VIDEOCOGQA: A CONTROLLABLE BENCHMARK FOR COGNITIVE ABILITIES IN VIDEO-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large video-language models (LVLMs) have made rapid progress on video understanding, but their ability to reason systematically about abstract, symbolic, and temporally evolving visual worlds remains poorly characterized. Existing benchmarks are commonly built from natural videos, where semantic context, uncontrolled complexity, and imperfect task isolation make it difficult to attribute errors to specific capabilities. We introduce **VideoCogQA**, a controllable benchmark for diagnosing cognition-inspired video reasoning. A programmatic game-based generator creates videos with explicit control over object attributes, temporal dynamics, scene structure, and difficulty. The resulting benchmark contains 800 videos and 3,280 multiple-choice questions spanning ten scenarios and six abilities: object perception, action perception, temporal reasoning, spatial reasoning, game-environment perception, and multimodal symbol grounding. Across 14 representative LVLMs, performance remains far below human accuracy (61.2% versus 90.9% for the strongest model, GPT-4o). Even Qwen2.5-VL-72B, a strong open model, obtains 54.1% overall accuracy; Qwen2-VL-72B falls from 63.3% on easy instances to 44.4% on hard instances. Analyses using regenerated data, varying frame budgets, and transfer experiments further support the benchmark's utility as a controlled diagnostic complement to natural-video evaluation. VideoCogQA reveals persistent weaknesses in symbolic perception and spatiotemporal reasoning, providing a scalable benchmark for progress.

1 INTRODUCTION

Large video-language models (LVLMs) have substantially improved video perception and question answering (Li et al., 2023; Zhang et al., 2023a; Lin et al., 2023; Li et al., 2024b; Ye et al., 2024; Tang et al., 2023; Hu et al., 2026; Li et al., 2026). However, high scores on natural-video benchmarks do not by themselves establish that a model can systematically reason about symbolic entities, abstract relations, or temporally composed events (Hagendorff et al., 2023). Natural videos couple many factors: semantic priors can reveal an action, scene composition cannot be independently varied, and the source of an error is often ambiguous. Consequently, a failure may reflect perception, memory, symbolic grounding, or question ambiguity; a success may exploit contextual shortcuts rather than the target capability. Cognitive science addresses this attribution problem with controlled tasks (Stillings, 1995). We adopt the same principle for video-language evaluation: hold task semantics fixed while systematically changing the visual configuration, temporal dynamics, and combinatorial load. This setting asks whether an LVLM can identify abstract object attributes (e.g., size, color, and shape), track motions and events, and reason about spatial and temporal relations when natural-language and world-knowledge shortcuts are unavailable.

We introduce **VideoCogQA**, a controllable and scalable benchmark for this purpose. Its fully programmatic generator creates ten game-inspired scenarios and varies their difficulty through interpretable parameters, such as grid size, object count, motion diversity, and path complexity. The benchmark covers object perception, action perception, temporal reasoning, spatial reasoning, game-environment perception, and multimodal symbol grounding. For example, an action-perception item requires recognizing an action from its visual dynamics, rather than inferring "cooking" simply from contextual cues such as the presence of a kitchen. This design makes the intended evidence and the complexity manipulation explicit. Table 1 positions VideoCogQA relative to existing benchmarks.

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.

Table 1: Comparison of video benchmarks in terms of evaluation scope, synthetic video generation, content controllability, and adjustable task difficulty.

Benchmark	Visual Understanding	Cognitive Reasoning	Synthetic Videos	Controllable Content	Adjustable Difficulty
Video-Bench (Ning et al., 2023)	✓	✗	✗	✗	✗
MMBench-Video (Fang et al., 2024)	✓	✗	✗	✗	✗
AutoEval-Video (Chen et al., 2023)	✓	✓	✗	✗	✗
MVBench (Li et al., 2024c)	✓	✓	✗	✗	✗
Video-MME (Fu et al., 2024)	✓	✓	✗	✗	✗
VideoNIAH (Zhao et al., 2024)	✓	✗	✓	✓	✗
VideoCogQA (Ours)	✓	✓	✓	✓	✓

Our evaluation reveals three consistent patterns. First, all evaluated models remain well below human performance, particularly on symbolic object perception and game-environment tasks. Second, accuracy decreases as controlled complexity increases: for Qwen2-VL-72B, mean accuracy drops by 18.9 points from the easy to the hard split. Third, replacing pixels with structured scene descriptions substantially improves performance, implicating visual extraction and structured grounding as important bottlenecks. Regenerated data and comparisons with natural-video benchmarks provide additional evidence that VideoCogQA is a useful diagnostic complement, while transfer experiments show modest gains after tuning on its training split.

Our main contributions are summarized as follows:

- We develop a programmatic video–QA generation pipeline that explicitly controls scene content, temporal dynamics, and difficulty, enabling scalable task variants.
- We introduce **VideoCogQA**: 800 videos and 3,280 questions across ten game-inspired scenarios and six cognition-inspired capabilities.
- We evaluate 14 LVLMs and identify large human–model gaps, difficulty-sensitive failures, and limitations in symbolic visual grounding. Additional analyses confirm the benchmark's robustness and practical relevance.

2 Related Work

2.1 Video-LMMs and Benchmarks

Recent advances in large multimodal models (Zhang et al., 2023b; Liu et al., 2024b;a; Wang et al., 2024a) have substantially improved multimodal understanding and reasoning, particularly for image-based tasks (Wu et al., 2023a; Fu et al., 2023; Zhang et al., 2024c; Tong et al., 2025). As the field progresses from static images to temporally dynamic videos (Li et al., 2024f), recent LVLMs equipped with visual encoders have demonstrated promising video understanding capabilities (Li et al., 2023; Zhang et al., 2023a; Lin et al., 2023; Xu et al., 2023; Li et al., 2024f; Song et al., 2024; Li et al., 2024b; Ye et al., 2024).

Alongside these developments, numerous benchmarks have been introduced to evaluate LVLMs (Ning et al., 2023; Chen et al., 2023; Fang et al., 2024; Li et al., 2024e; Fu et al., 2024; Li et al., 2024c). For example, MVBench (Li et al., 2024c) evaluates diverse video understanding skills using task-specific videos, while MMBench-Video (Fang et al., 2024) adopts long-form YouTube videos and free-form questions to approximate real-world video understanding. However, existing benchmarks primarily emphasize semantic and contextual understanding, providing limited evaluation of abstract cognitive abilities. For instance, action perception in MVBench may benefit from contextual cues, such as a playground scene suggesting that a person is running. In contrast, our setting uses abstract objects undergoing motions such as bouncing or rotating, requiring models to infer actions directly from visual dynamics rather than semantic context. Moreover, the limited controllability and scalability of real-world videos make it difficult to systematically vary task complexity. To address these limitations, we introduce **VideoCogQA**, a scalable and controllable benchmark for evaluating a broad range of cognition-inspired capabilities in LVLMs.

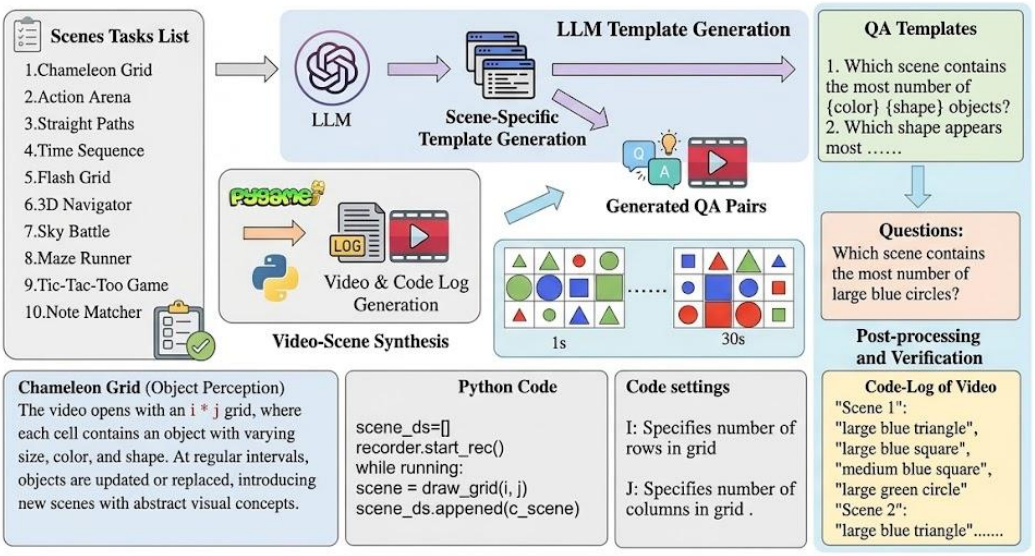


Figure 1: Pipeline for generating videos and corresponding QA templates. The variables i and j control video-scene complexity. A Python program executes the generation process and logs i and j, while GPT-4 produces scene-specific question templates. Finally, we collect the generated code, QA pairs, and videos. Additional cases are provided in Appendix A.4.

2.2 Synthetic Dataset

Synthetic datasets provide a cost-effective alternative to manual annotation (Grauman et al., 2022; Chen et al., 2023), extensive prompt engineering (Li et al., 2024c), and potential data leakage from pre-training corpora (Xu et al., 2024). More importantly, they enable controlled and scalable evaluation of AI models (Peng et al., 2024; Zhao et al., 2024). Synthetic data have been widely adopted in language model evaluation (Maheshwari et al., 2024) and vision-language research, where benchmarks such as CLEVR (Johnson et al., 2017), GQA (Hudson & Manning, 2019), and ARC (Chollet, 2019) use programmatically generated data to systematically assess reasoning abilities. In the video domain, early benchmarks such as CATER (Girdhar & Ramanan, 2019) and CLEVRER (Yi et al., 2019) employ 3D rendering engines to generate synthetic videos for controlled reasoning tasks. More recently, VideoNIAH (Zhao et al., 2024) incorporates textual and visual elements into videos to evaluate retrieval, ordering, and counting abilities. In contrast, **VideoCogQA** provides balanced coverage of six cognition-inspired abilities, with explicit control over task difficulty and fully programmatic, game-inspired scenarios. Our Python-based synthesis framework enables scalable generation of controlled videos, supporting fine-grained evaluation and analysis of LVLM understanding and reasoning beyond existing synthetic video benchmarks.

3 VideoCogQA

3.1 Dataset Design

Recent work on video-language models has increasingly emphasized the evaluation of cognition-inspired capabilities in video understanding (Li et al., 2024e). Common evaluation dimensions include Object Perception (OP), Action Perception (AP), Temporal Reasoning (TR), and Spatial Reasoning (SR). In VideoCogQA, we extend this scope with two additional dimensions: Game-environment Perception (GP) and Multimodal Symbol Grounding (MSG). Our synthesized videos incorporate symbolic objects and abstract concepts, including object attributes (e.g., color, shape, and size), motion properties (e.g., action type, speed, and direction), and spatial and temporal relationships in both 2D and 3D environments. Each dimension is evaluated under controllable levels of task difficulty. The six dimensions are defined as follows.

