

CVBench: Evaluating Cross-Video Synergies for Complex Multimodal Understanding and Reasoning

Nannan Zhu^{1†}, Yonghao Dong^{1†}, Teng Wang^{2*}, Xueqian Li¹, Shengjun Deng³, Yijia Wang¹
 Zheng Hong¹, Tiantian Geng⁴, Guo Niu³, Hanyan Huang¹, Xiongfei Yao³, Shuaiwei Jiao³
¹ Sun Yat-sen University ² University of Hong Kong
³ Foshan University ⁴ University of Birmingham

Abstract

While multimodal large language models (MLLMs) exhibit strong performance on single-video tasks (e.g., video question answering), their ability across multiple videos remains critically underexplored. However, this capability is essential for real-world applications, including multi-camera surveillance and cross-video procedural learning. To bridge this gap, we present CVBench, the first comprehensive benchmark designed to assess cross-video relational reasoning rigorously. CVBench comprises 1,000 question-answer pairs spanning three hierarchical tiers: *cross-video object association* (identifying shared entities), *cross-video event association* (linking temporal or causal event chains), and *cross-video complex reasoning* (integrating commonsense and domain knowledge). Built from five domain-diverse video clusters (e.g., sports, life records), the benchmark challenges models to synthesise information across dynamic visual contexts. Extensive evaluation of 10+ leading MLLMs (including GPT-4o, Gemini-2.0-flash, Qwen2.5-VL) under zero-shot or chain-of-thought prompting paradigms. Key findings reveal stark performance gaps: even top models, such as GPT-4o, achieve only 60% accuracy on causal reasoning tasks, compared to the 91% accuracy of human performance. Crucially, our analysis reveals fundamental bottlenecks inherent in current MLLM architectures, notably deficient inter-video context retention and poor disambiguation of overlapping entities. CVBench establishes a rigorous framework for diagnosing and advancing multi-video reasoning, offering architectural insights for next-generation MLLMs. The data and evaluation code are available at <https://github.com/Hokhim2/CVBench>.

1 Introduction

The ability to reason across multiple video streams is fundamental to real-world applications, such as multi-camera surveillance systems, distributed procedural learning, and multi-view activity analysis. However, despite significant progress in single-video understanding by multimodal large language models (MLLMs) [33, 60, 42, 15, 21, 26, 9], their capacity for cross-video relational understanding remains largely unexplored. This capability requires reasoning about entities, events, and their interrelationships across distinct spatiotemporal contexts. The research gap stems not from the insufficient use of single-video benchmarks [23, 10, 12], but from inherent architectural limitations in current evaluations that enforce isolated video processing.

Current MLLMs face fundamental barriers in cross-video scenarios: 1) *Inter-video context linking*: models lack mechanisms to persistently integrate and retain object/event states across videos separated by time and space; 2) *Entity disambiguation*: overlapping or visually similar entities across videos lack unique spatiotemporal grounding, causing identification errors; 3) *Temporal-causal modeling*: existing models poorly capture long-range dependencies and causal links spanning video sequences.

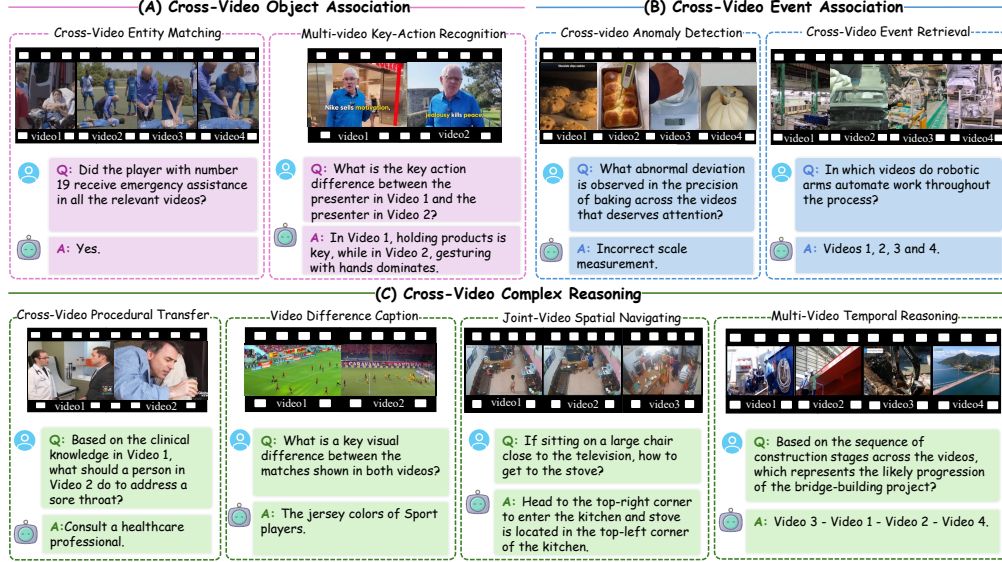


Figure 1: Illustration of CVBench task definition. CVBench focuses on three core tasks that are crucial for real-world applications: (A) cross-video object association (identifying shared entities across different videos), (B) cross-video event association (linking temporal or causal event chains that span across video sequences), and (C) cross-video complex reasoning (integrating commonsense to understand and reason about complex interrelationships between video content). The benchmark format is multi-choice QA, and here we only show the correct answer for visualization.

To bridge this gap and diagnose architectural limitations, we introduce CVBench, the first diagnostic benchmark for cross-video understanding and reasoning abilities. As shown in Fig. 1, it evaluates large multimodal models through three hierarchical tasks: *cross-video object association* for verifying entity persistence across varying visual contexts, *cross-video event association* for modeling temporal sequences and causal relationships across videos, and *cross-video complex reasoning* for integrating visual cues and external knowledge for hierarchical spatiotemporal understanding. CVBench comprises 1,000 QA pairs sampled from five diverse domains (artistic performances, sports competitions, films and television, life records, and knowledge) to ensure generalization. This domain diversity challenges models to reason across dynamic contexts, requiring spatio-temporal integration and multi-view synthesis. Expert annotations capture real-world complexities by modeling entities, events, and inter-video relationships. This rigorous foundation enables diagnostic evaluation of cross-video reasoning capabilities.

We evaluate over ten state-of-the-art closed-sourced and open-sourced MLLMs, including GPT-4o, Gemini series, and Qwen2.5-VL, under zero-shot and chain-of-thought prompting paradigms. Our evaluation reveals significant performance gaps in these models. Specifically, GPT-4o achieves only 60% accuracy on causal reasoning tasks, which is lower than the human-level performance of 91%. Additionally, these models struggle with inter-video context retention and difficulties in disambiguating overlapping entities, which hinders their ability to integrate information from multiple sources. These findings underscore the limitations of current MLLMs in handling complex cross-video tasks, highlighting the need for more advanced architectures to address challenges like long-term context retention and entity disambiguation.

The contributions of this paper can be summarized as:

- We introduce CVBench, a diagnostic benchmark explicitly targeting cross-video understanding through object association, event association, and complex reasoning.
- We release a carefully curated, multi-domain multiple-choice QA dataset spanning five settings to stress-test spatiotemporal integration and multi-view synthesis.
- We provide a systematic evaluation of leading MLLMs on CVBench, exposing performance gaps and critical challenges in cross-video reasoning, and further identifying key research directions to advance the field.

