VIDEOZEROBENCH: PROBING THE LIMITS OF VIDEO MLLMS WITH SPATIO-TEMPORAL EVIDENCE VERIFICATION

Anonymous authors
Paper under double-blind review

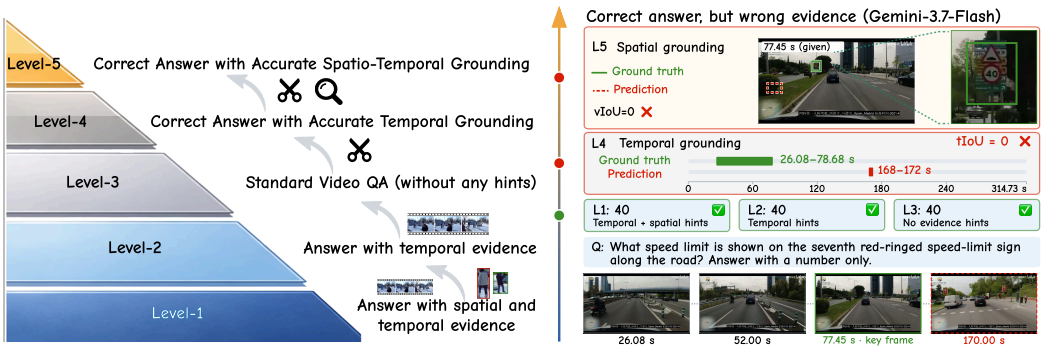

Figure 1: We introduce **VideoZeroBench**, a new challenging long-video understanding benchmark with hierarchical spatio-temporal evidence verification. The example highlights a reliability gap: even the frontier model Gemini-3.7-Flash answers correctly but may still fail to provide accurate spatio-temporal evidence.

ABSTRACT

Video multimodal large language models achieve strong results on existing benchmarks, but answer accuracy alone does not establish whether they can locate the evidence needed to answer a question. We introduce **VideoZeroBench**, a challenging long-video benchmark with manually annotated question–answer pairs spanning 13 video domains. Questions target fine-grained cues, fleeting events, and evidence distributed across multiple segments. Temporal intervals and key-frame boxes are annotated where applicable. All questions undergo two rounds of cross-verification for answer validity and evidence quality. Our **five-level diagnostic protocol** compares answering with and without evidence hints, then combines answer correctness with independently evaluated temporal and spatial grounding. Across 19 evaluated models, the best standard QA accuracy is 24.8% (Level-3), achieved by Gemini-3.7-Flash. No model exceeds 1.8% when correct answers and accurate spatio-temporal localization are jointly required (Level-5). Analyses of atomic abilities, evidence spans, input modalities, and thinking-with-videos inference further characterize where the evaluated systems struggle. These findings motivate more precise evidence search and localization for long-video question answering. Our code and data will be publicly released.

1 INTRODUCTION

Multimodal large language models (MLLMs) (Google, 2026; OpenAI, 2026) have made progress in video understanding, scoring over 80% on Video-MME (Fu et al., 2024a) and LongVideoBench (Wu et al., 2024) (Table 1). However, high QA accuracy can **obscure capability boundaries** in fine-grained perception and long-range reasoning, especially when supporting evidence is not verified. Two concerns remain: **capability**, whether models can locate critical visual details and reason over them, and **reliability**, whether their answers are supported by correct spatio-temporal evidence. Progress requires evaluating both answer correctness and evidence grounding.

Several recent video benchmarks, such as V-STaR (Cheng et al., 2025), ToG-Bench (Xu et al., 2025), Know-Show (Sugandhika et al., 2025), and Dr.V (Luo et al., 2025) incorporate temporal

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 with video and spatial benchmarks. TG/SG: temporal/spatial grounding; MCQ/OE: multiple-choice/open-ended; Anno: annotation source (M: manual, A: automatic). Scores use each benchmark's native protocol and are not directly comparable. The VideoZeroBench QA count denotes distinct questions, and its score is Level-5 accuracy.

Benchmark	Videos	Duration	QAs	Type	Anno	TG	SG	Hierarchic	SOTA(2026.08)
I. General Video MLLM Benchmarks									
Video-MME (Fu et al., 2024a)	900	1017.9s	2.7K	MCQ	M	✗	✗	✗	90.4%
MVBench (Li et al., 2024)	3.6K	16s	4K	MCQ	A&M	✗	✗	✗	78.1%
Video-MMMU (Hu et al., 2025)	300	506.2s	900	MCQ/OE	M	✗	✗	✓	88.7%
LongVideoBench (Wu et al., 2024)	3.8K	473s	6.6K	MCQ	M	✗	✗	✓	80.6%
CG-Bench (Chen et al., 2024)	1.2K	1624.4s	12K	MCQ/OE	M	✓	✗	✗	65.5%
II. Detailed Spatial Benchmarks									
ZeroBench (Roberts et al., 2025) (Image)	-	-	100	OE	M	✗	✗	✗	30%
BLINK (Fu et al., 2024b) (Image)	-	-	3.8K	MCQ	M	✗	✗	✗	81.4%
TreeBench (Wang et al., 2025a) (Image)	-	-	405	MCQ	A&M	✗	✓	✗	71.1%
MMSI-Video-Bench (Lin et al., 2025)	1.3K	72s	1.1K	MCQ	M	✗	✗	✓	38.0%
V-STaR (Cheng et al., 2025)	743	110.23s	2.1K	OE	A&M	✓	✓	✗	46.6%
ToG-Bench (Xu et al., 2025)	100	87.9s	2.7K	OE	A&M	✓	✓	✗	89.4%
VideoZeroBench	138	667.1s	500	OE	M	✓	✓	✓	1.8% (Level-5)

grounding and spatial grounding as additional evaluation dimensions. However, VideoZeroBench differs from these benchmarks in three aspects. First, VideoZeroBench decomposes difficult video QA into temporal evidence search, spatial evidence search, and answer reasoning through the hierarchical design. Second, VideoZeroBench emphasizes fine-grained cues such as fleeting frames, small objects, subtle spatial relations, and non-salient evidence. Third, its videos are longer on average than those in the compared spatio-temporal grounding benchmarks (Table 1), increasing the temporal search space. Detailed comparisons and representative cases are provided in Appendix A.3 and Figure 5. Meanwhile, a growing number of thinking-with-videos approaches (Zeng et al., 2026; Yan et al., 2025; Meng et al., 2025; Ding et al., 2025) adopt iterative evidence retrieval and reasoning. These approaches interleave reasoning with explicit visual grounding to seek key frames, timestamps, or spatial regions. However, existing benchmarks provide limited verification of their spatio-temporal evidence on difficult video QA. Overall, unlike ZeroBench (Roberts et al., 2025) and TreeBench (Wang et al., 2025a) in the image domain, the video domain still lacks a sufficiently challenging benchmark that probes fine-grained spatio-temporal reasoning and verifies evidence grounding. More related works are provided in Appendix A.9.

To address the limitations, we introduce **VideoZeroBench**, a capability-centric, evidence-grounded, and hierarchical benchmark for video understanding. It aims to explore the limits of leading video MLLMs by examining shared weaknesses in answering challenging questions and grounding answers in spatio-temporal evidence. VideoZeroBench contains manually annotated, open-ended questions across 13 video domains, including sports, instructional videos, life vlogs, and games. Its videos average 667.1 seconds, introducing complex scenes and long-range temporal dependencies. The benchmark covers 11 atomic abilities, including counting, small-object perception, spatial orientation discrimination, and multi-segment causal dependency, with many questions requiring multiple abilities. All questions require precise and verifiable answers (e.g., numbers, single words, or fixed short phrases), enabling direct answer comparison without an LLM judge. In summary, **VideoZeroBench has three key features: (1) High difficulty.** Questions are manually annotated and intentionally challenging, designed to expose the capability boundaries of contemporary video MLLMs under fine-grained and complex spatio-temporal conditions. **(2) Evidence-grounded evaluation.** We evaluate answer correctness together with temporal and spatial evidence localization on the same questions. **(3) Manually verified annotations.** PhD-level annotators construct the questions and evidence. Two-round cross-verification checks answer uniqueness and evidence completeness.

To distinguish answer correctness from correct evidence grounding, we also introduce a **five-level** hierarchical evaluation framework (Figure 1). The first three levels evaluate answering under different evidence conditions: **Level-1** provides both temporal and spatial evidence, **Level-2** provides temporal evidence only, and **Level-3** corresponds to standard video QA without hints. The remaining levels require grounding: **Level-4** requires correct answers with accurate temporal segments, while **Level-5** further requires both temporal and spatial localization. The five levels reuse the same question

set, with evidence annotations available where applicable (Section 2.2). This hierarchy compares answering with and without evidence hints and evaluates temporal and spatial predictions for the same questions. It reveals localization weaknesses that answer-only scores cannot distinguish. In Figure 1, Gemini-3.7-Flash answers correctly at Levels 1–3 but fails both temporal and spatial grounding.

