COGVID: A COMPREHENSIVE COGNITIVE BENCHMARK FOR EVALUATING VIDEO MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video understanding benchmarks have evolved from short clips to longer videos and from single to dozens of tasks, pushing multimodal large language models (MLLMs) toward long-form, multi-task comprehension. While existing benchmarks provide reasonable coverage of static perception, intricate tasks such as object permanence and episodic-like memory receive only partial coverage. In this paper, we propose **CogViD**, a cognitive theory-inspired systematic dataset that introduces new tasks and provides a comprehensive benchmark organized into 6 hierarchical levels. **CogViD** contains 25,505 QA pairs over 2,461 videos and 159.7 hours of footage, spanning 58 sub-tasks and questions in both multiple-choice and open-ended form. Complementary to this, we propose a template-driven multi-expert annotation pipeline in which level-specific templates drive QA generation, and verification is assigned to separate models to mitigate the bias of a model validating its own outputs, and blind-answerable and near-duplicate answers are filtered before human review. Our comprehensive evaluation across 25 state-of-the-art MLLMs reveals a large and consistent gap in maintaining quantities and states over time (including counting across frames), procedural memory to identify missing steps, and open-ended goal & error reasoning. **CogViD**, in the form of an extensive and robust benchmark, paves the way for future research toward human-like intelligence in video understanding.

1 INTRODUCTION

Recent advances in video multimodal large language models (MLLMs) (Bai et al., 2025a; Wang et al., 2025b; Zhang et al., 2025a; Team et al., 2026; Google DeepMind, 2025; Singh et al., 2025; Chen et al., 2025b; Shen et al., 2025) have produced strong results on tasks such as video captioning (Maaz et al., 2024; Zhang et al., 2023; Li et al., 2025; Lin et al., 2023), temporal grounding (Liu et al., 2024; Li et al., 2024c; Hong et al., 2025), and long-video QA (Fu et al., 2025; Wu et al., 2024b; Zhou et al., 2025; Wang et al., 2025a). Their capacity for temporal cognition, however, remains a significant limitation and is largely under-evaluated. Tracking an object once it is hidden, recalling when an event occurred, and reasoning about the causes and plans behind what happens are cornerstones of human video understanding and prerequisites for robotics, assistive agents, and safety monitoring, yet they remain a significant challenge and largely under-evaluated for MLLMs.

Current benchmarks for evaluating temporal cognition and reasoning of MLLMs have three major limitations. Firstly, existing benchmarks are **limited in tasks**, focusing predominantly on static perception questions that do not require spatial awareness and interaction between physical entities (Huang et al., 2026; Yu et al., 2019; Mangalam et al., 2023; Liu et al., 2024). Secondly, prior benchmarks lack a **systematic cognitive framework** for categorizing and evaluating different types of temporal reasoning ability, which leads to a fragmented, unbalanced list of tasks (Li et al., 2024c;d; Fu et al., 2025; Ning et al., 2025). Consequently, there is a noticeable gap between benchmarks to evaluate structural continuity and spatial arrangement, which are significantly under-tested. Thirdly, fewer benchmarks offer deeper reasoning questions and are often limited in the set of tasks, making them **insufficient in scale**. For example, some recent research works (Wang et al., 2025a; Yu et al., 2025; Ma et al., 2026) offer between 1K and 10K questions over natural videos, which are too few to robustly evaluate a model's capabilities or to motivate development of more sophisticated models.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

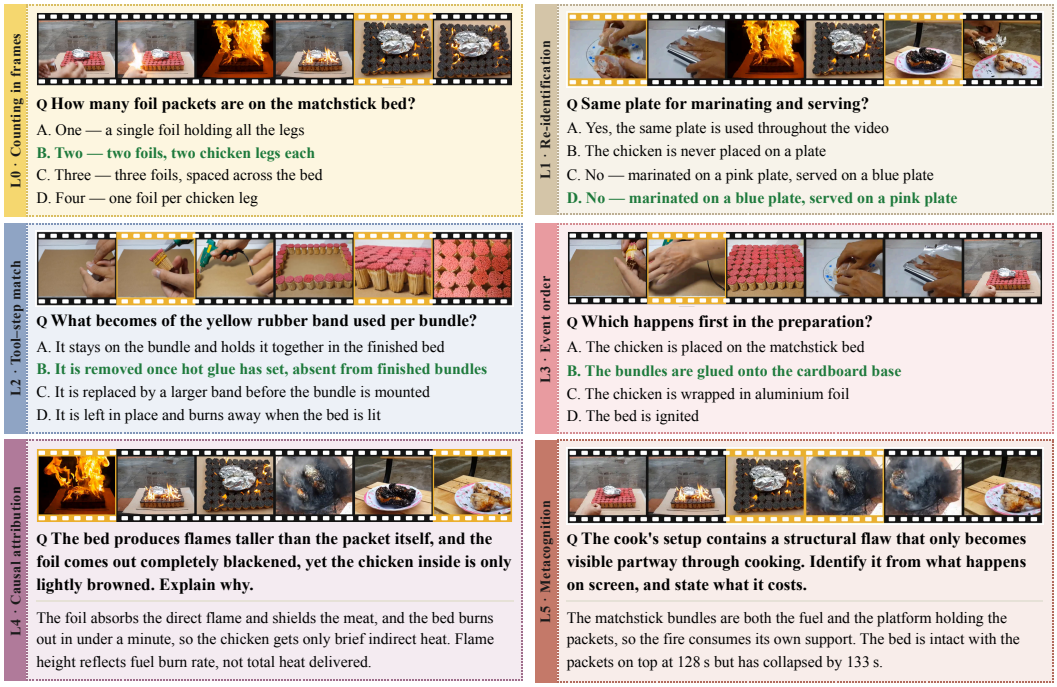


Figure 1: Overview of the proposed CogViD QA pairs drawn from a single source video.

To bridge this critical gap, we draw inspiration from cognitive science, which says intelligence grows through a series of stages (Piaget, 1954; Tulving, 1983; Anderson & Krathwohl, 2001). Also, unlike previous benchmarks that assemble a flat list of tasks or isolate a single cognitive faculty, we propose CogViD which adopts a unified six-level (L0-L5 as represented in Fig. 1). CogViD contains 25,505 verified QA pairs on 2,461 videos and 159.7 hours of footage, covering 58 sub-tasks at 6 cognitive levels. The videos cover diverse real-world scenarios collected from social platforms and existing video datasets, with clips ranging from 1 second to 21.8 minutes, a median duration of 2.6 minutes, and an interquartile range of 1.2 to 5.2 minutes. The benchmark comprises 13,732 multiple-choice and 11,773 open-ended questions, providing a well-balanced yet comprehensive coverage of video understanding capabilities from static perception to planning and metacognition.

To obtain annotations for CogViD, we develop a multi-expert annotation method that combines multiple LLMs with human verification, ensuring speed, quality, and reliability. To this end, first, we collect and curate videos using cognitive-level-specific keywords, followed by license verification, perceptual-hash deduplication, and manual inspection. Each video is then segmented and described by Gemini3Pro (Google DeepMind, 2025), providing detailed visual-semantic information for subsequent annotation. Second, Claude Opus generates candidate QA pairs from these descriptions using 58 level-specific question templates. Third, GPT-4o Hurst et al. (2024) independently verifies the candidate annotations by evaluating answer correctness, video groundedness, multi-moment evidence, and distractor plausibility. We further remove blind-answerable and near-duplicate questions through dedicated filtering stages. Finally, human annotators calibrate the annotation quality and resolve uncertain cases. By assigning different models to different stages, our framework leverages complementary capabilities while reducing model-specific bias and hallucination, resulting in reliable cognitive-level video annotations.

Our evaluation of mainstream MLLMs across 6 levels of cognitive video understanding reveals that current models have limited capabilities in maintaining and reasoning over structured information across time. Secondly, our fine-grained evaluation shows that models particularly struggle with recurrence counting and higher-level planning, while longer context and increased model scale do not consistently resolve these limitations. These findings underscore the difficulty of achieving robust cognitive understanding in video. These experiments also highlight that CogViD can serve as a reliable benchmark for diagnosing cognitive limitations in MLLMs and motivate future work towards models with stronger temporal memory, structured state tracking, and planning capabilities. The key contributions of our work are as follows:

- We introduce CogViD, designed to assess cognitive video understanding through progressively complex temporal reasoning across 6 levels comprising 58 sub-tasks. It provides $25,505$ QA pairs over $2,461$ videos and 159.7 hours, in both MCQ and OEQ formats.
- We propose a multi-expert annotation framework that utilized multiple expert models and human experts for video description, QA generation, verification, and filtering, effectively obtaining high-quality QA pairs at different cognitive levels.
- We conduct comprehensive experiments on proprietary and open-source MLLMs, revealing the challenges of video understanding, identifying notable gaps across different sub-tasks, and highlighting the limitations of current models.

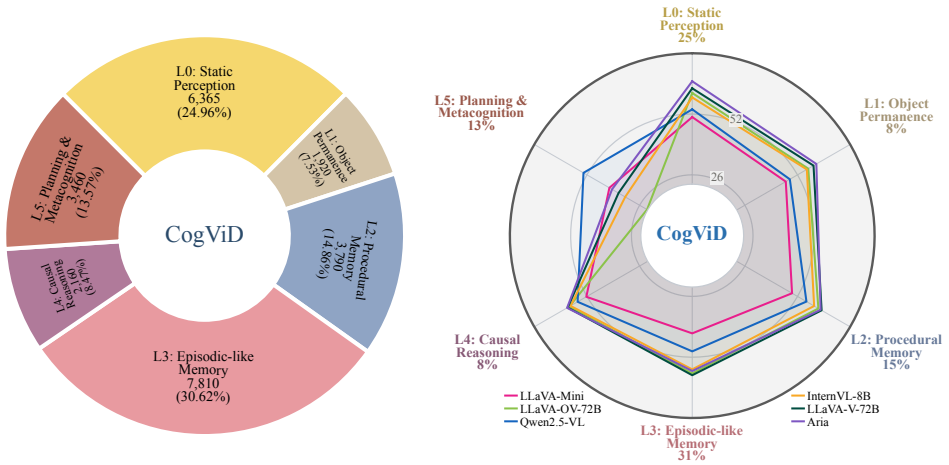
2 Related Work

General Video Understanding Benchmarks. Several benchmarks evaluate video MLLMs comprehensively. MVBench (Li et al., 2024c) assembles twenty temporal tasks converted from existing image and video datasets, each posed as a multiple-choice question with an automatically generated answer. Video-MME (Fu et al., 2025) spans 6 visual domains and three duration bands to measure general ability under a single protocol. VideoVista (Li et al., 2024d) separates understanding from reasoning across a large question pool covering fourteen video categories. MLVU (Zhou et al., 2025) and LongVideoBench (Wu et al., 2024b) extend the same format to long-form input by defining tasks over videos of up to an hour. LVBench (Wang et al., 2025a) pushes the duration further and requires evidence to be located across widely separated segments. EgoSchema (Mangalam et al., 2023) pairs three-minute egocentric clips with questions written so that no short window suffices, and MMBench-Video (Fang et al., 2024) collects free-form questions over multi-shot web video. CGBench (Chen et al., 2025a) requires each answer to be grounded in the specific segment that supports it. Video-MMMU (Hu et al., 2025), MMWorld (He et al., 2025) and MMVU (Zhao et al., 2025) shift the target from visual content to discipline knowledge acquired from video. CogViD is a general-purpose benchmark of comparable scope, covering 58 sub-tasks across 12 domains, but it is the first to arrange them as a progressive scale of temporal cognition in which each successive level requires evidence that the level below may not.

