Video-MME-Logical: A Controlled Diagnostic Benchmark for Video Temporal-Logical Reasoning

Hohin Kwan[1,2,*], Hongyu Li[1,*,†], Renrui Zhang[3], Manyuan Zhang[3], Xianghao Kong[2], Anyi Rao[2], Jiahao Xie[1,‡], Si Liu[1]

[1]Beihang University [2]HKUST [3]CUHK

Project page: https://mrakas.github.io/video-mme-logical/

[*]Equal contribution. [†]Project Leader. [‡]Corresponding Author.

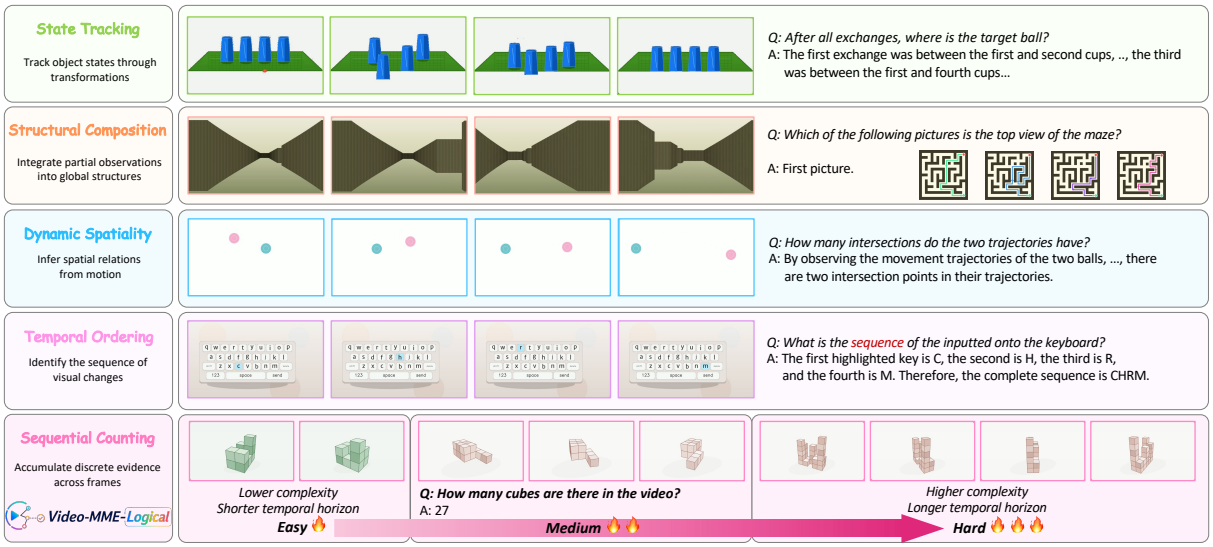


Figure 1: VIDEO-MME-LOGICAL is a controllable benchmark for video temporal-logical reasoning with 25 tasks. It evaluates whether models can reason over dynamic visual worlds through State Tracking, Structural Composition, Dynamic Spatiality, Temporal Ordering and Sequential Counting, spanning final-answer tasks, intermediate-state diagnostics, and difficulty-controlled settings.

Abstract

Recent interest in multimodal large language models (MLLMs) raises a central question: can they reason over dynamic visual evidence rather than merely recognize objects or events in individual frames? This ability, which we refer to as *video temporal-logical reasoning*, requires models to maintain, update, and compose evidence as visual states evolve across frames. Existing video benchmarks often conflate this capability with scene complexity, static recognition, or uncontrolled temporal variation. To isolate this capability, we introduce VIDEO-MME-LOGICAL, a controlled benchmark organized around five temporal-logical operations: state tracking, sequential counting, temporal ordering, dynamic spatiality, and structural composition. The benchmark contains 25 fine-grained task categories generated with controlled object states, transitions, temporal dependencies, and logical compositions. It enables difficulty-controlled final-answer evaluation by varying temporal horizon and reasoning complexity, and supports intermediate-state diagnostics by verifying whether models recover the required logical reasoning trace before producing the final answer. Experiments with state-of-the-art MLLMs reveal a substantial human-model gap, especially as temporal-logical complexity increases. Supervised fine-tuning on up to 500K generated samples improves performance but remains insufficient to close the reasoning gap, positioning VIDEO-MME-LOGICAL as a scalable testbed for analyzing and improving temporal-logical reasoning in MLLMs.

1 Introduction

Human visual cognition inherently relies on temporal-logical reasoning—the capacity to synthesize shifting visual sequences by maintaining, updating, and composing evidence as an event unfolds over time (Kahneman et al., 1992; Pylyshyn, 2001; Cavanagh, 2011). Currently, the quest to replicate this human ability is dominated by multimodal large language models (MLLMs) (Google DeepMind, 2025; OpenAI, 2026; Liu et al., 2024a;

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 8900–8919
October 24-29, 2026 ©2026 Association for Computational Linguistics

Team, 2026; Team et al., 2026; Li et al., 2025, 2026b), which have achieved impressive performance across standard video understanding benchmarks (Fu et al., 2025; Liu et al., 2026c; Fu et al., 2026; Liu et al., 2024b; Cheng et al., 2025; Liu et al., 2026a,b; Li et al., 2026c; Xiao et al., 2026; Li et al., 2026a; Xu et al., 2026). Despite their remarkable advancements, a critical conceptual gap remains: merely aggregating multi-frame inputs is not equivalent to executing logical reasoning over time. For instance, while a young child can effortlessly deduce the position of a hidden ball in a shell game by tracking motion through occlusion, contemporary MLLMs struggle with this foundational logic. Since existing benchmarks often conflate basic frame-stitching with robust temporal inference, the true logical faculties of current video models remain heavily overestimated.

This gap persists primarily because existing benchmarks fail to provide a controlled diagnostic setting that isolates temporal-logical reasoning from general temporal understanding. We identify three key limitations in existing evaluations: (1) The **categories** of temporal-logical reasoning remain under-specified. Many existing benchmarks are typically organized by data source, scene type, event category, or action class, rather than by temporal-logic operations, making it difficult to attribute model errors to specific reasoning capabilities. (2) Existing temporal benchmarks struggle to provide interpretable **difficulty levels**, since difficulty in natural videos often co-varies with scene complexity, annotation language, and dataset bias, rather than being controlled by temporal dependencies and logical complexity. (3) Most benchmarks evaluate only final answers and lack **verifiable intermediate states**, making it difficult to determine whether a model truly performs temporal-logical reasoning or merely relies on local cues.

To address these limitations, we propose VIDEO-MME-LOGICAL, a controlled diagnostic benchmark for video temporal-logical reasoning. Our design targets the three gaps above. **First**, to make temporal-logical reasoning explicit rather than loosely defined, we organize the benchmark around five temporal-logical operations: State Tracking, Sequential Counting, Temporal Ordering, Dynamic Spatiality, and Structural Composition. These operations specify what information must be maintained, accumulated, ordered, spatially inferred, or composed over time, turning a broad notion of video reasoning into an operation-centric evaluation framework. **Second**, to make difficulty interpretable and controllable, we construct 25 fine-grained task categories through procedural generation, which allows us to precisely control object states, state transitions, temporal dependencies, and logical compositions. Each task category is further divided into three difficulty levels according to temporal horizon and reasoning complexity. **Third**, to move beyond final-answer-only evaluation, we introduce VIDEO-MME-LOGICAL-S, an intermediate-state diagnostic subset whose intermediate evidence can be described and verified.

To study the effect of data scaling on video temporal-logical reasoning, we generate 500K procedurally created training samples and fine-tune Qwen3-VL-8B (Qwen Team, 2025b) with different data scales. The results show that supervised fine-tuning (SFT) improves performance, reaching 40% accuracy at 375K samples, but further scaling does not provide clear additional gains. A substantial gap to human experts remains, suggesting that data scaling alone is insufficient and that current models may still struggle with long temporal horizons and complex logical structures. We hope VIDEO-MME-LOGICAL, with its scalable training data and diagnostic evaluation setting, can support future research on video temporal-logical reasoning.

In summary, our contributions are as follows:

- We propose a taxonomy of temporal-logical reasoning and instantiate it as VIDEO-MME-LOGICAL, a controlled benchmark with five operation categories and 25 fine-grained tasks.
- We design difficulty-controlled evaluation settings and intermediate-state diagnostics, revealing a substantial human-model gap in temporal-logical reasoning.
- We build a large-scale training set and conduct scaling studies, showing that more data improves performance but remains insufficient to close the reasoning gap.

2 Related Work

General Video Understanding Benchmarks. Video-language benchmarks have evolved from short-clip question answering and action-level understanding toward broad evaluations of general video comprehension. Recent suites such as Video-MME (Fu et al., 2025), MLVU (Zhou et al., 2025), ALLVB (Tan et al., 2025), LongVideoBench (Wu et al., 2024), LVBench (Wang et al., 2025a),

Benchmark	#Tasks	#Videos	#Train	#Test	Control.	Difficulty.	Intermediate.
TOMATO (Shangguan et al., 2025)	6	1,417	0	1,417	×	×	×
TempCompass (Liu et al., 2024b)	5	410	0	410	×	×	×
ReXTime (Chen et al., 2024)	3	12,759	9,695	3,064	×	×	×
V-STaR (Cheng et al., 2025)	2	2,094	0	2,094	×	×	×
VIDEO-MME-LOGICAL	25	503,750	500,000	3,750	✓	✓	✓

Table 1: Comparison with recent video temporal reasoning benchmarks. Control. indicates whether the benchmark can programmatically control each element in the video; Difficulty. indicates whether it provides controlled difficulty settings; Intermediate. indicates whether it supports verifiable intermediate-state evaluation.