Experiments show that answering these questions correctly often does not coincide with accurate spatio-temporal localization. Gemini-3.7-Flash achieves the highest Level-3 accuracy at 24.8%, followed by GPT-5.6-Sol at 18.8%. **Under the strictest Level-5 requirement, which demands both correct answers and correct spatio-temporal grounding, all models score at most 1.8%, with many achieving zero.** Independent human participants reach 63.5% and 55.8% at Levels 4 and 5 on a 50-question subset (Appendix A.5).

We further analyze performance variations and error cases across video categories, atomic abilities, evidence span lengths, thinking-with-videos inference, and input modalities, leading to three conclusions. **First, models may answer challenging questions correctly yet fail to localize the supporting spatio-temporal evidence**. Incorrect grounding predictions remain common (Table 2, Figures 1 and 4). **Second, fine-grained spatial perception and temporal evidence search remain challenging** (Figure 3). **Third, thinking-with-videos inference** improves QA in the VideoChat-R1.5 ablation, but **does not improve grounding accuracy in this setting** (Table 3). These findings highlight the need to prioritize evidence-grounded perception and precise spatio-temporal reasoning as foundations of trustworthy video intelligence. We hope that VideoZeroBench can serve as a rigorous benchmark for diagnosing and advancing the next generation of video MLLMs toward reliable and verifiable understanding.

2 VIDEOZEROBENCH

2.1 DATASET CONSTRUCTION

Video Collection. We manually curate long-duration, structurally complex videos from YouTube and Bilibili, without using existing benchmark datasets. Annotators search by category and manually review candidates, prioritizing uploads from the preceding six months where possible. Videos are included only if they exhibit meaningful event progression, non-trivial scene transitions, and sufficient visual complexity such that critical evidence is not always salient. Short or visually simple clips are excluded. Most retained videos range from 5 to 20 minutes, increasing temporal depth and dispersing supporting evidence across time. As shown in Figure 2, the collection spans 13 domains, including Instructional, Gaming, Sports, Film&TV, Music, Daily Vlogs, Driving, News&Entertainment, Travel, Animals, Humor, Fashion&Beauty, and Animation. Both Chinese and English videos are incorporated to enhance linguistic diversity. The resulting benchmark comprises long, cluttered videos in which key evidence is subtle and temporally distributed.

Labeling Process. The annotation pipeline includes LLM-assisted reference generation followed by fully manual question and evidence construction. We first apply Gemini-2.5-Pro (Comanici et al., 2025) to produce dense captions and candidate question insights for each video. These outputs serve only as auxiliary references to help annotators understand long videos and expand the question design space, and they do not determine ground-truth labels. Annotators then watch each video in full and manually construct questions with precise and verifiable answers. All questions follow a fixed-output format, such as numbers, single words, or short phrases. Questions that can be answered by simple event recognition or straightforward object identification are avoided. Instead, annotators design problems that require subtle clue discovery or integration of non-salient evidence across frames.

For each question, temporal intervals are annotated when the question is explicitly associated with one or multiple distinct time spans in the video. Bounding boxes are annotated on key frames when the question clearly depends on specific small objects or spatial regions within those frames. Temporal evidence is stored as structured start–end timestamps, and spatial evidence is recorded as normalized bounding boxes associated with timestamps. This explicit representation enables systematic verification and quantitative evaluation of spatio-temporal grounding.

Quality Control. All questions undergo two-round cross-verification by annotators within the same team, in which each question is reviewed by members who do not participate in its original annotation. Reviewers rewatch the complete video to ensure that each question is fully answerable

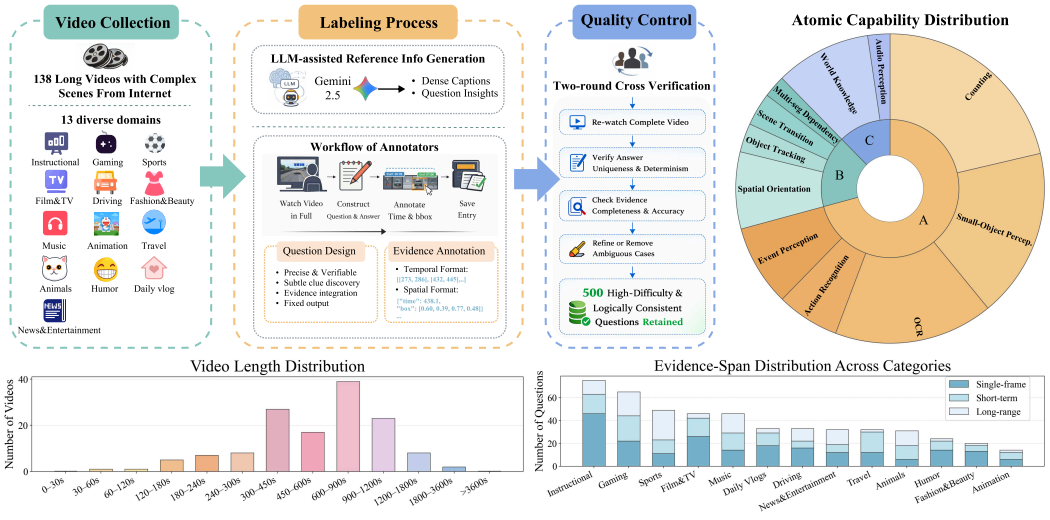

Figure 2: Data construction and statistics of VideoZeroBench. All questions and evidence are manually annotated and verified. The benchmark spans 13 video domains and covers 11 atomic capabilities grouped into Detailed Perception (A), Spatial&Temporal Reasoning (B), and Semantic&Cross-Modal Reasoning (C). The bottom plots show the distributions of video length and minimal evidence span across categories.

using the annotated evidence. They verify the uniqueness and determinism of answers under the specified format. Ambiguous, under-specified, or logically flawed questions are revised or removed. Evidence annotations are refined if incomplete or redundant. Atomic ability categories and minimal evidence span labels are also confirmed during verification. This targeted diagnostic set prioritizes detailed evidence annotation and repeated verification over maximizing question count, supporting focused analysis of answering and grounding failures. Annotation and filtering details are provided in Appendix A.4.

2.2 Benchmark Statistics

VideoZeroBench contains 500 distinct question–answer pairs from 138 manually curated videos across 13 domains, with 280 Chinese and 220 English questions. Of these questions, 442 have temporal intervals, 372 have key-frame boxes, and 320 have both. Appendix A.1 details annotation coverage. Figure 2 summarizes video categories, capabilities, durations, and evidence spans. Videos total 25.57 hours and range from 30 to 3,037 seconds. Their average duration is 667.1 seconds, and most last 5–20 minutes. Compared with recent grounding benchmarks (Cheng et al., 2025; Xu et al., 2025), the videos are longer on average (Table 1), requiring search over longer timelines. Annotated boxes cover 6.8% of the frame area on average. Fine-grained evidence includes small objects, spatial relations, fleeting actions, and object tracks.

Error patterns observed during annotation and cross-verification motivate a taxonomy of 11 atomic capabilities. **Detailed Perception** includes counting, small-object perception, OCR, action recognition, and event perception. These abilities test fine-grained visual perception in cluttered scenes. **Spatial & Temporal Reasoning** includes spatial orientation discrimination, object tracking, scene transition understanding, and multi-segment dependency. These abilities test how models integrate spatial relations and temporal cues. **Semantic & Cross-Modal Reasoning** includes world knowledge reasoning and audio perception. These abilities require combining visual evidence with contextual knowledge or auditory cues. This taxonomy supports the capability-level analysis in Figure 3.

We also group questions by their **minimal evidence span**: the shortest temporal coverage needed to answer them (Figure 2). The 216 single-frame questions need only one frame. The 155 short-term questions require continuous observation over less than 15 seconds. The 129 long-range questions require extended observation or evidence integration across temporally distant segments. This categorization describes evidence distribution and the information needed for reasoning.

2.3 FIVE-LEVEL VIEW OF VIDEOZEROBENCH

The five-level hierarchy diagnoses answering and evidence localization under different input conditions. Levels 1–3 compare QA with progressively fewer evidence hints. Levels 4–5 combine Level-3 answer correctness with independently prompted temporal and spatial predictions. The resulting scores measure success across tasks for the same question.

Level-1: Answer with Spatial and Temporal Evidence. In Level-1, the available temporal intervals and key spatial regions are provided in textual format, along with the entire video. The performance is evaluated using standard QA accuracy. Errors at this level may reflect difficulties in perceiving, integrating, or reasoning over the indicated evidence.

Level-2: Answer with Temporal Evidence. The input includes a query, a full video, and the available temporal annotations, without spatial region hints. By restricting the evidence to temporal cues only, this setting makes questions that depend on fine-grained spatial perception or small-object localization more difficult. Performance is measured by QA accuracy.

Level-3: Answer without Evidence Hints. The model receives only the video and the question. This is the standard end-to-end QA setting, and Level-3 performance is evaluated based on QA accuracy. Correct answers here are not required to be grounded in verified evidence.