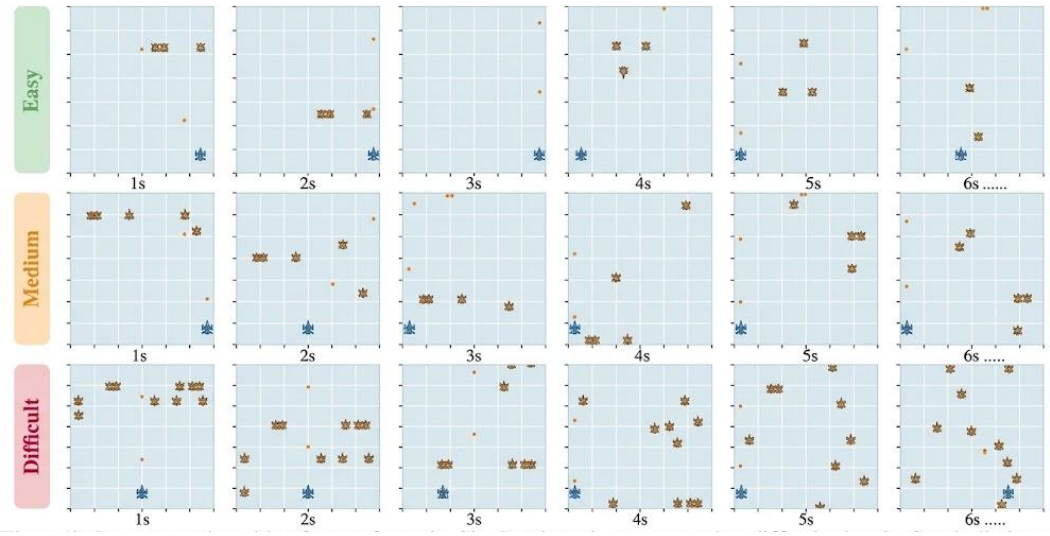


Figure 2: Representative video frames from the Sky Battle task across varying difficulty levels. Symbolic icons are used to represent players, bullets, and enemies. Difficulty is controlled by the varying number of enemies. Questions such as "*How many enemies are destroyed by the player?*" test the temporal counting capabilities of LVLMs in game scenes.

- **Object Perception (OP)**: This dimension evaluates the model's ability to recognize and distinguish symbolic objects with varying visual attributes, including color, shape, and size (Wang et al., 2025). It tests whether models can reliably identify abstract objects across diverse visual configurations.
- **Action Perception (AP)**: This dimension assesses the model's ability to perceive and interpret actions performed by symbolic objects (Chen et al., 2024). The tasks systematically vary motion properties such as action type, speed, and direction.
- **Temporal Reasoning (TR)**: This dimension evaluates the model's ability to understand event sequences and reason about temporal relationships in videos (Chu et al., 2023; Fatemi et al., 2024; Cai et al., 2024). It requires models to track events over time and infer their temporal order and dependencies.
- **Spatial Reasoning (SR)**: This dimension measures the model's ability to reason about spatial relationships in both 2D and 3D environments (Wu et al., 2024b; Tang & Kejriwal, 2024). It covers object position, orientation, distance, and relative spatial relationships.
- **Game-environment Perception (GP)**: This dimension evaluates the model's ability to understand and reason about simulated game environments involving abstract objects and structured dynamics (Wu et al., 2023b; Topsakal & Harper, 2024). It assesses the understanding of game states, interaction rules, action consequences, and overall game dynamics.
- **Multimodal Symbol Grounding (MSG)**: This dimension evaluates the model's ability to ground symbolic concepts across visual, textual, and auditory modalities (Li et al., 2024d). Tasks require models to associate symbolic representations, such as numbers and musical notes, with their corresponding auditory and visual signals, thereby assessing cross-modal understanding and reasoning beyond unimodal perception.

3.2 Automated Video and QA Generation

Guided by the six target abilities, we develop a Python-based synthesis pipeline inspired by game environments. The pipeline renders task-specific scenes with symbolic objects and records the latent scene state and event history. Randomized configurations provide visual diversity, while explicit parameters control video content, temporal dynamics, and difficulty. Question templates are instantiated from the structured logs, so the correct answer is determined by the recorded scene state rather than by a post-hoc language-model judgment. We describe the ten scenes below.

1. **Chameleon Grid (OP-S1)**: An $i \times j$ grid contains symbolic objects with varying attributes, including size (small, medium, large), color (red, green, blue), and shape (triangle, circle, square). The objects change periodically, creating dynamic visual patterns inspired by *Bejeweled* and *Candy Crush*. Task difficulty is controlled by the grid size and object configurations. This scene evaluates object recognition and counting under changing spatial layouts. **Example question:** How many red circles appear across all scenes?
2. **Action Arena (AP-S2)**: A scene contains n symbolic objects performing a types of actions, such as horizontal movement, jumping, scaling, and rotation. Task difficulty is controlled by the number of objects and the diversity of actions. This scene evaluates the ability to perceive and distinguish object actions in dynamic environments. **Example question:** What action is the red ball performing?
3. **Straight Paths (AP-S3)**: n symbolic objects move along linear trajectories, bouncing off boundaries and changing direction at different speeds. Task difficulty is varied by the number of objects and speed configurations. This scene evaluates motion tracking, speed comparison, direction understanding, and trajectory prediction. **Example question:** Which object moves faster, the green circle or the blue circle?
4. **Time Sequence (TR-S4)**: Symbolic objects appear and disappear sequentially at controlled time intervals, accompanied by a simulated clock display. Inspired by *Simon* and *Guitar Hero*, this scene varies task difficulty through the number of objects n and temporal interval t. It evaluates the ability to track event order and reason about temporal sequences. **Example question:** Which object appears immediately after the red circle?
5. **Flash Grid (SR-S5)**: An $i \times j$ grid presents symbolic objects that briefly appear at different locations, inspired by *Memory Matrix* and *Whac-A-Mole*. Task difficulty is primarily controlled by grid size and object configurations. This scene evaluates transient spatial memory and reasoning about relative object positions. **Example question:** For the second occurrence of the object, what is in the cell to its right?
6. **3D Navigator (SR-S6)**: A symbolic object moves along the edges of a 3D structure containing geometric elements such as cubes and pyramids, inspired by *Super Monkey Ball*. Task difficulty is controlled by motion speed t and path complexity e. This scene evaluates 3D spatial understanding, motion tracking, and trajectory reasoning. **Example question:** Which edge of the cube does the red ball traverse most frequently?
7. **Sky Battle (GP-S7)**: As shown in Figure 2, a player-controlled plane interacts with enemy planes and projectiles in a symbolic environment inspired by classic arcade shooters. Task difficulty is varied by the number n and speed a of enemies. This scene evaluates the understanding of game states, interactions, and action consequences. **Example question:** How many enemy planes were destroyed by the player?
8. **Maze Runner (GP-S8)**: A symbolic agent navigates an $i \times j$ maze toward a target location, inspired by classic puzzle games. Task difficulty is controlled by maze size and path complexity. This scene evaluates navigation understanding, path tracking, and reasoning about goal-directed actions. **Example question:** How many steps did the player take to solve the maze?
9. **Tic-Tac-Toe Game (GP-S9)**: A simulated tic-tac-toe game is played on a 3×3 grid. The scene requires models to track game states, understand symbolic interactions, and reason about game outcomes. **Example question:** How many moves did each player make?
10. **Note Matcher (MSG-S10)**: Visual symbols are paired with musical notes represented by numbers from 1 to 7, requiring models to associate auditory signals with their corresponding symbolic representations. Inspired by rhythm-based games such as *Patapon*, task difficulty is varied by the number of notes n and the frequency of visual changes t. This scene evaluates multimodal symbol grounding across visual and auditory information. **Example question:** Which musical note is associated with the appearance of the ball?

We generate videos for each predefined scene using Python, enabling fine-grained control of task difficulty through the parameters illustrated in Figure 1. Each video is approximately 30 seconds long. Repeated execution randomizes attributes such as color, size, and shape while retaining the requested difficulty setting. GPT-4 is used only to propose scene-specific question templates from

Table 2: Difficulty settings for representative scenes. Full configurations are provided in Appendix A.5.

Scene	Parameter	Easy	Medium	Hard
Chameleon Grid	Grid size ($I \times J$)	2×2	5×5	8×8
Action Arena	Objects (N)	3	6	9
	Action types (A)	3	6	8

the corresponding Pygame code (the prompt is given in Appendix A.5); structured scene logs instantiate those templates and determine the answer. Each item has 3–5 answer choices whose order is randomly shuffled. **VideoCogQA** contains 800 videos and 3,280 questions across ten scenarios and six abilities. Each scene contributes roughly 300 questions, with about 100 questions per difficulty level. This separation between generative variation, logged scene state, and answer construction supports reproducible controlled evaluation. Appendix A.4 provides additional examples and extensions enabled by the framework.

3.3 What the Controlled Setting Measures

The central design choice in VideoCogQA is to treat a video question as a composition of observable factors rather than a single opaque example. A scene specifies its entities, their attributes, the event sequence, and the relational query. Difficulty parameters then alter the amount of visual state that must be represented or composed. Increasing a grid size increases the number of candidate entities; increasing the number of action types makes motion discrimination less separable; shortening an event interval places a greater burden on temporal resolution; and increasing maze or trajectory complexity extends the state that must be tracked. The intended prediction target remains stable within a task family even as these sources of complexity change.

This formulation enables diagnostic comparisons that are rarely possible with natural videos. If performance falls as the number of distractor objects grows while the answer type remains unchanged, the change is evidence about robustness to visual set size rather than about a change in language prior. Likewise, comparisons between raw video and a text description of the same logged scene distinguish the need to extract visual state from the need to reason over that state. The generator also permits resynthesis: a task family can be regenerated with new attribute configurations without changing its underlying program or question semantics. We use this property in our regenerated-set analysis in Section 4.5.

The control is deliberately partial. Difficulty levels may change more than one factor in a given scene, so they should not be interpreted as a causal estimate for any individual parameter. Moreover, synthetic scenes do not reproduce the appearance variation or world knowledge required by natural video understanding. VideoCogQA therefore measures controlled symbolic and spatiotemporal competence, not general video intelligence. Its role is complementary: natural-video benchmarks measure performance in realistic settings, whereas VideoCogQA makes the evidence needed for a specific reasoning operation explicit and reproducible.

4 Experiment

4.1 Setup

We evaluate 12 open-source LVLMs—MiniCPM-V (Yao et al., 2024), Video-LLaMA2 (Cheng et al., 2024), InternVideo2 (Wang et al., 2024c), Video-LLaVA (Lin et al., 2023), Video-R1 (Feng et al., 2025), LongVILA-R1-7B (Chen et al., 2026), LLaVA-OneVision (Li et al., 2024a), LLaVA-NEXT-Video-7B/34B (Zhang et al., 2024b), InternLM-XComposer-2.5 (Zhang et al., 2024a), Qwen2-VL-72B (Wang et al., 2024a), and Qwen2.5-VL-72B (Bai et al., 2025)—and two proprietary models, Gemini-Flash-2.0 and GPT-4o. We use each model's official inference implementation and report the corresponding frame configuration; model-specific details are provided in Appendix A.6. All multiple-choice prompts append "`Best Option:`", following prior work (Li et al., 2024c). Accuracy is the percentage of questions answered with the correct option; invalid or free-form re-

Table 3: Comprehensive performance evaluation of Large Video-Language Models (LVLMs) across various scenes (S1–S10). **Bold** indicates the best performance among AI models.