CinePile (Rawal et al., 2024), VideoEspresso (Han et al., 2025) and MovieChat (Song et al., 2024) expand evaluation along multiple general-purpose axes, including video domains, durations, task formats, and open-ended instruction following. Other benchmarks improve realism or diagnostic coverage by focusing on egocentric activities, temporal relations, object interactions, and perception-oriented video understanding, as in Ego4D (Grauman et al., 2022), EgoSchema (Mangalam et al., 2023), MVBench (Li et al., 2024a), NExT-QA (Xiao et al., 2021), and the Perception Test (Patraucean et al., 2023). Together, these benchmarks are valuable precisely because they approximate diverse natural video-use scenarios and measure broad multimodal competence.

This breadth, however, leaves a more specific question under-specified: whether models can perform well-defined temporal-logical operations under controlled visual conditions. General video benchmarks are intentionally heterogeneous, and thus rarely organize questions around a fixed set of logical categories or difficulty-controlled traces. In contrast, VIDEO-MME-LOGICAL uses synthetic, controllable videos to isolate temporal-logical reasoning across multiple dimensions.

Video Temporal Reasoning Benchmarks. Recent benchmarks have moved beyond broad video comprehension toward targeted evaluations of temporal and spatiotemporal reasoning. VITATECS (Li et al., 2024b) probes fine-grained temporal concepts through counterfactual caption discrimination; ViLMA (Kesen et al., 2023) evaluates fine-grained linguistic and temporal grounding through controlled counterfactual tests; TemporalVQA (Imam et al., 2025) studies temporal order and time-lapse reasoning from image pairs; and TempCompass (Liu et al., 2024b) covers action, speed, direction, attribute change, and event order. Other benchmarks emphasize temporally grounded evidence: TOMATO (Shangguan et al., 2025) evaluates whether models can use ordered observations from continuous frames, ReXTime (Chen et al., 2024) links questions and answers across video segments, V-STaR (Cheng et al., 2025) studies chains that connect what, when, and where information, and VideoReasonBench (Liu et al., 2026c) evaluates vision-centric complex video reasoning. Together, these benchmarks show that current models still struggle with temporal cues, event ordering, cross-time relations, and spatiotemporal grounding.

Recent efforts have begun to examine video temporal-logical reasoning. Liu and Lee (2026) focuses on shell-game-style tracking, offering a controlled but task-specific test of logical reasoning, whereas Wang et al. (2026) studies video-generation-oriented reasoning at scale and emphasizes broad video-reasoning behaviors. These studies reveal important failures in temporal tracking and video reasoning, but they do not systematically distinguish which temporal-logical operation an MLLM must execute to solve a problem, such as maintaining hidden states, accumulating evidence, ordering visual changes, inferring dynamic spatial relations, or composing partial observations. As a result, final-answer accuracy alone makes it difficult to determine whether a model truly performs temporal-logical reasoning. VIDEO-MME-LOGICAL addresses this gap by providing a comprehensive MLLM benchmark organized around five temporal-logical operations, with controllable difficulty and verifiable intermediate states; Table 1 summarizes these distinctions.

3 VIDEO-MME-LOGICAL Benchmark

This section presents the construction of VIDEO-MME-LOGICAL, including its task architecture, dataset statistics, and data curation pipeline.

3.1 Temporal-Logical Reasoning Taxonomy

We organize VIDEO-MME-LOGICAL around five foundational temporal-logical reasoning abilities.

These categories are motivated by the observation that solving a video reasoning problem requires more than recognizing visible objects or events: a model must decide what information should be remembered, how it should be updated, and how multiple temporal states should be composed to support inference. Specifically, **State Tracking** refers to maintaining latent or hidden object states across visual transformations, especially when the target state is no longer directly visible. **Sequential Counting** refers to accumulating discrete evidence over time, where the answer depends on a temporal history rather than any individual frame. **Temporal Ordering** refers to identifying the order of state changes, revealed symbols, or event sequences that determine the final outcome. **Dynamic Spatiality** refers to inferring geometric and dynamic relations from continuous movement, including trajectories, rotations, intersections, and relative speeds. **Structural Composition** refers to composing spatial structures across viewpoints, occlusions, and partial observations. To operationalize these abilities, each category is implemented as a set of parameterized task generators that produce videos, questions, answers, and, when applicable, verifiable intermediate reasoning traces. Representative examples of these abilities are shown in Fig. 3.

3.2 Data Statistics

VIDEO-MME-LOGICAL contains 503,750 videos in total, including 500K training videos and 3,750 test videos. The test set covers 25 fine-grained task categories organized into five temporal-logical operation groups. Each category is evaluated in three difficulty levels (*i.e.*, *easy*, *medium*, and *hard*), which are defined by category-specific temporal horizons and reasoning complexity factors. Within the 25-category task space, an 8-category intermediate-state diagnostic subset provides intermediate-state annotations to verify whether the models maintain the correct temporal evidence trace rather than only producing the final answer. Detailed task definitions and examples are provided in Appendix A.

3.3 Programmatic Generation Pipeline

Inspired by synthetic benchmarks such as CLEVR (Johnson et al., 2017; Yi et al., 2020; Zhuo et al., 2025), VIDEO-MME-LOGICAL uses programmatic generation to decouple temporal-logical reasoning from visual noise and annotation ambiguity in natural videos. This design enables reproducible task generation, controllable difficulty, and verifiable intermediate evidence.

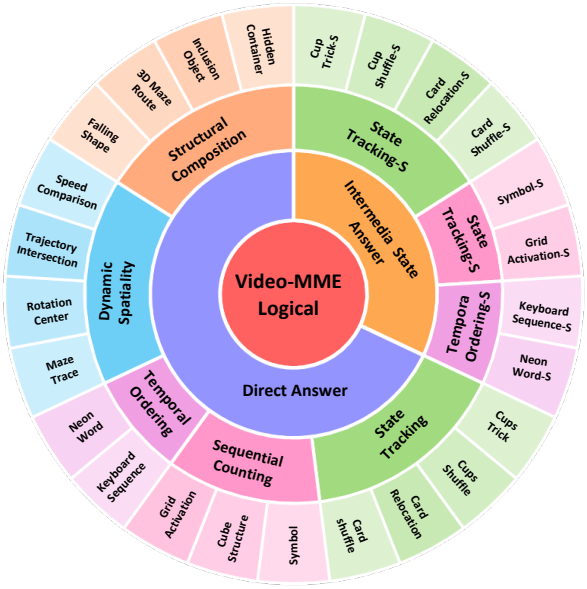


Figure 2: Taxonomy of VIDEO-MME-LOGICAL. The inner ring separates direct-answer tasks from the intermediate-state diagnostic subset, while the outer rings group fine-grained task categories under five temporal-logical operation groups.

Task Program Design. Each task category is implemented as an executable program composed of four core components: (1) *temporal transition*, which defines how the scene evolves over time, including the order in which objects move, swap, appear, or disappear; (2) *scene configuration*, which specifies the objects, visual attributes, spatial layout, and optional distractors in the video; (3) *metadata construction*, which records the complete temporal evidence trace for video rendering, question generation, answer computation, difficulty control, and intermediate-state supervision; and (4) *video rendering*, which converts the recorded metadata into an MP4 video at 30 FPS.

Difficulty Control. VIDEO-MME-LOGICAL defines three difficulty levels, *i.e.*, *easy*, *medium*, and *hard*, by increasing both temporal horizon and reasoning complexity. In cup-tracking tasks, for example, video duration instantiates temporal horizon, increasing from 10 seconds in *easy* videos to 20 seconds in *hard* videos, while the number of swaps instantiates reasoning complexity, increasing from 4 to 8 swaps.