Level-4: Correct Answer with Accurate Temporal Grounding. The model must both answer the question correctly and identify one or more supporting temporal intervals. At this level, we feed the original video and the question in a separate prompt and require the model to output only the key time intervals. We first compute multi-segment temporal IoU (tIoU). All five accuracy scores use the same denominator $N = 500$. Let $\mathcal{T}$ denote the 442 questions with temporal annotations. Combined with the Level-3 end-to-end answer correctness, the final Level-4 accuracy is computed as:

$$\text{Accuracy}_{L4} = \frac{1}{N} \sum_{i \in \mathcal{T}} \mathbb{I}(\hat{y}_i = y_i) \cdot \mathbb{I}(\text{tIoU}_i > 0.3). \tag{1}$$

where y_i is the ground-truth answer and $\hat{y}_i$ denotes the model's predicted answer from Level-3. A question passes when its answer and independently predicted temporal evidence are both correct.

Level-5: Correct Answer with Accurate Spatio-Temporal Grounding. The model must correctly answer the question, identify temporal intervals, and accurately localize bounding boxes to annotated timestamps. At this level, we input the original video, the question, and the key-frame timestamps and require the model to output 2D boxes at those timestamps. Fixed timestamps allow comparison with sparse manual boxes at the same frames despite object motion. For scoring, we first compute the visual IoU for one or multiple boxes at each frame, then average over frames, denoted by $vIoU$. Let $\mathcal{S}$ denote the 372 questions with spatial annotations. The final Level-5 accuracy is:

$$\text{Accuracy}_{L5} = \frac{1}{N} \sum_{i \in \mathcal{T} \cap \mathcal{S}} \mathbb{I}(\hat{y}_i = y_i) \cdot \mathbb{I}(\text{tIoU}_i > 0.3) \cdot \mathbb{I}(\text{vIoU}_i > 0.3). \tag{2}$$

Standalone tIoU and vIoU average over $\mathcal{T}$ and $\mathcal{S}$, respectively, including incorrect answers. Level-4/5 accuracy counts joint successes over all N questions, and questions without the required annotations contribute zero. Appendix A.1 gives the coverage-imposed bounds and scores normalized over annotated subsets. Appendix A.6 compares separate and joint prediction.

3 EVALUATION ON VIDEOZEROBENCH

We conduct a comprehensive evaluation of state-of-the-art video MLLMs on VideoZeroBench under the proposed five-level hierarchical protocol. We first present the overall experimental settings and main results across all evaluation levels, and we further perform detailed performance and error analyses from multiple perspectives to better understand capability gaps and failure patterns.

3.1 SETTINGS

In Table 2, we evaluate 19 representative models across three categories.

Table 2: Benchmark results on **VideoZeroBench** (%). All model accuracy columns use 500 questions. Standalone temporal IoU (tIoU) and visual IoU (vIoU) average over 442 and 372 annotated questions, respectively. "Frames" specifies sampling and frame limits. Bold and underlined values mark the best and second-best model scores. Human performance uses a randomly sampled 50-question subset. Annotation coverage and its effect on accuracy are detailed in Appendix A.1.

Models	Frames	Level-1	Level-2	Level-3	Level-4		Level-5	
		Acc	Acc	Acc	tIoU	Acc	vIoU	Acc
Human Reference	video	–	–	67.6	74.2	63.5	59.9	55.8
Proprietary Models								
Gemini-3.7-Flash (Google, 2026)	video	**40.0**	**33.0**	**24.8**	**42.9**	**14.2**	10.2	**1.8**
Gemini-3-Pro (Google, 2025)	video	23.6	20.6	17.0	<u>32.0</u>	<u>8.0</u>	9.6	1.0
Gemini-2.5-Pro (Comanici et al., 2025)	video	25.2	21.0	16.4	31.4	7.2	8.6	0.6
Gemini-2.5-Flash (Comanici et al., 2025)	video	17.0	11.2	11.4	27.9	5.4	4.9	0.0
GPT-5.6-Sol (OpenAI, 2026)	1fps, 96f	<u>32.0</u>	<u>22.2</u>	<u>18.8</u>	23.1	6.4	<u>16.6</u>	<u>1.6</u>
GPT-5.2 (OpenAI, 2025a)	1fps, 96f	17.4	14.6	8.0	21.9	1.8	12.6	0.4
Seed-2.0-Pro (Seed, 2026)	1fps, 160f	25.4	18.0	10.6	25.3	3.8	**21.8**	0.6
Open-Source Models								
Qwen3.5-397B-A17B (Qwen Team, 2026)	1fps, 250f	25.0	20.0	10.6	3.1	0.0	8.0	0.0
Qwen3-VL-235B-A22B (Bai et al., 2025a)	1fps, 384f	28.4	21.4	9.6	19.6	3.4	3.6	0.2
Qwen3-VL-8B (Bai et al., 2025a)	1fps, 384f	24.8	17.8	8.2	10.9	0.6	2.4	0.2
Qwen3-VL-4B (Bai et al., 2025a)	1fps, 384f	22.2	16.6	7.8	13.3	1.4	1.2	0.0
Qwen2.5-VL-7B (Bai et al., 2025b)	1fps, 384f	16.2	12.4	6.0	1.8	0.0	1.4	0.0
InternVL3.5-8B (Wang et al., 2025f)	1fps, 96f	19.4	18.0	9.8	2.9	0.2	0.8	0.0
InternVL3.5-4B (Wang et al., 2025f)	1fps, 96f	14.8	13.6	5.8	3.3	0.0	1.2	0.0
Video Reasoning Specialists								
Video-R1-7B (Feng et al., 2025)	1fps, 384f	17.2	13.4	7.8	2.3	0.0	5.5	0.0
VideoRFT-7B (Wang et al., 2025c)	1fps, 384f	12.0	11.2	5.8	3.6	0.0	4.2	0.0
Open-o3-Video-4B (Meng et al., 2025)	1fps, 384f	17.6	12.8	5.8	9.7	0.4	9.9	0.0
VideoChat-R1.5-7B (Yan et al., 2025)	1fps, 384f	17.4	12.2	8.6	3.3	0.0	2.9	0.0
Video-o3-7B (Zeng et al., 2026)	1fps, 384f	15.0	13.4	4.8	3.5	0.0	0.2	0.0

(1) Proprietary: Gemini-3.7-Flash (Google, 2026), Gemini-3-Pro (Google, 2025), Gemini-2.5 (Comanici et al., 2025), Seed-2.0-Pro (Seed, 2026), GPT-5.6-Sol and GPT-5.2 (OpenAI, 2026; 2025a).

(2) Open-source MLLMs: Qwen3.5 (Qwen Team, 2026), Qwen3-VL (Bai et al., 2025a), Qwen2.5 (Bai et al., 2025b), and InternVL3.5 (Wang et al., 2025f).

(3) Video-specific reasoning models: (a) textual reasoning models: Video-R1 (Feng et al., 2025) and VideoRFT (Wang et al., 2025c), (b) thinking-with-videos models: VideoChat-R1.5 (Yan et al., 2025), Video-o3 (Zeng et al., 2026), and Open-o3-Video (Meng et al., 2025).

We evaluate all models using the same questions and scoring rules (Section 2.3). Input formats and frame budgets follow each model's supported interface and serving constraints. Gemini models receive full videos with audio, whereas other models use uniform 1 FPS frame sampling with limits shown in Table 2. In addition, since Level-5 requires predicting bounding boxes at specified timestamps, we ensure that the corresponding frames are always included in the sampled frames, with each frame mapped to its timestamp in the input metadata. Results characterize these systems under their supported interfaces and reported budgets. Input-modality and frame-budget comparisons are provided in Tables 5 and 10. Further details are in Appendix A.1.

3.2 Benchmark Results

Table 2 reports evaluation results across five levels. Under the standard QA (Level-3) setting, Gemini-3.7-Flash achieves the highest accuracy at 24.8%, followed by GPT-5.6-Sol at 18.8% and Gemini-3-Pro at 17.0%. On a shared 50-question subset, humans reach 67.6% QA accuracy, compared with 22.0% and 20.0% for Gemini-3-Pro and Gemini-2.5-Pro (Table 11). Human Level-4/5 scores on that subset are 63.5% and 55.8%. The human study is described in Appendix A.5. Among open-source models, only Qwen3.5 surpasses 10% accuracy. Under the strict Level-5 requirement,

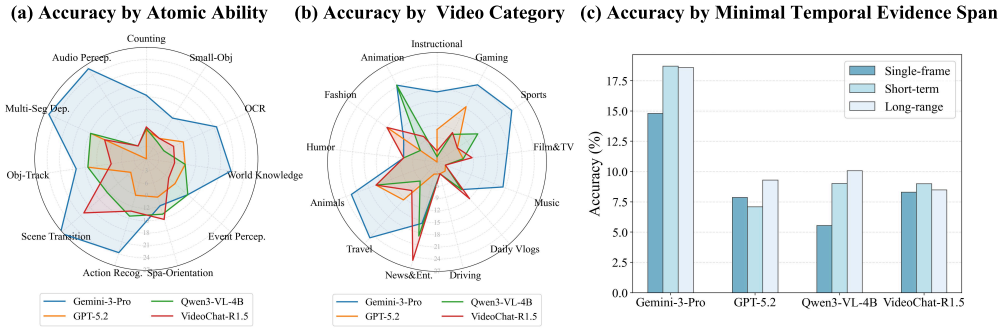

Figure 3: Performance across atomic abilities, video categories, and minimal evidence spans.