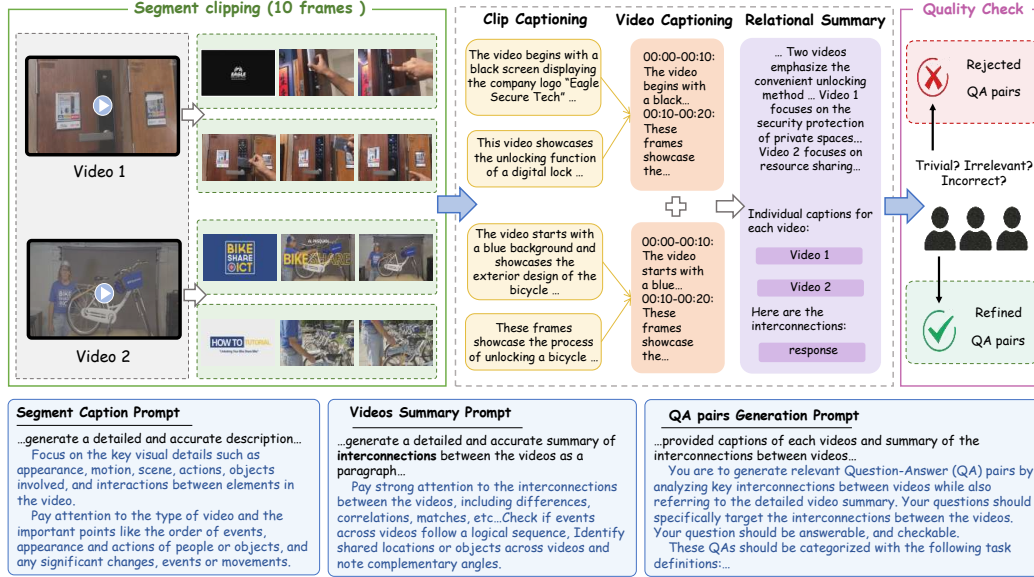


Figure 2: The pipeline for question-answer (QA) pairs annotations in CVBench. The QA generation pipeline is streamlined as four steps: video segment captioning, relational summary annotation, question-answer generation and quality control.

2 Related Work

2.1 Multimodal Large Language Models

The evolution of MLLMs has progressed from image-based models [30, 8, 64, 62, 19, 29, 46] to video-capable architectures [41, 4, 50, 22, 51, 1, 56, 63, 61, 14, 45, 65, 7], marked by two critical innovations: 1) Long-range context modeling that enables efficient handling excessively long temporal contexts [25, 42]; 2) Spatio-temporal fusion that integrates both appearance and motion dynamics with language reasoning [39, 34, 59]. Despite these advances, current MLLMs remain architecturally constrained to single-video scenario, lacking mechanisms for cross-video interactions, exhibiting inadequate spatiotemporal grounding for overlapping entities, and demonstrating limited modeling of temporal event chains. These fundamental limitations restrict their applicability to multi-camera surveillance and cross-video applications.

2.2 Video Benchmarks

A comprehensive and objective benchmark is essential for evaluating MLLMs, enabling the comparison and investigation of the performance of various models. Recent advances in multimodal learning have led to the development of diverse benchmarks [20, 24, 49, 53, 18, 58, 35, 5, 31, 11, 28, 38, 44, 13, 2, 16, 40, 52, 23, 37, 55] for video understanding tasks. However, existing video evaluation frameworks exhibit critical limitations in assessing cross-video reasoning capabilities. First, *general multimodal frameworks* are predominantly with broad domain coverage, like Video-MMMU [17] and Video-MME [10]), aiming to evaluating instruction-following capabilities but lack cross-video relational tasks. Second, *long-term video benchmarks* contains longer videos with multiple sub-events, however are confined to single source or similar creators, such as Egoschema [36], TempCompass [32], and LVBench [47]. These shortcomings significantly constrain the applicability of existing benchmarks to real-world settings, such as multi-camera surveillance and cross-video procedural learning, where robust cross-domain cross-video reasoning and relational understanding are critical. To address these gaps, we introduce CVBench, the first diagnostic framework for cross-video relational reasoning. CVBench pioneers a three-tiered evaluation of spatiotemporal integration capabilities, encompassing object-level association, event-level association and complex reasoning.

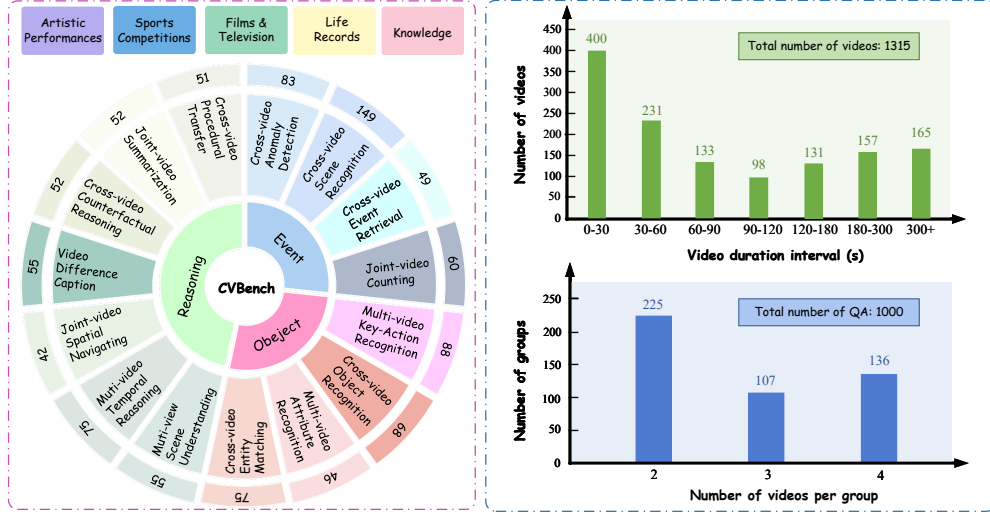


Figure 3: CVBench statistics and task categories. CVBench comprises three primary task categories and 15 video sub-categories across diverse video domains. We present the distribution of video durations, the number of videos, and representative domains.

3 CVBench: Cross-Video Reasoning Benchmark

3.1 Task Definition

CVBench pioneers the evaluation of cross-video relational reasoning through a three-tiered diagnostic framework, encompassing *object association*, *event association*, and *complex reasoning*. This taxonomy systematically addresses critical gaps in real-world video understanding by requiring models to establish inter-video relationships through spatiotemporal pattern recognition beyond single-video analysis. The framework comprises 15 subtasks (detailed in Appendix) categorized as follows:

- **Object association:** Cross-video object recognition, multi-video attribute recognition, joint-video counting, and cross-video entity matching. These require consistent identification of entities across domains or viewpoints despite occlusion or deformation.
- **Event association:** Cross-video anomaly detection, scene recognition, key-action recognition, event retrieval, multi-view scene understanding, and temporal reasoning. These demand fusion of spatiotemporal signatures across videos to reconstruct event dynamics.
- **Complex reasoning:** Joint-video spatial navigating, video difference captioning, cross-video counterfactual reasoning, joint-video summarization, and procedural transfer. These necessitate higher-order synthesis including spatial imagination, fine-grained reasoning, causal inference and adaptive knowledge transfer.

3.2 Dataset Collection

CVBench is constructed with three-step process: video curation, question-answer annotation, and quality control. First, diverse videos are selected to ensure broad scenario coverage. Next, question-answer pairs are crafted to capture complex reasoning and cross-video dependencies. Finally, rigorous quality control removes ambiguous or trivial entries, ensuring the dataset’s relevance and reliability for evaluation.

