features videos of varying lengths, with most being under 15 minutes and a small proportion exceeding 15 minutes, thereby ensuring coverage of both short and long videos. The number of dialogue turns typically ranges from 5 to 8, with an average of 5.88 turns per dialogue.



(a) Video Categories (b) Task Distribution (c) Video Duration Distribution (d) Dialogue Turn Distribution

Figure 3: Overview of MT-Video-Bench. (a) Video Categories. MT-Video-Bench includes videos spanning 5 major categories, ensuring diverse topical coverage. (b) Task Distribution. MT-Video-Bench consists of a total of 6 tasks with a relatively balanced distribution. (c) Video Duration Distribution. MT-Video-Bench includes both long and short videos. (d) Dialogue Turn Distribution. Multi-turn dialogues in MT-Video-Bench involve 5 to 8 rounds.

3.6 Evaluation Method

In multi-turn dialogues, each new turn depends on the interactions between users and assistants in previous turns. This dynamic is particularly crucial in tasks that involve high interactivity, such as proactive interactions. Therefore, we follow the multi-turn dialogue evaluation setup used in LLMs (Bai et al., 2024), leveraging our meticulously curated dataset as the golden context for dialogue history, rather than relying on self-predicted context from MLLMs.

For evaluation, we first use Gemini 2.5 Flash (Team, 2025) to construct a checklist for each QA pair. Specifically, each checklist consists of five yes/no questions designed to assess the accuracy of the model's responses and its performance on specific tasks. Then, manual validation is employed to filter out unqualified checklists. After filtering, each QA pair has an average of 3.29 questions in the final checklists, with 62.35% answered as yes and 37.65% as no. During the evaluation process, Gemini 2.5 Flash (Team, 2025) is used to answer each checklist question based on the model-generated answers. The evaluation metric is calculated as the accuracy (ACC), based on the proportion of correct answers across all checklists.

4 Experiments

4.1 Experimental Settings

For closed-source models, we evaluate Gemini 2.5 Pro (Team, 2025), Gemini 2.5 Flash (Team, 2025), and Doubao-Seed-1.6-vision (Seed, 2025). For open-source models, we select 18 representative MLLMs, including Qwen2.5 VL series (Bai et al., 2025), InternVL3.5 series (Wang et al., 2025c), LLaVA-Onevision series (Li et al., 2024b), InterVideo2.5 series (Wang et al., 2025d), LLaVA-Video series (Zhang et al., 2024b), LLaVA-NeXT-Video series (Zhang et al., 2024c), VideoChat-Flash series (Li et al., 2024c), VideoLlama3 series (Zhang et al., 2025a) and MiniCPM series (Yao et al., 2024).

Evaluation. For each model, we adopt a uniform sampling strategy to process video frames, setting the number of frames to 32. Each video is resized that the longer side is limited to 720 pixels and the other side is scaled proportionally. More details are described in Appendix B.1. For the prompts, we provide the evaluation prompts of six tasks of MT-Video-Bench in B.2.

4.2 Main Results

As shown in Table 2, we provide the performance results of different MLLMs on our MT-Video-Bench, and we have the following insightful and interesting observations:

- MT-Video-Bench is very challenging. Even the best-performing closed-source model, Gemini 2.5 Pro, only achieves 68.45% overall accuracy, which is inferior to the performance of human experts a lot.
- Among all evaluated models, Gemini 2.5 Pro consistently ranks first in both overall accuracy and every individual subtask. While closed-source systems still dominate overall performance, some open-source

Table 2: Evaluation results on MT-Video-Bench. **OR**: Object Reference. **MR**: Memory Recall. **CS**: Content Summary. **AR**: Answer Refusal. **TS**: Topic Shifting. **PI**: Proactive Interaction. The best performance and the second best performance are highlighted in green and blue, respectively.