Gemini-3.7-Flash reaches 1.8%, followed by GPT-5.6-Sol at 1.6%. Eight of the 19 models obtain non-zero scores at this level. These results indicate that current models rarely answer these challenging questions correctly while providing accurate supporting spatio-temporal evidence. **Answer accuracy alone does not capture the localization weaknesses observed here.**

Across evaluation levels, models achieve higher QA accuracy with spatio-temporal evidence hints (Level-1) than without hints (Level-3). When spatio-temporal evidence is provided (Level-1), Gemini-3.7-Flash reaches 40.0% and GPT-5.6-Sol reaches 32.0%, compared with their Level-3 scores of 24.8% and 18.8%. These gains show the value of evidence hints for answering. However, even the highest-scoring model answers 60% of questions incorrectly with evidence hints, suggesting continued difficulties in perceiving and reasoning over the indicated cues. Removing spatial hints (Level-2) lowers accuracy relative to Level-1 across all models, showing the value of region-level information. Notably, most current thinking-with-videos frameworks (Zeng et al., 2026; Ding et al., 2025) emphasize temporal retrieval and iterative zoom-in, while systematic spatial exploration remains limited. Our results suggest that fine-grained spatial search should also be incorporated as a core dimension alongside temporal reasoning.

Level-4 exposes a gap in temporal grounding. Gemini-3.7-Flash leads at 14.2%, followed by Gemini-3-Pro at 8.0% and Gemini-2.5-Pro at 7.2%. GPT-5.6-Sol reaches 6.4% despite ranking second in standard QA. Among open-source models, Qwen3-VL-235B achieves the highest Level-4 accuracy at 3.4%. Seven of the 19 models score 0% at this level. In terms of temporal grounding, Gemini-3.7-Flash reaches 42.9% tIoU, followed by Gemini-3-Pro at 32.0% and Gemini-2.5-Pro at 31.4%. Most models remain below 20%. These results indicate limited capability to identify correct temporal evidence in complex scenes. At Level-5, models are required to perform spatial grounding on key frames. Seed-2.0-Pro achieves the highest vIoU of 21.8%, followed by GPT-5.6-Sol at 16.6% and GPT-5.2 at 12.6%. Gemini-3.7-Flash reaches 10.2% vIoU despite leading Level-5 accuracy, showing that QA and standalone localization scores capture different strengths. On the 50-question human subset, the reference tIoU and vIoU are 74.2% and 59.9%, respectively. As shown in Figure 4 (2), the Gemini model produces the correct answer while providing inaccurate bounding boxes.

Across model categories, most proprietary systems outperform open-source models under the evaluated input settings. Gemini-3.7-Flash leads accuracy at all five levels and temporal IoU, while Seed-2.0-Pro leads spatial IoU. GPT-5.6-Sol ranks second at Levels 1–3 and 5. Among open-source models, Qwen3-VL-235B leads in tIoU (19.6%), Level-4 accuracy (3.4%), Level-1 (28.4%), and Level-2 (21.4%), while Qwen3.5 leads standard QA (10.6%). These differences motivate reporting answering and localization scores separately.

For video-specific reasoning models, none achieves non-zero performance at Level-5, and only Open-o3-Video obtains non-zero Level-4 accuracy. In terms of grounding capability, it also achieves higher temporal and spatial localization scores than other reasoning models, with 9.7% tIoU and 9.9% vIoU. For agentic thinking-with-videos models such as VideoChat-R1.5 and Video-o3, although they adopt multi-round reasoning and iterative zoom-in strategies, their performance remains 0 at both Level-4 and Level-5, showing no clear advantage on the leaderboard. For these tested systems, **iterative evidence seeking has yet to yield high grounding accuracy on this benchmark.**

Table 3: Effect of the thinking-with-videos paradigm. Here, L1–L4 denote accuracy under the first four levels on VideoChat-R1.5.

VideoChat-R1.5	L1	L2	L3	L4
Three Rounds	17.4	12.2	8.6	0.0
Single Round	16.0	11.2	6.8	0.0

Table 4: Comparison of visual input: original video, temporal segments only (zoom), and temporal segments with spatial crops (zoom&crop).

Model	Ori-Video	Zoom	Zoom&Crop
Gemini-3-Pro	17.0	21.6	41.2
Gemini-2.5-Pro	16.4	21.0	42.0
Qwen3-VL-4B	7.8	11.6	18.0

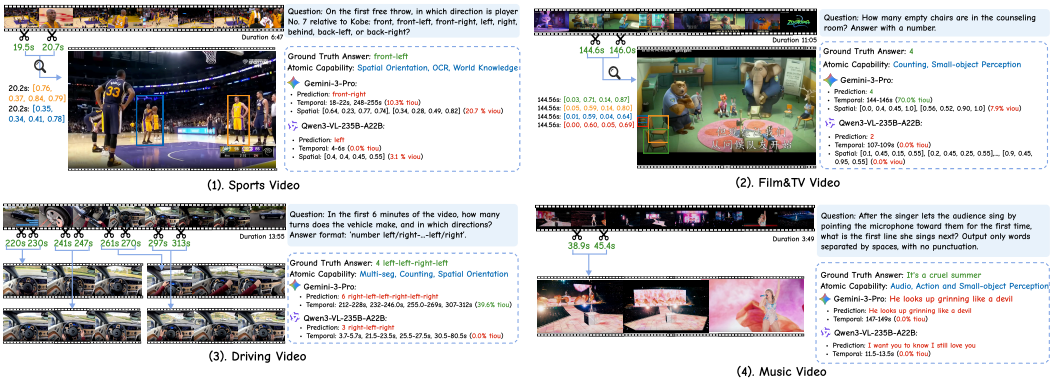

Figure 4: Examples from VideoZeroBench with annotated spatio-temporal evidence and model predictions. (1) Models localize redundant or incorrect evidence and misinterpret relative spatial relationships. (2) Correct answers may still lack precise spatial grounding for all relevant objects. (3) Models fail to integrate evidence across multiple temporal segments and correctly determine turning directions. (4) Models struggle to combine auditory cues with visual evidence for reasoning.

3.3 Findings and Discussions

Q1: Which atomic abilities remain challenging for current video MLLMs? We analyze atomic abilities for the models shown in Figure 3 (left). There are two key observations. **I. Small-object perception and spatial orientation discrimination remain particularly challenging.** For example, Gemini-3-Pro achieves 11.7% on small-object perception and 11.8% on spatial orientation, both below its overall QA accuracy. From Figure 3 (middle), the Driving category also exhibits low accuracy. The driving questions test subtle targets and relative positions in cluttered scenes. Their low accuracy is consistent with the spatial difficulties illustrated in the case studies. As shown in Figure 4 (2), models frequently fail to localize the correct targets, and in Figure 4 (1) and (3), models also misinterpret spatial direction in realistic environments. **II. Counting is consistently difficult across models.** Counting requires accurate localization of evidence and cross-frame aggregation. It involves both intra-frame object counting and inter-frame event counting. Models need to distinguish repeated appearances and handle occlusion. Empirically, most models shown achieve accuracies below 8% for counting. This reflects limited spatial precision and weak integration of multi-segment evidence. As shown in Figure 4 (2)(3), models often fail to fully identify small objects or short-lived actions.

Q2: Are single-frame answerable questions inherently easier? Some video questions can be answered from a single frame, requiring minimal temporal coverage. Without temporal reasoning, can models handle such problems more easily? To investigate this, we categorize questions by minimal temporal evidence span. As shown in Figure 3 (right), for the models included in this analysis, single-frame questions are not easier than short-term or long-range ones. Like Figure 4 (2), such questions in our benchmark often resemble needle-in-a-haystack scenarios. Although the question can be solved in a single frame, that frame may last only a few seconds in a long video. Uniform sampling may miss such fleeting frames, limiting the evidence available for answering. **Therefore, under limited visual tokens and sampling constraints, precise temporal localization of short-lived yet critical frames remains highly challenging.**

Q3: Does the agentic thinking-with-videos paradigm bring gains? We conduct an ablation study using VideoChat-R1.5 as a representative work in thinking-with-videos mode. This paradigm

Table 5: Effect of input modality on Gemini-3-Pro. Small-Obj and Audio denote Level-3 performance on small-object and audio perception, respectively.