Task-driven Video Understanding Benchmarks. Task-driven benchmarks evaluate a single capability in isolation. TempCompass (Liu et al., 2024) isolates temporal attributes such as direction and speed by pairing each video with conflicting candidate descriptions. MotionBench (Hong et al., 2025) targets fine-grained motion at the level of individual movements rather than whole actions. CLEVRER (Yi et al., 2020) builds synthetic scenes whose exact physical ground truth supports descriptive, predictive and counterfactual questions about collisions and stability. Several causal and episodic reasoning including NExT-QA (Xiao et al., 2021), STAR (Wu et al., 2024a), Causal-VidQA (Li et al., 2022), and Ego4D (Grauman et al., 2022). VSI-Bench (Yang et al., 2025) evaluates the spatial representation a model recovers from egocentric walkthrough of indoor scenes, and Spatial-

Table 1: Comparison of CogViD with existing video understanding benchmarks. ✓ (Benchmark has specialized sub-tasks in the dataset). ◎ (Some questions are covered but no sub-tasks designed). ✗ (Level and also sub-task is not presented).

Benchmark	Videos	Sub-tasks	QA Pairs	Cognitive Level					
				L0	L1	L2	L3	L4	L5
ActivityNet-QA (Yu et al., 2019)	800	4	8,000	✓	✗	✗	✗	✗	✗
EgoSchema (Mangalam et al., 2023)	5,063	–	5,063	✓	✗	✗	◎	✗	✗
MVBench (Li et al., 2024c)	3,641	20	4,000	✓	◎	✗	◎	✗	✗
LongVideoBench (Wu et al., 2024b)	3,763	17	6,678	✓	◎	✗	✗	◎	✗
TempCompass (Liu et al., 2024)	410	10	7,540	✓	✗	✗	✗	✗	✗
VideoVista (Li et al., 2024d)	894	27	24,906	✓	✗	✗	✗	✗	✗
MLVU (Zhou et al., 2025)	1,730	9	3,102	✓	✗	✗	◎	✗	✗
Video-Bench (Ning et al., 2025)	5,917	10	17,036	✓	✗	✗	✗	✗	✗
LVBench (Wang et al., 2025a)	103	6	1,549	✓	◎	✗	✗	◎	✗
CG-Bench (Chen et al., 2025a)	1,219	24	12,129	✓	◎	✗	✗	◎	✗
Video-MME (Fu et al., 2025)	900	12	2,700	✓	✗	✗	◎	✗	✗
Video-MMMU (Hu et al., 2025)	300	6	900	✓	✗	✗	✗	✗	✗
MMWorld (He et al., 2025)	1,910	7	6,627	✓	◎	✗	✗	◎	✗
MMVU (Zhao et al., 2025)	1,529	–	3,000	✓	◎	✗	✗	◎	✗
VRBench (Yu et al., 2025)	1,010	7	9,468	✓	✗	✗	✗	◎	✗
ScaleLong (Ma et al., 2026)	269	5	1,747	✓	✗	✗	◎	✗	✗
CogViD (Ours)	2,461	**58**	**25,505**	✓	✓	✓	✓	✓	✓



(a) Task Distribution (b) MLLMs performance on 6 cognitive levels

Figure 2: **(a)** Task distribution in CogViD, illustrating the proportion of QA pairs across the 6 hierarchical cognitive levels. **(b)** Cognitive level-wise model performance, comparing the evaluation results of SOTA MLLMs.

DISE (Huang et al., 2026) and SpatialBench (Xu et al., 2025) apply the same idea to further spatial question types. VUDG (Wang et al., 2026) aggregates videos from multiple sources to measure robustness under domain shift. VRBench (Yu et al., 2025) grades the intermediate steps of a multi-step reasoning chain instead of the final answer, and VideoReasonBench (Liu et al., 2026) constructs sequences whose state cannot be recovered from any single frame.

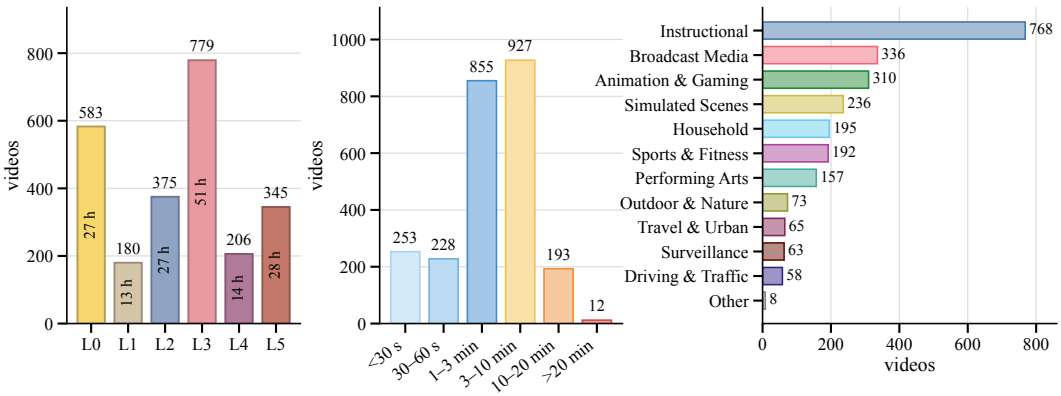
3 CogViD

Motivation: Despite significant recent progress in video understanding benchmarks, existing datasets evaluate isolated skills, with little account of how these skills depend on one another. While designing CogViD, we asked: *What structure must a benchmark capture to measure temporal cognition in video, rather than a flat list of tasks?* In pursuit of answering this question, we identified *three design principles* necessary in designing a structured benchmark: 1) The need for a *hierarchical taxonomy of temporal cognition*, ordered as a progression in which each level requires evidence similar to human understanding, from static perception to goal and error reasoning. 2) The need for *diverse video domains under consistent contexts*, so that the same cognitive questions can be asked across everyday footage rather than a single domain. 3) The need for *template-driven annotation without self-validation*, where generation and verification models are kept apart, and human review remains the last filter. These interconnected principles necessitate how we evaluate video MLLMs and directly motivate the design of CogViD.

CogViD consists of 2,461 source videos and 25,505 QA pairs, and 159.7 hours of footage, with clips from one second to 21.8 minutes (median 2.6 minutes), as summarized in Fig. 3. The questions span 58 sub-tasks, organized into 6 progressive levels of temporal cognition as mentioned in Fig. 2 and sub-task list of each level mentioned in the Appendix A. Each question is characterised along four dimensions: **(i)** ***Cognitive level***, covering perception (L0), object permanence (L1), procedural structure (L2), episodic memory (L3), causal reasoning (L4), and goal and error reasoning (L5); **(ii)** ***Subtask***, pairing each level with a template family such as counting, occlusion tracking, step ordering, recurrence counting, chain reconstruction, or goal inference; **(iii)** ***Video domain***, ranging across everyday activities so that a level is not confounded with a single content type; and **(iv)** ***Answer format***, multiple choice or open-ended. L0 is multiple choice only, L5 open ended only, and L1-L4 contain both; we score multiple choice on L0-L4 and open-ended generation on L5.

3.1 Cognitive Levels

We structure spatial and temporal cognitive abilities into 6 levels, each defined by the uniqueness of the questions from single frame to multiple frames across different tasks. CogViD organizes 58 sub-tasks into 6 cognitive levels, each defined by what a question requires the model to hold in mind beyond current frame. Fig. 3 shows how 2,461 videos distribute over 6 levels. Due to space limitations, definitions of each sub-task are provided in the Appendix B.



(a) Level assignment (b) Video duration (c) Content domains

Figure 3: Composition of the CogViD. **(a)** Videos per cognitive level **(b)** clip length, and **(c)** videos per content domain.

Static Perception (L0): This level supports 10 sub-tasks across 583 videos (27.3 hours) with 6,365 QA pairs. It focuses on what is directly observable on screen, including action snapshot, text/OCR, scene classification, object recognition, attribute recognition, and counting in frame.

Object Permanence (L1): This level executes on 8 sub-tasks across 180 videos (13.2 hours) with 1,920 QA pairs. It focuses on representing an object while it may not be visible, including occlusion tracking, counting under occlusion, re-identification, and trajectory plausibility.

Procedural Memory (L2): This level focuses on 11 sub-tasks across 375 videos (26.8 hours) with 3,790 QA pairs. It focuses on the state a procedure leaves behind rather than recall of the order of its steps, including step ordering, procedure-stage query, step-completion check, world-state query, tool-step match, and cross-segment dependency.

Episodic-like Memory (L3): This level maps 11 sub-tasks across 779 videos (51.3 hours) with 7,810 QA pairs. It focuses on binding what happened to when and where it happened, including agent history, event ordering, recurrence counting, revisit detection, and possession tracking.

Causal Reasoning (L4): This level supports 8 sub-tasks across 206 videos (14.2 hours) with 2,160 QA pairs. It focuses on linking an outcome to the events that produced it including causal attribution, chain reconstruction, forward prediction, and multi-step consequence.

Planning and Metacognition (L5): This level contains 10 sub-tasks across 345 videos (27.7 hours) with 3,460 open-ended QA pairs. It focuses on high-level goal inference, plan generation, error detection, and reflective reasoning about feasibility and alternative strategies.

3.2 Video Collection and Level Assignment

We curate video sources that capture the core challenges of temporal cognition, including object permanence under occlusion, long-range procedural and episodic structure, and causal or goal evidence that unfolds over time. We retrieve videos from social platforms and existing public datasets. Each of the 6 levels has its own human-assisted keyword list, written for the evidence that level needs, for example occlusion and hiding for object permanence, or step-by-step instruction for procedural memory, rather than a topic such as cooking or sport. Human assistance here mitigates the risk of overlooking highly challenging questions that MLLMs are unable to generate independently. Videos that can be answered from a single static frame, or whose content cannot support the evidence requirement of any level, are discarded. Every remaining candidate then undergoes *license verification, perceptual-hash deduplication* against the rest of the corpus, and manual inspection for visual clarity and information density. Once filtering is complete, each video is assigned to one of the 6 levels and 58 sub-tasks according to the temporal structure its content exposes.

3.3 Template-Driven QA Generation

Firstly, we divide each video into segments. Gemini 3 Pro (Google DeepMind, 2025) writes a summary of each segment, then synthesizes the ordered segment summaries into one full-video

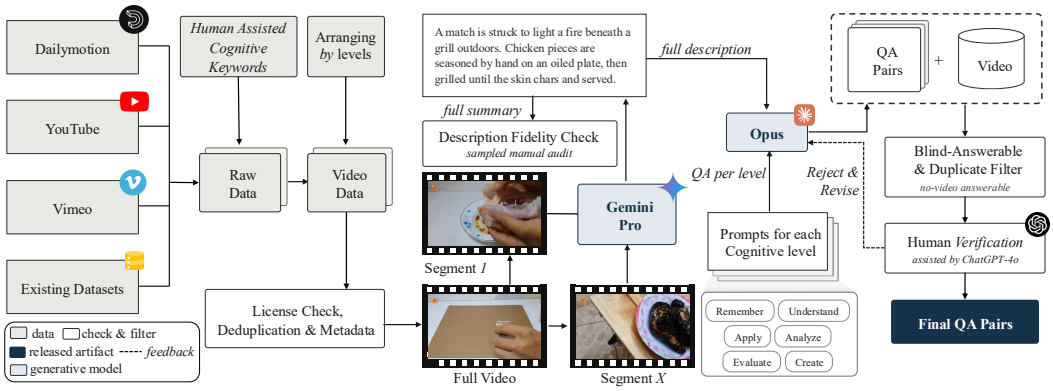


Figure 4: Overview of the CogViD dataset construction and annotation pipeline from video collection to human-verified QA pairs.

summary that preserves the timeline of the clip. We then audit 500 randomly sampled segments against the source for hallucinated objects, misattributed actions, and ordering errors, which yields a description fidelity of 94.8%. A video is re-described when an audited segment fails, and it is removed if the second description also fails.