Intermediate-State Annotation. For task categories with intermediate-state diagnostics, the program metadata provides supervision beyond the final answer. For example, in cup-tracking tasks, the metadata records the full state trajectory, includ-

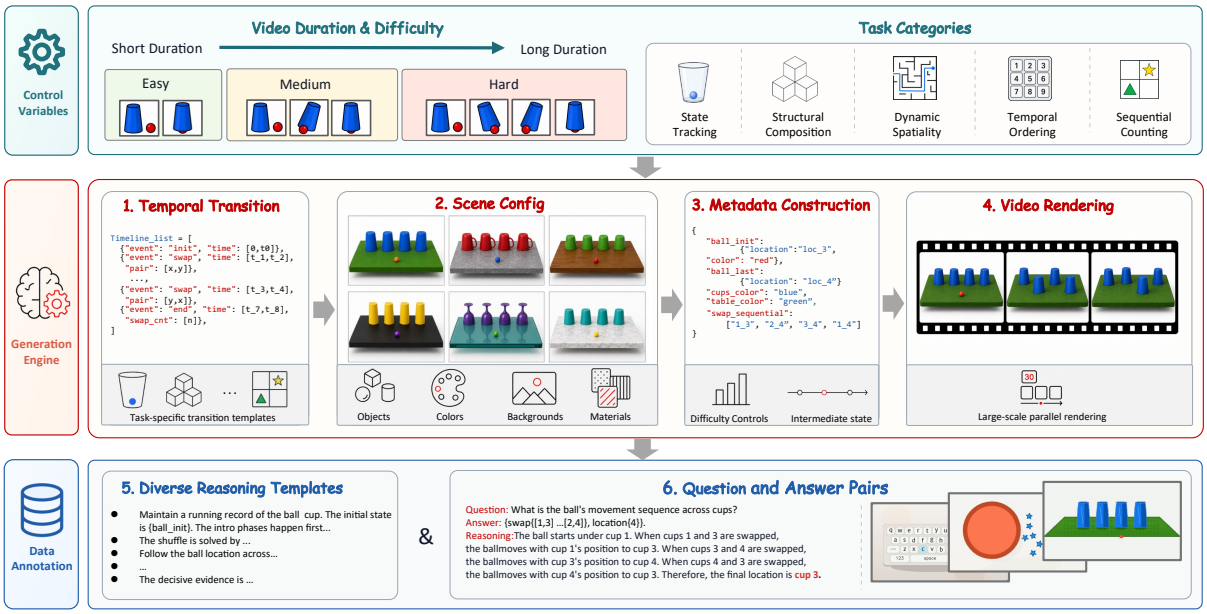


Figure 3: VIDEO-MME-LOGICAL combines controllable video generation, structured metadata, and diversified reasoning templates to build a 25-task temporal logical reasoning benchmark.

ing the initial ball position, each cup-swap pair with its timestamp, and the final ball position. These annotations allow us to evaluate whether a model follows the correct temporal evidence trace during reasoning, rather than merely producing the correct final answer.

QA and Training Annotation. We generate QA pairs directly from the program metadata using task-specific templates, ensuring that each question is paired with an exact, automatically computed answer. For the training split, we further generate reasoning traces by conditioning proprietary models (i.e.,GPT-5.4) on the video, metadata, question, and answer. We then manually curate ten reasoning templates for each task category to keep the generated reasoning aligned with the underlying temporal evidence trace while preserving controlled linguistic variation.

Metric Design. Since VIDEO-MME-LOGICAL provides program-derived ground-truth answers and states, we use exact-match accuracy as the main evaluation metric. *Multiple-choice questions.* We extract the selected option from the unified answer tag and count it as correct only if it exactly matches the ground-truth option. *Fill-in questions.* We extract the tagged answer, and count it as correct only if it exactly matches the ground-truth answer. *Intermediate-state questions.* In VIDEO-MME-LOGICAL-S, each example additionally contains program-recorded intermediate states for process verification. Models are required to output structured intermediate information in the same answer tag. We canonicalize the predicted structure and count it as correct only if it exactly matches the corresponding program-recorded state. Overall accuracy is computed as $\mathrm{Acc} = N_{\mathrm{correct}}/N_{\mathrm{total}}$.

4 Experiments

We study VIDEO-MME-LOGICAL from three aspects: final-answer evaluation, intermediate-state evaluation, and supervised fine-tuning scaling.

4.1 Implementation Details

Benchmark Models. We evaluate a diverse set of video-capable MLLMs on VIDEO-MME-LOGICAL, grouped into three categories according to model type. First, *open-source instruct models* include Qwen2.5-VL-3B/7B/72B (Qwen Team, 2025a), Qwen3-VL and Qwen3-Omni variants (Qwen Team, 2025b), InternVL3.5 variants (Wang et al., 2025b), LLaVA-Video-7B/72B-Qwen2 (Zhang et al., 2025), and KimiVL-16B-A3B-Instruct (Kimi Team et al., 2025). Second, *open-source thinking models* include Qwen3-VL, Qwen3-Omni, and KimiVL thinking variants, which are designed to produce explicit reasoning traces before answering. Third, *proprietary models* include GPT-5.4 (OpenAI, 2026) and gemini-3.1 Pro (Google DeepMind, 2025). All evaluations are conducted under zero-shot settings, with videos sampled at 2 FPS.

Models	Overall	Avg.			State.			Count.			Order.			Spat.			Struct.		
		E	M	H	E	M	H	E	M	H	E	M	H	E	M	H	E	M	H
Human Level	95.9	98.4	95.9	93.4	99.2	96.5	93.5	97.6	95.2	93.2	99.0	95.5	93.5	98.5	97.0	93.5	97.0	95.0	93.5
Open-source Instruct Models																			
Qwen3-VL-8B-Instruct	11.9	13.4	12.8	9.6	11.0	8.8	4.8	9.6	0.0	0.4	18.0	20.5	19.5	12.0	14.5	12.5	16.5	20.0	11.0
Qwen3-VL-30B-A3B-Instruct	11.8	14.5	12.4	8.7	11.8	8.8	**4.8**	11.6	0.4	0.0	17.0	20.5	14.0	18.5	20.5	12.5	13.5	12.0	12.0
Qwen3-Omni-30B-A3B-Instruct	5.8	6.3	6.1	4.9	3.8	4.0	1.0	2.8	0.4	0.0	6.5	9.0	10.5	9.5	6.5	3.5	9.0	10.5	9.5
Qwen2.5-VL-3B-Instruct	1.9	3.1	1.5	1.3	1.0	1.0	0.0	4.8	0.8	0.0	0.0	0.5	0.5	0.0	0.0	1.5	9.5	5.0	4.5
Qwen2.5-VL-7B-Instruct	7.4	10.3	7.7	4.3	9.5	6.8	2.2	7.2	0.0	0.4	0.0	0.0	0.0	23.0	26.0	10.5	12.0	5.5	8.5
Qwen2.5-VL-72B-Instruct	12.5	15.2	13.1	9.1	7.8	5.0	4.0	10.8	1.6	0.0	19.5	20.5	15.5	19.0	22.5	12.5	19.0	16.0	13.5
InternVL3.5-8B-Instruct	12.1	13.8	13.5	8.9	7.2	4.5	2.2	8.0	4.0	0.8	16.5	18.5	12.0	18.0	20.0	16.0	19.5	20.5	13.5
InternVL3.5-30B-A3B-Instruct	8.7	9.5	9.7	7.0	9.8	7.5	2.5	5.2	4.4	3.6	3.5	1.0	3.0	15.0	18.5	14.0	14.0	17.0	12.0
LLaVA-Video-7B-Qwen2	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LLaVA-Video-72B-Qwen2	2.4	4.7	1.9	0.7	4.8	5.2	2.8	13.2	1.2	0.8	0.0	0.0	0.0	5.5	3.0	0.0	0.0	0.0	0.0
KimiVL-16B-A3B-Instruct	2.9	5.4	2.2	0.9	7.0	4.8	1.5	11.6	0.0	0.0	2.5	3.0	1.5	1.0	0.0	0.0	5.0	3.5	1.5
Open-source Thinking Models																			
Qwen3-VL-8B-Think	6.6	7.9	5.8	6.0	0.8	0.8	1.8	4.4	3.6	1.6	5.5	4.0	6.0	19.5	15.5	13.0	9.5	5.0	7.5
Qwen3-VL-30B-A3B-Think	10.3	16.0	8.7	6.1	4.0	4.8	2.8	21.6	7.2	3.6	28.5	23.0	20.5	16.0	5.5	1.5	10.0	3.0	2.0
Qwen3-Omni-30B-A3B-Think	6.2	6.6	5.2	6.8	1.8	1.2	4.2	2.8	2.0	1.2	9.0	13.0	15.5	8.5	7.0	3.5	11.0	2.5	9.5
KimiVL-16B-A3B-Think	7.6	10.2	6.8	5.8	7.0	5.0	3.8	4.0	0.4	0.0	5.0	5.0	5.0	20.0	13.5	10.0	15.0	10.0	10.0
Proprietary Models																			
GPT-5.4	22.7	31.7	20.3	16.1	8.8	4.5	3.8	56.4	16.4	5.6	38.0	37.0	32.0	29.0	22.5	24.0	26.5	21.0	15.0
Gemini-3.1 Pro	**28.6**	**33.1**	**24.1**	**20.6**	**14.0**	**9.8**	3.0	**58.4**	**36.0**	**30.8**	**38.5**	**34.0**	**32.5**	**35.5**	**27.5**	**35.5**	**32.0**	**24.5**	**16.5**

Table 2: Main results on VIDEO-MME-LOGICAL. E/M/H denote easy, medium, and hard settings; State., Count., Order., Spat., and Struct. denote the five reasoning dimensions.

Human-Level Performance. We further estimate human performance on a sampled subset of VIDEO-MME-LOGICAL. Human evaluators independently answer each question under the same visual input setting as models, and their predictions are evaluated using the same metrics. This provides an approximate reference point for interpreting model performance and assessing whether the benchmark remains solvable for human annotators.

4.2 Final-Answer Evaluation

Table 2 reports final-answer accuracy over all 25 task categories and three difficulty levels. Overall, VIDEO-MME-LOGICAL remains highly challenging for current MLLMs. Human performance reaches 95.9% overall accuracy, whereas the strongest evaluated model, gemini-3.1 Pro, achieves only 28.6%. These results show that a large gap still remains between human performance and current models.

Thinking does not necessarily improve temporal-logical reasoning. Although KimiVL-16B-A3B improves from 2.9% in the instruct setting to 7.6% in the thinking setting, this trend is not consistent across model families. Qwen3-VL-8B drops from 11.9% to 6.6% when switching from instruct to thinking, and Qwen3-VL-30B-A3B drops from 11.8% to 10.3%. This suggests that generating a reasoning trace is not sufficient by itself; the trace must be grounded in the correct visual evidence.