Input	Level-3	Small-Obj	Audio
No vis.& aud.	6.8	5.4	0.0
Audio-only	8.4	4.9	7.4
Frames	17.0	15.6	7.4
Video	17.0	11.7	25.9

alleviates the limitations of uniform frame sampling by iteratively refining temporal focus. We compare three-round zoom-in reasoning with single-round direct answering. Under this paradigm, in round $(n-1)$ the model outputs its reasoning trace, predicted answer, and candidate temporal intervals, and in round n it performs dense sampling within the predicted interval and sparse sampling over the remaining video. As shown in Table 3, accuracy improves by 1.4, 1.0, and 1.8 percentage points at Level-1, Level-2, and Level-3, respectively. These QA gains do not translate into improved Level-4/5 accuracy. **Fine-grained grounding remains difficult for VideoChat-R1.5 under this thinking-with-videos setting.**

Q4: What is the gap between full-video and visual-evidence-only input? Beyond the standard Level-1 and Level-2 settings, we further compare different visual input configurations while controlling total visual tokens. As shown in Table 4, we evaluate three conditions. The first uses the original full video without hints (level-3). The second only uses annotated temporal segments with zoom-in sampling. The third uses annotated segments together with cropped key spatial regions. For Gemini models, zoom and crop substantially improve performance. **Gemini-3-Pro and Gemini-2.5-Pro increase from 17.0% and 16.4% to above 40% under the zoom&crop setting.** These gains show the value of direct access to fine-grained evidence, motivating better temporal search and spatial inspection.

Q5: How does input modality affect performance? We compare Gemini-3-Pro with no visual or audio input, audio-only input, frames sampled at 1 FPS with a 384-frame cap, and full video with audio (Table 5). Frames and full video both yield 17.0% overall QA accuracy, but their strengths differ across abilities. Full video raises audio accuracy from 7.4% to 25.9%, while frames yield higher small-object accuracy (15.6% versus 11.7%). Without visual or audio input, the model still answers 6.8% of questions correctly, suggesting that question priors or world knowledge can contribute to answer-only scores. These results reinforce the need to verify supporting evidence, not just answers.

More experiment results, visualizations, and details. Low grounding accuracy persists under relaxed thresholds for Gemini-3-Pro (Table 8) and under joint prediction for the three models tested on a shared subset (Table 12). Appendices A.2, A.5, A.7, and A.8 provide further ablations, the human study, repeated-run results, and error cases.

4 Conclusion

We present **VideoZeroBench**, a challenging hierarchical benchmark for evidence-grounded long-video question answering. Its five-level diagnostic protocol evaluates answering under different evidence hints and combines answer correctness with temporal and spatial localization scores. The best model reaches 24.8% in standard QA and 1.8% with correct answers and accurate spatio-temporal evidence. Our evaluation yields three takeaways: (1) correct answers do not ensure accurate evidence localization; (2) direct access to fine-grained visual evidence substantially improves QA; and (3) the tested thinking-with-videos approaches still struggle with precise grounding. These findings motivate methods that combine temporal search, spatial inspection, and explicit evidence verification. Limitations in coverage and evidence granularity, together with future benchmark extensions, are discussed in Appendix A.10. We hope VideoZeroBench can help advance video MLLMs that answer challenging long-video questions with precise, reliable, and verifiable evidence.

Reproducibility Statement

We take several steps to make our work reproducible. The dataset construction pipeline, annotation protocol, and quality-control procedure are described in Section 2 and Appendix A.4. The complete evaluation protocol, including video input and frame-sampling settings, inference setup, thinking-with-videos configurations, prompt templates, and score computation, is detailed in Section 3 and Appendix A.1. The five-level evaluation metrics are formally defined in Section 2.3. We will publicly release the benchmark data, annotations, and evaluation code to support reproduction of all results.

Ethics Statement

This work adheres to the ICLR Code of Ethics. All videos used in VideoZeroBench are collected from publicly accessible sources and are used solely for academic research. To mitigate privacy risks, identifiable information such as human faces is blurred or masked in the visualizations shown in this paper. The benchmark should not be used to develop systems that violate privacy, track individuals without consent, or enable harmful surveillance applications. Users of the released data and annotations are responsible for complying with applicable copyright, privacy, and data protection regulations.

AI Use Statement

For manuscript preparation, generative AI tools were used only for language editing and polishing. All benchmark questions, answers, and evidence annotations were created and verified entirely by human annotators. Gemini provided auxiliary reference information only to help annotators understand video content more quickly, not to generate or verify benchmark labels (Section 2 and Appendix A.4). The authors reviewed all AI-assisted text and take full responsibility for the paper.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, 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, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, et al. Cg-bench: Clue-grounded question answering benchmark for long video understanding. *arXiv preprint arXiv:2412.12075*, 2024.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Yang Ding, Yizhen Zhang, Xin Lai, Ruihang Chu, and Yujiu Yang. Videozoomer: Reinforcement-learned temporal focusing for long video reasoning. *arXiv preprint arXiv:2512.22315*, 2025.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, et al. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 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, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024a.

Shenghao Fu, Qize Yang, Yuan-Ming Li, Xihan Wei, Xiaohua Xie, and Wei-Shi Zheng. Love-r1: Advancing long video understanding with an adaptive zoom-in mechanism via multi-step reasoning. *arXiv preprint arXiv:2509.24786*, 2025.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024b.

Shida Gao, Feng Xue, Xiangfeng Wang, Anlong Ming, Teng Long, Yihua Shao, Haozhe Wang, Zhaowen Lin, Wei Wang, and Nicu Sebe. 1+ 1> 2: Detector-empowered video large language model for spatio-temporal grounding and reasoning. *arXiv preprint arXiv:2512.06673*, 2025.

Haonan Ge, Yiwei Wang, Kai-Wei Chang, Hang Wu, and Yujun Cai. Framemind: Frame-interleaved video reasoning via reinforcement learning. *arXiv preprint arXiv:2509.24008*, 2025.

Google. A new era of intelligence with gemini 3, 2025. URL https://blog.google/products/gemini/gemini-3.

Google. Introducing Gemini 3.7 Flash, 2026. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/introducing-gemini-3-7-flash/.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, et al. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.

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, 2024.

Jingli Lin, Runsen Xu, Shaohao Zhu, Sihan Yang, Peizhou Cao, Yunlong Ran, Miao Hu, Chenming Zhu, Yiman Xie, Yilin Long, et al. Mmsi-video-bench: A holistic benchmark for video-based spatial intelligence. *arXiv preprint arXiv:2512.10863*, 2025.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, et al. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024.

Meng Luo, Shengqiong Wu, Liqiang Jing, Tianjie Ju, Li Zheng, Jinxiang Lai, Tianlong Wu, Xinya Du, Jian Li, Siyuan Yan, et al. Dr. v: A hierarchical perception-temporal-cognition framework to diagnose video hallucination by fine-grained spatial-temporal grounding. *arXiv preprint arXiv:2509.11866*, 2025.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, et al. Open-o3 video: Grounded video reasoning with explicit spatio-temporal evidence. *arXiv preprint arXiv:2510.20579*, 2025.

OpenAI. Introducing gpt-5.2, 2025a. URL https://openai.com/index/introducing-gpt-5-2.

OpenAI. OpenAI-o3. https://openai.com/index/introducing-o3-and-o4-mini/, 2025b.

OpenAI. Previewing GPT-5.6 Sol: a next-generation model, 2026. URL https://openai.com/index/previewing-gpt-5-6-sol/.

Kun Ouyang, Yuanxin Liu, Linli Yao, Yishuo Cai, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Conan: Progressive learning to reason like a detective over multi-scale visual evidence. *arXiv preprint arXiv:2510.20470*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, et al. Zerobench: An impossible visual benchmark for contemporary large multimodal models. *arXiv preprint arXiv:2502.09696*, 2025.

Bytedance Seed. Seed2, 2026. URL https://seed.bytedance.com/en/seed2.

Chinthani Sugandhika, Chen Li, Deepu Rajan, and Basura Fernando. Know-show: Benchmarking video-language models on spatio-temporal grounded reasoning. *arXiv preprint arXiv:2512.05513*, 2025.

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

Yolo Y Tang, Daiki Shimada, Hang Hua, Chao Huang, Jing Bi, Rogerio Feris, and Chenliang Xu. Video-r4: Reinforcing text-rich video reasoning with visual rumination. *arXiv preprint arXiv:2511.17490*, 2025.

Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8238–8249, 2021.

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule Bai, Zijian Kang, Jiashi Feng, et al. Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology. *arXiv preprint arXiv:2507.07999*, 2025a.

Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025b.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *arXiv preprint arXiv:2505.12434*, 2025c.

Shijian Wang, Jiarui Jin, Xingjian Wang, Linxin Song, Runhao Fu, Hecheng Wang, Zongyuan Ge, Yuan Lu, and Xuelian Cheng. Video-thinker: Sparking" thinking with videos" via reinforcement learning. *arXiv preprint arXiv:2510.23473*, 2025d.

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, 2025e.