Each of the 58 sub-tasks has one plain-text generation template that fixes cognitive level, evidence type, question budget, and output schema. The complete QA generation pipeline is shown in Fig. 4. We release the 58 generation templates with the dataset so that the QA pipeline can be reproduced. Based on the template and the summaries, Claude Opus generates candidate questions with exactly four options when the question is multiple choice.

3.4 Multi-Model Verification and Threshold Calibration

To prevent self-validation bias, the tasks of description, generation, and verification are assigned to distinct models. The verifier is GPT-4o (Hurst et al., 2024), which shares no model family with the describer or the generator. Each candidate question is returned to the verifier with the source video and scored on four criteria: answer correctness, groundedness in the video, the multi-moment evidence requirement, and distractor plausibility. Questions that fall below the acceptance threshold are sent back for revision, with a maximum of three rounds before a question is discarded.

3.5 Screening and Human Review

Despite the structured pipeline for creating QA pairs, issues such as ambiguous answers and errors in open-ended responses can still occur and reduce quality. To mitigate this, we apply a human-based screening process that combines automated evaluation with expert review. First, GPT-5.2 (Singh et al., 2025), a more advanced proprietary model, reviews each QA pair while referring to the original video. It flags every pair into a category: (1) correct, (2) flawed. A human reviewer then audits these pairs by watching the source video and verifying the flags. Partially correctable QA pairs are fixed whereas invalid pairs are deleted. This ensures the final benchmark meets high standards of accuracy and clarity while largely reducing the cost of expensive human annotations.

3.6 Comparison with General Video Benchmarks

Existing video understanding benchmarks are broadly organised by clip duration and by sub-tasks. Short-video benchmarks like TGIF-QA (Jang et al., 2017) and MVBench (Li et al., 2024c) utilise sub-minute clips, which restrict the range of temporal evidence a question may require. Long-video benchmarks including CinePile (Rawal et al., 2024), EgoSchema (Mangalam et al., 2023), MLVU (Zhou et al., 2025), Video-MME (Fu et al., 2024), LongVideoBench (Wu et al., 2024b), and LVBench (Wang et al., 2025a) extended durations, while recent work has instead broadened the axis of difficulty of expert knowledge (Zhao et al., 2025; Hu et al., 2025), fine-grained motion (Hong et al., 2025), procedural structure (Tang et al., 2026), multi-step reasoning chains (Yu et al., 2025; Liu et al., 2026), and temporal scale (Ma et al., 2026).

Table 2: Performance of MLLMs on CogViD across 6 cognitive levels. Open-source models are arranged by parameter count. Darker blue represents better performance.

Model	Size	Date	Input	Cognitive Level						Overall
				L0	L1	L2	L3	L4	L5	
Human	-	-	Whole Video	92.8	91.3	88.9	91.0	90.7	91.2	91.0
Blind (Llama-2)	70B	2023-07	Text-Only	29.7	29.2	33.4	26.7	36.3	41.8	32.9
Blind (GPT-5.2)	-	2025-11	Text-Only	33.0	30.9	38.2	29.1	38.6	42.7	35.4
Proprietary Models										
Gemini 2.0 Flash	-	2025-02	256 frm	64.0	57.9	65.3	52.9	58.4	39.2	56.3
GPT-4o	-	2024-05	64 frm	60.3	54.8	62.6	50.0	63.7	35.4	54.5
Doubao 1.5-VL Pro	-	2025-01	256 frm	62.0	53.0	61.0	48.1	65.3	32.7	53.7
Gemini 2.5 Pro	-	2025-03	256 frm	68.7	65.0	69.4	60.0	68.7	48.5	63.4
GPT-Astra	-	2026-09	8 frm	81.2	77.8	73.8	78.6	74.8	41.6	71.3
Open-Source Models										
Qwen2-VL	2B	2024-09	8 frm	49.3	46.5	51.6	47.4	54.1	17.7	44.4
Phi-3.5-Vision	-	2024-04	64 frm	51.8	40.6	51.7	45.9	51.2	32.0	45.5
Phi-4-Multimodal	-	2025-03	128 frm	54.6	49.6	55.8	48.5	57.2	7.8	45.6
LongVILA	7B	2024-08	256 frm	60.4	56.0	60.1	55.7	60.1	31.7	54.0
LongVU-Qwen2	7B	2024-10	32 frm	55.0	51.8	56.9	50.7	56.5	17.9	48.1
LongVA	7B	2024-06	256 frm	52.1	45.1	53.2	45.1	52.4	36.6	47.4
Flash-VStream-Qwen	7B	2024-06	256 frm	54.0	47.5	53.0	49.3	55.8	0.0	43.3
LLaVA-OV (SI)	7B	2024-08	128 frm	56.3	49.9	56.7	50.9	56.4	33.3	50.6
Qwen2-VL	7B	2024-09	8 frm	56.0	53.1	58.2	53.3	58.0	19.2	49.7
LLaVA-Video	7B	2024-10	32 frm	58.2	50.1	58.0	52.0	56.8	22.4	49.6
Qwen2.5-VL	7B	2025-01	256 frm	54.0	48.2	56.4	49.5	56.6	53.7	53.1
LLaVA-Mini	8B	2025-01	256 frm	50.7	46.2	49.3	41.8	52.3	40.9	46.9
InternVL2.5	8B	2024-12	128 frm	59.1	56.6	60.2	57.2	60.1	33.1	54.4
Qwen3-VL	8B	2025-10	8 frm	61.2	54.8	62.8	54.2	61.8	40.0	55.8
MiniCPM-V-4.5	-	2025-08	64 frm	63.0	60.3	64.0	60.9	62.9	39.2	58.4
Aria	-	2024-10	256 frm	66.0	61.3	63.4	57.9	61.8	39.6	58.3
InternVL2.5	26B	2024-12	128 frm	63.4	60.9	64.4	61.8	62.1	42.3	59.2
InternVL2.5	38B	2024-12	256 frm	66.4	63.3	66.4	65.6	63.3	45.9	61.8
LLaVA-OV	72B	2024-08	64 frm	61.0	57.3	62.3	58.9	60.7	22.3	53.8
LLaVA-Video	72B	2024-10	128 frm	63.0	59.9	63.9	59.7	61.5	36.4	57.4
InternVL2.5	78B	2024-12	128 frm	69.3	65.7	68.7	69.5	65.2	50.1	64.7

Most of the datasets emphasize on static perception with only partial coverage of object permanence and causal reasoning, and no benchmark evaluates *procedural memory*, *episodic-like recall*, or *planning* and *metacognition* at all. Therefore, we ask two questions: (1) Are QAs queries predominantly about content present in individual frames rather than information that must be retained across them, such that overall performance remains high even when MLLMs fail to track occluded objects, count repeated actions, or recover event order? (2) Do benchmarks that probe higher capacities evaluate each capacity on a different set of videos, making weaker capacities and harder content indistinguishable?

CogViD is designed to close the gap by requiring every question to draw on at least multiple distinct frames, and place all 6 cognitive levels under one taxonomy and one collection protocol. Its 25,505 QA pairs span 58 sub-tasks. Each video is assigned to the level its content can support, so the typical clip contributes questions at a single level (2,468 level-assignments over 2,461 videos), cross-level comparison are between-video. All 6 levels are nonetheless collected and filtered under one protocol as mentioned in Appendix E.

4 Experiments

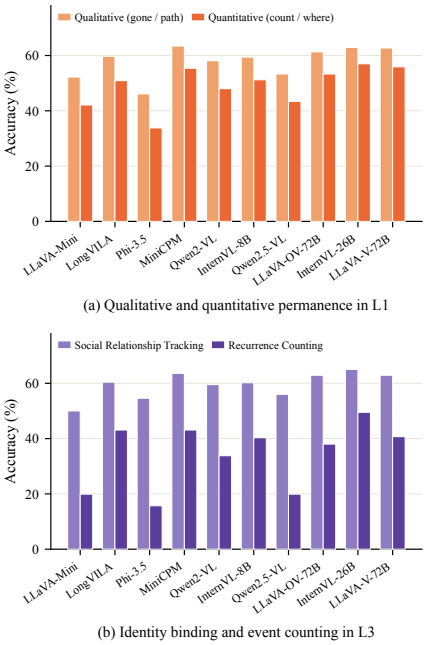
We evaluate a diverse set of open-source models as well as proprietary models, comprising Gemini 2.0 Flash and Gemini 2.5 Pro (Gemini Team, Google, 2025), GPT-4o (Hurst et al., 2024), and Doubao 1.5-VL Pro (ByteDance Seed, 2025). On the open-source side, we cover Qwen2-VL-2B and Qwen2-VL-7B (Wang et al., 2024) and Qwen2.5-VL-7B (Bai et al., 2025b); InternVL2.5 at 8B, 26B, 38B, and 78B (Chen et al., 2024); LLaVA-OneVision at 7B (SI) and 72B (Li et al., 2024a); LLaVA-Video at 7B and 72B (Zhang et al., 2025c); MiniCPM-V-4.5 (Yao et al., 2024); Phi-3.5-Vision (Abdin et al., 2024) and Phi-4-Multimodal (Abouelenin et al., 2025); LLaVA-Mini-8B (Zhang et al., 2025b); and SmolVLM2-500M (Marafioti et al., 2025). For long-video settings we include LongVILA-7B (Chen et al., 2025b), LongVU-Qwen2-7B (Shen et al., 2025), LongVA-7B (Zhang et al., 2024b), Flash-VStream-Qwen-7B (Zhang et al., 2024a), and Aria (Li et al., 2024b). For baselines, we include the three, following OpenEQA, Cambrian-S, and SAW-Bench (Majum-

dar et al., 2024; Yang et al., 2026; Li et al., 2026): (1) Human answers from the whole video and scores; (2) Blind (GPT-5.2) is a proprietary text-only baseline that sees the question and choices but no video. (3) Blind (Llama-2-7B) (Touvron et al., 2023) is the same text-only setting with an open-source LLM. To evaluate performance, we compute absolute accuracy by checking whether the predicted choice matches the ground-truth answer. For OEQs, we define several evaluation criteria and use GPT-4 to score how well the generated texts align with the reference answers.

4.1 Quantitative analysis

Table 2 reports the performance of proprietary and open-source MLLMs across the 6 cognitive levels of CogViD. Unless otherwise specified, every model is evaluated under the same protocol. Due to space constraints, the performance of each sub-task on each level is deferred to Appendix C and graphical representation in Appendix D. We highlight the following observations:

How does performance change across cognitive levels? Since each level of CogViD targets a different cognitive capability, we first examine which of these capabilities current video MLLMs possess and which they lack, reporting the level-wise results of all evaluated models in Tab. 2. From these results, we observe that the ordering of the levels is largely shared across models whose overall competence differs widely. Specifically, L0, L2, L4 remain the strongest levels for most models, whereas L1, L3 fall below them. L5 is the weakest level almost without exception, and its separation from the other levels persists even for the largest models we evaluate (Fig. 6b), indicating that capacity alone does not supply what the upper levels require. Notably, the strongest overall result comes from an open-source model, while most of the proprietary systems fall behind InternVL2.5-26B. Overall, MLLMs excel at recognizing content and at reusing script-like priors, while *they struggle to hold hidden state, to count events over time, and to write out a goal and a repair*. Even the strongest among them remains far below the human reference, which underscores the considerable room for improvement in current video MLLMs.