Models	Overall	Perceptivity			Interactivity		
		OR	MR	CS	AR	TS	PI
Closed-Sourced Models							
Gemini 2.5 Pro (Team, 2025)	68.45	66.13	67.80	80.49	67.50	73.67	55.12
Gemini 2.5 Flash (Team, 2025)	63.30	63.44	63.41	73.48	64.32	68.12	47.04
Doubao-Seed-1.6-vision (Seed, 2025)	58.55	66.19	60.85	68.95	43.84	65.99	45.50
Open-Sourced Models							
Model Size > 8B							
Qwen2.5-VL-72B (Bai et al., 2025)	58.48	60.60	56.40	74.20	57.07	64.27	38.35
InternVL3.5-38B (Think) (Wang et al., 2025c)	58.11	60.87	60.36	69.90	46.86	65.17	45.51
Qwen2.5-VL-32B (Bai et al., 2025)	57.88	60.20	59.63	74.88	50.71	63.41	38.47
InternVL3.5-38B (No Think) (Wang et al., 2025c)	50.04	52.51	46.37	61.86	44.24	58.78	36.46
4B < Model Size ≤ 8B							
InternVL3.5-8B (Think) (Wang et al., 2025c)	56.29	57.81	54.82	73.18	47.62	62.50	41.84
Qwen2.5-VL-7B (Bai et al., 2025)	53.12	56.18	49.99	67.21	52.20	57.20	35.92
InternVL3.5-8B (No Think) (Wang et al., 2025c)	49.35	51.71	46.95	61.50	40.83	57.23	37.85
LLaVA-Video-7B (Zhang et al., 2025b)	49.17	53.85	43.57	63.64	41.32	56.67	35.98
MiniCPM-o (Yao et al., 2024)	48.41	55.06	43.27	61.59	34.58	57.53	38.43
MiniCPM-V4.5 (Yao et al., 2024)	47.06	51.57	43.08	56.17	38.46	52.58	40.47
InternVideo2.5-8B (Wang et al., 2025e)	47.04	44.87	43.49	60.33	45.23	54.81	33.50
VideoLLaMA3-7B (Bai et al., 2025)	46.06	52.06	42.40	55.74	45.23	48.25	32.69
LLaVA-OneVision-7B (Li et al., 2024d)	45.75	50.01	43.36	59.34	32.79	55.44	33.56
VideoChat-Flash-7B (Li et al., 2024e)	41.11	47.92	39.33	51.14	28.02	48.27	32.01
LLaVA-NeXT-Video-7B (Zhang et al., 2024d)	38.04	43.05	36.04	48.58	27.60	42.94	30.00
Model Size ≤ 4B							
InternVL3.5-4B (Think) (Wang et al., 2025c)	52.25	54.94	53.78	67.50	37.74	54.67	44.89
Qwen2.5-VL-3B (Bai et al., 2025)	48.07	50.64	43.54	65.82	46.80	50.33	31.30
InternVL3.5-4B (No Think) (Wang et al., 2025c)	45.90	46.03	46.19	61.30	30.41	55.72	35.74

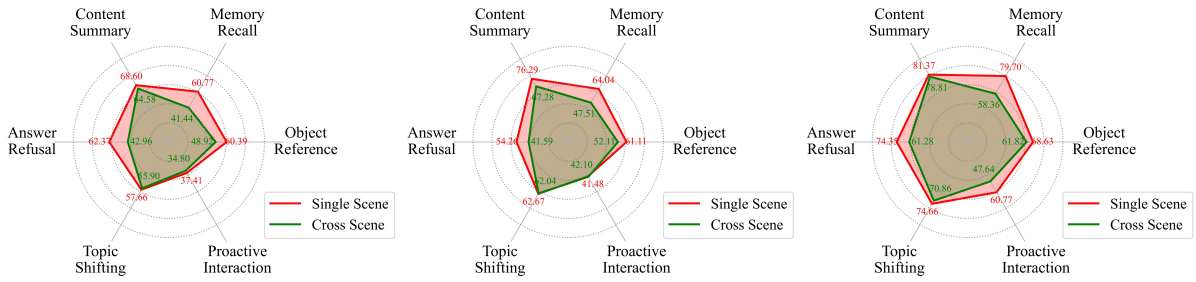
models demonstrate competitive results in specific dimensions. For example, Qwen2.5-VL-72B shows strong ability in MR, narrowing the gap with Gemini 2.5 Pro. However, on interaction-related subtasks such as AR, the performance difference between open-source and closed-source models remains substantial.