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*, 2025f.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024.

Yuxi Xiao, Longfei Li, Shen Yan, Xinhang Liu, Sida Peng, Yunchao Wei, Xiaowei Zhou, and Bingyi Kang. Spatialtree: How spatial abilities branch out in mllms. *arXiv preprint arXiv:2512.20617*, 2025.

Qi'ao Xu, Tianwen Qian, Yuqian Fu, Kailing Li, Yang Jiao, Jiacheng Zhang, Xiaoling Wang, and Liang He. Tog-bench: Task-oriented spatio-temporal grounding in egocentric videos. *arXiv preprint arXiv:2512.03666*, 2025.

Ziang Yan, Xinhao Li, Yinan He, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *arXiv preprint arXiv:2509.21100*, 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025a.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Chengwei Qin, Shijian Lu, Xingxuan Li, and Lidong Bing. Longvt: Incentivizing" thinking with long videos" via native tool calling. *arXiv preprint arXiv:2511.20785*, 2025b.

Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Yan Shu, Nicu Sebe, Ji-Rong Wen, and Zhicheng Dou. Videoexplorer: Think with videos for agentic long-video understanding, 2025. URL https://arxiv.org/abs/2506.10821.

Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Zikang Wang, Changlian Ma, Qingyu Zhang, Zizheng Huang, Kun Ouyang, Tianxiang Jiang, et al. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning. *arXiv preprint arXiv:2601.23224*, 2026.

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. *arXiv preprint arXiv:2509.23652*, 2025a.

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. *arXiv preprint arXiv:2508.04416*, 2025b.

Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10668–10677, 2020.

Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, et al. Mmvu: Measuring expert-level multi-discipline video understanding. *arXiv preprint arXiv:2501.12380*, 2025.

A Appendix For VideoZeroBench

Overview. This appendix provides additional implementation details, annotation descriptions, and extended analyses to complement the main paper.

- Section A.1 presents additional experimental details.
- Section A.2 provides further analyses on frame sampling strategies, test-time scaling, the effect of video length and threshold sensitivity of tIoU and vIoU.
- Section A.3 provides a detailed comparison with recent video spatio-temporal grounding benchmarks.
- Section A.4 describes additional annotation details.
- Section A.5 reports human performance at Levels 3–5 on a subset of VideoZeroBench.
- Section A.6 explains the rationale behind our evaluation protocol.
- Section A.7 examines the robustness of our findings.
- Section A.8 presents additional qualitative visualizations and error cases to illustrate typical failure patterns of current video MLLMs.
- Section A.9 gives related works discussion.
- Section A.10 discusses the limitations and future works.

A.1 Additional Experimental Details

This section provides additional experimental details for reproducibility.

Questions and annotation coverage. The benchmark contains 500 distinct question–answer pairs. Table 6 reports the availability of sparse manual evidence annotations. There are 442 questions with temporal intervals and 372 with key-frame boxes, of which 320 have both. The reported 2,314 task instances count the three QA settings over the shared question set, plus 442 temporal and 372 spatial annotation targets: $3 \times 500 + 442 + 372 = 2{,}314$. This instance count distinguishes repeated tasks from distinct questions.

Table 6: Sparse manual annotation coverage in VideoZeroBench.

	Spatial boxes	No spatial boxes	Total
Temporal intervals	320	122	442
No temporal intervals	52	6	58
Total	372	128	500

Video Input and Frame Sampling. For the Gemini series, the raw video is base64-encoded and sent directly to the API as input. Due to API size limits, videos are compressed to less than 100MB before encoding. All other models require sampled frames as input. Frames are uniformly sampled at 1 FPS with a predefined maximum number of frames. As shown in Table 2, most models use a limit of 384 frames. Some models use smaller limits due to context length or memory constraints, including Seed-2.0-Pro (Seed, 2026) (160 frames), Qwen3.5-397B (Qwen Team, 2026) (250 frames), and GPT-5.6-Sol (OpenAI, 2026), GPT-5.2 (OpenAI, 2025a), and InternVL3.5 (Wang et al., 2025f) (96 frames each). Each sampled frame is resized while preserving the original aspect ratio. For Level-5, the visual input explicitly includes every designated key frame and its timestamp, including when raw video is supplied. These settings aim to make use of each model's supported input capabilities. We use native video with audio for Gemini and sampled frames for the other models. Section A.2 (Table 10) reports frame-rate and frame-budget ablations. Table 5 in the main paper reports the input-modality ablation.

Inference Setup. Proprietary models and Qwen3.5-397B (Qwen Team, 2026) are evaluated through API calls. All other open-source models are evaluated using the vLLM inference framework. The decoding temperature is set to 0 for the main evaluation. Repeated-run results at temperature 0.7 are reported in Section A.7. Models with native chain-of-thought generation keep their default reasoning style during evaluation, while other models are prompted to output only the final answer.

Thinking-with-videos Paradigm Configurations. For VideoChat-R1.5 (Yan et al., 2025), frames within predicted key intervals are sampled more densely than other parts of the video. The density ratio is controlled by the parameter `key_ratio`. We follow the original setting and use `key_ratio` $= 1.5$. Multi-round reasoning is fixed at three rounds. For Video-o3 (Zeng et al., 2026), following the original paper, dynamic sampling with adaptive pixel budgets is used. The reasoning process runs for at most eight rounds. If the maximum round limit is reached, the model is forced to output the final answer.

Prompt Templates Across Five Levels. Prompt templates are adapted to match the timestamp and coordinate formats required by different models. For example, for Qwen3-VL models, bounding boxes are represented using normalized coordinates in the range 0–1000. For models based on Qwen2.5-VL, bounding boxes are represented using absolute pixel coordinates. The output parser is implemented accordingly to correctly extract temporal intervals and spatial boxes. Figure 8 illustrates example prompt templates used for Qwen3-VL models (without chain-of-thought) across different evaluation levels.

Score Computation. All five accuracy scores use the full question pool as a common denominator ($N = 500$). This choice expresses Level-4/5 joint success and Level-1–3 answering success on the same benchmark-wide scale, making their gaps easier to compare without changing the reference population. For Level-4/5 accuracy, questions lacking the required annotations contribute no verified joint success. Standalone tIoU and vIoU instead average over the 442 temporally and 372 spatially annotated questions, respectively, regardless of answer correctness. The annotation coverage therefore bounds Level-4 accuracy by $442/500 = 88.4\%$ and Level-5 accuracy by $320/500 = 64.0\%$. Cross-level accuracy gaps reflect both annotation coverage and prediction failures.

Re-normalizing the best model's existing success counts over the relevant annotated subsets gives $71/442 = 16.1\%$ for Level-4 and $9/320 = 2.8\%$ for Level-5, compared with the reported 14.2% and 1.8% on the full question pool. Thus, low joint success persists on the annotated subsets. This change of denominator also preserves model rankings within each level.

We use an IoU threshold of 0.3 for both temporal and spatial grounding. This threshold is chosen to account for the inherent ambiguity of video evidence annotation: supporting temporal intervals may have soft boundaries, and spatial boxes may vary slightly across annotators, especially for small objects, motion blur, and cluttered scenes. At the same time, the threshold still requires the predicted evidence to substantially overlap with the annotated supporting region, preventing models from receiving credit for coarse or irrelevant localization. If a model fails to output a valid temporal interval or spatial box, the accuracy score for this query is set to 0 in Level-4 or Level-5. Temporal IoU measures precision at the annotated temporal granularity: a substantially wider prediction can cover the event yet fall below the threshold. Accordingly, low tIoU can reflect imprecise boundaries as well as an incorrect event location. Table 8 examines threshold sensitivity.

A.2 Additional Analysis

This section examines frame sampling, test-time scaling, video length, and grounding thresholds. The input-modality experiment is reported in the main paper (Table 5).

Effect of Frame Sampling Strategy. We analyze frame sampling for Qwen3-VL-4B and Qwen3-VL-235B by varying the maximum number of input frames. Table 10 reports the Level-3 accuracy under different frame budgets. The first four settings use 1 FPS with different frame caps, while the last setting uses 2 FPS with a maximum of 768 frames. The results show that increasing the frame budget does not consistently improve performance. Both models reach their highest tested accuracy at 96 frames, with no consistent gains at larger budgets. This motivates more selective temporal sampling for these models rather than relying only on more frames. Future work should instead focus on identifying informative temporal segments through temporal grounding, which increases sampling density in key segments while reducing redundant frames.

Effect of Test-Time Scaling. We investigate whether allocating more inference compute at test time improves long-video understanding for Qwen3-VL-4B (Bai et al., 2025a) under the standard Level-3 setting. We evaluate two distinct paradigms. The first is *parallel scaling*, where we sample $K = 5$ independent responses per question at a higher temperature ($\tau = 0.7$) to measure if the model can generate the correct answer among multiple attempts. The second is a two-round *sequential*

Table 7: **Effect of test-time scaling strategies on Qwen3-VL-4B under the Level-3 setting.** The baseline utilizes single-round greedy decoding. Pass@5 (Any) indicates that at least one of the 5 generated answers is correct, while Pass@5 (All) requires all 5 answers to be correct.