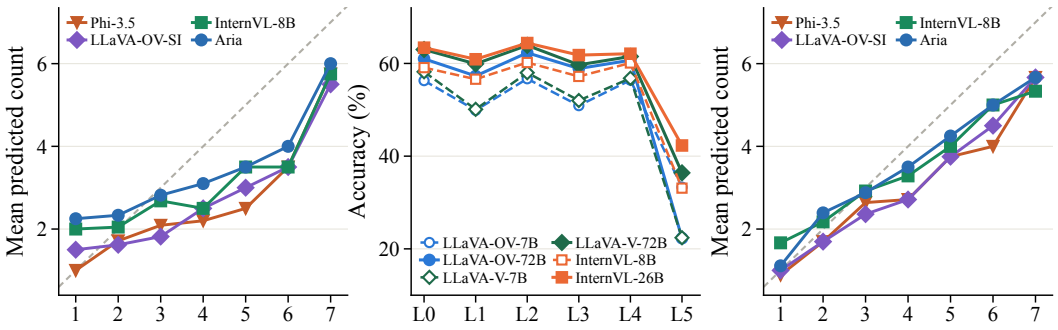
(a) Qualitative and quantitative permanence in L1

(b) Identity binding and event counting in L3

Figure 5: **(a)** Qualitative permanence is easier than quantitative permanence. **(b)** Models struggle on recurrence counting.

What do object-permanence errors reveal? Given that object permanence is where performance first drops for every model (Tab. 2), we examine which of its questions models fail on, using the sub-task breakdown in Tab. A10. We find that this drop is concentrated rather than spread across the level, as models struggle most when they must carry a quantity or a location through an occlusion (Counting under Occlusion, Occlusion Tracking, and Final-location Prediction). In contrast, questions that only ask whether something has changed (Disappearance Type, Existence Check, and Trajectory Plausibility) remain comparatively easy. Also, this contrast holds for every model in the table (Fig. 5a), indicating that it reflects the demand of the question rather than the capacity of the model. Lastly, these failures suggest that *models can tell when an object goes away or when a path looks wrong, but they cannot keep track of which object it was or how many are left once it is out of sight*.

Where does episodic memory break? Beyond object permanence, a natural question is whether models also fail at remembering events in general, or only at particular kinds of recall. To examine this, we compare how models perform on the individual sub-tasks of episodic memory (Tab. A12). We find that models break down on specific kinds of questions, namely those that require counting how often an event recurs (Recurrence Counting) or recalling when it occurred (First/Last Occurrence and Event-time Binding). In contrast, questions about who did what (Social Relationship Tracking) or whether a scene has changed since an earlier visit (Revisit Detection and Change across Visits) remain comparatively easy, and this contrast holds for every model in the table (Fig. 5b). No-



(a) Recurrence count (b) Accuracy of models (c) Frame counting

Figure 6: **(a)** Mean predicted true count for recurrence counting **(b)** Performance comparison on each level. **(c)** Counting in frame; the dashed line marks perfect prediction.

tably, Recurrence Counting is the hardest multiple-choice sub-task in entire benchmark, which we examine more closely next. Therefore, memory in CogViD is not a single skill: *models can recall what happened, but they cannot say how many times an event happened or when it first or last occurred*.

Probing counting via predicted counts. Since counting emerges as the common weakness of both object permanence and episodic memory, we further examine why models fail at it by discarding the multiple-choice scoring and reading out the number a model commits to. Specifically, we re-prompt four MLLMs for a bare integer on recurrence counting and counting in frame, and compare the mean prediction for each ground-truth count against the identity line. As shown in Fig. 6a, the curves rise far more slowly than the true count and fall below the identity line for all but the smallest counts. The four models differ in level but share this shape, suggesting that *MLLMs do not enumerate what they see; they emit a prior over small counts that is only weakly modulated by the evidence*. Crucially, the same compression appears for counting in frame (Fig. 6c), which requires no memory across time, suggesting that the bottleneck lies in maintaining a count rather than in forgetting over long videos.

Do long-video models remember better? Several of the evaluated models are designed specifically for long-video inputs, for instance by compressing visual tokens or maintaining memory over past frames. To examine whether such designs lead to stronger memory, we revisit the results in Tab. 2 and compare these models with general-purpose models of similar scale on object permanence and episodic memory. We do not observe a consistent advantage for the long-video models on either level, despite their typically larger frame budgets, and their drops from perception to these two levels are comparable to those of the general-purpose models. For example, Aria shows one of the largest drops to episodic memory, whereas Qwen2-VL-7B, which observes far fewer frames, shows one of the smallest. Consistent with the counting probe, these results suggest that *stronger long-video input handling alone does not necessarily translate into better memory* and that a significant uphill effort is required to build the advanced models capable of handling such intricate tasks.

5 Conclusion

In this work, we introduce CogViD, a comprehensive benchmark to evaluate temporal cognition in real-world videos. It provides curated QA pairs with a six-level hierarchy of temporal skills across diverse domains and clip durations, addressing the lack of benchmarks for temporally extended, hierarchical reasoning in multimodal models. Unlike prior benchmarks based on short clips and isolated recognition, CogViD requires models to integrate visual evidence across multiple moments to track hidden state, reconstruct events, and demonstrate deep temporal comprehension. Our evaluation of proprietary and open-source video MLLMs shows that recognition is close to saturated while counting, object permanence, episodic memory, and open-ended goal reasoning remain far below human performance, which underscores the considerable room for improvement in current video MLLMs. We discuss the limitations of CogViD in Appendix E.

AI USE STATEMENT

We used generative AI tools to edit the writing of the paper for grammar and formatting. We have not used generative AI tools to invent ideas or any kind of involvement in data set and for theoretical claims. Also, dataset construction and evaluation were done without the aid of generative AI tools. Human annotators audited 500 video summaries and a random subset of 100 questions against the source footage.

ETHICS STATEMENT

All authors have read and adhere to the ICLR Code of Ethics. CogViD is built from publicly posted videos collected from social platforms and from existing public datasets, after license verification, perceptual-hash deduplication, and manual inspection. This work does not recruit human subjects for an interventional study, and it does not collect private recordings. Human annotators reviewed source videos only to check that questions are video-grounded and that answers and distractors are valid. The dataset includes everyday public scenes, including a small amount of already public surveillance footage, which we treat as evaluation content rather than as a tool for identification or monitoring. We will release the benchmark with source attribution and license constraints, and we will not redistribute footage whose license forbids it. The intended use is scientific evaluation of video MLLMs.

REPRODUCIBILITY STATEMENT

We document six-level taxonomy and all 58 sub-tasks are described with video collection, template-driven generation, multi-model verification, and human review. The evaluation protocol includes uniform frame sampling at each model's maximum budget, including MCQs and OEQs. The individual level scores appear in Table 2 in the main paper, and the per-subtask scores appear in the appendix. We will release the 58 templates, the QA split, and the evaluation scripts with the benchmark so that others can rerun the reported protocol.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Lorin W Anderson and David R Krathwohl. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc., 2001.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

ByteDance Seed. Doubao-1.5-pro. Technical Report, 2025. URL `https://seed.bytedance.com/en/special/doubao_1_5_pro`.

Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. In *International Conference on Learning Representations*, volume 2025, pp. 45647–45682, 2025a.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. In *International Conference on Learning Representations*, volume 2025, pp. 18227–18246, 2025b.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124, 2024.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 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*, pp. 24108–24118, 2025.

Gemini Team, Google. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Google DeepMind. Introducing gemini 3: our most intelligent model that helps you bring any idea to life. Google Blog, 2025. URL `https://blog.google/products-and-platforms/products/gemini/gemini-3/#gemini-3`.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, Kevin Lin, William Yang Wang, Lijuan Wang, and Xin Eric Wang. MMWorld: Towards multi-discipline multi-faceted world model evaluation in videos. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=tRNKe2Vgqt`.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8450–8460, 2025.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.

Xinmiao Huang, Qisong He, Zhenglin Huang, Boxuan Wang, Zhuoyun Li, Guangliang Cheng, Yi Dong, and Xiaowei Huang. Spatial-dise: A unified benchmark for evaluating spatial reasoning in vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 135833–135865, 2026.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2758–2766, 2017.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Chuhan Li, Rilyn R. Han, Joy Hsu, Yongyuan Liang, Rajiv Dhawan, Jiajun Wu, Ming-Hsuan Yang, and Xin Eric Wang. SAW-bench: Learning situated awareness in the real world. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=8lwrYjv6r7.

Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Fan Zhou, Chengen Huang, Yanpeng Li, et al. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024b.

Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21241–21250. IEEE, 2022.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024c.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *Science China Information Sciences*, 68(10):200102, 2025.

Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. Videovista: A versatile benchmark for video understanding and reasoning. *arXiv preprint arXiv:2406.11303*, 2024d.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

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

Yuanxin Liu, Kun Ouyang, Haoning Wu, Yi Liu, Lin Sui, Xinhao Li, Yan Zhong, Y. Charles, Xinyu Zhou, and Xu Sun. Videoreasonbench: Can MLLMs perform vision-centric complex video reasoning? In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=1Mblo6U8kp.

David Ma, Huaqing Yuan, Xingjian Wang, Qianbo Zang, Tianci Liu, Xinyang He, Yanbin Wei, Jiawei Guo, Zhenzhu Yang, Meng Cao, et al. Scalelong: A multi-timescale benchmark for long video understanding. In *International Conference on Learning Representations*, volume 2026, pp. 154328–154347, 2026.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16488–16498. IEEE, 2024.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *Computational Visual Media*, 2025.

Jean Piaget. *The Construction of Reality in the Child*. Basic Books, 1954.

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

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. LongVU: Spatiotemporal adaptive compression for long video-language understanding. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54582–54599. PMLR, 13–19 Jul 2025.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Zitian Tang, Rohan Myer Krishnan, Zhiqiu Yu, and Chen Sun. Spacewalk-18: A benchmark for multimodal and long-form procedural video understanding in novel domains. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4472–4482, 2026.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Endel Tulving. *Elements of Episodic Memory*. Oxford University Press, 1983.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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

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

Ziyi Wang, Zhi Gao, Boxuan Yu, Zirui Dai, Peiyao Wang, Yuxiang Song, Qingyuan Lu, Jin Chen, and Xinxiao Wu. Vudg: A dataset for video understanding domain generalization. In *International Conference on Learning Representations*, volume 2026, pp. 65064–65091, 2026.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024a.

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

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9772–9781. IEEE, 2021.

Peiran Xu, Sudong Wang, Yao Zhu, Jianing Li, Gege Qi, and Yunjian Zhang. Spatialbench: Benchmarking multimodal large language models for spatial cognition. *arXiv preprint arXiv:2511.21471*, 2025.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643. IEEE, 2025.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. In *International Conference on Learning Representations*, volume 2026, pp. 78185–78225, 2026.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: collision events for video representation and reasoning. In *ICLR*, 2020.

Jiashuo Yu, Yue Wu, Meng Chu, Zhifei Ren, Zizheng Huang, Pei Chu, Ruijie Zhang, Yinan He, Qirui Li, Songze Li, et al. Vrbench: A benchmark for multi-step reasoning in long narrative videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21655–21666. IEEE, 2025.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 9127–9134, 2019.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pp. 543–553, 2023.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*, 2024a.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024b.

Shaolei Zhang, Qingkai Fang, Yang Yang, and Yang Feng. Llava-mini: Efficient image and video large multimodal models with one vision token. In *International Conference on Learning Representations*, volume 2025, pp. 53285–53310, 2025b.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun MA, Ziwei Liu, and Chunyuan Li. LLaVA-video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025c. ISSN 2835-8856. URL https://openreview.net/forum?id=EElFGvt39K.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, et al. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8475–8489, 2025.

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

Appendix Contents

1. **Dataset Statistics:** Discuss the statistics and distribution of the dataset, including the number of videos, QA pairs, and sub-tasks.
2. **Question Templates:** Describe the templates used to generate questions across the different cognitive levels and sub-tasks.
3. **Annotation and Evaluation Protocol:** Specify the verification thresholds, the blind answerable filter, and human baseline protocols.
4. **Experimental Results:** Present detailed experimental results and additional analyses across the evaluated MLLMs.
5. **Limitations:** Discuss the limitations of CogViD and the benchmark construction and evaluation process.