Controlled difficulty exposes different degradation patterns. As difficulty increases, GPT-5.4 drops from 31.7% on easy tasks to 16.1% on hard tasks, a degradation of 15.6%, while gemini-3.1 Pro drops from 33.1% to 20.6%, a degradation of 12.5%. The smaller degradation of gemini-3.1 Pro suggests stronger robustness on harder examples. However, its hard-task performance still remains far below human level, indicating that longer temporal horizons and higher reasoning complexity remain challenging for current models.

4.3 Intermediate-State Evaluation

Table 3 evaluates VIDEO-MME-LOGICAL-S. This subset tests whether models can generate accurate intermediate logical reasoning rather than only produce final answers.

Final-answer accuracy can hide intermediate-state failures. Although gemini-3.1 Pro outperforms GPT-5.4 on easy final-answer State, Count., and Order. tasks (14.0% vs. 8.8%, 58.4% vs. 56.4%, and 38.5% vs. 38.0%), GPT-5.4 is stronger on the easy step-subset categories: 63.0% vs. 35.0% on Count.-S and 27.0% vs. 13.0% on Order.-S. This mismatch shows that a higher final-answer score does not necessarily imply that the model can generate the process for the final answer.

Proprietary models lead on intermediate-state evaluation. GPT-5.4 and gemini-3.1 Pro achieve 17.4% and 10.8% on the intermediate-state eval-

Models	Overall	Avg.			State.-S			Count.-S			Order.-S		
		E	M	H	E	M	H	E	M	H	E	M	H
Human Level	96.1	98.5	96.0	93.8	99.0	96.5	93.5	97.0	95.0	94.0	99.0	96.0	94.0
Open-source Instruct Models													
Qwen3-VL-8B-Instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen3-VL-30B-A3B-Instruct	0.1	0.3	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
Qwen3-Omni-30B-A3B-Instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen2.5-VL-3B-Instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen2.5-VL-7B-Instruct	0.1	0.2	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen2.5-VL-72B-Instruct	0.1	0.2	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
InternVL3.5-8B-Instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
InternVL3.5-30B-A3B-Instruct	0.1	0.3	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LLaVA-Video-7B-Qwen2	0.1	0.2	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LLaVA-Video-72B-Qwen2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
KimiVL-16B-A3B-Instruct	0.1	0.2	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Open-source Thinking Models													
Qwen3-VL-8B-Think	0.6	1.3	0.3	0.0	0.0	0.0	0.0	1.0	0.0	0.0	3.0	1.0	0.0
Qwen3-VL-30B-A3B-Think	3.6	9.0	1.3	0.3	0.0	0.0	0.0	11.0	0.0	0.0	16.0	4.0	1.0
Qwen3-Omni-30B-A3B-Think	1.2	1.7	1.3	0.7	0.0	0.0	0.0	0.0	0.0	0.0	5.0	4.0	2.0
KimiVL-16B-A3B-Think	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Proprietary Models													
GPT-5.4	**17.4**	**30.8**	**13.7**	**7.7**	2.5	0.0	0.0	**63.0**	**20.0**	9.0	**27.0**	**21.0**	**14.0**
Gemini-3.1 Pro	10.8	18.7	8.5	5.2	**8.0**	**3.5**	**0.5**	35.0	12.0	**10.0**	13.0	10.0	5.0

Table 3: Main results on VIDEO-MME-LOGICAL-S.

Models	Overall	Avg.			State.			Count.			Order.			Spat.			Struct.		
		E	M	H	E	M	H	E	M	H	E	M	H	E	M	H	E	M	H
Base Model																			
Qwen3-VL-8B-Instruct	11.9	13.4	12.8	9.6	11.0	8.8	4.8	9.6	0.0	0.4	18.0	20.5	19.5	12.0	14.5	12.5	16.5	20.0	11.0
SFT Scaling – Thinking																			
Ours-25K-Thinking	36.8	46.7	**35.2**	**28.4**	37.8	**13.8**	6.5	**43.6**	**28.4**	**20.4**	50.5	43.5	34.5	60.0	59.0	53.0	41.5	31.5	27.5
Ours-125K-Thinking	36.9	52.5	32.6	25.6	39.0	9.8	4.2	42.4	12.0	1.6	54.5	43.0	33.0	75.0	66.0	60.0	**51.5**	32.5	29.0
Ours-250K-Thinking	36.9	53.6	33.2	24.0	**40.8**	7.5	3.0	43.2	8.4	1.6	57.5	**51.0**	31.5	83.0	62.0	56.0	43.5	37.0	28.0
Ours-375K-Thinking	**39.2**	**54.8**	34.7	28.1	40.2	7.5	4.2	41.6	6.0	3.2	**60.5**	49.0	**37.0**	82.5	**73.0**	**69.5**	49.0	**38.0**	26.5
Ours-500K-Thinking	37.7	52.0	34.0	27.0	33.5	6.0	3.5	40.0	9.6	2.4	57.0	48.0	33.0	**87.0**	71.0	66.0	42.5	35.5	**30.0**
SFT Scaling – Instruct																			
Ours-25K-Instruct	23.9	33.2	21.2	17.3	23.5	7.0	4.5	32.0	4.4	0.8	19.5	10.0	6.0	60.5	51.5	48.0	30.5	33.0	27.0
Ours-125K-Instruct	28.9	40.6	26.7	19.4	24.8	6.8	5.0	29.2	4.8	1.2	43.0	30.5	13.5	74.5	63.5	55.5	31.5	28.0	22.0
Ours-250K-Instruct	30.4	41.5	26.9	22.6	28.8	10.8	**7.2**	30.0	1.6	0.0	44.5	37.5	30.5	75.0	63.0	55.5	29.5	21.5	20.0
Ours-375K-Instruct	30.6	41.9	27.3	22.6	27.5	6.8	6.8	30.4	2.8	0.4	45.0	35.5	29.5	77.5	69.0	57.0	29.0	22.5	19.5
Ours-500K-Instruct	27.3	38.5	25.1	18.2	27.2	9.2	6.5	32.0	4.0	1.2	22.5	18.0	3.0	80.5	69.5	54.5	30.5	25.0	26.0

Table 4: SFT scaling results on VIDEO-MME-LOGICAL.

uation subset, respectively, while the strongest open-source model, Qwen3-VL-30B-A3B-Think, reaches only 3.6%. GPT-5.4 is therefore 4.8× higher than the strongest open-source model, and gemini-3.1 Pro is 3.0× higher. This large gap highlights the value of VIDEO-MME-LOGICAL-S.

4.4 SFT Scaling Analysis

Table 4 studies SFT using Qwen3-VL-8B as the base model. We sample training data from the 500K training split at five sizes: 25K, 125K, 250K, 375K, and 500K, with balanced proportions across task categories. *Ours-*K-Instruct* denotes models trained with answer supervision, while *Ours-*K-Thinking* denotes models trained with reasoning trajectories. Since the training data is constructed from easy-level instances, evaluation on medium and hard settings further tests whether the learned temporal-logical reasoning behaviors can generalize to longer duration and higher complexity. We provide training details in Appendix B.

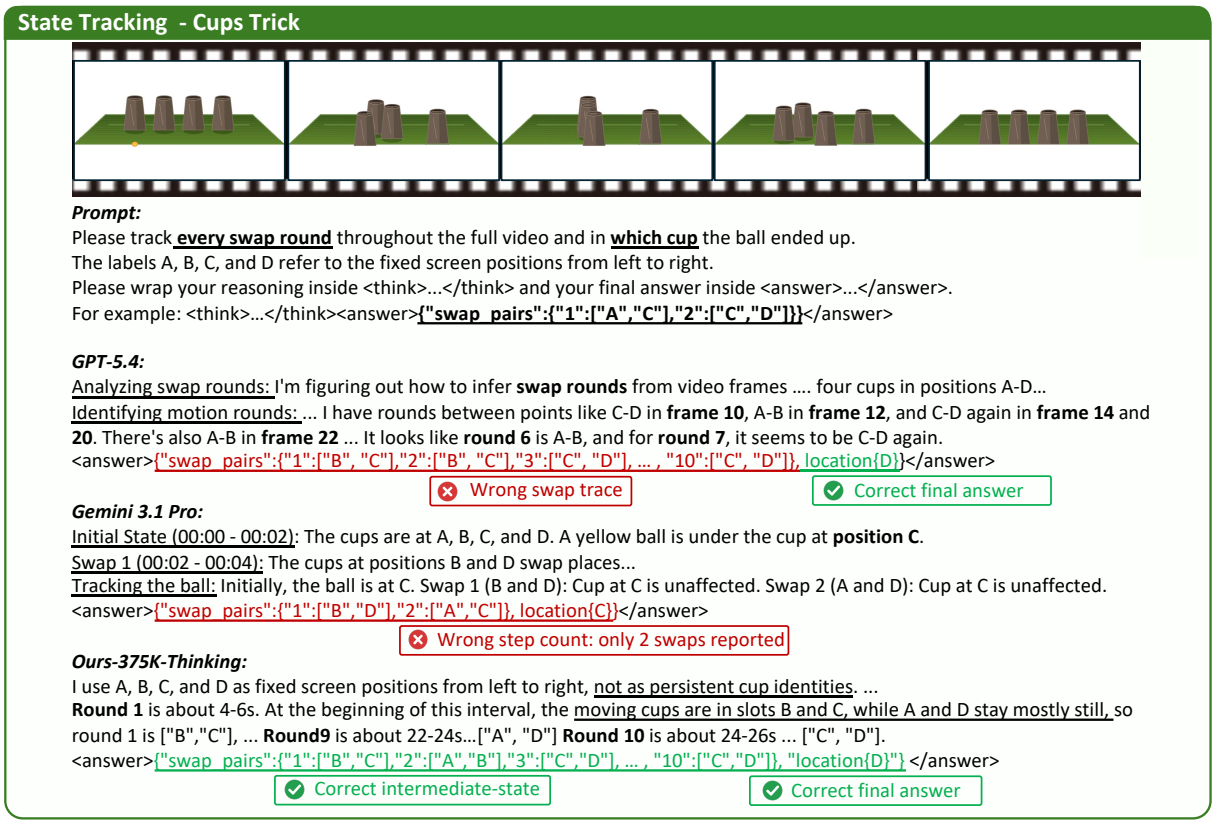


Figure 4: A qualitative example of intermediate-state evaluation on a state-tracking task.