Method	Accuracy (%)
Baseline (Greedy Decoding)	7.8
Parallel Scaling (Pass@5, $\tau = 0.7$)	
Pass@5 (Any)	15.0
Pass@5 (All)	1.8
Sequential Scaling	
Localize-then-Answer	7.0

Table 8: **Threshold sensitivity analysis on Gemini-3-Pro.** We report Level-5 accuracy under different combinations of temporal IoU and visual IoU thresholds. The default setting used in the main paper is tIoU > 0.3 and vIoU > 0.3.

tIoU Threshold	vIoU Threshold		
	> 0.1	> 0.3	> 0.5
> 0.1	3.4	1.8	0.6
> 0.3	2.0	**1.0**	0.2
> 0.5	1.2	0.6	0.2

scaling approach (localize-then-answer): In the first round, the model predicts key temporal evidence windows. In the second round, the video is dynamically resampled with higher density within these predicted windows to generate the final answer.

As shown in Table 7, parallel scaling notably increases the upper bound of correctness, with Pass@5 (Any) reaching 15.0% compared to the 7.8% greedy decoding baseline. However, the extremely low Pass@5 (All) score of 1.8% reveals high uncertainty in the model's reasoning paths. Conversely, the sequential reasoning paradigm degrades performance to 7.0%. This decline aligns with our previous findings regarding the temporal grounding bottleneck: Because the model's intrinsic localization ability is weak, the first round frequently predicts incorrect evidence windows. The second round may concentrate its sampling budget on irrelevant frames, filtering out genuine visual cues.

Effect of Video Length. We further analyze model performance under different video durations. Videos are grouped into three categories based on their length: short (<5 minutes), medium (5–15 minutes), and long (>15 minutes). Figure 7 shows the Level-3 accuracy for each group. For most models shown, accuracy decreases as the video duration increases. This trend suggests that long videos introduce additional challenges for current video MLLMs, including a larger search space, more distractors, and more complex reasoning logic. There are also some exceptions. Gemini-3-Pro (Google, 2025) and video reasoning models such as Video-R1 (Feng et al., 2025) and VideoChat-R1.5 (Yan et al., 2025) demonstrate stronger robustness on long videos. This behavior indicates that advanced reasoning capabilities may help mitigate the difficulty of long-video comprehension.

Threshold Sensitivity of tIoU and vIoU. We use tIoU > 0.3 and vIoU > 0.3 as the default setting because this threshold provides a moderate balance between evidence specificity and localization tolerance, especially for short temporal evidence and small spatial regions, as mentioned in Section A.1. As shown in Table 8, relaxing the thresholds increases Level-5 accuracy, while stricter thresholds further reduce the score. However, Gemini-3-Pro still achieves only 3.4% accuracy under the most relaxed setting, showing that its grounding weakness persists across the tested thresholds.

A.3 Detailed Comparisons with Related Spatio-temporal Grounding Benchmarks

Recent video understanding benchmarks with spatio-temporal grounding provide valuable evidence annotations, but their evaluation goals and difficulty profiles differ from those of VideoZeroBench,

Table 9: Comparison with recent spatio-temporal grounding benchmarks. The VideoZeroBench query count denotes distinct questions, and its five-level protocol yields 2,314 task instances.

Benchmark	Primary goal	Video/Data source	#Videos	Dur.	#Queries	Key distinctions / features
V-STaR (Cheng et al., 2025)	Reverse spatial and temporal reasoning over "what–when–where" chains	VidSTG, TVQA+, GOT-10K, and inserted YouTube videos	743	110.23s	2.1K	Focuses on grounding over relatively salient object-event relations; lacks evaluation on difficult QA in long, complex videos and does not provide hierarchical analysis.
ToG-Bench (Xu et al., 2025)	Task-oriented STVG for embodied agents	ScanNet egocentric indoor trajectories	100	87.88s	2.7K	Focuses on task-relevant indoor object localization rather than difficult open-ended video QA; the scenes are mostly egocentric indoor environments with limited duration.
Dr.V (Luo et al., 2025)	Hallucination diagnosis across perceptive, temporal, and cognitive levels	15 existing public video QA datasets	4,974	0–600s	10K	Targets hallucination diagnosis with many MCQ/binary/caption-style samples; most videos are relatively short and transformed from existing datasets, and the overall QA difficulty is lower.
Know-Show (Sugandhika et al., 2025)	Action-conditioned grounded reasoning over persons, objects, hands, and interactions	Charades, Action Genome, Ego4D test splits	675	47.97s	2.5K	Evaluates action-conditioned entity grounding, but videos are much shorter and scenarios are tied to established action datasets with constrained action-role bindings.
VideoZeroBench (ours)	Evidence-grounded long-video QA with strict answer and evidence verification	YouTube and Bilibili videos across 13 domains	138	667.1s	500	Long and complex Internet videos; subtle evidence such as fleeting frames, small objects, OCR cues, and multi-segment dependencies; five-level diagnostic hierarchy; human annotations with two-round cross-verification.

as summarized in Table 9. V-STaR (Cheng et al., 2025) studies chain-style "what–when–where" reasoning, ToG-Bench (Xu et al., 2025) focuses on task-level object localization in egocentric indoor scenes, Dr.V (Luo et al., 2025) uses spatial-temporal evidence for hallucination diagnosis, and Know-Show (Sugandhika et al., 2025) evaluates action-conditioned entity grounding. As illustrated in Figure 5, examples from V-STaR and Dr.V often involve salient targets or simple actions over relatively short videos, making the required evidence generally more accessible once the relevant frames are observed. In contrast, VideoZeroBench targets hard long-video QA and explicitly separates three abilities that are often mixed together in prior settings: retrieving the correct temporal evidence, localizing the relevant spatial region, and reasoning from the retrieved evidence. This is achieved through a five-level hierarchy that starts from answer correctness under different evidence hints and gradually tightens the requirement to temporal and spatial grounding, so the gaps across levels reveal where models fail. VideoZeroBench also introduces greater video-level complexity: its newly collected Internet videos average 667.1 seconds and contain more scene changes, distractors, and long-range dependencies than most existing spatio-temporal grounding benchmarks. Separately, the required evidence is often intrinsically harder to perceive. Instead of relying mainly on salient large objects or obvious actions (Cheng et al., 2025), many questions depend on fleeting frames, small objects, subtle spatial relations, OCR cues, object tracks, or evidence distributed across multiple segments. Finally, all final questions, answers, temporal intervals, and spatial boxes are manually created or verified, improving label quality and reducing the risk of contamination from older public datasets or machine-generated labels. As a result, VideoZeroBench exposes a much sharper capability gap: even state-of-the-art models perform poorly under standard QA and nearly fail when correct answers must be supported by accurate spatio-temporal evidence.

A.4 Additional Annotation Details

This section covers annotation costs, filtering statistics, and the reference-generation interface.

Annotation Effort and Filtering. Annotating each QA pair and its spatio-temporal evidence takes 5–10 minutes. Each question undergoes two verification rounds, taking 3–5 minutes per round. We remove ambiguous or unanswerable questions and those with non-unique answers or annotation errors. Of 608 candidate questions, 108 (17.8%) are discarded, leaving 500. The total annotation effort is approximately 200 person-hours.

LLM-Assisted Reference Information Generation. Before manual annotation, we use Gemini to generate two types of auxiliary information for each video: dense captions and question insights. These outputs help annotators quickly understand the video content and explore potential question