A Dataset Statistics and Taxonomy

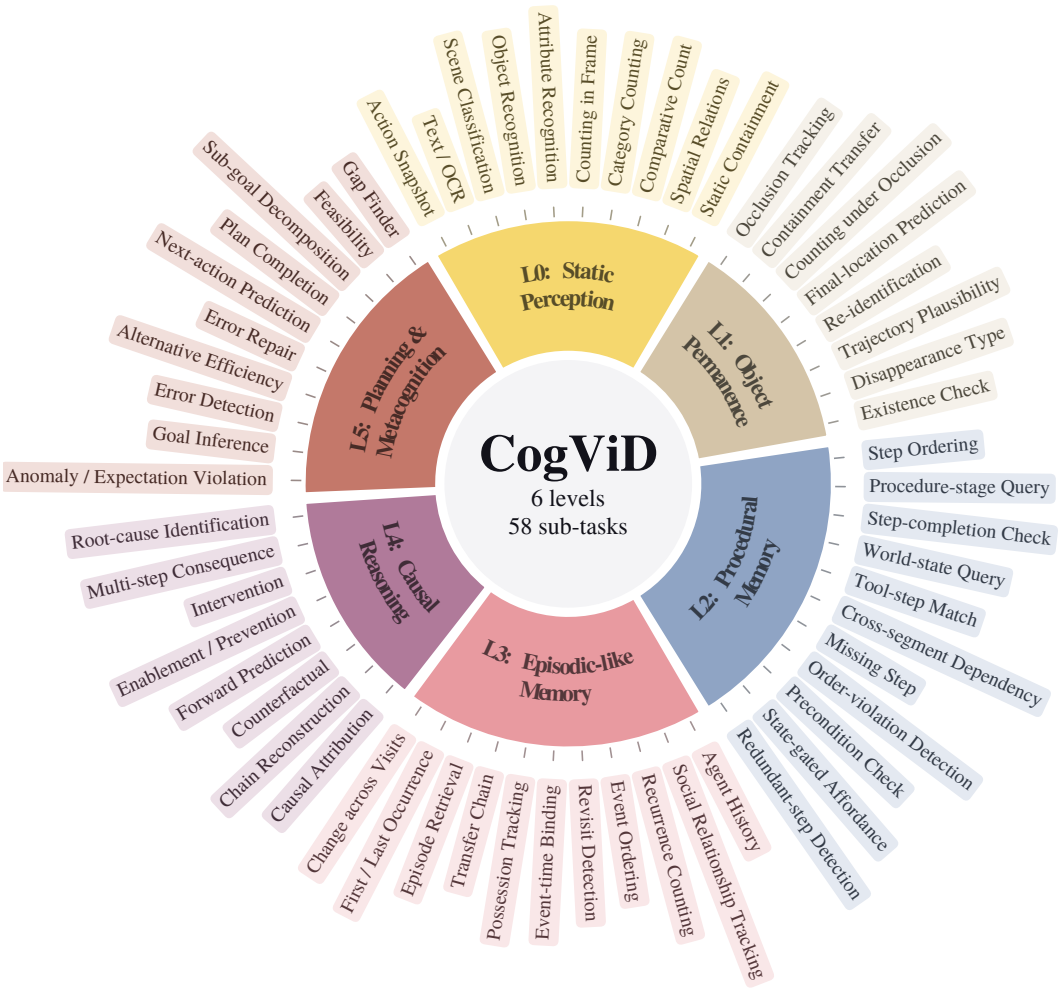


Figure A1: The CogViD cognitive taxonomy: six levels (inner ring) decomposed into 58 sub-tasks (outer ring).

B Question Generation Templates

Each of the 58 sub-tasks in CogViD is defined by a single plain-text template supplied verbatim to the generator. A template fixes the cognitive level, the sub-task, the evidence requirement, the

Table A1: Level 0 with sub-tasks and their definitions. The level focuses on directly observable content in the video.

Level	Subtask	Definition
Static Perception	Action Snapshot	Identify the action a visible agent is performing at a specified moment in the video.
	Text / OCR	Read and transcribe on-screen or scene text visible in a specified frame or segment.
	Scene Classification	Classify the type of location or setting depicted in the video segment.
	Object Recognition	Identify a specific object visible in the frame by its category or name.
	Attribute Recognition	Describe a visible attribute such as color, size, shape, or material of a named object.
	Counting in Frame	Count the number of instances of a specified object visible within a single frame.
	Category Counting	Count how many distinct object categories appear within a specified frame or segment.
	Comparative Count	Compare the counts of two specified object categories and determine which is greater.
	Spatial Relations	Determine the spatial relationship, such as left, right, above, or inside, between two objects.
	Static Containment	Determine whether one visible object is contained within or resting on another at a moment.

Table A2: Requires tracking objects through occlusion, disappearance and re-appearance rather than reading a single frame.

Level	Subtask	Definition
L1 Object Permanence	Occlusion Tracking	Track a specific object's position across a period during which it becomes occluded.
	Containment Transfer	Determine which container holds a specified object after a transfer between containers.
	Counting under Occlusion	Count objects present in a location while accounting for those temporarily hidden by occlusion.
	Final-location Prediction	Predict the final resting location of an object after it leaves view.
	Re-identification	Determine whether an object seen at two separate moments is the same physical instance.
	Trajectory Plausibility	Judge whether a proposed continuation of an object's motion path is physically plausible.
	Disappearance Type	Classify how an object left the scene, such as by occlusion, removal, or exiting frame.
	Existence Check	Determine whether a specified object is still present in the scene at a given moment.

item budget and the output schema. Every template shares the seven rules and the output contract reproduced below; only the task specification differs between sub-tasks. We reproduce one template in full, `comparative_count`, which asks for two counts and a comparison between them rather than one count in isolation, and which is not evaluated by any of the sixteen benchmarks in Table 1. All 58 templates are released with the benchmark.

Shared rules (all levels)

1. **Multi-moment evidence:** every question must draw on at least two distinct moments, frames, shots, or segments.
2. **Temporal dependence:** the answer must depend on temporal change, persistence, ordering, motion, transfer, comparison, causal linkage, plan state, or state update across the video.
3. **Single-frame exclusion:** no question may target an object, action, text, count, or scene attribute that is answerable from one still frame alone.
4. **Grounding:** only facts supported by the provided segment summaries may be used.
5. **Candidates:** multiple-choice items carry exactly four plausible candidates, each grounded in the video.
6. **Open-ended form:** open-ended items carry no candidates, options, or labels.
7. **Labelling:** every item carries `"cognitive_level": "Ln"` and `"subtask": "subtask_name"` verbatim.

Table A3: Subtasks of level 2 with their definition

Level	Subtask	Definition
L2 Procedural Memory	Step Ordering	Arrange a set of depicted procedure steps into their correct chronological sequence.
	Procedure-stage Query	Identify which overall stage of the procedure the video is depicting at a given moment.
	Step-completion Check	Determine whether a specified procedure step was fully completed or left unfinished.
	World-state Query	Describe the state of the task-relevant environment at a specified point in the procedure.
	Tool-step Match	Identify which tool was used to carry out a specified procedure step.
	Cross-segment Dependency	Determine how a step performed in one segment depends on a step from another segment.
	Missing Step	Identify a step that is absent from an otherwise depicted procedure sequence.
	Order-violation Detection	Detect a point where two procedure steps were performed in the wrong sequence.
	Precondition Check	Determine whether the condition required before a given step could be performed was satisfied.
	State-gated Affordance	Determine whether an object's current state permits a specified action to be performed on it.
	Redundant-step Detection	Identify a step that was performed unnecessarily or repeated without changing the outcome.

Table A4: Level 3 sub-tasks and their definitions. This level is 40% multiple choice and 60% open-ended.

Level	Subtask	Definition
L3 Episodic-like Memory	Agent History	Summarize the sequence of actions a specified agent performed earlier in the video.
	Social Relationship Tracking	Infer the relationship between two agents based on their interactions across the video.
	Recurrence Counting	Count how many times a specified event or action recurred across the video.
	Event Ordering	Determine the chronological order in which two or more distinct events occurred.
	Revisit Detection	Determine whether an agent returned to a previously visited location or object.
	Event-time Binding	Determine the approximate time or segment at which a specified event occurred.
	Possession Tracking	Track which agent currently possesses a specified object across multiple scene changes.
	Transfer Chain	Reconstruct the sequence of agents or locations an object passed through over time.
	Episode Retrieval	Retrieve details of a specific earlier episode matching a given description or cue.
	First / Last Occurrence	Identify the first or last moment at which a specified event or object appeared.
	Change across Visits	Identify what changed in a location between two separate visits to it.

Shared output contract (all levels)

- Valid JSON only, with no surrounding prose.
- Each item carries `question_format`, set to either `mcq` or `open_ended`.

Template: `comparative_count` (L0, Static Perception)

Task. From the segment summaries below, write multiple-choice questions that compare the number of instances of two named object categories observed at two different moments of the video, and ask which is greater.

Evidence. The two counts must come from two different segments. A question whose two counts can both be read off the same frame is invalid.

Budget. Up to 10 items per video.

Rules.

1. **Multi-moment evidence:** every question must draw on at least two distinct moments, frames, shots, or segments.
2. **Temporal dependence:** the answer must depend on temporal change, persistence, ordering, motion, transfer, comparison, causal linkage, plan state, or state update across the video.
3. **Single-frame exclusion:** no question may target an object, action, text, count, or scene attribute that is answerable from one still frame alone.
4. **Grounding:** only facts supported by the provided segment summaries may be used.
5. **Candidates:** multiple-choice items carry exactly four plausible candidates, each grounded in the video.

Table A5: This level contributes 2,160 QA pairs over 206 videos (14.2 h), 20% multiple choice and 80% open-ended.

Level	Subtask	Definition
L4 Causal Reasoning	Causal Attribution	Identify which earlier action directly caused a specified later outcome to occur in the video.
	Chain Reconstruction	Reconstruct the full causal chain of actions linking an initial cause to its final effect.
	Counterfactual	Infer what would have happened if a specified action had not occurred.
	Forward Prediction	Predict the next event likely to occur immediately following the observed footage.
	Enablement / Prevention	Determine whether a given action enabled or prevented a subsequent event from occurring.
	Intervention	Determine the outcome that would result from a hypothetical external intervention on the scene.
	Multi-step Consequence	Trace how a single action propagated through multiple downstream consequences over time.
	Root-cause Identification	Identify the earliest underlying cause responsible for an observed chain of events.

Table A6: Level 5 sub-tasks and their definitions. This level is 0% multiple choice and 100% open-ended.

Level	Subtask	Definition
L5 Planning & Metacognition	Anomaly / Expectation Violation	Identify the moment at which an event violated a reasonable expectation for the scene.
	Goal Inference	Infer the underlying goal that an agent's observed actions are directed toward achieving.
	Error Detection	Identify a point where an agent's action deviated from the correct procedure.
	Alternative Efficiency	Judge whether an alternative course of action would have achieved the goal more efficiently.
	Error Repair	Identify the action an agent took to correct a previously made error.
	Next-action Prediction	Predict the specific action the agent is most likely to perform next.
	Plan Completion	Determine whether the agent's plan was fully carried out by the end of the video.
	Sub-goal Decomposition	Break down the agent's overall goal into its constituent sub-goals as observed.
	Feasibility	Judge whether the agent's apparent goal is achievable given the current scene conditions.
	Gap Finder	Identify a missing step or resource needed to complete the agent's plan.

6. **Open-ended form:** open-ended items carry no candidates, options, or labels.
7. **Labelling:** every item carries `"cognitive_level": "L0"` and `"subtask": "comparative_count"` verbatim.

Output.

- Valid JSON only, with no surrounding prose.
- Each item carries `question_format`, set to either `mcq` or `open_ended`.