Model	**OP**	**AP**		**TR**	**SR**		**GP**			**MSG**	**Avg.**
	S1	*S2*	*S3*	*S4*	*S5*	*S6*	*S7*	*S8*	*S9*	*S10*	
Open-Source Models											
MiniCPM-V	28.2	49.5	39.3	47.8	32.2	34.7	28.9	26.7	46.0	54.4	38.8
Video-LLaMA2	31.3	50.5	33.5	48.3	36.4	18.7	26.7	27.8	52.0	52.2	37.7
InternVideo2	35.3	36.5	32.5	43.3	33.4	21.7	23.4	23.8	46.0	51.5	34.7
Video-LLaVA	40.4	21.0	40.8	23.2	37.5	21.3	16.7	25.5	38.0	60.0	32.4
Video-R1-7B	46.0	52.0	39.3	60.8	55.8	54.7	46.7	36.7	56.7	72.2	52.1
LongVILA-R1-7B	33.6	38.8	42.3	42.7	34.7	34.7	32.2	30.0	39.3	41.1	36.9
LLaVA-OneVision-7B	42.0	49.2	41.8	51.2	46.4	48.0	46.3	37.8	49.3	56.7	46.9
LLaVA-NEXT-Vid-7B	20.4	22.5	30.7	21.0	33.8	18.7	12.2	14.4	15.3	46.7	23.6
LLaVA-NEXT-Vid-34B	28.4	42.0	42.7	39.0	22.9	58.0	37.8	12.2	33.3	58.9	37.5
InternLM-XComp-2.5	36.0	38.2	45.5	44.5	43.1	20.0	8.9	25.6	35.3	61.1	35.8
Qwen2-VL-72B	**51.8**	58.2	60.0	56.8	60.7	42.0	32.2	37.8	62.0	76.7	53.8
Qwen2.5-VL-72B	48.8	57.2	**61.0**	56.8	61.7	42.0	34.2	38.8	62.5	**78.2**	54.1
Closed-Source Models											
Gemini-Flash-2.0 (32f)	41.2	45.2	41.8	**68.5**	50.0	37.3	27.8	40.0	55.5	71.1	47.8
GPT-4o (32f)	44.2	**64.8**	51.5	61.3	**67.3**	**68.7**	**60.0**	**52.2**	**72.0**	70.0	**61.2**
Average models	37.7	44.7	43.1	46.5	44.0	37.2	31.0	30.7	47.4	60.8	42.3
Human	89.8	91.1	90.8	93.1	88.1	85.7	90.2	94.3	95.9	89.8	90.9

Table 4: Human reference accuracy (%) by programmed difficulty, averaged across annotators and scenes.

Difficulty	**Easy**	**Medium**	**Hard**
Human accuracy	98.2	91.6	81.1

sponses are counted as incorrect. We also report the mean accuracy of five independent human annotators. Humans view continuous videos, whereas LVLMs are evaluated from sampled frames at 2 FPS; human results should therefore be interpreted as a reference rather than a strictly matched comparison. The human-evaluation protocol is detailed in Appendix A.8.

4.2 Human Reference and Difficulty Check

The human study provides an independent check that the programmed complexity settings correspond to perceptual task difficulty. Five annotators completed each sampled item twice with a 48-hour interval; they could replay each video up to three times, and we average their two sessions. Table 4 shows a monotonic decline from 98.2% on Easy to 81.1% on Hard examples. Thus, the difficulty manipulation does not merely produce adversarial failures for models: it also increases human error, while leaving a substantial margin between humans and LVLMs. Because human participants view continuous video and can replay it, this result is a reference for task solvability rather than a directly matched comparison to frame-sampled models. Per-annotator and per-scene results are reported in Appendix A.8.

4.3 Main Results

VideoCogQA exposes large gaps in controlled symbolic video reasoning. Table 3 shows that performance varies sharply across abilities. The average model score is lowest on game-environment

Table 5: Performance across different difficulty levels. Models generally show a pronounced decline as task complexity increases.

Model	Diff.	OP	AP		TR	SR		GP			MSG	Avg.
		S1	*S2*	*S3*	*S4*	*S5*	*S6*	*S7*	*S8*	*S9*	*S10*	
MiniCPM-V	Easy	34.7	54.5	50.5	53.5	47.3	36.0	43.3	30.0	46.0	70.0	46.3
	Medium	26.0	48.5	36.5	50.5	30.0	40.0	23.3	26.7	–	43.3	35.8
	Hard	24.0	45.5	31.0	39.5	19.3	28.0	20.0	23.3	–	50.0	31.0
Video-LLaMA2	Easy	41.3	61.0	46.5	50.0	49.3	20.0	26.7	33.3	52.0	70.0	44.8
	Medium	29.3	52.0	26.5	53.0	31.3	18.0	23.3	20.0	–	56.7	34.2
	Hard	23.3	38.5	27.5	42.0	28.7	18.0	30.0	30.0	–	30.0	29.6
LLaVA-Next-34B	Easy	29.3	48.5	53.5	44.0	42.0	62.0	56.7	10.0	33.3	70.0	44.7
	Medium	26.0	40.0	42.5	45.0	18.7	54.0	36.7	6.7	–	56.7	36.9
	Hard	30.0	37.5	32.0	28.0	8.0	58.0	20.0	20.0	–	50.0	31.4
Qwen2-VL-72B	Easy	**61.3**	**65.0**	**67.0**	**64.5**	**76.0**	**52.0**	**40.0**	**63.3**	**62.0**	**83.3**	**63.3**
	Medium	48.7	59.5	62.5	56.0	56.0	34.0	33.3	23.3	–	80.0	50.1
	Hard	45.3	50.0	50.5	50.0	50.0	40.0	23.3	26.7	–	66.7	44.4

Table 6: Performance under varying frame budgets. The last column reports descriptive Pearson correlations with VideoCogQA over the four frame-budget settings.

Dataset	8 Frames	16 Frames	32 Frames	64 Frames	Corr. Coeff. ↑
VideoCogQA (Ours)	50.9	53.0	54.9	55.1	1.0000
LongVideoBench	52.6	56.8	58.5	59.0	0.9836
VideoMME	53.6	58.8	60.6	63.2	0.9775
LVBench	33.7	36.5	39.6	40.0	0.9984

perception (36.4% across S7–S9) and spatial reasoning (40.6% across S5–S6), while multimodal symbol grounding is comparatively easier (60.8% on S10). GPT-4o achieves the strongest overall result (61.2%), but still trails the human reference by 29.7 points. Among open models, Qwen2.5-VL-72B is best overall (54.1%); its aggregate scores are 48.8% on OP, 59.1% on AP, 56.8% on TR, 51.9% on SR, 45.2% on GP, and 78.2% on MSG. Video-R1-7B reaches 52.1%, suggesting that reasoning-oriented training may be promising, though this comparison does not isolate architecture or training data as causal factors. Overall, the results show that strong general video models remain unreliable when success requires extracting and composing abstract visual state.

More experimental results are in the Appendix. We present a case study of the Maze Runner task in Appendix A.1, report model performance across video difficulty levels in Appendix A.2, and analyze the impact of model size in Appendix A.3.

4.4 What Limits Performance?

The text-description intervention separates access to structured scene state from visual extraction. When GPT-4 receives a description derived from scene logs rather than raw video, its accuracy increases substantially (Figure 3 in Appendix). This result does not prove that perception is the only limiting factor, but it identifies structured visual extraction as an important bottleneck. We probe this issue with Chameleon Grid (S1): given a single frame containing an $i \times j$ grid, the model identifies each object's size, color, and shape in reading order. The models frequently confuse size (Figure 4 in Appendix), despite the absence of temporal demands. Together, these diagnostics suggest that failures on complex video tasks can originate before high-level reasoning, in the representation of symbolic attributes and structured visual state.

Table 7: Original versus regenerated-set performance by ability. "New" uses 20 newly synthesized videos per scene; correlations summarize agreement across the three Qwen2.5-VL scales with 128-frame sampling.

Models	OP		AP		TR		SR		GP		MSG	
	Orig.	New	Orig.	New	Orig.	New	Orig.	New	Orig.	New	Orig.	New
Qwen2.5-VL-3B	46.8	47.5	42.5	41.0	50.2	51.0	45.2	46.5	42.3	41.5	52.2	52.5
Qwen2.5-VL-7B	45.6	46.5	48.3	49.0	56.1	57.5	50.2	50.5	44.4	42.5	57.2	57.0
Qwen2.5-VL-72B	51.8	52.5	59.1	61.0	56.8	56.0	51.3	50.0	44.8	43.5	76.7	75.5
Corr. Coeff.	0.9996		0.9984		0.9496		0.9597		0.9308		0.9996	

Table 8: Transfer performance after LoRA tuning on VideoCogQA.

Model	VideoMME	LVB	MLVU	Avg.
Qwen2.5-VL-7B	60.5	57.2	48.2	55.3
+ VideoCogQA tuning	**61.7**	**58.5**	**50.6**	**56.9**
Qwen2.5-VL-3B	54.7	**51.6**	46.2	50.8
+ VideoCogQA tuning	**59.8**	51.0	**48.5**	**53.1**

4.5 Benchmark Evaluation and Analysis

We evaluate whether the benchmark is stable under resynthesis and whether its trends align with natural-video evaluation. We compare VideoCogQA with VideoMME (Fu et al., 2024), LongVideoBench (Wu et al., 2024a), and LVBench (Wang et al., 2024b), and generate a fresh subset of 20 videos per scene (200 videos and 1,000 questions) with newly sampled object configurations. Scores increase with the frame budget on every benchmark (Table 6), and their frame-budget trends are closely aligned. Because these correlations use only four frame budgets, they are descriptive rather than a statistical validation of predictive power. The regenerated set also agrees closely with the original set across the three Qwen2.5-VL scales, supporting robustness to random scene instantiation. Appendix A.10 reports complementary benchmark- and ability-level comparisons, and Appendix A.11 evaluates transfer after fine-tuning on the VideoCogQA training split.

4.6 Transfer to Natural-Video QA

To assess whether skills learned from controlled data transfer beyond synthetic settings, we LoRA-tune Qwen2.5-VL-3B and Qwen2.5-VL-7B on the VideoCogQA training split and evaluate them on three natural-video benchmarks without further tuning. As shown in Table 8, the 7B model improves by 1.2, 1.3, and 2.4 points on VideoMME, LongVideoBench, and MLVU, respectively, yielding a 1.6-point average gain. The 3B model also improves overall, despite a slight drop on LongVideoBench. While limited to two model scales and one tuning setup, these results suggest that controlled symbolic and spatiotemporal training can transfer to natural-video QA.