Video curation. We employ task-driven sampling from YouTube to construct diverse video groups (at most four videos per group). Detailed statistic are shown in Fig. 3. We ensure source diversity through the inclusion of multi-creator channels (videos with similar or same creators), cross-domain associations (videos across different domain but with shared topics or intrinsic relations), and the segmentation of long videos (sequential occurrence of event-level activities). Specifically, we measure video quality under the following four curation criteria:

Benchmark	#Video	#QA	#Video Per QA	Duration (s)	Multi-level	CV.U	CV.R
TempCompass [32]	410	500	1	11.4	✗	✗	✗
MSVD-QA [53]	504	13,157	1	9.8	✗	✗	✗
Video-MME [10]	900	2,700	1	1017.9	✓	✗	✗
NExT-QA [52]	1,000	8,564	1	39.5	✗	✗	✗
AutoEval-Video [5]	327	327	1	14.6	✗	✗	✗
EgoSchema [36]	5,063	5,063	1	180.0	✗	✗	✗
ActivityNet-QA [57]	800	8,000	1	111.4	✗	✗	✗
CVBench (Ours)	1,315	1,000	2 ~ 4	106.6	✓	✓	✓

Table 1: Comparison of video understanding benchmarks. CV.U and CV.R denote cross-video understanding and cross-video reasoning, respectively. “Multi-level” indicates videos with multiple duration levels.

- Multi-domain coverage (daily life, sports, surveillance);
- Intrinsic inter-video relationships (differences, similarities, co-occurrence, causality causal, temporal order, etc);
- Duration constraints (primarily within 10 min, with videos exceeding 10 min comprising less than 5% of the dataset);
- Visual quality and availability (at least 720p 24fps videos).

QA generation. The multi-choice QA generation pipeline comprises four sequential stages (see Fig. 2). First, videos are segmented using an adaptive frame sampling strategy, where the sampling interval increases with video length to optimize both semantic coverage and annotation efficiency. Next, detailed segment-level captions are generated, emphasizing key visual elements such as appearance, actions, and interactions. These captions are then aggregated into comprehensive video-level descriptions, followed by a relational summary that explicitly captures interconnections across videos, including entity state transitions, causal relationships, and contextual complementarities or contradictions. Leveraging these multi-level annotations, we synthesize at least two multi-choice QA (with one correct choice and three disturbance) pairs per group, each designed to require cross-video dependencies, incorporate multiple adversarial distractors (entity, temporal, or causal), and involve multi-hop reasoning. These constraints are enforced through prompt engineering with GPT-4o to ensure the QA pairs are non-trivial and task-aligned. Finally, all QA pairs undergo rigorous quality control, where trivial, irrelevant, or incorrect pairs are filtered out, resulting in a challenging and reliable QA dataset. The prompts used for caption generation, relational summary construction, and QA pair synthesis with GPT-4o are illustrated in Fig. 2.

Quality control. In terms of video collection, we hire eight annotators with bachelor or master degrees follow a rigorous three-stage collection protocol. First, an initial keyword-based retrieval process yields 1,500 candidate videos. Next, a gap-filling stage augments underrepresented domains to enhance coverage. Finally, a quality-controlled replacement procedure is activated when QA pair quality are not met, ensuring the reliability of the dataset. The resulting curated dataset comprises 1,315 videos with a stratified duration distribution: short, mid-short, mid, and long videos constitute 30.4%, 17.5%, 27.5%, and 24.4% samples, respectively.

For the quality control of QA pairs, we follow a two-stage human checking protocol. First, eight annotators verify QA pairs against a set of 15 task-specific criteria, initiating corrective actions as needed: manual revision for low-quality questions, option correction for unclear disturbance term, recollect videos for topic inconsistencies, etc. Then, five new annotators conduct a comprehensive re-examination (200 QA pairs per annotator), verifying question compliance, answer correctness, and the absence of hallucinations.

3.3 Comparison with Previous Benchmarks

CVBench differs significantly from traditional single-video and multi-image benchmarks, as shown in Table 1. While single-video benchmarks focus on isolated content, they fail to evaluate models’

Categories (QA pair numbers)	Average (382)	M. SU (55)	M. TR (75)	J. SN (42)	VDC (55)	C. CR (52)	J. S (52)	C. PT (51)
Human	91.3	80.0	100.0	83.3	100.0	100.0	100.0	83.3
Random Choice	26.2	21.8	30.7	21.4	21.8	30.8	25.0	29.4
<i>Closed-source MLLMs</i>								
GPT-4o-mini	64.4	87.3	41.3	47.6	58.2	67.3	80.8	70.6
GPT-4o	69.1	89.1	58.7	61.9	61.8	63.5	82.7	66.7
Gemini-1.5-flash	62.8	85.5	49.3	38.1	61.8	57.7	80.8	64.7
Gemini-2.0-flash	69.4	89.1	58.7	50.0	70.9	67.3	76.9	66.7
<i>Open-source MLLMs</i>								
Qwen2.5-Omni-7B [54]	52.4	70.9	22.7	31.0	56.4	57.7	75.0	60.8
Qwen2.5-VL-7B [3]	51.3	80.0	22.7	26.2	50.9	55.8	69.2	60.8
LLaVA-OneVision-7B [27]	52.6	83.6	40.0	38.1	45.5	42.3	61.5	52.9
VideoLLaMA3-7B [4]	57.6	78.2	29.3	35.7	60.0	59.6	80.8	64.7
InternVL2.5-8B [6]	59.4	83.6	26.7	50.0	60.0	69.2	67.3	68.6
Phi-4-Multimodal-5B [1]	49.7	69.1	26.7	31.0	43.6	57.7	65.4	58.8
Internvideo2.5-8B [48]	57.3	85.5	33.3	47.6	56.4	55.8	63.5	64.7
<i>RL-based Thinking MLLMs</i>								
Video-R1-7B [9]	49.2	74.5	22.7	28.6	47.3	48.1	76.9	52.9

Table 2: Performance of MLLMs on CVBench in cross-video complex reasoning tasks, evaluated across closed-source and open-source MLLMs. The tasks include: multi-view scene understanding (M. SU), multi-video temporal reasoning (M. TR), joint-video spatial navigation (J. SN), video difference captioning (VDC), cross-video counterfactual reasoning (C. CR), joint-video summarization (J. S), and cross-video procedural transfer (C. PT). For human evaluation, we employed five annotators and reported the average accuracy.

ability to integrate information across multiple videos and understand temporal and spatial relationships. Multi-image benchmarks, although involving multiple inputs, do not address the dynamic, temporal aspects of videos or the complex reasoning required for cross-video understanding. In contrast, CVBench is explicitly designed to assess cross-video relationships, incorporating tasks that require models to reason across multiple video streams. With diverse video domains and task types, CVBench provides a comprehensive evaluation framework that advances cross-video understanding and reasoning. Unlike existing benchmarks, CVBench evaluates cross-video understanding (CV.U) and reasoning (CV.R), ensuring a more holistic assessment of multimodal models’ capabilities in real-world applications.

4 Benchmarked Results

4.1 Experimental Settings

Models. Our evaluation comprehensively assesses various MLLMs, including closed-source and open-source models. A prominent model in our suite is Gemini [43], a closed-source model known for its sophisticated multimodal processing capabilities. On the open-source side, Qwen2.5-VL [9] stands out, offering transparency with its accessible codebase for further exploration and customization. We also evaluate advanced multimodal reasoning models such as Video-R1 [9], which leverages reinforcement learning to integrate multimodal information and perform complex cross-video reasoning tasks. For ablation study, we choose Qwen2.5VL-7B as the default model.