C Main Results

Perception (L0): For most models, recognition in L0 is already strong, while counting remains consistently weak (Tab. A9). Scene Classification sits between 89.5% and 96.9% for every model, whereas Category Counting spans 24.2% (LongVA-7B) to 54.2% (Aria) and Counting in Frame spans 31.2% (Phi-3.5-Vision) to 65.0% (Aria), with LongVA-7B at 38.1% on Counting in Frame. Aria (66.0%), InternVL2.5-26B (63.4%), and MiniCPM-V-4.5 (63.0%) lead the level, yet even these models lose more than 10 points once the question asks how many objects are present. This consistent gap highlights that perception is not one skill, and it suggests that MLLMs excel at naming what is in a frame while they fail to count or compare it. *Further progress on L0 will come from counting and comparison, not from naming scenes.*

Object permanence (L1): Object permanence drops below perception for every model, and the loss is largest on questions that keep a count through an occlusion (Tab. A10). Counting under Occlusion ranges from 30.0% to 59.0%, followed by Occlusion Tracking and Final-location Prediction, while Trajectory Plausibility, Disappearance Type, and Existence Check remain the easiest columns for every model. Phi-3.5-Vision falls to a 40.6% mean, whereas Aria (61.3%), InternVL2.5-26B (60.9%), and MiniCPM-V-4.5 (60.3%) stay close to their L0 scores. This consistent gap highlights that qualitative permanence is easier than quantitative permanence. This suggests that MLLMs excel at saying that an object went away or that a path looks wrong, while *they cannot keep track of which*

Table A7: CogViD statistics by cognitive level and sub-task. **Videos** counts video slots and so double-counts a video used by several sub-tasks; the italic subtotals give distinct videos per level. $|q|$ and $|a|$ are mean question and answer length in words.

Subtask	QA	Videos	MCQ	Open	Med. len.	$\|q\|$	$\|a\|$
L0 — Perception							
Action Snapshot	3033	258	3033	0	1.1m	20.3	9.5
Text / OCR	1127	111	1127	0	1.4m	29.3	6.7
Scene Classification	534	48	534	0	1.8m	21.2	7.9
Object Recognition	340	33	340	0	3.9m	23.1	5.8
Attribute Recognition	291	29	291	0	2.9m	22.0	7.6
Counting in Frame	260	26	260	0	3.2m	26.0	1.7
Category Counting	240	24	240	0	2.8m	22.6	1.1
Comparative Count	240	24	240	0	1.8m	26.8	6.8
Spatial Relations	230	23	230	0	8.5m	19.7	14.6
Static Containment	70	7	70	0	3.9m	24.6	6.4
subtotal	*6365*	*583*	*6365*	*0*	*1.8m*	*22.8*	*8.0*
L1 — Object Permanence							
Occlusion Tracking	260	26	208	52	4m	25.4	9.3
Containment Transfer	250	25	200	50	3m	23.6	11.5
Counting under Occlusion	250	25	200	50	4.2m	29.5	2.7
Final-location Prediction	250	25	200	50	4.1m	29.3	8.1
Re-identification	250	25	200	50	4.7m	26.6	10.5
Trajectory Plausibility	250	25	200	50	2.5m	27.7	14.7
Disappearance Type	220	22	176	44	2.6m	24.1	11.3
Existence Check	190	19	152	38	4.1m	27.2	11.1
subtotal	*1920*	*180*	*1536*	*384*	*3.4m*	*26.7*	*9.8*
L2 — Procedural Structure							
Step Ordering	970	97	582	388	2.2m	22.4	15.2
Procedure-stage Query	390	39	234	156	4.7m	26.9	18.1
Step-completion Check	370	37	222	148	4.1m	25.5	16.2
World-state Query	370	37	222	148	2.1m	24.2	18.2
Tool-step Match	270	27	162	108	3.7m	27.0	9.3
Cross-segment Dependency	250	25	150	100	4m	29.0	19.8
Missing Step	250	25	150	100	3.7m	40.3	10.5
Order-violation Detection	250	25	150	100	3.2m	28.6	18.5
Precondition Check	250	25	150	100	4.8m	24.7	12.5
State-gated Affordance	250	25	150	100	3.7m	26.2	14.9
Redundant-step Detection	170	17	102	68	2.8m	39.8	26.2
subtotal	*3790*	*375*	*2274*	*1516*	*3.2m*	*26.9*	*16.0*
L3 — Episodic Memory							
Agent History	3770	377	1508	2262	1.5m	22.0	17.8
Social Relationship Tracking	970	97	388	582	2.6m	23.8	34.9
Recurrence Counting	560	56	224	336	1.9m	21.5	1.6
Event Ordering	540	54	216	324	3.1m	23.7	23.4
Revisit Detection	360	36	144	216	3m	28.2	20.7
Event-time Binding	320	32	129	191	4.8m	23.3	19.5
Possession Tracking	290	29	116	174	2.2m	24.8	19.2
Transfer Chain	260	26	104	156	2.2m	27.4	31.3
Episode Retrieval	250	25	100	150	6.2m	26.2	22.9
First / Last Occurrence	250	25	100	150	3.4m	20.1	7.6
Change across Visits	240	24	96	144	9.2m	22.7	21.0
subtotal	*7810*	*779*	*3125*	*4685*	*2m*	*23.0*	*19.8*
L4 — Causal Reasoning							
Causal Attribution	390	39	78	312	3.5m	24.8	29.3
Chain Reconstruction	260	26	52	208	3.4m	25.7	39.8
Counterfactual	260	26	52	208	3.5m	39.1	24.7
Forward Prediction	260	26	52	208	3.3m	38.2	18.3
Enablement / Prevention	250	25	50	200	3.2m	37.3	36.1
Intervention	250	25	50	200	4.6m	46.3	32.4
Multi-step Consequence	250	25	50	200	4.5m	30.4	41.8
Root-cause Identification	240	24	48	192	3.6m	34.8	31.4
subtotal	*2160*	*206*	*432*	*1728*	*3.7m*	*33.9*	*31.5*
L5 — Goal & Error Reasoning							
Anomaly / Expectation Violation	840	84	0	840	2.2m	38.6	35.8
Goal Inference	570	57	0	570	1.3m	39.2	34.4
Error Detection	370	37	0	370	6.3m	45.0	37.0
Alternative Efficiency	250	25	0	250	5.1m	46.3	38.4
Error Repair	250	25	0	250	4.5m	45.9	45.8
Next-action Prediction	250	25	0	250	5.3m	49.7	30.9
Plan Completion	250	25	0	250	6m	51.4	51.1
Sub-goal Decomposition	240	24	0	240	5.9m	46.6	52.3
Feasibility	220	22	0	220	3m	50.0	40.1
Gap Finder	220	22	0	220	3.4m	48.6	29.8
subtotal	*3460*	*345*	*0*	*3460*	*3.5m*	*44.1*	*38.4*
Total	**25,505**	**2,461**	**13,732**	**11,773**	**2.6m**	**27.6**	**19.1**

Table A8: Content-domain taxonomy used to annotate the source videos. Each video receives exactly one domain label.

Domain	Definition
Instructional & Product Demo	Tutorials, product demonstrations, reviews and promotional material that walk through a procedure or an object.
Conversational & Broadcast Media	Interviews, talk shows, scripted drama and other broadcast material centred on people in conversation.
Animation & Gaming Media	Animated cartoons, machinima and recorded video-game sessions, including on-screen interface content.
Household & Everyday Life	Unscripted domestic and family activity: chores, cooking, childcare and everyday indoor routine.
Performing Arts	Dance, music, stage and street performance recorded for an audience.
Simulated & Synthetic Scenes	Game-engine, simulated-home and procedurally generated footage in which an avatar acts in a synthetic environment.
Sports & Fitness	Athletic and fitness activity, from gym training to competitive and recreational sport.
Driving & Traffic	Dash-cam and in-vehicle footage of driving, traffic and road incidents.
Outdoor Recreation & Nature	Camping, hiking, fishing and other outdoor recreation in natural settings.
Public-Space Surveillance	Fixed-camera footage of shops, streets and transport, including the incidents such cameras are installed to capture.
Travel & Urban Exploration	Walking tours, market visits and travel vlogs that move through an unfamiliar public space.
Unresolved	Videos whose generated context was empty, leaving the content domain undetermined.

Table A9: Performance of open-source MLLMs on **perception (L0)** subtasks of CogViD. Each score is MCQs accuracy (%) and Mean is unweighted average across subtasks.

Model	Action Snapshot	Text / OCR	Scene Classification	Object Recognition	Attribute Recognition	Counting in Frame	Category Counting	Comparative Count	Spatial Relations	Static Containment	Mean
LLaVA-Mini-8B	50.0	50.0	89.7	50.0	52.5	32.4	31.7	50.4	50.0	50.0	*50.7*
LongVILA-7B	60.5	59.1	95.8	56.8	58.8	51.5	47.5	56.2	59.3	58.6	*60.4*
LongVU-Qwen2-7B	56.0	56.9	92.9	55.6	55.4	36.9	32.5	52.1	55.7	55.9	*55.0*
Phi-3.5-Vision	54.1	53.8	91.8	53.1	52.7	31.2	25.4	51.7	53.7	50.5	*51.8*
LongVA-7B	54.7	53.4	89.5	53.8	50.0	38.1	24.2	53.7	51.9	51.6	*52.1*
Flash-VStream-Qwen-7B	55.1	56.2	92.9	52.5	53.6	39.2	29.6	53.7	55.3	52.1	*54.0*
Phi-4-Multimodal	58.7	56.8	93.3	55.7	57.2	31.9	26.3	54.2	57.3	54.3	*54.6*
LLaVA-OV-7B (SI)	57.5	57.0	93.5	55.7	55.6	40.4	35.8	56.7	55.3	55.4	*56.3*
MiniCPM-V-4.5	64.9	62.4	96.6	60.1	61.9	52.4	50.4	55.8	65.0	60.7	*63.0*
Qwen2-VL-7B	57.9	58.0	94.1	55.9	56.7	41.5	33.8	49.6	58.5	54.3	*56.0*
LLaVA-Video-7B	59.6	57.4	94.5	56.6	58.1	40.8	43.3	57.9	56.9	57.0	*58.2*
InternVL2.5-8B	62.9	60.4	96.0	56.9	59.2	50.0	41.2	50.0	58.1	55.9	*59.1*
Qwen2.5-VL-7B	56.0	58.2	92.6	54.9	57.3	33.1	35.0	44.2	52.2	56.4	*54.0*
Aria	62.6	65.0	94.9	65.0	65.0	65.0	54.2	58.8	64.2	65.0	*66.0*
LLaVA-OV-72B	63.8	61.5	96.0	59.6	60.2	50.3	43.3	57.1	60.5	57.5	*61.0*
InternVL2.5-26B	65.0	62.5	96.9	59.7	62.8	55.5	53.3	56.2	62.8	59.6	*63.4*
LLaVA-Video-72B	64.3	62.1	96.2	58.7	61.0	52.9	50.8	61.3	64.2	58.0	*63.0*

object it was or how many are left once it is out of sight. Counting under occlusion is the hardest test at this level.

Procedural structure (L2). Procedural structure separates models far less than the memory levels do (Tab. A11). The model means cover only 15.1 points from LLaVA-Mini-8B (49.3%) to InternVL2.5-26B (64.4%), and the three leaders sit within half a point of one another (InternVL2.5-26B 64.4%, MiniCPM-V-4.5 64.0%, LLaVA-Video-72B 63.9%). Order-violation Detection is the hardest subtask at a 55.9% average, only a few points below the rest of the table. For most models this narrow spread suggests that script-like language priors still help on multiple choice after the blind-answerable filter, such as mixing before baking, so a model need not track the step graph closely to sit near the leaders. *High L2 accuracy should not be read as evidence that a model understands procedures.*

Episodic memory (L3): Episodic memory fails on the questions that ask a model to count or time an event rather than across the level as a whole (Tab. A12). Recurrence Counting has the lowest average of any multiple-choice subtask at 31.0%, where Phi-3.5-Vision scores 15.7%, LLaVA-Mini-8B and Qwen2.5-VL-7B score 19.9%, and even InternVL2.5-26B reaches only 49.5%. First/Last Oc-

Table A10: Performance of open-source MLLMs on **object permanence (L1)** subtasks of CogViD. Each score is MCQs accuracy (%) and Mean is unweighted average across subtasks.