5 Conclusion

We introduced VideoCogQA, a programmatic benchmark for controlled evaluation of cognition-inspired video reasoning. By varying scene content, temporal dynamics, and complexity independently, the benchmark makes it possible to localize failures that are difficult to identify in natural-video evaluation alone. Across ten scenarios and six abilities, current LVLMs exhibit a substantial gap to the human reference and degrade as task complexity increases. Our interventions further indicate that extracting symbolic visual state is an important obstacle to end-to-end performance. VideoCogQA is intended to complement, rather than replace, natural-video benchmarks: its synthetic setting enables controlled diagnosis, while its transfer and alignment analyses motivate further study of how controlled skills contribute to real-world video understanding. We hope it supports models with more reliable symbolic grounding and structured spatiotemporal reasoning.

AI USE STATEMENT

AI tools were used to assist with language editing and proofreading, including improving grammar, clarity, and overall readability. The ideas, arguments, and final content remain my own, and I reviewed and approved all AI-assisted revisions.

REFERENCES

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*, 2025.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, et al. TemporalBench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024.

Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Xiangdi Meng, Tianyu Liu, and Baobao Chang. Pca-bench: Evaluating multimodal large language models in perception-cognition-action chain. *arXiv preprint arXiv:2402.15527*, 2024.

Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. *arXiv preprint arXiv:2311.14906*, 2023.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, et al. Scaling rl to long videos. *Advances in Neural Information Processing Systems*, 38:172842–172870, 2026.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. URL https://arxiv.org/abs/2406.07476.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. TimeBench: A comprehensive evaluation of temporal reasoning abilities in large language models. *arXiv preprint arXiv:2311.17667*, 2023.

Yang Ding, Xin Lai, Yizhen Zhang, Wei Li, Ruihang Chu, and Yujiu Yang. Videozoomer: Reinforcement-learned temporal focusing for long video reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 20087–20111, 2026.

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. *arXiv preprint arXiv:2406.14515*, 2024.

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. Test of time: A benchmark for evaluating LLMs on temporal reasoning. *arXiv preprint arXiv:2406.09170*, 2024.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. URL https://arxiv.org/abs/2306.13394.

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. *arXiv preprint arXiv:2405.21075*, 2024.

Rohit Girdhar and Deva Ramanan. CATER: A diagnostic dataset for compositional actions and temporal reasoning. *arXiv preprint arXiv:1910.04744*, 2019.

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.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, 3(10):833–838, 2023.

Yupeng Hu, Zixu Li, Zhiwei Chen, Qinlei Huang, Zhiheng Fu, Mingzhu Xu, and Liqiang Nie. Refine: Composed video retrieval via shared and differential semantics enhancement. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2026.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2024a.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024b.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.

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.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Jian Yang, Siwei Wu, et al. OmniBench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272*, 2024d.

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*, 2024e.

Yunxin Li, Xinyu Chen, Baotian Hu, and Min Zhang. LLMs meet long video: Advancing long video comprehension with an interactive visual adapter in LLMs. *arXiv preprint arXiv:2402.13546*, 2024f.

Zixu Li, Yupeng Hu, Zhiwei Chen, Qinlei Huang, Guozhi Qiu, Zhiheng Fu, and Meng Liu. Retrack: Evidence-driven dual-stream directional anchor calibration network for composed video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23373–23381, 2026.

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.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b.

Gaurav Maheshwari, Dmitry Ivanov, and Kevin El Haddad. Efficacy of synthetic data as a benchmark. *arXiv preprint arXiv:2409.11968*, 2024.

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. *arXiv preprint arXiv:2311.16103*, 2023.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.

Kun Ouyang, Yuanxin Liu, Linli Yao, Yishuo Cai, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Conan: Progressive learning to reason like a detective over multi-scale visual evidence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41089–41099, 2026.

Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, and Zuxuan Wu. Synthesize diagnose and optimize: Towards fine-grained vision-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13279–13288, 2024.

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

Neil A. Stillings. *Cognitive Science: An Introduction*. MIT Press, 1995.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*, 2023.

Zhisheng Tang and Mayank Kejriwal. GRASP: A grid-based benchmark for evaluating commonsense spatial reasoning. *arXiv preprint arXiv:2407.01892*, 2024.

Jingqi Tong, Jixin Tang, Hangcheng Li, Yurong Mou, Ming Zhang, Jun Zhao, Yanbo Wen, Fan Song, Jiahao Zhan, Yuyang Lu, et al. Code2Logic: Game-code-driven data synthesis for enhancing VLMs general reasoning. *arXiv preprint arXiv:2505.13886*, 2025.

Oguzhan Topsakal and Jackson B. Harper. Benchmarking large language model (LLM) performance for game playing via tic-tac-toe. *Electronics*, 13(8):1532, 2024.

Han Wang, Yongjie Ye, Yanjie Wang, Yuxiang Nie, and Can Huang. Elysium: Exploring object-level perception in videos via mllm. In *European Conference on Computer Vision*, pp. 166–185. Springer, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024b.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. InternVideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024c.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-Bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023a.

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, 2024a.

Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind's eye of LLMs: Visualization-of-thought elicits spatial reasoning in large language models. *arXiv preprint arXiv:2404.03622*, 2024b.

Yue Wu, Xuan Tang, Tom M. Mitchell, and Yuanzhi Li. SmartPlay: A benchmark for LLMs as intelligent agents. *arXiv preprint arXiv:2310.01557*, 2023b.

Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, and Yan Lu. Retrieval-based video language model for efficient long video question answering. *arXiv preprint arXiv:2312.04931*, 2023.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*, 2024.

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.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 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. *arXiv preprint arXiv:1910.01442*, 2019.

Congzhi Zhang, Zhibin Wang, Yinchao Ma, Jiawei Peng, Yihan Wang, Qiang Zhou, Jun Song, and Bo Zheng. Rewatch-r1: Boosting complex video reasoning in large vision-language models through agentic data synthesis. In *International Conference on Learning Representations*, volume 2026, pp. 140958–140985, 2026.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023a.

Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Haodong Duan, Songyang Zhang, Shuangrui Ding, et al. InternLM-XComposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023b.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Hang Yan, Conghui He, Xingcheng Zhang, Kai Chen, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024a.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024b. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

Zicheng Zhang, Haoning Wu, Erli Zhang, Guangtao Zhai, and Weisi Lin. A benchmark for multi-modal foundation models on low-level vision: From single images to pairs. *arXiv preprint arXiv:2402.07116*, 2024c.

Zijia Zhao, Haoyu Lu, Yuqi Huo, Yifan Du, Tongtian Yue, Longteng Guo, Bingning Wang, Weipeng Chen, and Jing Liu. Needle in a video haystack: A scalable synthetic framework for benchmarking video mllms. *arXiv preprint arXiv:2406.09367*, 2024.

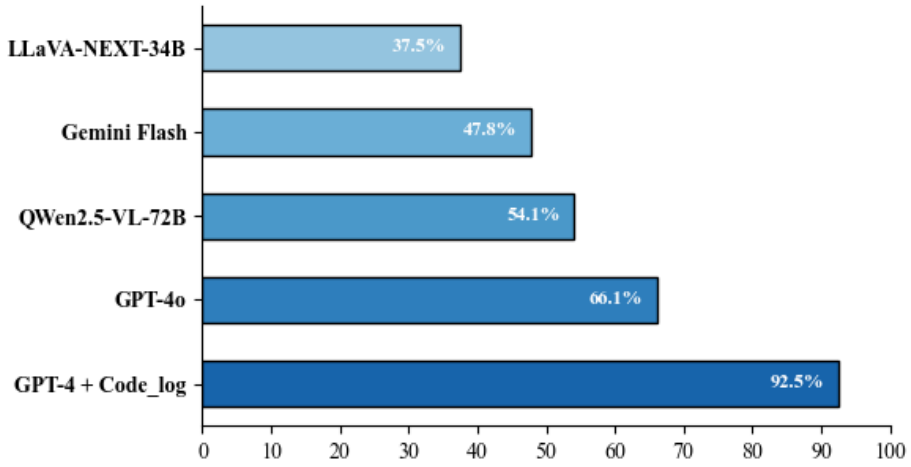


Figure 3: GPT-4 accuracy with raw video input versus a structured text description derived from the corresponding scene logs.

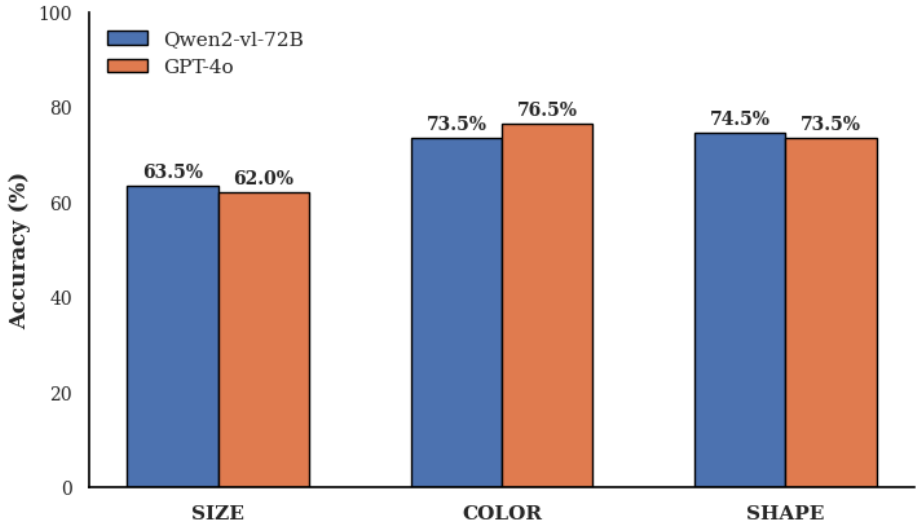


Figure 4: Single-frame S1 performance on object size, color, and shape.

A Appendix

A.1 Case Study

Figure 7 presents a representative example from the *Maze Runner* scene, where the model must track the trajectory of a moving green block and reason about its path. Although straightforward for humans, the task requires accurate spatial perception and temporal tracking. Among the evaluated models, only Qwen2-VL-72B selects the correct answer (option D), while Qwen2-VL-7B and LLaVA-NEXT-Video-34B select option C, suggesting incomplete tracking of the movement trajectory. Notably, Qwen2-VL-2B performs comparably to the larger LLaVA-NEXT-Video-7B in this example, illustrating that model architecture and training can also play an important role beyond parameter scale.

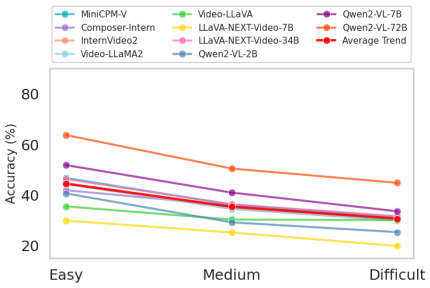


Figure 5: Performance of LVLMs across different levels.