Settings. Unless otherwise specified, all open-source models are evaluated using 8 frames per video (e.g., totally 32 frames for a four-video QA example). For closed-source APIs, we use the default frame sampling rate. To enable multi-video inputs for mainstream models primarily trained on single-video tasks, we insert prompt tokens (i.e., “The video [video_index].”) before the visual tokens of each video to indicate their order and identifiers. The default resolution is set to 448×448 for all open-source models that support dynamic resolution; for models without such support, we use their default resolution as specified in this paper.

Categories (QA pair numbers)	Average (249)	C. OR (68)	M. AR (46)	J. C (60)	C. EM (75)
Human	88.9	85.7	83.3	80.0	100.0
Random Choice	27.4	25.0	32.6	25.0	28.4
<i>Closed-source MLLMs</i>					
GPT-4o-mini	56.5	58.8	60.9	36.7	58.1
GPT-4o	66.9	66.2	76.1	51.7	66.2
Gemini-1.5-flash	60.5	63.2	69.6	35.0	60.8
Gemini-2.0-flash	64.5	67.6	73.9	38.3	63.5
<i>Open-source MLLMs</i>					
Qwen2.5-Omni-7B [54]	54.4	52.9	60.9	35.0	59.5
Qwen2.5-VL-7B [3]	53.2	52.9	63.0	33.3	51.4
LLaVA-OneVision-7B [27]	48.0	39.7	65.2	35.0	43.2
VideoLLaMA3-7B [4]	53.6	45.6	73.9	35.0	50.0
InternVL2.5-8B [6]	60.9	58.8	76.1	43.3	54.1
Phi-4-Multimodal-5B [1]	48.8	55.9	56.5	30.0	40.5
Internvideo2.5-8B [48]	59.3	60.3	67.4	43.3	56.8
<i>RL-based Thinking MLLMs</i>					
Video-R1-7B [9]	54.4	54.4	63.0	30.0	40.5

Table 3: Performance of MLLMs on CVBench regarding cross-video object association, evaluated across closed-source and open-source MLLMs. Tasks include: cross-video object recognition (C. OR), multi-video attribute recognition (M. AR), joint-video counting (J. C), and cross-video entity match (C. EM).

Categories (QA pair numbers)	Average (369)	C. AD (83)	C. SR (149)	M. KAR (88)	C. ER (49)
Human	92.7	100.0	91.7	83.3	83.3
Random Choice	33.8	25.0	53.0	14.8	24.5
<i>Closed-source MLLMs</i>					
GPT-4o-mini	60.0	56.0	60.4	58.0	69.4
GPT-4o	70.8	66.7	74.5	65.9	75.5
Gemini-1.5-flash	68.7	64.3	70.5	64.8	77.6
Gemini-2.0-flash	67.0	65.5	67.8	67.0	67.3
<i>Open-source MLLMs</i>					
Qwen2.5-Omni-7B [54]	58.9	53.6	61.1	58.0	63.3
Qwen2.5-VL-7B [3]	54.0	53.6	49.7	58.0	61.2
LLaVA-OneVision-7B [27]	51.6	33.3	65.8	54.5	34.7
VideoLLaMA3-7B [4]	60.3	56.0	61.1	63.6	59.2
InternVL2.5-8B [6]	62.4	57.1	61.7	64.8	69.4
Phi-4-Multimodal-5B [1]	51.9	44.0	54.4	59.1	44.9
Internvideo2.5-8B [48]	64.0	60.7	62.4	68.2	67.3
<i>RL-based Thinking MLLMs</i>					
Video-R1-7B [9]	61.6	59.5	65.8	55.7	63.3

Table 4: Performance of MLLMs on CVBench regarding cross-video event association, evaluated across closed-source and open-source MLLMs. Tasks include: cross-video anomaly detection (C. AD), cross-video scene recognition (C. SR), multi-video key-action recognition (M. KAR), and cross-video event retrieval (C. ER).

4.2 Results and Findings

Overall performance. As presented in Table 2, Table 3 and Table 4, our evaluation on CVBench reveals a significant performance gap between current MLLMs and human-level proficiency in cross-video reasoning. While human evaluators achieve an average accuracy of 91.3%, the best-performing model, Gemini-2.0-flash, reaches only 69.4%, leaving a substantial 21.9 point deficit. The results highlight a clear hierarchy of task difficulty. Models demonstrate pronounced weaknesses in tasks demanding precise temporal and causal alignment across videos, such as multi-video temporal reasoning (M. TR) and joint-video spatial navigation (J. SN), where even top models fail to surpass 60% accuracy. In contrast, they exhibit relative competence in tasks that rely on aggregating high-level semantic information, like multi-view scene understanding (M. SU) and joint-video summarization (J. S), with several models achieving over 80% accuracy.

Closed-source vs. open-source models. Closed-source models consistently outperform open-source models. This suggests that proprietary architectures may utilize more refined optimization techniques or superior capabilities in handling complex cross-video relationships, revealing a substantial gap between commercial-grade and publicly available MLLMs in this domain.

Sensitivity to input perturbations. For multi-video understanding, the way videos are presented may impact model performance. Table 5 illustrates the performance differences when input videos are presented in regular order, disordered, or randomly selected. In the disordered input scenario, where the video order is randomized, the accuracy drops from 46.8% to 42.3%, with an average change of 4.5%. For random single inputs, when models are presented with only a single video, performance decreases more sharply, from 46.8% to 37.5%, with an average change of 9.3%. These findings highlight the sensitivity of models to input order, which further implies our benchmark focuses on fine-grained, temporal-aware relations between videos instead of merely global coarse perception.

Run ID	#1	#2	#3	Avg.
<i>Disordered Input</i>				
Normal	46.8	46.8	46.8	46.8
Disordered	43.0	41.3	42.7	42.3
Δ	3.8 ↓	5.5 ↓	4.1 ↓	4.5 ↓
<i>Random Single Input</i>				
Normal	46.8	46.8	46.8	46.8
Single	37.2	37.5	37.8	37.5
Δ	9.6 ↓	9.3 ↓	9.0 ↓	9.3 ↓

Table 5: Ablation on input perturbations.

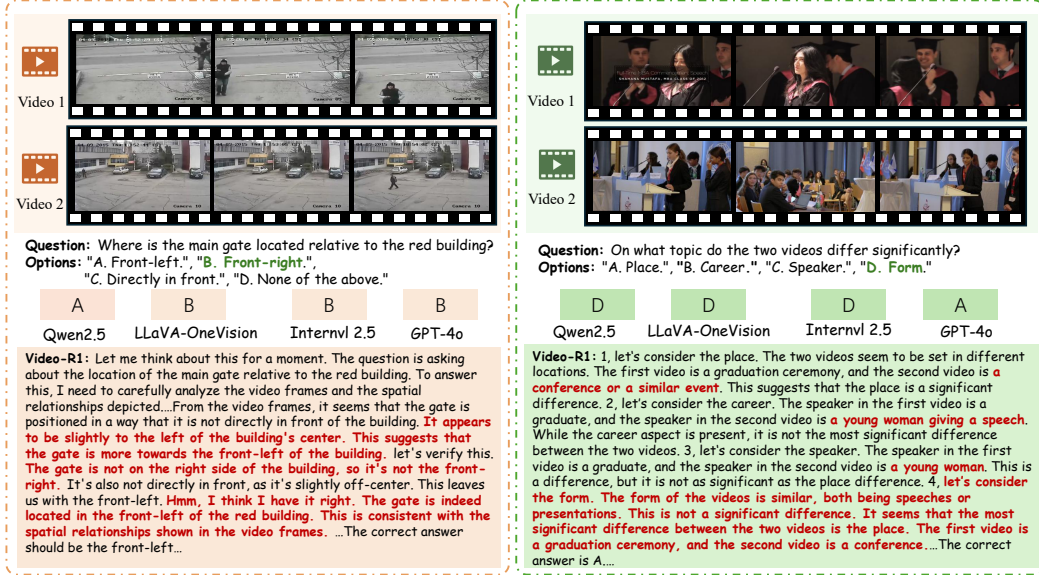


Figure 5: Qualitative cases of multi-video reasoning on CVBench based on Video-R1.