Data scaling brings clear but saturating gains. Increasing the training size to 375K improves overall accuracy to 39.2%, the best result in the scaling table. However, using the full 500K training set reduces performance to 37.7%. This trend suggests that the current model and SFT recipe can learn useful temporal-logical reasoning behaviors, but simply adding more supervised data does not bring sustained improvement.

Generalization to harder settings remains limited. Ours-375K-Thinking reaches 54.8% Avg. E, indicating that the model learns transferable temporal-logical behaviors from easy-level training data. However, medium and hard performance do not improve consistently with scale: Ours-25K-Thinking is slightly higher than Ours-375K-Thinking by 0.5% on Avg. M and 0.3% on Avg. H. These results suggest that easy-level supervision can generalize to some more complex settings, but it does not provide stable gains under longer temporal horizons and higher reasoning complexity.

Visualization Analysis. Figure 4 illustrates the role of VIDEO-MME-LOGICAL-S in diagnosing intermediate-state reasoning, using state tracking as an example. GPT-5.4 predicts the correct final location but produces an incorrect swap trace, showing that final-answer accuracy can hide process-level errors. Gemini 3.1 Pro reports an incorrect number of reasoning steps, indicating incomplete temporal state updates. In contrast, Ours-375K-Thinking recovers both the intermediate swap sequence and the final answer. This demonstrates that verifiable intermediate states can distinguish genuine temporal-logical reasoning from superficially correct final answers but flawed intermediate reasoning.

5 Conclusion

We introduced VIDEO-MME-LOGICAL, a controlled diagnostic benchmark for video temporal-logical reasoning. By organizing 25 tasks around five temporal-logical operations and providing difficulty-controlled settings with intermediate-state diagnostics, our benchmark isolates whether MLLMs can maintain, update, and compose visual evidence over time. Experiments reveal a substantial human-model gap, especially under longer temporal horizons, higher reasoning complexity, and process-level evaluation. Our SFT scaling study further shows that supervised data improves performance but quickly saturates, suggesting that naive supervised scaling in our setting is insufficient for robust temporal-logical reason-

ing. We hope VIDEO-MME-LOGICAL will support more precise diagnosis of video reasoning failures and encourage future models with stronger temporal-logical reasoning capabilities.

Limitations

This work has several limitations. First, VIDEO-MME-LOGICAL is built from procedurally generated videos. This design enables scalable data generation, controllable difficulty, and verifiable intermediate states, but it also introduces a gap from natural videos in visual appearance, scene diversity, and real-world ambiguity. However, natural videos are difficult to annotate at a large scale with reliable temporal states, exact answers, and process-level supervision, which motivates our controlled diagnostic setting. Second, our scaling experiments are conducted with an 8B MLLM. Larger models, such as 72B-scale MLLMs, may exhibit different scaling behaviors, especially in long-horizon state maintenance and compositional reasoning, and we leave a broader model-scale study to future work. Third, our intermediate-state evaluation relies on task-specific structured outputs and exact-match scoring. While this makes the evaluation reproducible and directly tied to program-derived ground truth, it may penalize semantically valid reasoning traces that use different surface forms or alternative but equivalent descriptions.

References

Patrick Cavanagh. 2011. Visual cognition. *Vision research*, 51(13):1538–1551.

Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Yu-Chiang Frank Wang. 2024. Rextime: A benchmark suite for reasoning-across-time in videos. In *Advances in Neural Information Processing Systems*, volume 37.

Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. 2025. V-star: Benchmarking video-llms on video spatio-temporal reasoning. *arXiv preprint arXiv:2503.11495*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24108–24118.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, Yongkang Xie, Xiawu Zheng, Xue Yang, Haoyu Cao, Yunsheng Wu, Ziwei Liu, Xing Sun, Caifeng Shan, and Ran He. 2026. Video-mme-v2: Towards the next stage in benchmarks for comprehensive video understanding. *Preprint*, arXiv:2604.05015.

Google DeepMind. 2025. Gemini 3 pro model card. https://deepmind.google/models/model-cards/gemini-3-pro/. Model card update: December 2025.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, and 1 others. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012.

Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. 2025. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26181–26191.

Mohamed Fazli Imam, Chenyang Lyu, and Alham Fikri Aji. 2025. Can multimodal llms do visual temporal understanding and reasoning? the answer is no! *arXiv preprint arXiv:2501.10674*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Daniel Kahneman, Anne Treisman, and Brian J Gibbs. 1992. The reviewing of object files: Object-specific integration of information. *Cognitive psychology*, 24(2):175–219.

Ilker Kesen, Andrea Pedrotti, Mustafa Dogan, Michele Cafagna, Emre Can Acikgoz, Letitia Parcalabescu, Iacer Calixto, Anette Frank, Albert Gatt, Aykut Erdem, and Erkut Erdem. 2023. Vilma: A zero-shot benchmark for linguistic and temporal grounding in video-language models. *Preprint*, arXiv:2311.07022.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, and 1 others. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Hongyu Li, Songhao Han, Yue Liao, Junfeng Luo, Jialin Gao, Shuicheng Yan, and Si Liu. 2025. Reinforcement learning tuning for videollms: Reward design and data efficiency. *arXiv preprint arXiv:2506.01908*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024a. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22195–22206.

Shicheng Li, Lei Li, Yi Liu, Shuhuai Ren, Yuanxin Liu, Rundong Gao, Xu Sun, and Lu Hou. 2024b. VITATECS: A diagnostic dataset for temporal concept understanding of video-language models. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXX*.

Songze Li, Xiaoke Guo, Tianqi Liu, Biao Yi, Zhaoyan Gong, Zhiqiang Liu, Huajun Chen, and Wen Zhang. 2026a. What's missing in screen-to-action? towards a ui-in-the-loop paradigm for multimodal gui reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 17674–17690.

Songze Li, Yarong Lan, Zhongpu Bo, Zhaoyang Wang, Zhiqiang Liu, Yuan Yuan, Chengtao Gan, Menghao Qian, Enpei Niu, Xiaoke Guo, and 1 others. 2026b. Scaling llm knowledge boundaries via distribution-optimized synthesis. *arXiv preprint arXiv:2606.23271*.

Yifei Li, Pengyiang Liu, Yuhang Zang, Zhongyue Shi, Qi Fu, Hongye Hao, and Jiwen Lu. 2026c. OVO-S-Bench: A hierarchical benchmark for streaming spatial intelligence in multimodal LLMs. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Pengyiang Liu, Junbo Niu, Xiaoyang Hu, Zhongyue Shi, Zitian Wang, Linjiang Huang, and Si Liu. 2026a. TRACE: Temporal retrieval with anchored and convergent evidence for long-horizon video understanding. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*.

Pengyiang Liu, Zhongyue Shi, Hongye Hao, Qi Fu, Xueting Bi, Siwei Zhang, Xiaoyang Hu, Zitian Wang, Linjiang Huang, and Si Liu. 2026b. SVCBench: A streaming video counting benchmark for spatial-temporal state maintenance. In *European Conference on Computer Vision*.

Tiedong Liu and Wee Sun Lee. 2026. Can vision-language models solve the shell game? *arXiv preprint arXiv:2603.08436*.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024b. TempCompass: Do video LLMs really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8731–8772, Bangkok, Thailand. Association for Computational Linguistics.

Yuanxin Liu, Kun Ouyang, Haoning Wu, Yi Liu, Lin Sui, Xinhao Li, Yan Zhong, Y.Charles, Xinyu Zhou, and Xu Sun. 2026c. Videoreasonbench: Can MLLMs perform vision-centric complex video reasoning? In *The Fourteenth International Conference on Learning Representations*.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems*, volume 36.

OpenAI. 2026. Introducing gpt-5.4. `https://openai.com/index/introducing-gpt-5-4/`. Published March 5, 2026.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, joseph heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, and 5 others. 2023. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, volume 36, pages 42748–42761. Curran Associates, Inc.

Zenon W Pylyshyn. 2001. Visual indexes, preconceptual objects, and situated vision. *Cognition*, 80(1-2):127–158.