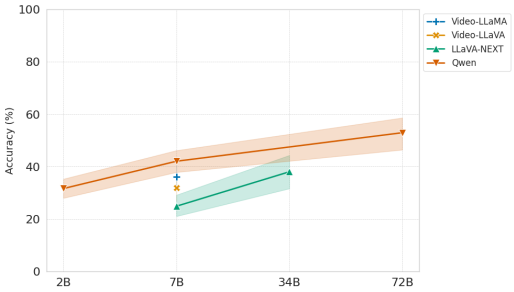
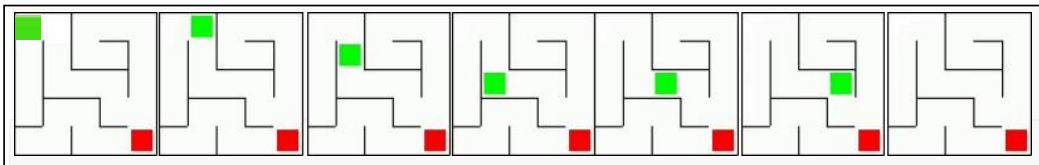


Figure 6: Performance of LVLMs under different parameters.



This is a simulated maze navigation video where the green square represents the player, and the red square marks the maze's endpoint. The player's objective is to navigate through the maze and reach the endpoint, with four possible movement directions: up, down, left, and right.

Question: Describe the player's full sequence of movements, which option is correct ?

A. Down, Left, Left, Down, Left, Down, Left, Down

B. Right, Down, Right, Right, Down, Right, Down, Right, Down

C. Right, Down, Down, Right, Right,Up, Left, Up

D. Right, Down, Down, Right, Right, Down, Right, Down

Best Option:

Model_Output:

Qwen2-VL-72B: The best option is: D. Qwen2-VL-7B: The best option is: C. Qwen2-VL-2B: The best option is: A.

LLaVA-NEXT-Video-34B: The best option is: C. LLaVA-NEXT-Video-7B: The best option is: A.

Figure 7: Maze Runner cases. The task requires describing the trajectory of the green square player navigating the maze to reach the red square goal. Only Qwen2-VL-72B produced the correct answer.

A.2 Performance Across Difficulty Levels

Increasing task difficulty reveals the limited generalization of current LVLMs and enables more fine-grained model comparison. As shown in Table 5 and Figure 5, model performance consistently declines as task difficulty increases, highlighting the challenges of complex cognition-inspired video tasks. On average, most models drop by approximately 10 percentage points from Easy to Medium, followed by a further 5-point decrease at the Hard level. Moreover, difficulty-aware evaluation reveals differences that are less apparent on easier tasks. For example, Video-LLaMA2 and LLaVA-NEXT-Video-34B achieve comparable performance at the Easy level, whereas LLaVA-NEXT-Video-34B increasingly outperforms Video-LLaMA2 as task difficulty grows. These results demonstrate the value of controllable difficulty levels for exposing differences in model robustness and generalization.

A.3 Impact of Model Size on Performance

Larger model scale generally leads to stronger performance on VideoCogQA. As shown in Figure 6, performance increases consistently with model size. For example, scaling Qwen2-VL from 2B to 7B and 72B parameters improves the average accuracy from 31.9% to 42.5% and 53.7%, respectively. This trend suggests that increased model capacity contributes substantially to cognition-inspired video understanding and reasoning.

{} The above is the code to generate a game video using Pygame. The QA template is: {} We want to test {} ability in this scene. Please refine and filter about 10 non-redundant QA templates.

Figure 8: The prompt used to refine QA templates.

1. How many enemies are destroyed by the player's plane in the sky battle scenes video?
2. At what timestamp does the player defeat the first enemy in the sky battle scenes video?
3. What is the maximum number of enemies visible on screen at any time in the sky battle scenes video? .....

Figure 9: Automatically Generated Questions by GPT-4 in the Sky Battle Scene.

A.4 Additional Cases

We provide additional examples of the benchmark scenes in Figures 10, 11, 12, 14, 13, 15, 16, 17, and 18. These examples illustrate the diversity of our synthetic environments and the range of cognition-inspired abilities evaluated by VideoCogQA.

Beyond the ten predefined scenarios, our Python-based framework can be readily extended to synthesize new tasks for evaluating additional capabilities of LVLMs. Representative examples are shown in Figures 19, 20, and 21. These scenarios support controllable difficulty through parameters such as the number and configuration of objects, enabling systematic evaluation of capabilities including pattern recognition, object tracking, and fine-grained visual perception.

Scene	Parameters	Easy	Medium	Hard
Chameleon Grid	I: Number of cells per row J: Number of cells per column	I=2 J=2	I=5 J=5	I=8 J=8
Action Arena	N: Number of objects A: Number of action types	N=3 A=3	N=6 A=6	N=9 A=8
Straight Paths	N: Number of objects A: Number of speed types	N=3 A=3	N=6 A=5	N=9 A=8
Time Sequence	T: Time interval of changes N: Number of objects	T=5 N=3	T=3 N=5	T=1 N=8
Flash Grid	I: Number of cells per row J: Number of cells per column	I=2 J=2	I=5 J=5	I=8 J=8
3D Navigator	T: Time to traverse each edge E: Number of edges	T=2 E=5	T=1 E=8	T=0.5 E=12
Sky Battle	N: Number of enemy planes A: Enemy speed types	N=3 A=2	N=5 A=5	N=10 A=8
Maze Runner	I: Maze length J: Maze width	I=3 J=3	I=5 J=5	I=8 J=8
Note Matcher	T: Time interval of changes N: Number of notes	T=5 N=3	T=3 N=5	T=1 N=7
Tic-Tac-Toe Game	No difficulty levels	None	None	None

Table 9: Detailed parameters for all scene types across difficulty levels.

A.5 Detailed Parameters and QA Templates

Table 9 summarizes the configurable parameters for each scene, while Table 24 presents the corresponding QA templates, covering diverse tasks such as object perception, action understanding, temporal reasoning, and spatial reasoning. To generate QA templates in a scalable yet controllable manner, we adopt a fixed prompt with scene-specific inputs. For each scene, the prompt is instantiated with three components: (i) the complete Python/Pygame code, (ii) a set of seed QA templates, and (iii) the target cognitive ability. As shown in Figure 8, these components fill the three {} placeholders in the core user prompt. GPT-4 is instructed to generate questions grounded strictly in code-defined objects and events, avoid semantic redundancy, and remain aligned with the specified cognitive ability.

A.6 Models

We evaluate a diverse set of LVLMs that span different architectures, model scales, and paradigms.

- **MiniCPM-V** (Yao et al., 2024): MiniCPM-V 2.6 is an 8B-parameter vision-language model supporting single-image, multi-image, and video understanding. It combines efficient multimodal processing with strong OCR, multilingual, and on-device video understanding capabilities.
- **Video-LLaMA2** (Cheng et al., 2024): A 7B-parameter video-language model equipped with a Spatial-Temporal Convolution (STC) connector to capture spatial and temporal dynamics in videos.
- **InternVideo2** (Wang et al., 2024c): A unified video representation framework trained with masked video modeling, video-audio-text contrastive learning, and next-token prediction. We evaluate its 7B configuration for video-language understanding.
- **Video-LLaVA** (Lin et al., 2023): A 7B-parameter model jointly trained on image and video data. It maps visual representations into a shared language feature space, enabling unified image and video understanding.
- **Video-R1** (Feng et al., 2025): A 7B-parameter reasoning-oriented video-language model designed to enhance spatial and temporal video understanding. It follows the R1 paradigm and applies rule-based reinforcement learning to elicit explicit reasoning for video tasks.
- **LongVILA-R1-7B** (**?**): A 7B-parameter model designed for long-video understanding. It adopts a two-stage training pipeline based on chain-of-thought supervised fine-tuning (CoT-SFT) and reinforcement learning (RL) to improve long-range temporal reasoning.
- **LLaVA-OneVision** (Li et al., 2024a): A 7B-parameter multimodal model supporting single-image, multi-image, and video inputs within a unified architecture. Its training strategy facilitates transfer across visual modalities and tasks.
- **LLaVA-NEXT-Video** (Zhang et al., 2024b): A video-language model that extends LLaVA-NEXT to video understanding and uses CLIP ViT-L/14 for visual feature extraction. We evaluate both the 7B and 34B variants.
- **InternLM-XComposer-2.5** (Zhang et al., 2024a): A 7B-parameter multimodal model that represents videos as high-resolution composite visual inputs, enabling unified understanding of images, text, and videos.
- **Qwen-VL** (Wang et al., 2024a): A multimodal model that introduces Multimodal Rotary Position Embedding (M-RoPE) to jointly encode positional information across text, images, and videos. We evaluate the 2B, 7B, and 72B variants to study the effect of model scale on VideoCogQA.

Table 10: Correlation coefficients across benchmarks.

Dataset	VideoCogQA	LongVideoBench	VideoMME	LVBench
VideoCogQA	1.0000	0.9836	0.9775	0.9984
LongVideoBench	0.9836	1.0000	0.9951	0.9902
VideoMME	0.9775	0.9951	1.0000	0.9857
LVBench	0.9984	0.9902	0.9857	1.0000

Table 11: Additional model results on VideoCogQA. Accuracy (%) is reported by difficulty level and overall. The highest score in each column is bolded.

Model	Access	Easy	Medium	Hard	Overall
Gemini-3.1-Pro	Closed	**78.7**	**72.9**	**66.4**	**73.0**
Doubao-Seed-2.0	Closed	71.5	64.3	58.7	65.1
Kimi-K2.6†	Closed	35.8	22.2	16.6	25.4
VideoZoomer	Open	58.5	57.6	54.3	57.0
Conan	Open	59.4	58.1	51.5	56.6
SpaceR	Open	56.3	52.2	45.2	51.4
Rewatch-R1	Open	50.3	48.3	47.5	48.8

†For Kimi-K2.6, answers can be extracted from approximately 41% of responses on Medium and Hard examples because of output-format deviations. The reported score therefore reflects both task accuracy and answer-format compliance.

A.7 Evaluation on Additional Recent Models

To further assess the generalizability of **VideoCogQA**, we extend our evaluation to several recent open- and closed-source models, including Gemini-3.1-Pro, Doubao-Seed-2.0, VideoZoomer (Ding et al., 2026), Conan (Ouyang et al., 2026), SpaceR (Ouyang et al., 2025), Rewatch-R1 (Zhang et al., 2026), and Kimi-K2.6.

The additional models also show lower accuracy on more difficult instances, extending the main difficulty trend to a broader model set. Gemini-3.1-Pro obtains the highest score in this set (73.0%), followed by Doubao-Seed-2.0 (65.1%). Among open models, VideoZoomer is strongest overall (57.0%), closely followed by Conan (56.6%). VideoZoomer decreases from 58.5% on Easy to 54.3% on Hard examples, whereas SpaceR decreases from 56.3% to 45.2%. These differences illustrate why reporting difficulty-stratified results is informative in addition to a single aggregate score.

Kimi-K2.6 obtains substantially lower measured accuracy, particularly on Medium and Hard tasks. We find that this is largely affected by answer-format deviations: its valid answer extraction rate drops to approximately 41% on Medium and Hard examples. Therefore, its reported accuracy should be interpreted with caution and may underestimate its actual reasoning capability. These additional results further demonstrate that VideoCogQA provides a fine-grained diagnostic of both model capability and robustness under increasing cognitive complexity.