Model	Occlusion Tracking	Containment Transfer	Counting under Occlusion	Final-location Prediction	Re-identification	Trajectory Plausibility	Disappearance Type	Existence Check	Mean
LLaVA-Mini-8B	40.4	41.5	41.5	44.5	45.0	52.7	53.1	50.7	*46.2*
LongVILA-7B	50.1	57.3	50.0	52.5	59.2	61.6	59.5	58.1	*56.0*
LongVU-Qwen2-7B	45.2	53.0	44.5	49.5	52.9	58.3	56.9	53.9	*51.8*
Phi-3.5-Vision	37.5	39.5	30.0	34.0	46.0	50.0	50.0	38.2	*40.6*
LongVA-7B	41.8	42.5	38.5	37.5	41.0	54.6	53.8	50.7	*45.1*
Flash-VStream-Qwen-7B	39.4	50.0	41.0	39.0	50.0	56.1	54.3	50.0	*47.5*
Phi-4-Multimodal	44.2	44.5	38.5	48.5	51.2	58.6	55.0	56.0	*49.6*
LLaVA-OV-7B (SI)	50.0	52.1	39.0	45.5	48.0	56.1	56.7	51.8	*49.9*
MiniCPM-V-4.5	53.2	61.1	53.5	59.5	65.0	65.0	62.4	62.7	*60.3*
Qwen2-VL-7B	50.6	54.3	46.5	47.0	52.5	61.0	56.9	56.3	*53.1*
LLaVA-Video-7B	49.0	49.5	38.5	45.5	51.2	56.1	56.9	53.9	*50.1*
InternVL2.5-8B	51.0	59.9	48.0	54.5	61.2	60.7	59.3	58.1	*56.6*
Qwen2.5-VL-7B	45.7	44.0	34.5	50.0	51.7	56.4	51.7	51.8	*48.2*
Aria	65.0	62.1	55.5	59.5	63.8	54.9	65.0	65.0	*61.3*
LLaVA-OV-72B	52.9	57.7	49.5	57.5	57.1	63.5	60.9	59.5	*57.3*
InternVL2.5-26B	53.5	65.0	59.0	58.5	62.9	64.4	61.9	62.3	*60.9*
LLaVA-Video-72B	53.2	60.3	54.0	60.5	63.3	63.5	61.6	63.0	*59.9*

Table A11: Performance of open-source MLLMs on **procedural structure (L2)** subtasks of CogViD. Each score is MCQs accuracy (%) and Mean is unweighted average across subtasks.

Model	Step Ordering	Procedure-stage Query	Step-completion Check	World-state Query	Tool-step Match	Cross-segment Dependency	Missing Step	Order-violation Detection	Precondition Check	State-gated Affordance	Redundant-step Detection	Mean
LLaVA-Mini-8B	41.2	50.0	50.0	50.0	53.9	51.7	50.0	46.0	50.0	50.0	50.0	*49.3*
LongVILA-7B	59.0	62.2	60.1	61.8	59.3	60.8	59.3	55.4	60.7	61.3	61.0	*60.1*
LongVU-Qwen2-7B	53.5	55.9	57.7	57.6	58.8	56.2	58.8	53.6	56.7	59.6	57.5	*56.9*
Phi-3.5-Vision	48.8	51.5	52.4	52.7	50.0	50.0	54.6	50.0	50.0	52.5	56.0	*51.7*
LongVA-7B	50.8	50.5	55.5	53.4	50.0	55.0	55.2	50.0	53.4	55.8	55.5	*53.2*
Flash-VStream-Qwen-7B	50.0	53.8	54.9	52.5	51.8	55.8	54.2	52.4	52.0	53.3	52.5	*53.0*
Phi-4-Multimodal	51.8	56.3	57.7	57.6	53.1	57.1	57.7	51.8	58.7	55.8	56.5	*55.8*
LLaVA-OV-7B (SI)	54.0	55.1	56.7	58.1	54.9	58.3	60.8	50.6	54.9	60.0	60.0	*56.7*
MiniCPM-V-4.5	63.0	65.0	65.0	64.5	64.1	64.2	63.9	62.6	63.6	64.1	64.0	*64.0*
Qwen2-VL-7B	55.0	58.7	58.1	59.6	57.5	58.8	59.3	53.6	59.6	59.1	61.0	*58.2*
LLaVA-Video-7B	56.3	55.3	57.7	59.3	56.1	60.4	60.8	57.2	57.8	57.9	59.5	*58.0*
InternVL2.5-8B	60.8	62.0	60.5	63.3	58.4	59.2	57.7	57.2	61.1	60.4	62.0	*60.2*
Qwen2.5-VL-7B	53.0	57.9	57.3	56.2	52.6	59.6	58.8	54.2	53.8	56.6	60.0	*56.4*
Aria	65.0	61.2	64.4	65.0	57.1	64.3	63.8	63.8	64.3	64.4	63.8	*63.4*
LLaVA-OV-72B	61.4	60.7	63.0	60.6	64.6	59.6	65.0	63.2	64.0	62.5	60.5	*62.3*
InternVL2.5-26B	64.4	63.7	63.8	64.7	65.0	65.0	62.9	65.0	64.5	64.1	65.0	*64.4*
LLaVA-Video-72B	63.0	62.7	65.0	62.5	65.0	63.8	64.5	63.8	65.0	65.0	62.5	*63.9*

currence and Event-time Binding follow at 49.2% and 51.6%, whereas Social Relationship Tracking, Revisit Detection, and Change across Visits remain consistently between 55% and 65% for the stronger models. This consistent gap highlights that memory in CogViD is not one skill. This suggests that MLLMs excel at tracking who did what and at noticing that a scene has changed, while they cannot say how many times an event happened or when it first or last occurred. *Recurrence Counting remains the clearest unsolved multiple-choice task in CogViD.*

Table A12: Performance of open-source MLLMs on **episodic memory (L3)** subtasks of CogViD. Each score is MCQs accuracy (%) and Mean is unweighted average across subtasks.

Model	Agent History	Social Relationship Tracking	Recurrence Counting	Event Ordering	Revisit Detection	Event-time Binding	Possession Tracking	Transfer Chain	Episode Retrieval	First / Last Occurrence	Change across Visits	Mean
LLaVA-Mini-8B	42.7	50.0	19.9	33.5	44.4	34.9	50.0	47.1	50.0	35.0	52.7	*41.8*
LongVILA-7B	57.5	60.4	43.1	55.4	59.3	52.5	59.8	55.8	58.2	53.7	57.0	*55.7*
LongVU-Qwen2-7B	51.1	56.6	31.5	49.5	57.9	48.8	57.5	56.9	53.0	41.0	53.8	*50.7*
Phi-3.5-Vision	46.3	54.6	15.7	51.3	50.0	45.7	53.5	51.7	43.0	37.0	56.4	*45.9*
LongVA-7B	48.7	51.7	23.1	43.4	51.4	45.7	53.5	52.3	46.0	30.0	50.0	*45.1*
Flash-VStream-Qwen-7B	49.2	55.3	31.5	48.6	53.2	50.3	53.5	56.4	53.8	39.0	52.1	*49.3*
Phi-4-Multimodal	50.0	60.6	18.1	50.3	52.5	50.0	47.4	56.9	49.0	44.0	54.8	*48.5*
LLaVA-OV-7B (SI)	50.0	57.2	19.9	51.6	59.7	54.1	52.9	53.5	55.2	47.0	59.1	*50.9*
MiniCPM-V-4.5	63.8	63.6	43.1	62.4	62.8	56.0	64.5	63.9	62.8	65.0	61.8	*60.9*
Qwen2-VL-7B	52.3	59.5	33.8	51.6	58.2	52.5	58.6	59.2	58.2	47.0	55.4	*53.3*
LLaVA-Video-7B	54.0	56.6	29.6	53.9	59.3	51.3	54.0	51.7	56.0	50.0	55.9	*52.0*
InternVL2.5-8B	59.0	60.2	40.3	58.0	63.6	52.9	63.9	59.8	55.2	58.2	58.0	*57.2*
Qwen2.5-VL-7B	49.9	56.0	19.9	50.0	51.1	45.7	60.4	54.6	54.5	48.0	54.8	*49.5*
Aria	61.7	59.9	28.8	65.0	64.3	65.0	55.6	50.0	61.0	62.3	63.2	*57.9*
LLaVA-OV-72B	60.5	62.9	38.0	60.8	61.0	56.4	61.5	62.7	65.0	57.5	61.8	*58.9*
InternVL2.5-26B	65.0	65.0	49.5	63.4	65.0	57.3	63.9	65.0	60.5	59.8	65.0	*61.8*
LLaVA-Video-72B	60.5	62.9	40.7	62.8	61.0	58.6	65.0	60.4	59.8	62.0	62.9	*59.7*

Causal reasoning (L4). Causal reasoning is the level on which models look most alike (Tab. A13). The model means cover only 11.7 points from Phi-3.5-Vision (51.2%) to MiniCPM-V-4.5 (62.9%), and InternVL2.5-26B (62.1%) stays within 1 point of the leader. Chain Reconstruction remains consistently the hardest subtask at a 55.2% average, yet it still falls only a few points below the other causal columns. This narrow spread suggests that event priors still help on multiple choice after the blind-answerable filter, such as a glass breaking after it falls, so a model need not follow the causal chain closely to sit near the leaders. *High L4 accuracy should not be read as evidence that a model understands causality.*

Planning and Metacognition (L5). Moving from multiple choice to open-ended generation inverts the ranking of the other five levels (Tab. A14). Qwen2.5-VL-7B leads at 53.7% after sitting mid-pack on L0 to L4, and it is strongest on Feasibility (65.0%) relative to Gap Finder (50.5%) and Next-action Prediction (43.2%). InternVL2.5-26B (42.3%), Aria (39.6%), and MiniCPM-V-4.5 (39.2%) follow, while LLaVA-OV-72B drops to 22.3% and LongVU-Qwen2-7B to 17.9%. Flash-VStream-Qwen-7B (0.0%) and Phi-4-Multimodal (7.8%) are generation failures rather than cognitive scores. This consistent gap highlights that neither scale nor strong multiple-choice perception transfers to stating a goal, naming an error, and proposing a repair. *This underscores the considerable room for improvement that remains in CogViD.*

D Qualitative analysis

To complement the aggregate accuracies, we inspect individual model answers, specifically, we use 3 MLLMs chosen to span the leaderboard, namely MiniCPM-V 4.5 (74.7%), LLaVA-Video-7B (63.6%) and LLaVA-Mini-8B (53.2%).

In order to isolate the effect of the cognitive level while holding the video fixed, we build the case study in Fig. A2 around a single clip. Of the 2,461 source videos, only seven questions at two levels, and we therefore use one of them so that the comparison is within-video rather than between-video. Both L1 items come from Trajectory Plausibility, where the first asks for the identity of the ball that starts the chain and the second for the crater it ends in, whereas the L4 item asks for the shape of the causal chain over the same footage. Note, however, that the L4 stem itself names the blue ball and

Table A13: Performance of open-source MLLMs on **causal reasoning (L4)** subtasks of CogViD. Each score is MCQs accuracy (%) and Mean is unweighted average across subtasks.