Qwen Team. 2025a. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Qwen Team. 2025b. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. 2024. Cinepile: A long video question answering dataset and benchmark. *arXiv preprint arXiv:2405.08813*.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. 2025. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models. In *The Thirteenth International Conference on Learning Representations*.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, and 1 others. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232.

Xichen Tan, Yuanjing Luo, Yunfan Ye, Fang Liu, and Zhiping Cai. 2025. Allvb: All-in-one long video understanding benchmark. In *Proceedings of the AAAI*

Conference on Artificial Intelligence, volume 39, pages 7211–7219.

Qwen Team. 2026. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 74 others. 2026. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Maijunxian Wang, Ruisi Wang, Juyi Lin, Ran Ji, Thaddäus Wiedemer, Qingying Gao, Dezhi Luo, Yaoyao Qian, Lianyu Huang, Zelong Hong, and 1 others. 2026. A very big video reasoning suite. *arXiv preprint arXiv:2602.20159*.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, and 1 others. 2025a. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025b. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems*, volume 37, pages 28828–28857. Curran Associates, Inc.

Canran Xiao, Tianxiang Xu, Siyuan Ma, Yiyang Jiang, Haoyu Gao, and Yuhan Wu. 2026. Reversible primitive–composition alignment for continual vision–language learning. In *International Conference on Learning Representations*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Mengzhu Xu, Hanzhi Liu, Ningkang Peng, Qianyu Chen, and Canran Xiao. 2026. Affordance-first decomposition for continual learning in video-language understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3908–3919.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. 2020. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations*.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2025. Llava-video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2025. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13691–13701.

Le Zhuo, Songhao Han, Yuandong Pu, Boxiang Qiu, Sayak Paul, Yue Liao, Yihao Liu, Jie Shao, Xi Chen, Si Liu, and 1 others. 2025. Factuality matters: When image generation and editing meet structured visuals. *arXiv preprint arXiv:2510.05091*.

A Benchmark Details and Task Taxonomy

Task	Format	Description
State Tracking (8)		
Cup Trick	MC	Locate hidden ball after swaps
Cup Trick-S	Fill-in	Recover cup swap sequence
Cup Shuffle	MC	Track empty cup through shuffles
Cup Shuffle-S	Fill-in	Recover three-cup shuffle trace
Card Relocation	MC	Locate target card after moves
Card Relocation-S	Fill-in	Recover card position history
Card Shuffle	Fill-in	Count target card relocations
Card Shuffle-S	Fill-in	Recover card move sequence
Structural Composition (4)		
Falling Shape Count	Fill-in	Count falling target shapes
3D Maze Route	MC	Match route through 3D maze
Occlusion Object Count	Fill-in	Count objects under occlusion
Hidden Container Inference	MC	Infer hidden container shape
Dynamic Spatiality (4)		
Maze Trace	Fill-in	Count turns along route
Rotation Center	MC	Locate image rotation center
Trajectory Intersection	MC	Count trajectory intersections
Speed Comparison	MC	Compare object motion speeds
Temporal Ordering (4)		
Keyboard Sequence	Fill-in	Read ordered letter sequence
Keyboard Sequence-S	Fill-in	Recover letter reveal order
Neon Word	MC	Identify word from sequential flashes
Neon Word-Step	Fill-in	Recover word formation sequence
Sequential Counting (5)		
Symbol	Fill-in	Count matching symbols over time
Symbol-S	Fill-in	Recover symbol reveal sequence
Cube Structure Count	Fill-in	Count cubes in 3D structure
Grid Activation	Fill-in	Count unique activated cells
Grid Activation-S	Fill-in	Recover cell activation trace

Table 5: Task categories grouped by cognitive category. Format denotes the answer type: MC for multiple-choice and Fill-in for open-ended numeric, string, or JSON answers.

Figure 5: Word cloud of VIDEO-MME-LOGICAL.

B Training Configuration for the Qwen3-VL-8B SFT Experiments

Item	Value
Base model	Qwen3-VL-8B-Instruct
Training stage	Supervised fine-tuning (SFT)
Fine-tuning type	Full-parameter fine-tuning
Frozen modules	Vision tower; multimodal projector
DeepSpeed	ZeRO-2
Flash attention	FlashAttention-2
Template	Qwen3-VL no-thinking
Dataset mixture	Video-only logic SFT + image logic SFT
Maximum samples	1,000,000
Cutoff length	16,384 tokens
Image max pixels	802,816
Video max pixels	100,352
Maximum video length	128 frames
Video sampling rate	2 FPS
Epochs	2
Batch size per device	4
Gradient accumulation	1
Learning rate	2.0×10^{-5}
Weight decay	1.0×10^{-6}
Scheduler	Cosine
Warmup ratio	0.1
Precision	BF16
Save interval	1,000 steps
Checkpoint limit	5
GPU hours	H800 x 8 x 48h

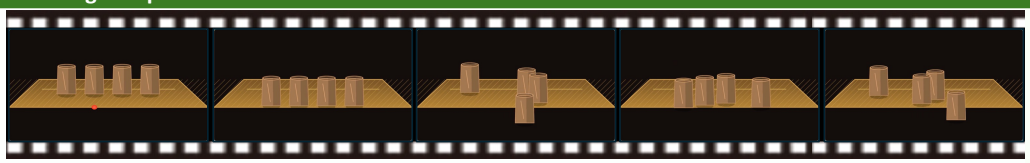
Table 6: Training configuration for the Qwen3-VL-8B SFT experiments.

C Human Evaluation Protocol

To estimate human-level performance, we sampled 3750 examples from VIDEO-MME-LOGICAL. The sampled examples were evaluated by three human annotators under the same visual-input setting used for model evaluation: annotators were shown the video and the corresponding question, but were not given access to program metadata, ground-truth answers, intermediate-state annotations, or model predictions. The annotators were instructed to answer each question according to the same output format used in our automatic evaluation, including structured answers for intermediate-state tasks when required. Human responses were then scored with the same evaluation script used for model outputs. Annotators were compensated at a rate of 50 USD per hour. The resulting human-level score is used only as a reference for benchmark solvability and for contextualizing the gap between human performance and current MLLMs.

D More Visualization Analysis

State Tracking - Cups Trick



Prompt:
Watch the full video. Which cup is the blue ball under at the end? The cups are labeled A, B, C, and D from left to right.
Please wrap your reasoning inside <think>...</think> and your final answer inside <answer>...</answer>.
Example:
<answer>A</answer>
GT: A

GPT-5.4:
Tracking cup movements
I need to infer the sequence of movements from the frames. The ball starts in front of the leftmost cup A, maybe just under its edge. As I look through the frames, the first few are static, but by frame four, the cups shift while the ball is still in front of A. By frame five, the ball disappears, likely because cup A moved over it. Then I observe the cups in different positions by frame ten, making the shuffling process complex. I'll need to analyze previous movements to deduce their final placements.
…
Evaluating cup positions and identity
From frames ten to twelve, it seems like cup A possibly moved from the left/front to the back-left cluster. But maybe it's simpler to think that the final positions of the cups retain their left-right identities as A, B, C, D after the shuffle—if there were no swaps with the ball cup. However, I suspect there were swaps. It's common for the ball to end up under cup B or C. In frame twelve, the front low center cup appears to be the one that came from the right/front, likely cup D moving left, while the right solo might be cup C. It's tricky!
<answer>B</answer>

Gemini 3.1 Pro:
Based on the movements of the cups in the video, we can track the position of the blue ball step-by-step:
1. **Initial State (00:00 - 00:03):** The blue ball is placed under the first cup from the left (Cup A).
2. **First Move (00:03 - 00:05):** The first two cups swap positions. The cup with the ball moves from position 1 to position 2.
3. **Second Move (00:05 - 00:07):** There is a three-way rotation among the cups at positions 2, 3, and 4. The cup at position 2 moves to position 3, the cup at position 3 moves to position 4, and the cup at position 4 moves all the way to position 2. Because the ball was at position 2, it moves to position 3.
4. **Third Move (00:12 - 00:14):** The cups swap in adjacent pairs simultaneously: position 1 swaps with position 2, and position 3 swaps with position 4. The ball, currently at position 3, swaps with the cup at position 4.

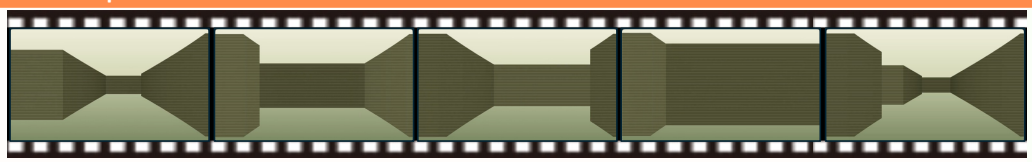
After the final move, the cup containing the blue ball is at the 4th position from the left, which is Cup D.
<answer>D</answer>

Qwen3-VL-8B-Think:
<answer>A</answer>

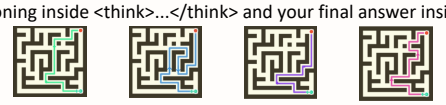
Ours-375K-Thinking:
The shuffle is solved by updating the current cup after each swap interval. From 0-2s the tracked ball starts under cup A, and from 2-4s it is still under cup A. After the 4-6s A/D swap it becomes D. After the 6-8s B/D swap it becomes B. After the 8-10s C/D swap it is still B. After the 10-12s A/C swap it is still B. After the 12-14s B/C swap it becomes C. After all 5 swaps, the ball ends under cup C. Slot coordinates are A=`xy=[289,699]`, B=`xy=[429,699]`, C=`xy=[569,699]`, D=`xy=[708,699]`. The timed ball trace with coordinates is 0-2s A at `xy=[289,699]`, 2-4s A at `xy=[289,699]`, 4-6s D at `xy=[708,699]`, 6-8s B at `xy=[429,699]`, 8-10s B at `xy=[429,699]`, 10-12s B at `xy=[429,699]`, 12-14s C at `xy=[569,699]`.
<answer>C</answer>

Figure 6: Additional visual examples from VIDEO-MME-LOGICAL.