A.8 Human Reference and Difficulty Validation

To establish a human reference and assess the difficulty design of VideoCogQA, we conduct a controlled human evaluation.

Participants and evaluation protocol. We recruit five annotators (H1–H5), each of whom completes the full evaluation twice, with a 48-hour interval between sessions. The evaluation covers all 10 scenes (S1–S10), with 20 videos and 2–3 questions per video, resulting in approximately 50 questions per scene. Each scene is divided into three difficulty levels—Easy, Medium, and Hard—with approximately 6–7 videos per level. We report each annotator's performance averaged over the two sessions.

Viewing procedure. Annotators may inspect the questions before or during video playback. Each video can be replayed up to three times, and submitted answers cannot be revised. Unlike LVLMs, annotators view the full continuous videos without frame subsampling.

Performance across scenes and difficulty levels. Table 12 reports the accuracy of each annotator across individual scenes, together with their overall averages. Table 13 summarizes performance across difficulty levels. Human annotators achieve approximately 90.0% overall accuracy on VideoCogQA, substantially exceeding the evaluated LVLMs under their respective input settings. Human performance also decreases from Easy to Hard, consistent with the intended difficulty ordering.

Table 12: Human accuracies (%) by scene and annotator. The last column (AVG) is the per-annotator average over all scenes, and the last row reports the average across annotators for each scene.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	**AVG**
H1	91.2	93.1	88.4	90.7	87.3	86.5	92.8	96.1	97.0	90.4	91.3
H2	88.6	94.0	92.4	94.8	89.6	83.4	88.2	92.4	96.6	88.6	90.9
H3	90.2	89.4	87.6	95.2	90.2	82.6	91.6	91.2	93.8	87.4	89.9
H4	92.2	88.6	93.4	90.2	85.4	87.2	89.6	97.0	97.8	92.4	91.4
H5	86.8	90.6	92.0	94.4	88.2	88.6	89.0	95.0	94.4	90.2	90.9
AVG	89.8	91.1	90.8	93.1	88.1	85.7	90.2	94.3	95.9	89.8	90.9

Table 13: Human accuracies (%) by difficulty level, averaged over all scenes and annotators.

Difficulty	**Easy**	**Medium**	**Hard**
Avg human accuracy	98.2	91.6	81.1

The monotonic performance decline from Easy to Hard suggests that the programmatic difficulty controls in VideoCogQA—such as increasing the number of objects, trajectory complexity, and temporal dependencies—align well with human-perceived task difficulty. Human performance remains below 100%, primarily due to occasional counting errors, fast or ambiguous motion trajectories, and visually similar grid configurations.

Table 14: **Data Distribution.** Distribution of videos and QA pairs across abilities and scenes. The benchmark is balanced across difficulty levels (Easy/Medium/Hard).

Ability	**Scene**	**#Vid**	**#QA Pairs (E/M/H)**
Object (OP)	S1	80	328 (109/109/110)
Action (AP)	S2, S3	160	656 (218/218/220)
Temporal (TR)	S4	80	328 (109/109/110)
Spatial (SR)	S5, S6	160	656 (218/218/220)
Game (GP)	S7–S9	240	984 (327/327/330)
Multimodal Symbol Grounding (MSG)	S10	80	328 (109/109/110)
Total	–	**800**	**3,280**

A.9 Dataset Statistics and Domain-Specific Evaluations

Table 14 provides a comprehensive breakdown of video distributions and QA pairs across the six cognitive dimensions and their respective scenes in VideoCogQA. To further assess model proficiency in dynamic environments, we evaluate three critical game-related capabilities—temporal reasoning, spatial reasoning, and counting—within scenes S7–S9. As summarized in Table 15, even state-of-the-art VideoLLMs encounter significant challenges in these scenarios. These results underscore a substantial performance gap, suggesting that current models still struggle with fine-grained multimodal understanding in complex game-centric tasks.

A.10 Correlation with Real-World Benchmarks

To assess both the diagnostic value and practical relevance of VideoCogQA, we examine its relationship with natural-video benchmarks at two levels: (i) *benchmark-level*, comparing overall scores on VideoCogQA with those on VideoMME, LongVideoBench (LVB), and MLVU; and (ii) *ability-level*, comparing per-ability scores on VideoCogQA (OP, AP, TR, SR, and GP) with the corresponding categories in VideoMME. We evaluate three Qwen2.5-VL models (3B, 7B, and 72B) under a consistent setting of 2 FPS and a maximum of 64 frames per video.

Table 15: Accuracies on three game-related capabilities (temporal reasoning, spatial reasoning, counting) for representative video-language models on the GameWorld scenes (S7–S9) in VideoCogQA.

Model	Temporal	Spatial	Counting
VideoLLaVA	34.4	33.7	25.4
Qwen2-VL-7B	34.4	47.6	33.3
Qwen2-VL-2B	25.0	31.3	28.6
LLaVA-NeXT-Video-34B	37.5	39.9	14.3
Qwen2-VL-72B	28.1	52.9	39.6
Qwen2.5-VL-72B	32.0	53.9	42.0
Gemini	56.0	47.6	25.7
GPT-4o	67.0	59.6	55.7

Table 16: Comparative performance across Qwen2.5-VL model scales. We report the ability-based breakdown on VideoCogQA and VideoMME (top), along with average scores across multiple benchmarks (bottom).

	VideoCogQA			**VideoMME**		
Ability	3B	7B	72B	3B	7B	72B
OP (Object Perception)	43.1	47.8	**48.8**	55.0	62.5	**67.0**
AP (Action Perception)	46.2	53.5	**59.1**	52.0	58.5	**60.0**
TR (Temporal Reasoning)	51.2	**57.0**	56.8	48.0	56.0	**57.0**
SR (Spatial Reasoning)	45.9	50.7	**51.9**	68.1	71.5	**77.0**
GP (Game-environment Perception)	43.9	44.3	**45.2**	30.6	**36.2**	34.7
Overall Average	**46.3**	**50.9**	**54.1**	**54.7**	**60.5**	**64.2**
LVB (Avg)	51.6 (3B)		57.2 (7B)		**61.6 (72B)**	
MLVU (Avg)	46.2 (3B)		48.2 (7B)		**48.2 (72B)**	

Table 18 reports benchmark-level Pearson correlations between VideoCogQA and the three natural-video benchmarks across model scales. Table 16 presents paired ability-wise accuracies on VideoCogQA and VideoMME, while Table 17 summarizes the corresponding ability-level correlations. The correlations are positive at both levels: the three Qwen2.5-VL scales with stronger VideoCogQA scores also tend to score better on the natural-video benchmarks, and the ability-wise trends align with VideoMME. Because these analyses contain only three model scales, they should be treated as descriptive evidence rather than a conclusive estimate of generalization. Nevertheless, the agreement motivates using VideoCogQA as a controlled complement to natural-video evaluation.

Table 17: Pearson correlation coefficients between ability scores on VideoCogQA and VideoMME across three Qwen2.5-VL model scales.

Ability	OP	AP	TR	SR	GP
Correlation r	0.997	0.998	0.998	0.999	0.995

A.11 Transfer to Natural-Video QA After Tuning on VideoCogQA

To examine whether training on the synthetic tasks in VideoCogQA transfers to natural-video QA, we fine-tune Qwen2.5-VL-7B and Qwen2.5-VL-3B on the training split of VideoCogQA using LoRA, denoted as "Qwen2.5-VL-7B + VideoCogQA Tuning" and "Qwen2.5-VL-3B + VideoCogQA Tuning," respectively.

Table 18: Pearson correlation coefficients between average scores on VideoCogQA and three natural-video benchmarks (VideoMME, LongVideoBench, MLVU) using three Qwen2.5-VL scales.

Pair	**Correlation r**
VideoCogQA Avg ↔ VideoMME Avg	0.995
VideoCogQA Avg ↔ LongVideoBench Avg	0.997
VideoCogQA Avg ↔ MLVU Avg	0.913

Table 19: Overall accuracies (%) before and after VideoCogQA tuning. "+T" denotes models fine-tuned on our synthetic benchmark.

Model	**V-MME**	**LVB**	**MLVU**	**Avg**
Qwen2.5-7B	60.5	57.2	48.2	55.3
Qwen2.5-7B + T	**61.7**	**58.5**	**50.6**	**56.9**
Qwen2.5-3B	54.7	51.6	46.2	50.8
Qwen2.5-3B + T	**59.8**	51.0	**48.5**	**53.1**

At the *benchmark level*, Table 19 shows overall improvements after tuning. Qwen2.5-VL-7B improves from 60.5 to 61.7 on VideoMME, from 57.2 to 58.5 on LongVideoBench, and from 48.2 to 50.6 on MLVU, yielding an average gain of $+1.6$ points.

At the *ability level*, Table 21 and Table 20 break down the VideoMME results by category and shows that the largest gains occur on temporal, spatial, and counting questions. For example, Qwen2.5-VL-3B improves from 40.1 to 47.8 on temporal questions and from 30.6 to 35.8 on counting, while Qwen2.5-VL-7B improves from 36.2 to 45.2 on counting. Taken together, these benchmark- and ability-level improvements suggest that training on VideoCogQA transfers beyond its synthetic environments, with particularly clear gains in spatiotemporal and symbolic reasoning.

A.12 Full Results

We report the random-guessing accuracy for each scene in VideoCogQA in Table 22, assuming that a model always selects a valid option uniformly at random. In practice, some LVLMs occasionally fail to produce a valid answer choice, for example by returning free-form text instead of one of the provided options. We treat such responses as incorrect. Consequently, the observed accuracy of a model can fall below the nominal random baseline, which assumes that every response corresponds to a valid option.

Table 23 presents the complete results across all scenes and difficulty levels. These fine-grained results reveal substantial variation across cognitive dimensions and task complexities, providing a more detailed view of each model's strengths and limitations.

Table 20: Category-level accuracies on VideoMME (3B model). Tuning on synthetic data yields consistent gains.

Category	Base	+ VideoCogQA
Object	57.7	**63.0**
Action	52.5	**56.5**
Temporal	40.1	**47.8**
Spatial	68.2	**73.6**
Counting	30.6	**35.8**
Information	71.5	**75.9**
OCR	64.8	**67.6**

Table 21: Category-level accuracies on VideoMME (7B model). The largest gain is observed in the *Counting* category.

Category	Base	+ VideoCogQA
Object	64.5	**65.2**
Action	**59.0**	58.2
Temporal	44.0	**45.7**
Spatial	71.8	**72.7**
Counting	36.2	**45.2**
Information	**76.8**	74.9
OCR	65.5	**69.1**

Table 22: **Random Guessing Accuracies.** Baseline performance (%) across ten scenes. All values approximate 1/3 based on question options.

Scene	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Random	33.3	33.3	33.3	30.3	32.7	28.9	28.9	30.3	33.3	33.3

Table 23: Performance of more models across scenes and difficulty levels.