Frame resolution. We investigate the impact of frame resolution using Qwen2.5-VL, considering different resolutions and varying numbers of input frames. With 8 frames, accuracy increases slightly from 46.0% at 224×224 to 46.8% at 448×448 (+0.8%), indicating that higher resolution leads to better performance. Similar trends are observed with 16 and 32 frames. Additionally, increasing the number of input frames consistently improves performance across both resolutions.

Number of video instances & duration. Fig. 4 illustrates how video duration and the number of videos affect model performance (Qwen2.5-VL). Accuracy improves with video length up to 900 seconds (48.8% for 0–300s), but declines for longer videos (45.6% for 900–1200s), likely due to increased sequence complexity and context management challenges. We also highlight the difficulty of multi-video reasoning. Models that perform well on single-video tasks struggle with cross-video tasks. Shown in Fig. 4, the model achieves 52.2%/47.9%/44.8% accuracy with two/three/four videos, showing increased complexity of multi-video tasks.

Resolution	224 × 224	448 × 448	Δ
8f	46.0	46.8	0.8 ↑
16f	47.3	47.6	0.3 ↑
32f	49.4	51.0	1.6 ↑
Δ	1.7 ↑	2.1 ↑	-

Table 6: Impact of frame number & resolution.

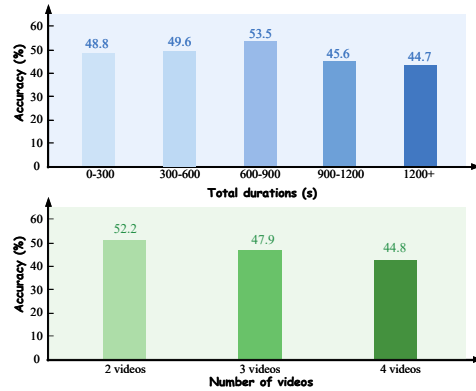


Figure 4: Impact of video number and duration.

Qualitative analysis of reasoning-based MLLMs. Fig. 5 illustrates the reasoning process of Video-R1, emphasizing the implicit cues that shed light on the model’s behavior. The left sub-figure depicts hallucination in reasoning, where the model generates responses that are logically disconnected. In contrast, the right sub-figure highlights the model’s limitations in achieving deeper understanding and reasoning compared to human cognition. Video-R1 frequently produces logically incorrect responses and struggles to reach human-level reasoning. This shows the need for ongoing advancements in multimodal models to better emulate human-like reasoning, particularly for tasks that require cross-modal integration and contextual understanding in real-world scenarios.

5 Conclusion

This paper introduces CVBench, a novel benchmark for evaluating the performance of MLLMs in cross-video understanding and reasoning tasks. Our evaluation demonstrates that, while MLLMs excel in single-video tasks, they encounter significant challenges in spatiotemporal alignment, entity recognition, and temporal reasoning when processing multiple video streams. CVBench provides a comprehensive dataset and clearly defined tasks, highlighting the unique demands of cross-video scenarios, which require the synthesis of information across both temporal and spatial dimensions. We present a systematic evaluation of leading MLLMs on CVBench, revealing notable performance gaps and critical challenges in cross-video reasoning, and further identifying key research directions to advance the field. Addressing these challenges is essential for improving multi-video reasoning and expanding the applicability of MLLMs to real-world problems.

References

- [1] Abdelrahman Abouelenin et al. “Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras”. In: *arXiv preprint arXiv:2503.01743* (2025).
- [2] Kirolos Ataallah et al. “Infinibench: A comprehensive benchmark for large multimodal models in very long video understanding”. In: *arXiv preprint arXiv:2406.19875* (2024).
- [3] Shuai Bai et al. *Qwen2.5-VL Technical Report*. 2025. arXiv: 2502.13923 [cs.CV]. URL: <https://arxiv.org/abs/2502.13923>.
- [4] Zhang Boqiang et al. “VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding”. In: *arXiv preprint arXiv:2501.13106* (2025). URL: <https://arxiv.org/abs/2501.13106>.
- [5] Xiuyuan Chen et al. “Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering”. In: *European Conference on Computer Vision*. Springer. 2024, pp. 179–195.
- [6] Zhe Chen et al. “Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling”. In: *arXiv preprint arXiv:2412.05271* (2024).
- [7] Gheorghe Comanici et al. “Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities”. In: *arXiv preprint arXiv:2507.06261* (2025).
- [8] Wenliang Dai et al. “Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023”. In: *arXiv preprint arXiv:2305.06500* 2 (2023).
- [9] Kaituo Feng et al. “Video-R1: Reinforcing Video Reasoning in MLLMs”. In: *arXiv preprint arXiv:2503.21776* (2025).
- [10] Chaoyou Fu et al. “Video-mme: The first-ever comprehensive evaluation benchmark of multimodal llms in video analysis”. In: *arXiv preprint arXiv:2405.21075* (2024).
- [11] Chaoyou Fu et al. “Video-mme: The first-ever comprehensive evaluation benchmark of multimodal llms in video analysis”. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 24108–24118.
- [12] Tiantian Geng et al. “Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos”. In: *arXiv preprint arXiv:2411.19772* (2024).
- [13] Ridouane Ghermi et al. “Short film dataset (sfd): A benchmark for story-level video understanding”. In: *arXiv e-prints* (2024), arXiv:2406.
- [14] Dong Guo et al. “Seed1. 5-vl technical report”. In: *arXiv preprint arXiv:2505.07062* (2025).
- [15] Bo He et al. “Ma-lmm: Memory-augmented large multimodal model for long-term video understanding”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 13504–13514.
- [16] Xuehai He et al. “Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos”. In: *arXiv preprint arXiv:2406.08407* (2024).
- [17] Kairui Hu et al. “Video-MMMU: Evaluating Knowledge Acquisition from Multi-Discipline Professional Videos”. In: *arXiv preprint arXiv:2501.13826* (2025).
- [18] Yunseok Jang et al. “Tgif-qa: Toward spatio-temporal reasoning in visual question answering”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2758–2766.