- Results vary significantly across different dimensions, and models generally perform better on perception-related subtasks, where large-scale models generally achieve stronger scores, sometimes exceeding 60. For example, the average score of OR is 54.55, while for PI is 38.60.
- Larger models tend to achieve higher accuracy. For instance, within the Qwen2.5-VL series, the 72B and 32B models significantly outperform the 7B and 3B variants across nearly all subtasks. Similarly, larger InternVL3.5 models achieve better results than their smaller counterparts. However, sometimes small MLLMs can lead to higher scores. For instance, the AR scores for Qwen2.5-VL-7B, Qwen2.5-VL-32B, and InternVideo2.5-8B are 52.20, 50.71, and 45.23, respectively. In addition, enabling *thinking mode* within the same model variant leads to significant performance improvements, suggesting that inference strategies, beyond model size, can substantially affect benchmark outcomes.

4.3 Further Analysis

4.3.1 Performance comparison between single scene and cross scene

Based on the selected three models in Figure 4, we summarize the following conclusions: (1) Across almost all abilities, model performance under the cross-scene setting is worse than under the single-scene setting. (2) Regardless of the setting, Gemini 2.5 Pro consistently outperforms Qwen2.5-VL-7B and InternVL3.5-8B across all abilities, particularly in Content Summary and Memory Recall, while also sustaining relatively high performance under the cross-scene condition. In comparison, InternVL3.5-8B performs comparably to Gemini 2.5 Pro in the single-scene setting but suffers from substantial degradation in the cross-scene setting. Meanwhile, Qwen2.5-VL-7B shows severe performance drops in Proactive



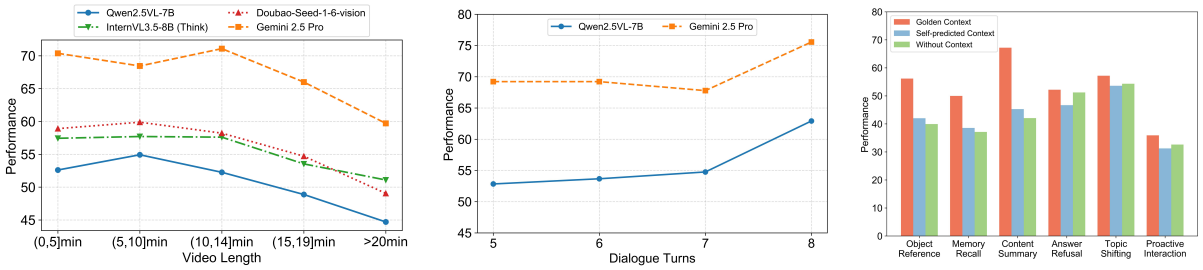
(a) Qwen2.5-VL-7B (b) InternVL3.5-8B (Think) (c) Gemini 2.5 Pro

Figure 4: Performance comparison of Qwen2.5-VL-7B, InternVL3.5-8B (Think), and Gemini 2.5 Pro across various tasks under single-scene and cross-scene settings.

Interaction and Memory Recall under cross-scene evaluation.

4.3.2 Performance of different video lengths

To study the impact of video length on model performance, videos are grouped into different length ranges. From Figure 5 (a), we find that: (1) Model performance generally decreases as video length increases, suggesting that longer videos pose greater challenges for capturing and reasoning over multi-turn dialogue content. (2) Higher-capacity models, such as Gemini 2.5 Pro, tend to achieve higher overall scores across all video lengths compared to smaller models like Qwen2.5VL-7B. However, all models exhibit noticeable performance drops for very long videos. (3) The performance gap between models is more pronounced for shorter videos, while for longer videos, the performance difference narrows.



(a) Performance of different video lengths (b) Performance across dialogue turns (c) Effect of dialogue context

Figure 5: Performance of different video lengths, dialogue turns, and settings of dialogue context.

4.3.3 Model performance across dialogue turns

To evaluate the impact of dialogue length on model performance, we conduct experiments with dialogues of varying total turn numbers with Gemini-2.5-Pro and Qwen2.5VL-7B. Several key observations can be drawn from the results shown in Figure 5 (b): Model performance tends to improve as the total number of turns increases, although the degree and stability of this improvement vary across models. This suggests that dialogue length plays a dual role in multi-turn video understanding: offering more contextual cues beneficial for reasoning while increasing the burden of sustaining coherent dialogue states. One possible reason for this pattern is that larger models are generally able to integrate contextual information more efficiently, leveraging additional turns to further improve. Smaller models, on the other hand, tend to rely more heavily on the accumulation of dialogue context across multiple turns.