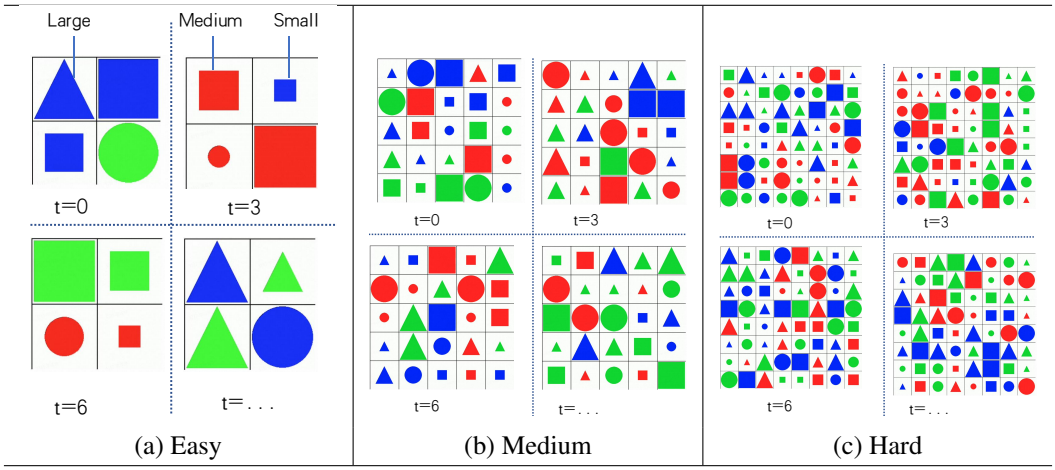
Method	Difficulty	OP	AP		TR	SR		GP			MSG
		S1	*S2*	*S3*	*S4*	*S5*	*S6*	*S7*	*S8*	*S9*	*S10*
MiniCPM-V	Easy	34.7	54.5	50.5	53.5	47.3	36.0	43.3	30.0	46.0	70.0
	Medium	26.0	48.5	36.5	50.5	30.0	40.0	23.3	26.7	—	43.3
	Hard	24.0	45.5	31.0	39.5	19.3	28.0	20.0	23.3	—	50.0
Video-LLaMA2	Easy	41.3	61.0	46.5	50.0	49.3	20.0	26.7	33.3	52.0	70.0
	Medium	29.3	52.0	26.5	53.0	31.3	18.0	23.3	20.0	—	56.7
	Hard	23.3	38.5	27.5	42.0	28.7	18.0	30.0	30.0	—	30.0
Video-LLaVA	Easy	42.7	21.0	36.0	27.5	40.0	28.0	23.3	33.3	38.0	66.7
	Medium	37.3	19.5	46.5	19.0	41.3	18.0	16.7	20.0	—	53.3
	Hard	41.3	22.5	40.0	23.0	31.3	18.0	10.0	23.3	—	60.0
LLaVA-NEXT-Video-34B	Easy	29.3	48.5	53.5	44.0	42.0	62.0	56.7	10.0	33.3	70.0
	Medium	26.0	40.0	42.5	45.0	18.7	54.0	36.7	6.7	—	56.7
	Hard	30.0	37.5	32.0	28.0	8.0	58.0	20.0	20.0	—	50.0
LLaVA-NEXT-Video-7B	Easy	31.3	25.5	38.0	22.5	38.0	16.0	6.7	6.7	15.3	56.7
	Medium	17.3	21.5	32.5	25.5	39.3	16.0	16.7	16.7	—	43.3
	Hard	12.7	20.5	21.5	15.0	24.0	24.0	13.3	13.3	—	40.0
InternLM-XComposer-2.5	Easy	53.3	47.0	52.0	46.5	34.0	20.0	13.3	40.0	35.3	70.0
	Medium	26.7	36.0	46.5	48.5	51.3	22.0	10.0	16.7	—	63.3
	Hard	28.0	31.5	38.0	38.5	44.0	18.0	3.3	20.0	—	50.0
Qwen2-VL-72B	Easy	61.3	65.0	67.0	64.5	76.0	52.0	40.0	63.3	62.0	83.3
	Medium	48.7	59.5	62.5	56.0	56.0	34.0	33.3	23.3	—	80.0
	Hard	45.3	50.0	50.5	50.0	50.0	40.0	23.3	26.7	—	66.7
Gemini-Flash-2.0	Easy	52.5	52.0	52.5	71.5	53.3	38.0	30.0	56.7	55.5	76.7
	Medium	37.6	41.5	41.5	70.0	51.3	38.0	26.7	40.0	—	73.3
	Hard	32.7	42.0	31.5	64.0	45.3	36.0	26.7	20.0	—	63.3
GPT-4o	Easy	48.0	72.0	55.0	67.0	71.3	74.0	70.0	60.0	72.0	76.6
	Medium	45.3	66.5	54.5	64.0	70.7	66.0	56.7	53.0	—	73.3
	Hard	39.3	56.0	45.0	53.0	62.6	66.0	53.3	43.0	—	60.0

Table 24: QA templates for the ten video scenarios in VideoCogQA.

Scene 1: Chameleon Grid
Which scene contains the largest number of {color} {shape} objects?
Which shape appears most frequently in the first column of Scene {i}?
How many {size} {color} {shape} objects appear across all scenes?
What is the total count of {size} {color} {shape} objects across all scenes?
In Scene {i}, which {attribute} is most common among the {size} objects?
In which scene does the grid contain the highest number of objects with the same {attribute}?
In Scene {i}, are there more {size_1} objects or {size_2} objects?
Is there a row in Scene {i} where all objects share the same {size} but vary in {color} or {shape}?
Between Scene {i} and Scene {j}, which shape appears more frequently: {shape_1} or {shape_2}?
Does Scene {i} contain any rows where all objects share the same {color} but differ in {shape}?
Scene 2: Action Arena
How many objects are performing the {action_type} action?
How many distinct actions are performed by the {object}?
Which action has more objects performing it: {action_type1} or {action_type2}?
Which action is most frequently performed across all objects?
Which objects are currently performing the {action_type} action?
Which {action_type} actions appear more than {n} times?
What action is the {object} performing?
Are there any {shape} objects performing the {action_type} action?
Is there any {color} object performing the {action_type} action?
Scene 3: Straight Paths
Which object moves the fastest: {object_a}, {object_b}, {object_c}, or {object_d}?
Which object moves more slowly: {object_a} or {object_b}?
How many times does the {object} change direction after hitting a wall?
How many objects initially move upward or downward when the movement begins?
How many objects initially move left or right when the movement begins?
Is the {object} moving left or right before it first hits a wall?
Is the {object} moving up or down before it first hits a wall?
Scene 4: Time Sequence
What object appears immediately after the object {object}?
How long does the {object} appear?
At what timestamp does the first {object} appear?
What is the color of the object displayed after object {object}?
What is the total duration for which {object} is displayed after {time}?
At what time does the red circle first appear?
In what sequence do the objects appear after {time}, including {time} itself?
For how long does the blue circle appear after {time}, including {time} itself?
Did the blue circle appear at any time after {time}?
Did any object appear after {time}?

Are there any objects that appear both before and after {time}?
Scene 5: Flash Grid
What is the shape and color of the first object that appears?
In which cell position does the first object appear?
Is there any {shape} located in row {i}?
Is there any {shape} located in column {j}?
Is there any {color} object located in row {i}?
Is there any {color} object located in column {j}?
In which row does {object} appear most frequently?
In which column does {object} appear most frequently?
How many unique cell positions are occupied by objects throughout the video?
For the {i}-th occurrence of the object, what is the object in the adjacent cell to the {direction}?
Scene 6: 3D Navigator
Which edge of the cube does the red ball traverse most frequently?
Describe the path of movement taken by the ball.
On which edge does the small ball start moving?
On which edge does the ball conclude its movement?
How many times does the ball reach a vertex during its movement?
How many unique edges does the ball traverse throughout the video?
Scene 7: Sky Battle
How many enemies are destroyed by the player?
At what timestamp does the player defeat the first enemy?
What is the movement pattern of the player's plane?
What is the maximum number of enemies visible on screen at any time?
How many enemies appear throughout the video?
How many times does the player's plane change direction?
Does the player's plane survive until the end of the scene?
Scene 8: Maze Runner
Describe the player's full sequence of movements.
How many steps did the player take to solve the maze?
How many times did the player move upward or downward?
How many times did the player move left or right?
What is the length of the shortest path to solve the maze?
Is there a scene where the player fails to solve the maze within a specified number of steps?
Scene 9: Tic-Tac-Toe Game
Which player made the first move in the game?
Which grid positions remained unplayed by the end of the game?
Which player won the game?
What was the final move made in the game?
What is the total number of moves made during the game?
How many moves did each player make?

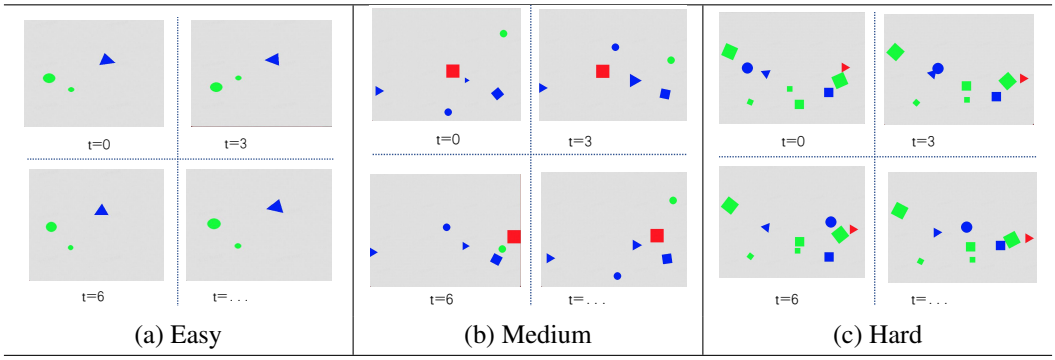
Did either player win by completing a diagonal sequence?
Scene 10: Note Matcher
Which musical note is associated with the appearance of the {object}?
How many different notes were played in total?
Which object appeared most frequently?
What was the first note played?
Was the note "{note}" played at any point?
How many times did the {note} occur alongside a {color} {shape}?
How many times does a {color} object appear alongside a {note}?



(a) Easy (b) Medium (c) Hard

Question: In all the scenes, how many times did the red circle appear?
(A) 0 (B) 14 (C) 26 (D) 18

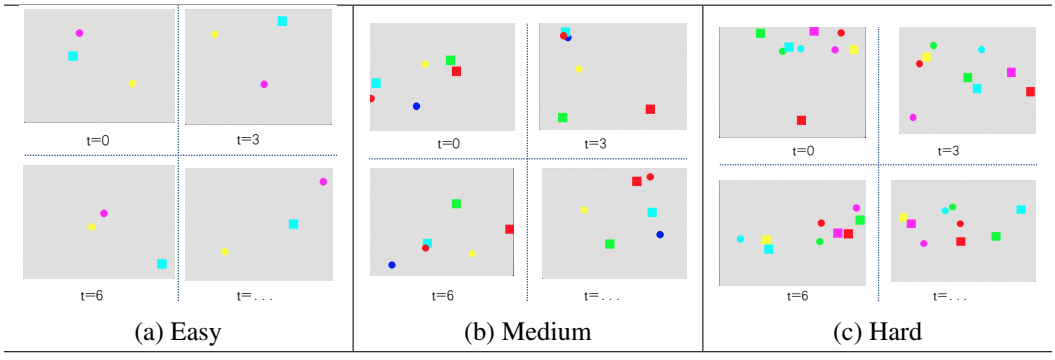
Figure 10: Case of the 'Chameleon Grid' scene at different difficulty levels with the corresponding QA example.