- [19] Xin Lai et al. “Lisa: Reasoning segmentation via large language model”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 9579–9589.
- [20] Jie Lei et al. “Tvqa: Localized, compositional video question answering”. In: *arXiv preprint arXiv:1809.01696* (2018).
- [21] Bo Li et al. “Llava-onevision: Easy visual task transfer”. In: *arXiv preprint arXiv:2408.03326* (2024).
- [22] Bo Li et al. “Otter: A multi-modal model with in-context instruction tuning”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).
- [23] Kunchang Li et al. “Mvbench: A comprehensive multi-modal video understanding benchmark”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 22195–22206.
- [24] Linjie Li et al. “Hero: Hierarchical encoder for video+ language omni-representation pre-training”. In: *arXiv preprint arXiv:2005.00200* (2020).
- [25] Xinhao Li et al. “Videochat-flash: Hierarchical compression for long-context video modeling”. In: *arXiv preprint arXiv:2501.00574* (2024).
- [26] Xinhao Li et al. “VideoChat-R1: Enhancing Spatio-Temporal Perception via Reinforcement Fine-Tuning”. In: *arXiv preprint arXiv:2504.06958* (2025).
- [27] Yanwei Li, Chengyao Wang, and Jiaya Jia. “Llama-vid: An image is worth 2 tokens in large language models”. In: *European Conference on Computer Vision*. Springer. 2024, pp. 323–340.
- [28] Yunxin Li et al. “Videovista: A versatile benchmark for video understanding and reasoning”. In: *arXiv preprint arXiv:2406.11303* (2024).
- [29] Haotian Liu et al. “Improved baselines with visual instruction tuning”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, pp. 26296–26306.
- [30] Haotian Liu et al. “Visual instruction tuning”. In: *Advances in neural information processing systems* 36 (2024).
- [31] Yuanxin Liu et al. “Tempcompass: Do video llms really understand videos?” In: *arXiv preprint arXiv:2403.00476* (2024).
- [32] Yuanxin Liu et al. “Tempcompass: Do video llms really understand videos?” In: *arXiv preprint arXiv:2403.00476* (2024).
- [33] Muhammad Maaz et al. “Video-chatgpt: Towards detailed video understanding via large vision and language models”. In: *arXiv preprint arXiv:2306.05424* (2023).
- [34] Muhammad Maaz et al. “Videogpt+: Integrating image and video encoders for enhanced video understanding”. In: *arXiv preprint arXiv:2406.09418* (2024).
- [35] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. “Egoschema: A diagnostic benchmark for very long-form video language understanding”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 46212–46244.
- [36] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. “Egoschema: A diagnostic benchmark for very long-form video language understanding”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 46212–46244.
- [37] Munan Ning et al. “Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models”. In: *arXiv preprint arXiv:2311.16103* (2023).
- [38] Ruchit Rawal et al. “Cinepile: A long video question answering dataset and benchmark”. In: *arXiv preprint arXiv:2405.08813* (2024).
- [39] Shuhuai Ren et al. “Timechat: A time-sensitive multimodal large language model for long video understanding”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 14313–14323.
- [40] Darshana Saravanan et al. “VELOCITI: Benchmarking Video-Language Compositional Reasoning with Strict Entailment”. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 18914–18924.
- [41] Fangxun Shu et al. “Audio-visual llm for video understanding”. In: *arXiv preprint arXiv:2312.06720* (2023).
- [42] Enxin Song et al. “Moviechat: From dense token to sparse memory for long video understanding”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 18221–18232.

- [43] Gemini Team et al. “Gemini: a family of highly capable multimodal models”. In: *arXiv preprint arXiv:2312.11805* (2023).
- [44] Andong Wang et al. “Sok-bench: A situated video reasoning benchmark with aligned open-world knowledge”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 13384–13394.
- [45] Peiyu Wang et al. “Skywork r1v2: Multimodal hybrid reinforcement learning for reasoning”. In: *arXiv preprint arXiv:2504.16656* (2025).
- [46] Weihang Wang et al. “Cogvlm: Visual expert for pretrained language models”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 121475–121499.
- [47] Weihang Wang et al. “Lvbench: An extreme long video understanding benchmark”. In: *arXiv preprint arXiv:2406.08035* (2024).
- [48] Yi Wang et al. “InternVideo2.5: Empowering Video MLLMs with Long and Rich Context Modeling”. In: *arXiv preprint arXiv:2501.12386* (2025).
- [49] Bo Wu et al. “Star: A benchmark for situated reasoning in real-world videos”. In: *arXiv preprint arXiv:2405.09711* (2024).
- [50] Shengqiong Wu et al. “Next-gpt: Any-to-any multimodal llm”. In: *Forty-first International Conference on Machine Learning*. 2024.
- [51] Zhiyu Wu et al. “Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding”. In: *arXiv preprint arXiv:2412.10302* (2024).
- [52] Junbin Xiao et al. “Next-qa: Next phase of question-answering to explaining temporal actions”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 9777–9786.
- [53] Dejing Xu et al. “Video question answering via gradually refined attention over appearance and motion”. In: *Proceedings of the 25th ACM international conference on Multimedia*. 2017, pp. 1645–1653.
- [54] Jin Xu et al. *Qwen2.5-Omni Technical Report*. 2025. arXiv: 2503.20215 [cs.CL]. URL: <https://arxiv.org/abs/2503.20215>.
- [55] Lu Xu et al. “Beyond raw videos: Understanding edited videos with large multimodal model”. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 503–512.
- [56] Cilin Yan et al. “Visa: Reasoning video object segmentation via large language models”. In: *European Conference on Computer Vision*. Springer. 2024, pp. 98–115.
- [57] Zhou Yu et al. “ActivityNet-QA: A Dataset for Understanding Complex Web Videos via Question Answering”. In: *Computer Vision and Pattern Recognition* (2019).
- [58] Zhou Yu et al. “Activitynet-qa: A dataset for understanding complex web videos via question answering”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 9127–9134.
- [59] Haobo Yuan et al. “Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos”. In: *arXiv preprint arXiv:2501.04001* (2025).
- [60] Hang Zhang, Xin Li, and Lidong Bing. “Video-llama: An instruction-tuned audio-visual language model for video understanding”. In: *arXiv preprint arXiv:2306.02858* (2023).
- [61] Xiaoying Zhang et al. “Towards Self-Improving Systematic Cognition for Next-Generation Foundation MLLMs”. In: *arXiv preprint arXiv:2503.12303* (2025).
- [62] Yanzhe Zhang et al. “Llavar: Enhanced visual instruction tuning for text-rich image understanding”. In: *arXiv preprint arXiv:2306.17107* (2023).
- [63] Yipeng Zhang et al. “LLaVA-UHD v2: an MLLM Integrating High-Resolution Feature Pyramid via Hierarchical Window Transformer”. In: *arXiv preprint arXiv:2412.13871* (2024).
- [64] Deyao Zhu et al. “Minigt-4: Enhancing vision-language understanding with advanced large language models”. In: *arXiv preprint arXiv:2304.10592* (2023).
- [65] Jinguo Zhu et al. “Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models”. In: *arXiv preprint arXiv:2504.10479* (2025).

A Dataset Collection

A.1 Task Definitions

While single-video tasks assess a model’s comprehension of individual videos, CVBench evaluates cross-video relational reasoning. This critical capability enables MLLMs to advance from controlled lab settings to robust real-world deployment. CVBench introduces 15 carefully curated evaluation dimensions that collectively measure a model’s capacity for inter-video relational reasoning, each targeting a unique but complementary aspect of cross-video understanding. An overview of these tasks, their concepts, differences, and example questions is provided in Table 7.

Task 1: Cross-video anomaly detection

Concept: The comparative analysis of multi-source video data identifies anomalous segments deviating from normal patterns through global spatiotemporal feature and behavioral pattern comparisons.

Difference: Analyzes anomaly saliency through multi-video event richness and viewpoint coherence rather than single-video perspectives.

Example Questions:

- *Question:* Which video features individuals directly engaged in construction work rather than showcasing materials or interior design?
Option: A. Video 1, B. Video 2, C. Video 3, D. Video 4

Task 2: Cross-video scene recognition

Concept: Fusion of semantic features from multiple videos to locate scene segments matching target spatiotemporal attributes.

Difference: Focuses on cross-video differences rather than single-video similarities.

Example Questions:

- *Question:* “Did both videos feature a team successfully scoring a decisive goal or point during the match’s closing moments?”
Option: Yes/No.