4.3.4 Effect of dialogue context

To investigate the role of contextual information, we design three experimental settings:

Without Context. The model answers each question solely based on the video.

With Self-predicted Context. The model is provided with its own generated dialogue history.

With Golden Context. The model is provided with our meticulously curated golden dialogue history.

As shown in Figure 5 (c), the golden context yields the highest performance across all abilities. However,

the accuracy of the context is more critical than its mere presence. Self-predicted context does not always lead to performance gains and remains roughly on par with the no-context setting, as the model-generated dialogue history may contain factual errors or semantic drift. These inconsistencies may accumulate over multiple rounds, causing the model to be misled by "incorrect memories" in subsequent responses.

4.3.5 Effect of different numbers of frames

In Figure 6, results of Qwen2.5-VL-7B are grouped according to the number of frames, with the resolution fixed at 720p and the number of frames varying from 4 to 64. Several distinct trends emerge from the results:

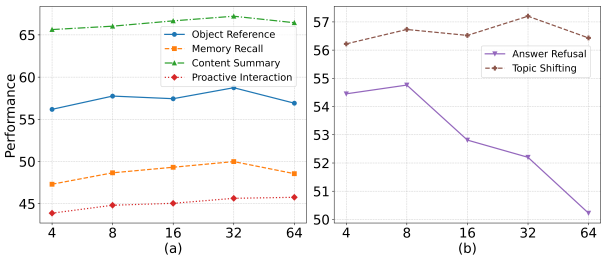


Figure 6: Ablation results of frames on different abilities. (a) Results of Object Reference, Memory Recall, Content Summary, and Proactive Interaction; (b) Results of Answer Refusal and Topic Shifting.

(1) **Topic Shifting.** The performance on topic shifting remains largely unaffected by the number of frames. This suggests that the ability to adapt to unexpected user queries and maintain coherent responses is primarily dependent on dialogue-level reasoning rather than fine-grained visual information.

(2) **Anwser Refusal.** Models perform better on answer refusal cases when fewer frames are provided. With limited visual evidence, the model becomes more cautious in generating answers and is less likely to hallucinate unsupported content, while when more frames are provided, the model may overfit to irrelevant visual cues and produce unwarranted responses, leading to decreased performance on this ability.

(3) **Long Context Benefits.** For the other four abilities, as shown in Figure 6 (a), models' performance consistently improves with more frames, because longer visual evidence provides richer contextual signals, which support more accurate reasoning.

4.3.6 Effect of different resolutions

To further analyze the impact of video input quality on model performance, we evaluate the performance of Qwen2.5-VL-7B under different resolutions, with the number of frames fixed at 32. The input video frames are set to 120p, 240p, 480p, 720p, and 960p.

In Figure 7, the scores of nearly all abilities continue to improve from 120p to 720p, while a slight decline is observed when the resolution further increases to 960p. This suggests that, within a certain range, higher resolution indeed enhances the model's ability to capture visual details, but excessive resolution may lead to a decline in performance, primarily due to the increased number of input tokens that exceeds the model's optimal processing capacity.

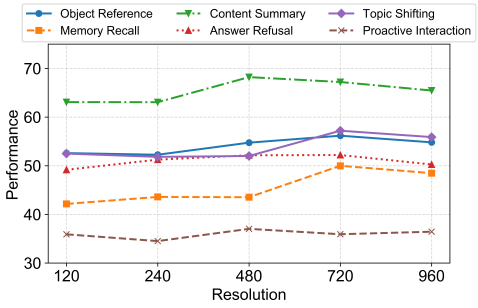


Figure 7: Ablation results of resolutions on different abilities.

5 Conclusion

In this paper, we presented MT-Video-Bench, a holistic benchmark for evaluating MLLMs in multi-turn video dialogues. Unlike prior video understanding benchmarks that primarily focus on single-turn factual perception, MT-Video-Bench jointly assesses perceptivity and interactivity through six carefully defined capabilities, covering tasks such as memory recall, topic shifting, and proactive interaction. Our evaluation of 20 state-of-the-art models provides insightful findings, and we hope our MT-Video-Bench can establish a rigorous foundation for future research, highlighting the need for models that can reason over long contexts while engaging in natural, adaptive conversations.