Model	Causal Attribution	Chain Reconstruction	Counterfactual	Forward Prediction	Enablement / Prevention	Intervention	Multi-step Consequence	Root-cause Identification	Mean
LLaVA-Mini-8B	54.8	51.4	52.4	51.8	52.5	50.0	50.0	55.8	*52.3*
LongVILA-7B	61.0	55.7	60.2	60.6	62.5	62.0	58.4	60.4	*60.1*
LongVU-Qwen2-7B	52.3	54.3	59.5	57.1	60.0	53.0	61.2	54.6	*56.5*
Phi-3.5-Vision	50.0	50.0	53.2	50.0	50.0	50.0	53.8	52.3	*51.2*
LongVA-7B	52.3	52.2	50.0	53.5	53.8	51.0	56.6	50.0	*52.4*
Flash-VStream-Qwen-7B	50.7	52.9	57.9	55.3	57.5	58.0	58.4	55.8	*55.8*
Phi-4-Multimodal	55.5	55.0	59.5	57.9	58.8	61.0	56.6	53.5	*57.2*
LLaVA-OV-7B (SI)	53.1	55.0	58.7	53.5	57.5	59.0	57.5	56.9	*56.4*
MiniCPM-V-4.5	65.0	57.9	65.0	61.5	65.0	63.0	61.2	65.0	*62.9*
Qwen2-VL-7B	59.5	53.6	57.9	57.9	58.8	58.0	59.4	59.2	*58.0*
LLaVA-Video-7B	56.3	53.6	57.1	57.9	57.5	57.0	56.6	58.1	*56.8*
InternVL2.5-8B	60.2	54.3	61.0	58.8	60.0	63.0	65.0	58.1	*60.1*
Qwen2.5-VL-7B	52.3	57.1	57.9	56.2	60.0	58.0	57.5	53.5	*56.6*
Aria	53.7	65.0	64.2	65.0	63.8	62.3	55.6	65.0	*61.8*
LLaVA-OV-72B	60.2	55.7	64.2	60.6	62.5	65.0	59.4	58.1	*60.7*
InternVL2.5-26B	64.2	56.4	62.6	63.3	63.8	62.0	62.2	62.7	*62.1*
LLaVA-Video-72B	61.9	57.9	65.0	61.5	62.5	62.0	59.4	61.6	*61.5*

Table A14: Performance of open-source MLLMs on L5 sub-tasks of CogViD. Each score is open-ended accuracy (%) and Mean is unweighted average across subtasks.

Model	Anomaly / Expectation Violation	Goal Inference	Error Detection	Alternative Efficiency	Error Repair	Next-action Prediction	Plan Completion	Sub-goal Decomposition	Feasibility	Gap Finder	Mean
LLaVA-Mini-8B	42.3	45.6	47.0	38.8	40.8	38.4	35.2	43.3	39.5	37.7	*40.9*
LongVILA-7B	32.1	29.1	31.1	29.2	30.4	24.4	34.8	38.3	41.8	25.5	*31.7*
LongVU-Qwen2-7B	24.2	16.3	23.5	14.4	13.2	16.0	16.4	14.2	29.5	11.4	*17.9*
Phi-3.5-Vision	33.3	31.9	32.7	29.6	30.0	24.0	35.2	40.8	42.7	19.5	*32.0*
LongVA-7B	38.1	30.9	40.0	38.4	36.4	28.8	38.4	42.5	45.9	26.4	*36.6*
Flash-VStream-Qwen-7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	*0.0*
Phi-4-Multimodal	8.6	6.0	10.0	7.2	6.4	8.4	6.0	10.4	9.5	5.9	*7.8*
LLaVA-OV-7B (SI)	33.2	31.6	36.5	34.4	26.4	25.2	36.0	34.6	47.3	28.2	*33.3*
MiniCPM-V-4.5	46.1	37.4	42.4	41.6	34.8	29.2	39.2	38.8	49.5	32.7	*39.2*
Qwen2-VL-7B	25.8	18.1	20.0	16.4	17.6	17.6	19.2	23.3	24.1	10.0	*19.2*
LLaVA-Video-7B	26.8	23.5	23.0	16.4	18.8	16.8	23.6	22.9	36.8	15.5	*22.4*
InternVL2.5-8B	33.1	33.0	39.5	30.4	28.8	28.0	35.2	36.7	40.5	25.5	*33.1*
Qwen2.5-VL-7B	54.5	55.8	58.4	50.8	54.8	43.2	51.2	52.5	65.0	50.5	*53.7*
Aria	42.5	40.0	41.1	41.6	36.4	30.8	40.4	42.4	48.6	32.0	*39.6*
LLaVA-OV-72B	25.5	15.8	22.7	17.2	20.4	18.0	26.0	24.6	39.1	13.6	*22.3*
InternVL2.5-26B	44.0	40.5	45.1	48.8	39.6	31.2	40.8	46.7	49.1	36.8	*42.3*
LLaVA-Video-72B	37.4	32.5	35.9	34.4	34.8	32.0	41.2	41.7	47.3	26.4	*36.4*

every option begins with it, so the identity that L1 probes is supplied to the model for free. This is consistent with the higher text-only Blind baseline at L4 than at L1.

Turning to the per-level view in Fig. A3, we show one item per level over separate clips. Since content variation would otherwise confound the comparison, we hold the domain fixed at Instructional across the five examples. Furthermore, we pick items that span the range of outcomes rather than sampling at random, and the figure should therefore be read as an illustration of what each level



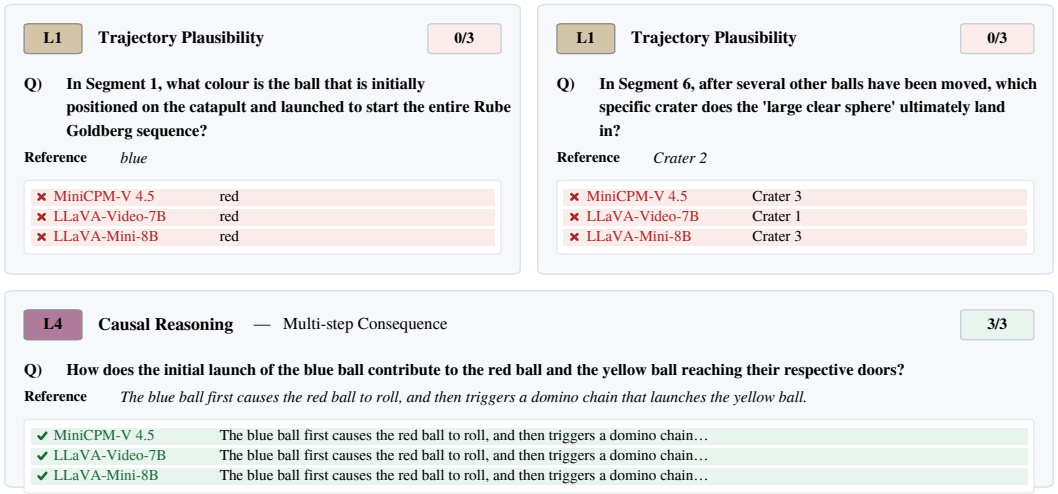


Figure A2: Two object-permanence questions and one causal question asked on the same Rube Goldberg clip, with the answers of three video MLLMs.

asks for and what the errors look like, not as a difficulty ordering. Finally, L5 is omitted because the stored per-question outputs cover L0 to L4 only, with the OEQs level being scored through a separate protocol.

E Limitations and Future Work

CogViD annotates temporal questions on real-world videos through a two-model pipeline. Gemini 3 Pro first writes per-segment and full-video summaries, and Claude Opus then formulates each question from those summaries and a subtask template. The question writer therefore never sees the pixels, only the text that the summarizer chose to record. We audit 500 segments and find that 94.8% of the written descriptions match the source, but that check asks whether the text is correct, not whether every visible motion or spatial arrangement was written down. Fine-grained kinematics and subtle layouts that a caption rarely names can therefore be missing from the question pool even when they are present in the clip. Human reviewers still watch the video and drop questions that the footage cannot support, yet what gets asked is filtered by this text stage. We identify this as a limitation of our work and plan to explore video-native question generation, so that later versions of CogViD can probe visual detail that does not first have to survive a summary.

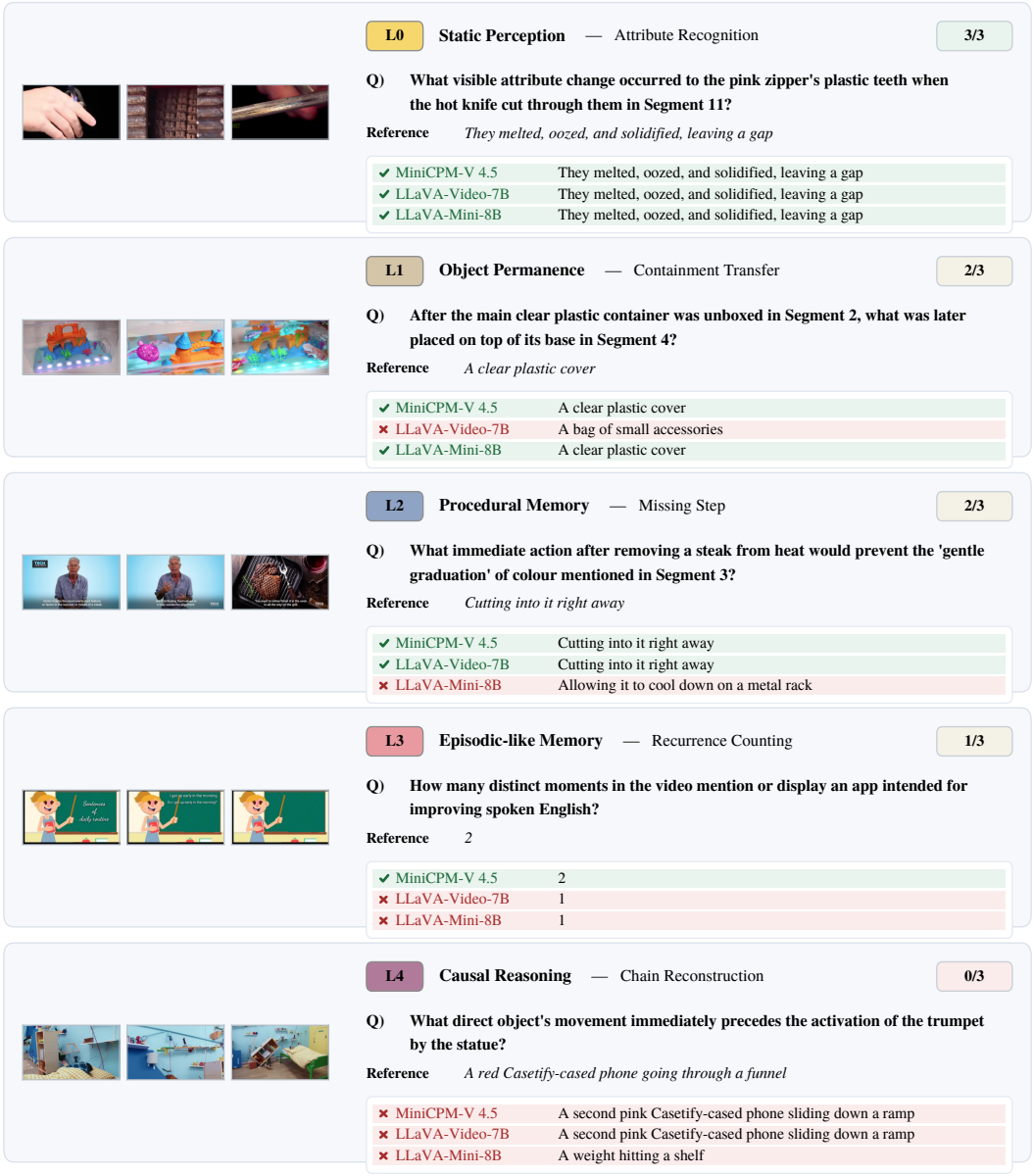


Figure A3: One Instructional clip per cognitive level with the question, the reference answer and the predictions of three video MLLMs.