Task 3: Multi-video key-action recognition

Concept: This category involves spatiotemporal alignment of motion trajectories across videos to identify differences in action execution, supporting action quality assessment and optimization.

Difference: It emphasizes modeling the spatiotemporal consistency of actions across multiple videos, rather than classifying actions within a single video.

Example Questions:

- *Question:* What is the key action difference between the presenter in Video 1 and the presenter in Video 2?
Option: A. In Video 1, holding products is key, while in Video 2, gesturing with hands dominates., B. In Video 1, gesturing is emphasized more, while in Video 2, interacting with products is dominant., C. In both, sitting is a primary action, but outdoor walking is unique to Video 2., D. Video 2 uses textual prompts for interactions, while Video 1 does not.

Task 4: Cross-video procedural transfer

Concept: This category focuses on cross-video dynamic knowledge transfer capabilities, this approach analyzes the knowledge in the source video (such as first aid steps or mathematical problem solving methods) and adaptively refines or supplements the execution logic of similar tasks in the target video, while overcoming the interference caused by scene differences.

Difference: Associating procedural semantics across videos, as a single video can only capture isolated processes and cannot achieve knowledge generalization or adaptation.

Example Questions:

- *Question:* What decor themes from Video 2 can be transferred to Video 4 to enhance the bar’s elegance?
Option: A. Floral arches and greenery., B. Heart motifs on walls., C. Outdoor nature-inspired pieces., D. Grand tents and fairy lights.

Task 5: Joint-video summarization

Concept: This category involves extracting semantic information from multiple videos to generate event logic and descriptions that satisfy specific query requirements.

Difference: Cross-video information fusion and semantic compression, rather than simple fragment stitching.

Example Questions:

- *Question:* How can the overall narrative of Videos 1, 2, and 3 be summarized?
Option: A. A guide on photo gallery exploration, cloud synchronization, and DSLR photography tips., B. A sequential explanation of transferring photos, securing them, and organizing them within a smartphone., C. Three independent tutorials on iPhone photo management, Google Photos app features, and manual gallery browsing., D. A step-by-step guide on the physical, cloud, and app-based photo storage and transfer across platforms.

Task 6: Multi-video attribute recognition

Concept: This category involves confirming and extracting the attributes (such as texture, color, function, relationship, characteristic and shape) of a specific target across multiple videos, capturing the same target in different states.

Difference: It focuses on cross-video attribute state transitions, rather than static attribute recognition within a single frame.

Example Questions:

- *Question:* What color predominantly features in the attire of winning players across both sporting events?
Option: "A. White.", "B. Red.", "C. Blue.", "D. Green."

Task 7: Joint-video counting

Concept: This category involves the precise identification and statistical analysis of the same target across multiple videos.

Difference: The targets are distributed across multiple videos rather than a single video, requiring cross-video unified authentication of the targets, while eliminating limitations of single-video perspectives (such as occlusion). Targets may also be abstract, such as events.

Example Questions:

- *Question:* How many monitors other than cell phones are in all the videos?
Option: A. 1, B. 2, C. 3, D. 4.

Task 8: Cross-video entity matching

Concept: This category involves making similarity judgments of entities across multiple videos with varying spatiotemporal conditions (different space and time, same space and different times, different space and different times), such as identifying criminal suspects.

Difference: It emphasizes identity association across non-overlapping fields of view, overcoming the limitations of single-view tracking.

Example Questions:

- *Question:* Which brand is consistently featured in both videos, reinforcing the high-performance gaming theme?"
Option: "A. ASUS.", "B. Corsair.", "C. Gigabyte.", "D. RGB."

Task 9: Multi-view scene understanding

Concept: This category involves integrating spatiotemporal clues from multiple perspectives to reconstruct the complete causal chain and logical relationships of event development.

Difference: It constructs a global event knowledge graph, rather than understanding event fragments from a single perspective.

Example Questions:

- *Question:* What is the common theme across the videos that ties together the movie date, romantic dinner, family daily, and cobblestone alley walk?"
Option: "A. Solitude.", "B. Companionship.", "C. Adventure.", "D. Conflict."

Task 10: Multi-video temporal reasoning

Concept: This category involves integrating multiple videos and making judgments about hidden logical relationships at specific times, such as predicting the future or sorting events.

Difference: It addresses the asynchronous timestamp issue in single videos, enabling global temporal modeling.

Example Questions:

- *Question:* Based on the interconnections, what is the correct chronological order of the criminal case depicted across the videos?
Option: A. 1-2-3-4., "B. 4-1-2-3.", "C. 1-2-4-3.", "D. 3-1-4-2.

Task 11: Joint-video spatial navigating

Concept: This category involves fusing multi-view geometric information to construct a 3D spatial semantic map, supporting cross-view path planning.

Difference: Spatial registration and joint reasoning of multi-source visual data.

Example Questions:

- *Question:* Where are the girl's two wardrobes located in the room across all videos?
Option: A. Head to the top-right corner to enter the kitchen and stove is located in the top-right corner of the kitchen, B. Head to the top-left corner to enter the kitchen and stove is located in the top-left corner of the kitchen, C. Head to the

Task 12: Video difference caption

Concept: This category involves fine-grained cross-video comparison, identifying differences across multiple videos in dimensions such as event progression and object states.

Difference: It emphasizes the identification of differences in dynamic processes, rather than static image comparison.

Example Questions:

- *Question:* What differentiates Video 1's organization method from Video 2 and Video 3?
Option: A. Uses a rotating tray., B. Combines different gadgets., C. Focuses on vertical space., D. Emphasizes utensil storage.

Task 13: Cross-video counterfactual reasoning

Concept: Based on the spatiotemporal factual foundation from multiple videos, this category constructs a causal inference chain for a virtual scenario.

Difference: Causal effect estimation based on multi-source observational data, overcoming the limitations of single-video causal inference.

Example Questions:

- *Question:* What element remains unchanged across the videos, regardless of the setting?
Option: A. Clothing style., B. Dance type., C. Footwear focus., D. Presence of music.

Task 14: Cross-video object recognition

Concept: This category involves fusing multi-view object features to address challenges such as occlusion and deformation, enabling consistent identity recognition of objects across videos.

Difference: It leverages multi-source information to compensate for the limitations of single-video perspectives.

Example Questions:

- *Question:* Which item is commonly recognized across both Videos 1 and 2 despite differences in design aesthetic?
Option: A. Bamboo plants., "B. Distressed wooden door.", "C. Umbrella stands.", "D. Metalwork tools.

Task 15: Cross-video event retrieval

Concept: This category involves rapidly locating segments across multiple videos that meet specific event elements, enabling effective filtering and selection.

Difference: The events are distributed across different videos, rather than being contained within a single video.

Example Questions:

- *Question:* Which video shows the celebration of a team after successfully leading or winning in the match?
Option: A. Video 1., "B. Video 2.", "C. Both of them.", "D. None of them.

Table 7: Overview of cross-video tasks, concepts, differences, and example questions.

You are an expert in analyzing videos. You will receive the video consisting of a series of frames. Your task is to carefully observe the details of the video frames and generate a detailed and accurate description of a video in English based on your observations.