(a) Easy (b) Medium (c) Hard

Question: How many objects are performing the 'bounce' action in this scene?
(A) 0 (B) 2 (C) 6 (D) 8

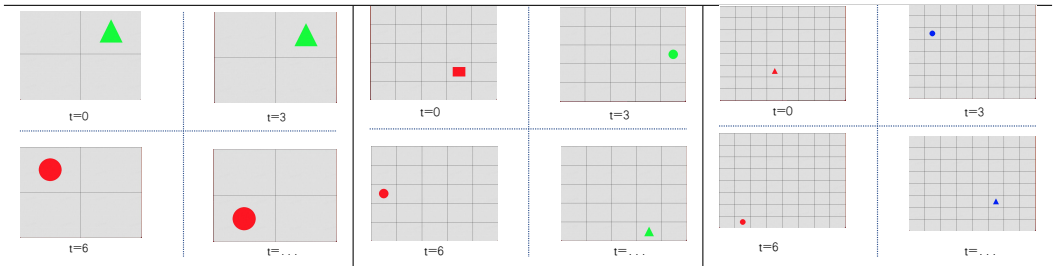
Figure 11: Case of the 'Action Arena' scene at different difficulty levels with the corresponding QA example.



(a) Easy (b) Medium (c) Hard

Question: Which object moves faster: the Green circle or the Blue circle?
(A) Green circle (B) Blue circle (C) Both are equal

Figure 12: Case of the 'Straight Paths' scene at different difficulty levels with the corresponding QA example.

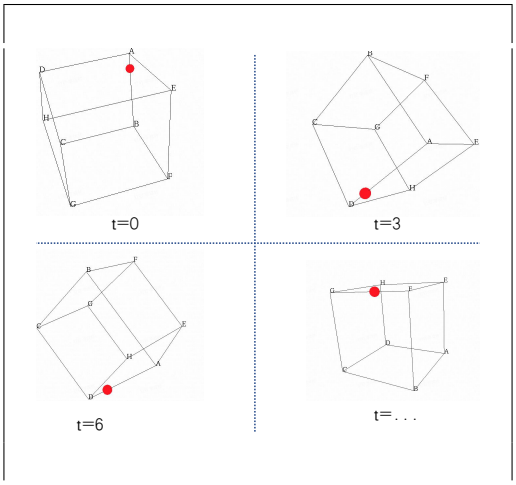


(a) Easy (b) Medium (c) Hard

Question: For the second occurrence of the object, what is the object in the adjacent cell to the left?

(A) Green circle (B) Red circle (C) Blue circle (D) Blue square

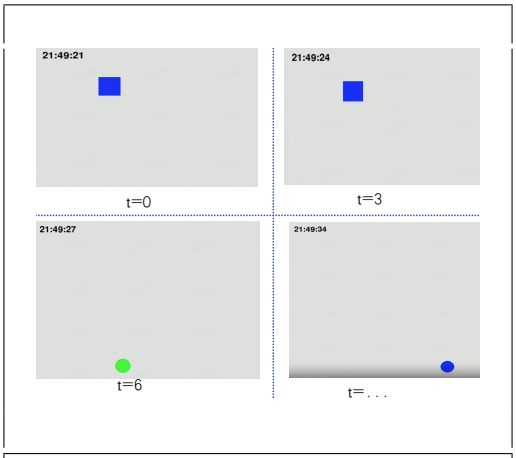
Figure 13: Case of the 'Flash Grid' scene at different difficulty levels with the corresponding QA example.



Question: Which edge has been traversed the most times?

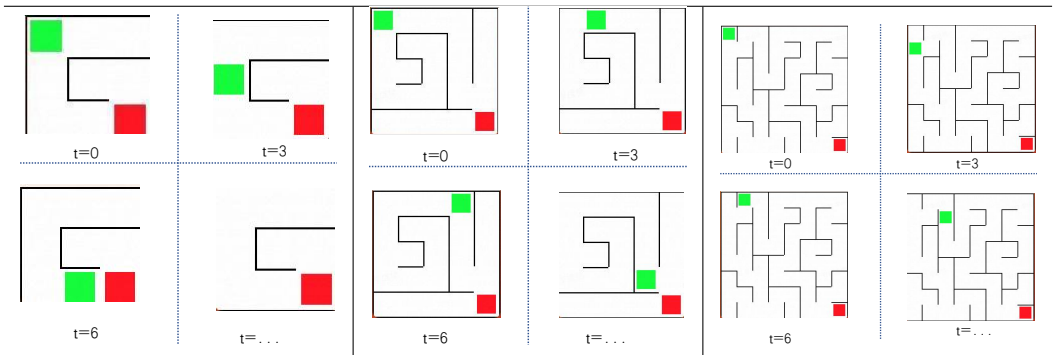
(A) AB (B) BC (C) CD (D) EF

Figure 14: Example of the '3D Navigator' scene at the Easy difficulty level with the corresponding QA example.



Question: What object is after the green circle?
(A) Green circle (B) Blue circle
(C) Blue Square (D) Red Square

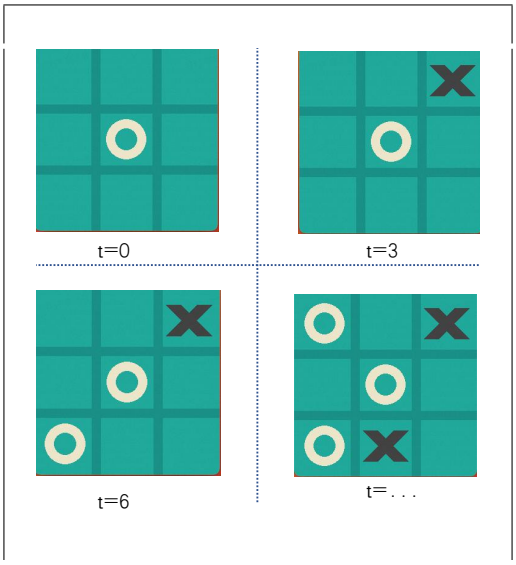
Figure 15: Example of the 'Time Sequence' scene at the Easy difficulty level with the corresponding QA example.



(a) Easy (b) Medium (c) Hard

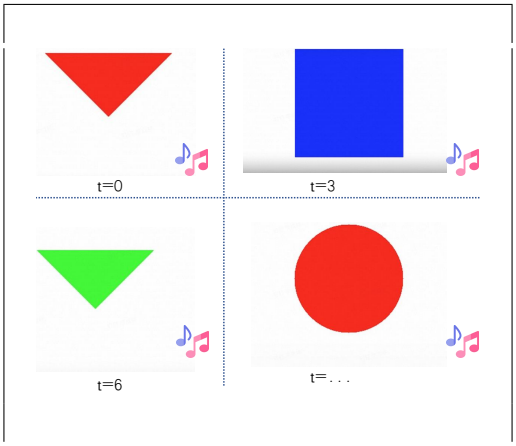
Question: How many steps did the player take to solve the maze?
(A) 8 (B) 9 (C) 4

Figure 16: Example of the 'Maze Runner' scene at different difficulty levels, along with the corresponding QA example.



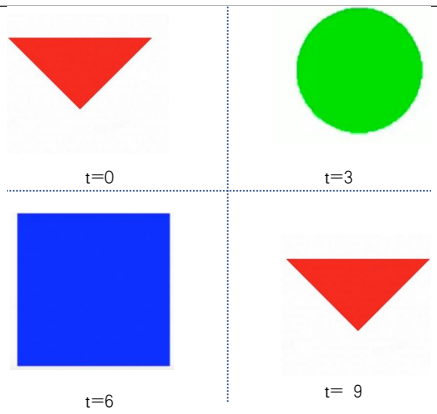
Question: How many moves did each player make?
(A) 2,1 (B) 4,3 (C) 5,2 (D) 3,2

Figure 17: Example of the 'Tic-Tac-Toe Game' scene at the Easy difficulty level with the corresponding QA example.



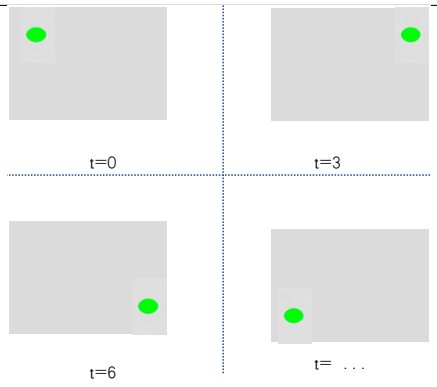
Question: How many different notes were played in total?
(A) 4 (B) 3 (C) 5 (D) 2

Figure 18: Example of the 'Note Matcher' scene at the Easy difficulty level with the corresponding QA example.



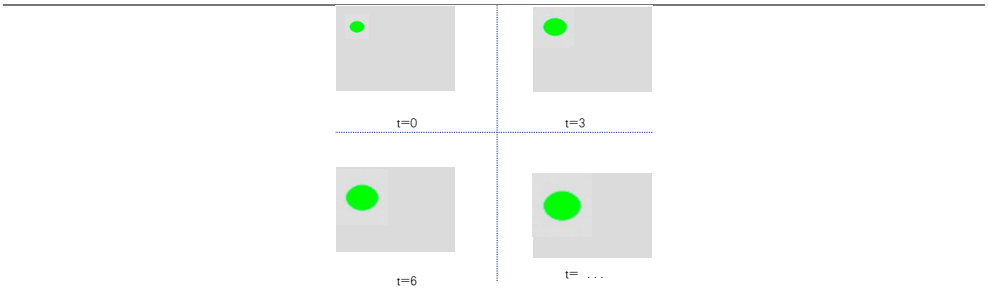
Question: What object will appear on the screen in the next 3 seconds after the entire video?
(A) Green circle (B) Blue circle (C) Green square (D) Blue square

Figure 19: Automatically generated test video and QA example. The model is asked to predict the next object based on previously observed patterns, assessing its temporal pattern recognition and next-event prediction ability.



Question: What is the motion trajectory of the object?
(A) Circle (B) Square (C) Triangle (D) Line

Figure 20: Automatically generated test video and QA example. The question requires inferring the object's motion trajectory from the video, evaluating object tracking and motion understanding ability.



Question: What changes occur in the video content?
(A) The green ball gets bigger (B) The green ball gets smaller (C) No change (D) The green ball gets bigger and then smaller

Figure 21: Automatically generated test video and QA example. The model must identify subtle changes in the object over time (e.g., size variation), assessing fine-grained visual perception of dynamic attributes.