Structural Composition – 3D Maze Route



Prompt:
Watch the full video. The video shows a first-person traversal through a colored maze. The reference image has four panels, an each panel has its option label A, B, C, or D shown below it. Each panel shows one candidate route for the same maze.
Which option matches the path shown in the video?
Please wrap your reasoning inside <think>...</think> and your final answer inside <answer>...</answer>.
For example:
<answer>A</answer>
GT: A



GPT-5.4:
Analyzing corridor options
I'm looking at the visible final frames of the corridor, which narrow into a central tunnel with symmetric walls. This seems to match a horizontal corridor that's vertically centered. Option C features a long horizontal corridor in the middle-lower area, and so does Option D. However, C starts simply, likely reflecting early movement through a top corridor before going down. Those initial frames show walls slanting towards the center, suggesting a forward perspective. I think C is the most likely option.
<answer>C</answer>

Gemini 3.1 Pro:
Based on a careful analysis of the video, we can trace the path by identifying the direction of each turn.
1. **Turn 1 (00:05):** The camera approaches the first wall and turns **Left**. This eliminates options B and D, which both start with a right turn.
2. **Turn 2 (00:09):** The camera turns **Right**.
…
The complete sequence of turns in the video is a perfectly alternating pattern: **Left, Right, Left, Right, Left, Right, Left, Right**. Therefore, the map that accurately represents the path shown in the video is Option A.
<answer>A</answer>

Qwen3-VL-8B-Think:
So, let's try to figure out which option matches the video. First, I need to understand the video's path. The video shows a first-person view moving through a maze with purple walls. The maze has corridors that expand and contract, and the movement is through these corridors.
- Frame 0: narrow corridor (like a tunnel)
- Frame 1-2: same
…
- Frame 36: the screen turns purple with a circle
<answer>B</answer>

Ours-375K-Thinking:
The decisive evidence comes from when the path changes heading, because those turns separate visually similar candidates. The traversal starts at top_right and ends at top_left. The intro spans 0-1s and the outro spans 36-38s. The move intervals are 1-2s left -> prefix L, 2-3s left -> prefix LL, 3-4s left -> prefix LLL, 4-5s … prefix LLLLDDLUULLLULLLLLLULLLUU, 33-34s up -> prefix LLLLDDLUULLLULLLLLLULLLUUUU, 35-36s left -> prefix LLLLDDLUULLLULLLLLLULLLUUUUL. The heading changes are 5-6s left to down, 8-9s down to left … 18-18s up to left, 24-25s left to up, 29-30s up to left, 34-35s left to up….
Once those turns are aligned, B=LLLLDDLUULLLULLLUURUULL (22 steps, 9 turns) is the only route left.</think>
<answer>C</answer>

Figure 7: Additional visual examples from VIDEO-MME-LOGICAL.

Structural Composition – Occlusion Object Count

Prompt:
Watch the full video. A moving foreground object covers different small objects at different times.
How many small objects are there in total?
Please wrap your reasoning inside <think>...</think> and your final answer inside <answer>...</answer>.
For example:
<answer>1</answer>
GT: 9

GPT-5.4:
Counting yellow objects
I need to count the total number of small yellow plus-like objects around a big circle, taking into account that some might be occluded in different frames. I've got 6 montage frames, likely from a video.
In frame 1, I see 6; frame 2 shows 4; frame 3 might have 7 or possibly 8 with some hidden; frame 4 matches frame 1 with 6 again; frame 5 has 3; and frame 6 has 6 again. So, I'm concluding there could be around 8 unique objects total. I'll consider their movement too, since the big ball covers them.
Counting objects in frames
In frame 2, I see a left and top object, along with partial views of the lower-left and lower-right objects....
<answer>7</answer>

Gemini 3.1 Pro:
To find the total number of small objects, let's observe the scene frame by frame and map the positions of all visible objects. The small objects are stationary throughout the video.
1. **Frame 00:00**: We can clearly see **3** small objects in the lower half of the screen.
   * Object 1: Far left
   * Object 2: Middle left
   * Object 3: Middle right
2. **Frame 00:01**: We can clearly see **1** small objects in the lower half of the screen

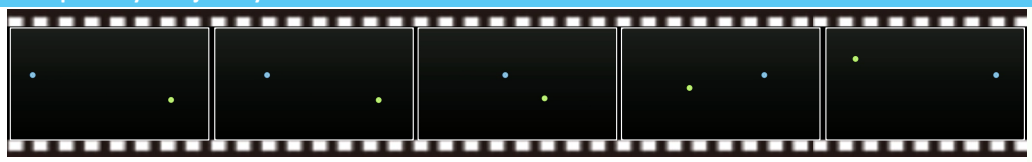
Summing them all up, there are 3 + 1 + 2 + 2 = 8 distinct small objects in total.
<answer>8</answer>

Qwen3-VL-8B-Think:
First, I need to look at each frame and see how many plus signs there are.
Frame 0: There are 3 plus signs (one on the left, one in the middle, one on the right) around the red circle.
Frame 1: The red circle is moving, but the plus signs are still there. Wait, no, in frame 1, the circle is moving, but the plus signs are the same? Wait, no, let's check the frames carefully.
Wait, the video has frames 0 to 5. Let's list each frame...
<answer>9</answer>

Ours-375K-Thinking:
I track how the visible subset changes over time instead of recounting from scratch. Early confirmations are 1 at 0-1.64s near xy=[612,720], 2 at 0-1.64s near xy=[558,720], 3 at 0-0.837s near xy=[630,539], 4 at 0-1.64s near xy=[501,757].
The visible-count timeline begins like this: 0-0.837s: 6 visible, 0.837-1.038s: 5 visible, 1.038-1.172s: 6 visible, 1.172-1.305s: 7 visible, 1.305-1.439s: 6 visible. Since the count oscillates but never reaches 10, I keep adding only newly exposed positions.
After merging all first-time exposures, the total number of distinct crosss is 10.
<answer>10</answer>

Figure 8: Additional visual examples from VIDEO-MME-LOGICAL.

Dynamic Spatiality - Trajectory Intersection



Prompt:
Watch the full video. Two colored points move back and forth along two fixed trajectories.
How many distinct intersection points are there between the two full trajectories? A. 0 B. 1 C. 2 D. 3
Please wrap your reasoning inside <think>...</think> and your final answer inside <answer>...</answer>.
For example:
<answer>A</answer>
GT: B

GPT-5.4:
Analyzing path intersections
I'm considering whether the paths intersect, with blue on the right descending and green on the left ascending. Their sampled positions suggest blue is mostly above and to the left of green, then switches horizontally. Midway, blue is on the right and still above green, then later ends up below it. It's interesting that they could cross twice! I think I just need to tag this information concisely.
<answer>C</answer>

Gemini 3.1 Pro:
To determine the number of distinct intersection points between the two trajectories, we need to carefully trace the path of each colored point over the entire video.
1. **Analyze the Blue Point's Trajectory:**
   * The blue point starts on the left side, moving downwards and to the left....
2. **Analyze the Green Point's Trajectory:**
   * The green point starts on the right side, moving downwards and to the right.
3. **Determine Intersections:**
   * We need to see how many times the green diagonal line (`/`) crosses the blue inverted V-shape (`/\`).
   * **Left side:** The left leg of the blue V-shape goes from the middle-left (low) to the upper-middle (high). ...(You can visually confirm this intersection point roughly one-third of the way from the left edge).
   * **Right side:** The right leg of the blue V-shape goes from the upper-middle (high) to the middle-right (low)... Since both are straight segments and the blue segment is always above the green segment on the right side, they never cross there.

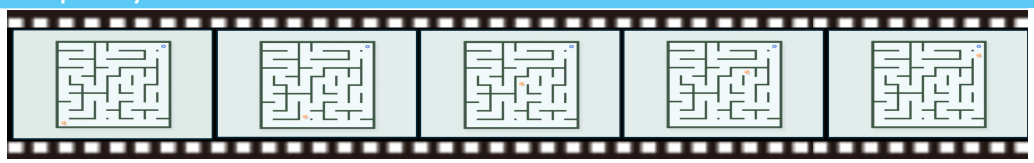
Therefore, there is exactly one point where the two full trajectories intersect.
<answer>B</answer>

Qwen3-VL-8B-Think:
I need to figure out how many distinct intersection points there are between the two full trajectories of the colored points. The video shows two points: one blue and one green, moving back and forth along their own paths...
[0.0] : blue at left, green at right (same height)
[1.3] : blue left, green right (but maybe green moved right more)
Wait, no, the key is to see the paths. Let's list the positions of each point over time...
Frame 0: blue at (x1, y), green at (x2, y), x1 < x2
Frame 1: blue at (x1', y), green at (x2', y), x1' < x1 < x2 < x2' (since blue moved left, green moved right)
Frame 2: blue at (x1'', y), green at (x2'', y), x1'' < x1' < x1 < x2 < x2' < x2''...
<answer>A</answer>

Ours-375K-Thinking:
I read the whole video like a drawing process: each dot gradually reveals its path, and the answer comes from where those two drawings overlap.
One trajectory is straight while the other has a single bend. The important caution is that one visible turn in a path is not automatically an intersection.
With that distinction in mind, the overlap pattern gives one distinct shared location: `xy=[222,424]`.
So the final option is B.
<answer>B</answer>

Figure 9: Additional visual examples from VIDEO-MME-LOGICAL.