Instructions for writing the detailed description:

1. Focus on the key visual details such as appearance, motion, scene, actions, objects involved, and interactions between elements in the video.
2. Pay attention to the type of video and the important points like the order of events, appearance and actions of people or objects, and any significant changes, events or movements.
3. Note every detail in the video and give a clear and easy to understand description.
4. Indicators like ([Video number], Timestamp: [start time]-[end time]/[total time]), for example, (Video 1, Timestamp: 00:00-00:10/01:20), mean that the video is divided into parts, and the timestamps represent the start time and end time of the part. Please note that the change in Video number means that a new video has been inputted, the description of the new video has to be restarted, and the flow between the different parts of the same video has to ensure narrative continuity and avoid abrupt breaks in the description.
5. There is no need for markdown titles or emphasis such as "***XX**" or words such as "(Frame X)" or "(Video X, Timestamp: 00:XX-00:XX/0X:XX)", emphasize the time in a uniform way, using a simple (00:XX-00:XX) will do!

Figure 6: Prompt template for generating cohesive video summaries with GPT-4o. generate a cohesive summary of the following video by merging the segment-wise captions into a comprehensive temporal description that fully encapsulates the video’s content.

A.2 QA Generation

We propose a hierarchical framework for generating cross-video question-answer pairs with three key stages for comprehensive evaluation. In the first stage, adaptive video segmentation is combined with GPT-4o to generate fine-grained spatiotemporal annotations, capturing critical elements such as object interactions and action sequences. The prompt template for generating cohesive video summaries is shown in Figure 6. In the second stage, a relationship extraction module constructs a cross-video association graph, identifying multidimensional relationships such as object evolution and event continuity. The cross-video relation extraction prompt is depicted in Figure 7. Based on 15 professional templates, the system generates diverse questions, ranging from basic matching to advanced reasoning, ensuring quality through multi-model validation and difficulty calibration. The QA pair prompt template is shown in Figure 8.

This framework integrates visual content analysis with logical reasoning, featuring a multi-layered quality control system that includes multi-model consensus validation and hallucination detection to ensure the verifiability and distinctiveness of the generated questions. Ultimately, the framework provides a comprehensive evaluation spectrum, ranging from single-hop to commonsense reasoning, with multidimensional metrics such as accuracy and reasoning depth. Figures 6, 7, and 8 present standardized prompt templates, ensuring the systematization and reproducibility of the evaluation process.

A.3 More CVBench Examples Visualization

CVBench’s 15 carefully designed subcategories of videos and questions evaluate key dimensions such as spatiotemporal correlation, entity-event association, and causal-logical reasoning. These tasks cover various challenges, from tracking object consistency across cameras to inferring causal relationships between multiple videos, detecting spatial annotation conflicts, and performing cross-modal data fusion. Each subcategory presents complex tasks that closely resemble real-world scenarios. More examples of these tasks are provided in Figure 9.

You are an expert in analyzing interconnections of multiple videos. We extracted detailed captions (per timestamp) for each video. For example, 'Video 1 : caption: \n00:00-00:10: ...'\n' corresponds to the caption generated for the first video between 00:00 and 00:10.

Your task is to generate a detailed and accurate summary of interconnections between the videos (Video 1, Video 2, etc.) as a paragraph, based on all the given video captions and videos. Make sure not to lose any important information. Use the following details to create a clear and complete narrative:

Instructions for writing the detailed summary:

1. Pay strong attention to the interconnections between the videos, including differences, correlations, matches, etc. For instance, the correlations can be that the videos are split by a complete video, just in a different order; The differences can be different types of videos, or the same objects in them doing different actions; Matches can be the same person appearing in different videos. Check if events across videos follow a logical sequence, Identify shared locations or objects across videos and note complementary angles.
2. Identify the theme of each video and use your intelligence to learn to infer hidden links.
3. Note the commonalities and contradictions between videos. The timestamps are just more detailed to allow you to understand the complete video, the focus is between the videos and not between the timestamps.
4. Do not mention that the information comes from captions.
5. Also, use your common sense to judge interconnections between the videos.

Figure 7: Cross-video relation extraction prompt. Analyze the interconnections between multiple videos by combining their captions.

I would like you to act as a quizmaster who designs questions based on provided captions of each videos and summary of the interconnections between videos that should be grounded in real-life contexts and made more relevant to everyday realities. You are provided with individual video captions (per timestamps) and a summary of the interconnections between the videos.

You are to generate relevant QA pairs by analyzing key interconnections between videos while also referring to the detailed video summary. Your questions should specifically target the interconnections between the videos. Your question should be answerable, and checkable. Guidelines for Response:

1. The questions can be set into two types: multiple-choice questions (containing four options where only one correct answer is given and the others are confusing and incorrect, do not give only one correct option), and yes-no questions (only yes or no answers are required). If descriptions are ambiguous, (e.g., timeline or content), prioritize answer uniqueness. Please remember that multiple-choice questions must have four options, and yes-no questions have two options.
2. Avoid excessive similarity between options, and be more careful to avoid answers that can be inferred directly from the options.
3. When involving multiple subjects, use subject attributes (e.g., "adult in red") for reference. Avoid abstract identifiers like "subject 1" or "person A".
4. Different questions should cover more videos content, avoiding repetitive questioning. 4. Questions should avoid sentences such as "based on captions" or "baesd on descriptions" indicating that the questions is derived from captions.
5. Avoid questions that may prompt to seek solutions from specific videos, the questions should focus on cultivating a comprehensive understanding and critical thinking across all videos, ensuring that answers can only be derived after viewing the complete content series.
6. The questions should be focused on the interconnections between the videos, including differences, correlations, matches, etc. Questions should avoid absolute timestamps and absolute spatial locations and use relative terms. For example, after drinking water, next to the TV, etc.
7. You should try to be innovative, and you may increase the difficulty of the questions, the questions should avoid being simple, but the questions should be grounded in real-life contexts and made more relevant to everyday realities.

Figure 8: QA pair prompt. Create quiz questions based on the interconnections between video captions.

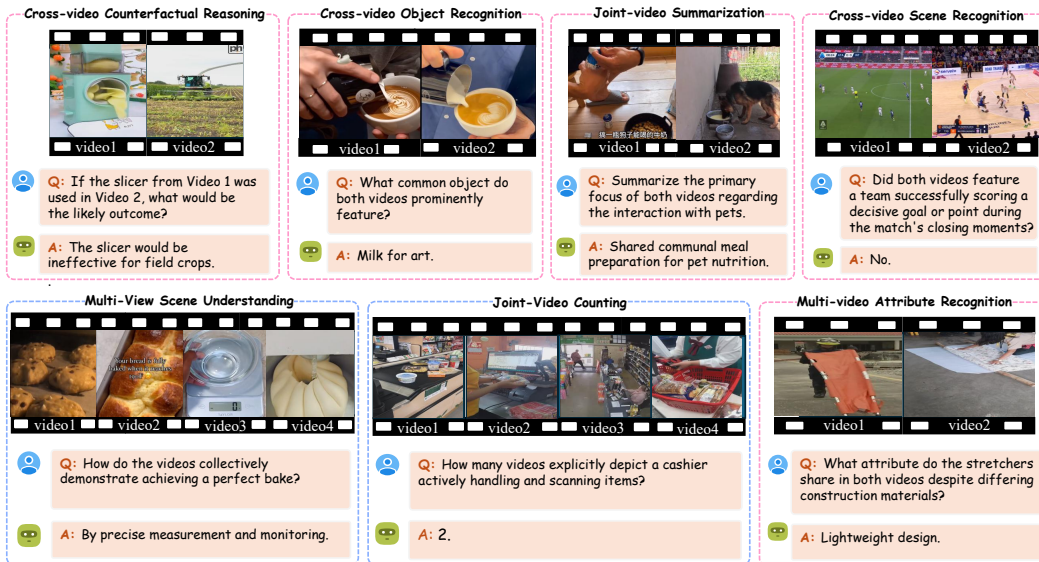


Figure 9: More examples of CVBench.