Dynamic Spatiality - Maze Trace



Prompt:
Watch the full video. Focus on the moving ball as it travels through the maze to the goal.
Count how many times the ball changes direction before it reaches the goal.
Please wrap your reasoning inside <think>...</think> and your final answer inside <answer>...</answer>.
For example:
<answer>1</answer>
GT: 9

GPT-5.4:
Analyzing ball movement
I'm figuring out the path of the ball based on its frames. It starts at the bottom-left orange and aims for the top-right blue goal. I'll trace the motion through the screenshots. It seems the ball moves from the bottom-left corner, moving upwards and…
Analyzing maze navigation
I'm thinking about a maze where I start from the bottom-left opening. To reach the goal at the top-right, I would likely make 3 turns: first going up the left side to an opening around the lower third, then moving right along a horizontal corridor…
Clarifying the ball's movement
I think if the ball starts moving to the right first before going up, it might require one extra turn. However, since the starting position is at the bottom-left against the inner wall, it's likely that the initial movement is up along that left edge…
<answer>C</answer>

Gemini 3.1 Pro:
<answer>11</answer>

Qwen3-VL-8B-Think:
So, let's watch the video and count the direction changes. The ball starts at the bottom left, moves through the maze to the top right (goal). Let's track each segment:
1. First movement: from start (bottom left) moving right. Then, when it hits a wall, it turns. Let's list the path segments:
- Start at bottom left. First, it moves right (along the bottom row). Then, it hits a wall and turns up. That's a direction change (1).
- Then, moving up, hits a wall and turns right (2).
- Then, moving right, hits a wall and turns up (3)…
- Wait, let's check the video frames. Let's see the movement step by step:
Frame 0: ball at start (bottom left corner).
Frame 1: ball moves right (first segment)….
<answer>A</answer>

Ours-375K-Thinking:
The right audit is prefix-based: after each move, update the route prefix, then compare neighboring directions to decide whether a turn happened. The counting rule is: add one turn only when two consecutive movement directions differ; the initial start and the final stop do not count. Prefix checkpoints are step 1 ends with prefix U, step 2 ends with prefix UU, step 3 ends with prefix UUU, step 15 ends with prefix UUURRRUULLURRU, step 16 ends with prefix UUURRRUULLURRUU. Full move trace: 1.102-1.657s up -> prefix U, 1.657-2.212s up -> prefix UU, 2.212-2.767s up -> prefix UUU, 2.767-3.322s right -> prefix UUUR, 3.322-3.877s right -> prefix UUURR, 3.877-4.432s right -> prefix UUURRR, 4.432-4.987s up -> prefix UUURRRU, 4.987-5.542s up -> prefix UUURRRUU, 5.542-6.097s left -> prefix UUURRRUUL, 6.097-6.652s left -> prefix UUURRRUULL, 6.652-7.207s up -> prefix UUURRRUULLU, 7.207-7.762s right -> prefix UUURRRUULLUR, 7.762-8.317s right -> prefix UUURRRUULLURR, 8.317-8.872s up -> prefix UUURRRUULLURRU, 8.872-9.427s up -> prefix UUURRRUULLURRUU. Counted turns: 2.644-2.89s up to right at cell (4,0) -> turn count 1, 4.309-4.555s right to up at cell (4,3) -> turn count 2, 5.419-5.665s up to left at cell (2,3) -> turn count 3, 6.529-6.775s left to up at cell (2,1) -> turn count 4, 7.084-7.33s up to right at cell (1,1) -> turn count 5, 8.194-8.44s right to up at cell (1,3) -> turn count 6. The completed route is UUURRRUULLURRUU (up up up right right right up up left left up right right up up) with 16 steps and 6 counted turns, so the answer is 6.
<answer>6</answer>

Figure 10: Additional visual examples from VIDEO-MME-LOGICAL.

Temporal Ordering - Keyboard Sequence-S

Prompt:
Watch the full video. The video shows keyboard key presses. Report the full order in which letters appear in the video. Do not add spaces, punctuation, or separators.
Please wrap your reasoning inside <think>...</think> and your final answer inside <answer>...</answer>.
For example:
<answer> *ABC* </answer>
GT: **MQWERQWERF**

GPT-5.4:
Analyzing pressed keys
I need to identify the keys pressed across the frames:
- Frame 1: no highlight (initial state)
- Frame 2: m highlighted
…
- Frame 10: f highlighted again (last two frames are identical).
<answer> MQWERQWERF </answer>

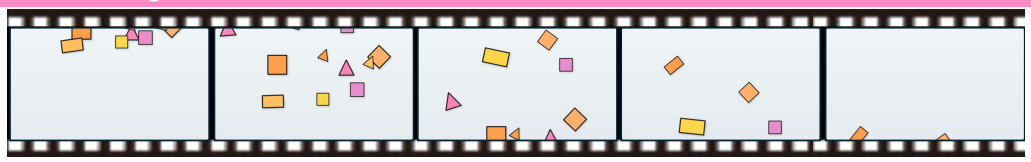
Gemini 3.1 Pro:
To determine the sequence of key presses, we can observe which key is highlighted in each frame of the video:
- **00:00 - 00:01**: No keys are highlighted.
- **00:02**: The **M** key is highlighted.
…
- **00:10**: The **F** key is highlighted again.
Consolidating the consecutive highlighted frames as single key presses, we get the resulting sequence.
<answer> MQWERQWEFF</answer>

Qwen3-VL-8B-Think:
Got it, let's watch the video and track the key presses. The video shows a keyboard with keys being highlighted in order. Let's list each key press step by step.
First, the video starts with the keyboard. Then:
- At some point, 'm' is highlighted (key 2). Then 'v' (key 4), then 'p' (key 6), then 'u' (key 7), then 'q' (key 1), then 's' (key 2), then 'n' (key 13), then 'v' again (key 4), then 'q' again (key 1), then 'v' (key 4), then 'k' (key 19).
Wait, let's check each frame:
…
Let me confirm once more. The video shows the keys being highlighted in that order. Each highlight corresponds to a key press, so the letters are Q, X, V, P, U, Q, S, N, V, Q, V, K.
<answer>QXVPUQSNVQVK</answer>

Ours-375K-Thinking:
The answer is the full keypress trace, not just distinct letters. The trace is 0-1s key M, 1-2s key Q, 2-4s key W, 4-5s key E, 5-6s key R, 6-7s key Q, 7-8s key W, 8-10s key E, 10-11s key R. So the correct full sequence is MQWERQWER.
<answer>MQWERQWER</answer>

Figure 11: Additional visual examples from VIDEO-MME-LOGICAL.

Sequencial Counting - Cube Structure Count



Prompt:
Watch the full video. Count how many triangles fall from the top across the whole video.
Please wrap your reasoning inside <think>...</think> and your final answer inside <answer>...</answer>.
For example:
<answer>1</answer>
GT: 8

GPT-5.4:
Calculating visible triangles
I'm counting visible triangles in the frames. In frame 3, there's one triangle under the top rectangle, two in the top right cluster, and one lone triangle on the right, bringing the total to five. The left object adds another triangle, so it seems there are possibly five triangles entering from the top. I wonder if the cluster in frame 2 contains similar objects....
<answer>4</answer>

Gemini 3.1 Pro:
To determine the total number of triangles that fall from the top, we can carefully examine each frame and track the unique triangles based on their appearance, position, and orientation. We will assume objects falling downwards originated from above the top edge, even if they first appear mid-screen due to the frame rate, provided they have a distinct orientation from previously seen objects. We will exclude objects that enter horizontally from the sides.
1. **00:01**: Three triangles fall from the top right. A small one pointing down-left, and two large overlapping ones. (**Count: 3**)
...
10. **00:11**: A large triangle pointing left appears at the bottom-center. Its orientation is different from the bottom-left one in the previous frame, indicating it is a distinct new object falling from the top. (**Count: 1**)
Adding them all up: 3 + 1 + 4 + 1 + 1 + 2 + 1 + 1 + 2 + 1 = 17.
<answer>17</answer>

Qwen3-VL-8B-Think:
<answer>9</answer>

Ours-375K-Thinking:
I use the full clip rather than a snapshot, because a single frame can hide whether two nearby polygons are separate falls or one continuing motion. So I count only distinct `triangle` events across time and ignore duration after entry.
Chronologically, the evidence is:
0-1s: triangle falls from the top; this is a distinct queried event, so the count becomes 1.
0-2s: triangle falls from the top; this is a distinct queried event, so the count becomes 2.
...
Across the whole clip, that yields 8 `triangle` events.</think>
<answer>8</answer>

Figure 12: Additional visual examples from VIDEO-MME-LOGICAL.