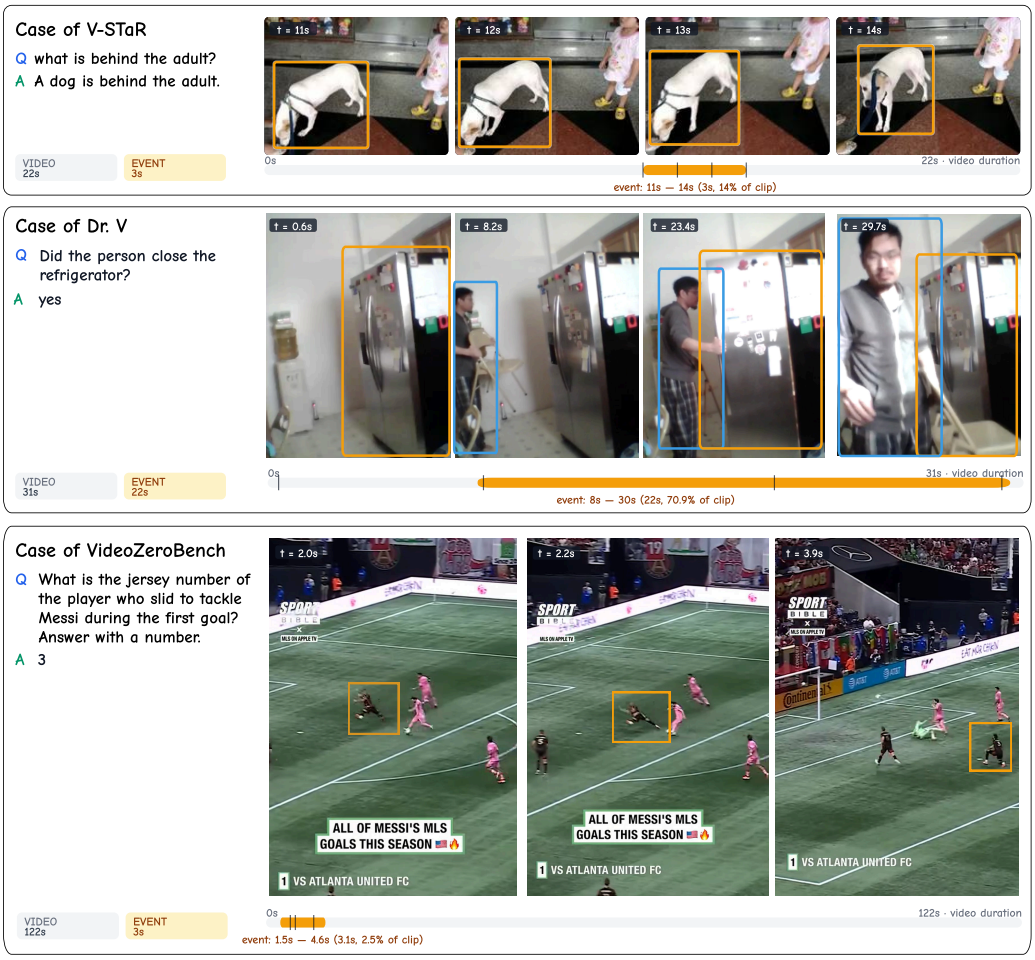

Figure 5: Qualitative comparison with representative spatio-temporal grounding benchmarks. We randomly select examples from V-STaR, Dr.V, and VideoZeroBench for visualization. In V-STaR, many questions focus on relatively salient targets and explicit object relations, such as a dog behind a person. In Dr.V, many cases also examine simple actions involving obvious objects, such as whether a person closes a refrigerator. Moreover, these videos are relatively shorter and the supporting evidence is temporally denser. By contrast, VideoZeroBench often requires identifying events involving small objects within fleeting moments, such as a jersey number during a short tackle event, and further combines such visual evidence with higher-level semantic reasoning. The example illustrates that VideoZeroBench presents higher video complexity, harder questions, and sparser evidence.

directions. They serve only as reference information, and all final questions and answers are designed and verified by human annotators. Figure 9 shows the prompt for these auxiliary descriptions.

Annotation Interface. As shown in Figure 6, the human annotation interface includes the video player, Gemini-generated reference information, annotation guidelines, and input fields for questions, answers, and evidence. The interface also supports direct API calls to Gemini models, allowing annotators to query the model during annotation. This helps annotators understand the capability boundaries of video MLLMs and design challenging questions that expose potential weaknesses.

Table 10: Ablation on the frame sampling strategy. Results are reported on Level-3 accuracy.

Max Frames	48	96	192	384	768 (FPS=2.0)
Qwen3-VL-4B	7.0	**8.8**	7.0	7.8	7.0
Qwen3-VL-235B-A22B	7.0	**10.4**	8.6	9.6	10.0

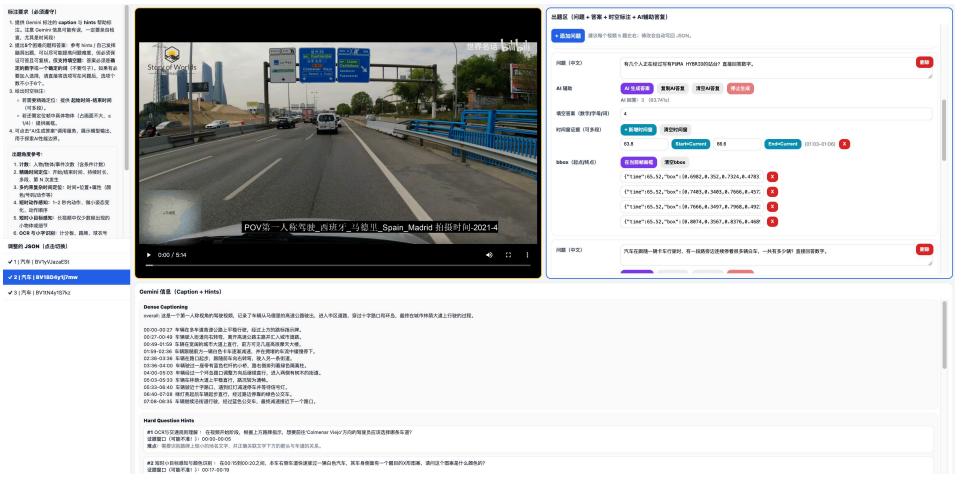

Figure 6: Annotation interface used for question construction and evidence labeling.

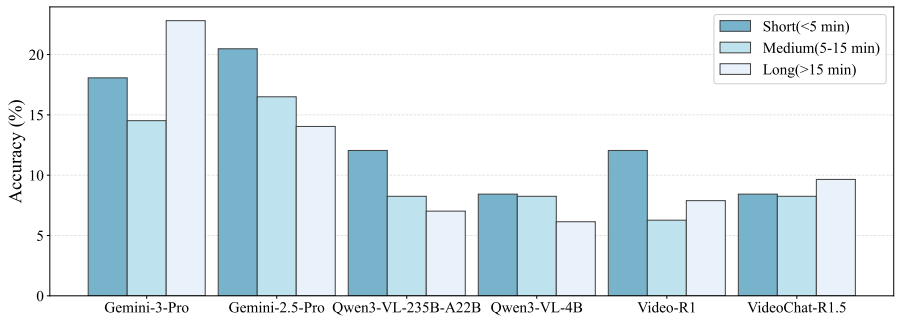

Figure 7: Level-3 accuracy grouped by video duration. Videos are divided into short (<5 minutes), medium (5–15 minutes), and long (>15 minutes).

Evaluation Prompt of Qwen3-VL w/o CoT

System Prompt:
You are a video understanding assistant. Based on the user's question, answer according to the video content and strictly follow the required output format specified by the user.

Level-1 QA:
Question: {question}
The temporal evidence for answering the question is: From <{start_1} seconds> to <{end_1} seconds>; From <{start_2} seconds> to <{end_2} seconds>; ...
The spatial evidence for answering the question is: (time=<{time_1} seconds>, normalized box=[{x1},{y1},{x2},{y2}]); (time=<{time_2} seconds>, normalized box=[{x1},{y1},{x2},{y2}]); ...
Please directly output the final answer.

Level-2 QA:
Question: {question}
The temporal evidence for answering the question is: From <{start_1} seconds> to <{end_1} seconds>; From <{start_2} seconds> to <{end_2} seconds>; ...
Please directly output the final answer.

Level-3 QA:
Question: {question}
Please directly output the final answer.

Level-4 Temporal Grounding:
Question: {question}
Task: Find one or more the most important key time ranges (no more than 20 segments) in the video that provide sufficient evidence to answer the question.
Output format:
From <start_timestamp_1 seconds> to <end_timestamp_1 seconds>. From <start_timestamp_2 seconds> to <end_timestamp_2 seconds>.
Rules:
- Output ONLY the time ranges in the specified sentence format. No explanation, no extra words. Do not need to answer the original question.
- Each segment must follow exactly: 'From <X seconds> to <Y seconds>.'
- Use absolute time in seconds (NOT MM:SS format).
- Ensure each start_timestamp < end_timestamp.

Level-5 Spatial Grounding:
Question: {question}
Given key time points (absolute seconds): <{t_1} seconds>, <{t_2} seconds>, <{t_3} seconds>, ...
Task: For each provided time point, output 1 or more 2D bounding boxes that are relevant evidence for answering the question.
Output format: a JSON array of objects.
[{"time": ..., "bbox_2d":[[...],[...],...]}, {"time": ..., "bbox_2d":[[...],...]}, ...]
Rules:
- Output ONLY valid JSON. No markdown fences, no explanation text. Do not need to answer the original question.
- The length of the json data should be consistent with the number of key time points provided.
- Each object's 'time' MUST be one of the provided time points (in seconds).
- 'bbox_2d' MUST be a list of one or more boxes.
- Each box is normalized coordinates in [0,1000]: [x_min, y_min, x_max, y_max].

Figure 8: Example prompt templates used for Qwen3-VL models across different evaluation levels.

System Prompt (Dense Captioning)

You are a video understanding and data annotation assistant. Follow all formatting and uncertainty rules strictly. The output must be valid JSON.

User Prompt (Dense Captioning)

Please generate dense captioning for the video as an event log to assist human annotators.

Requirements:
1. Divide the video into temporal segments based on events or actions.
2. The total number of segments should not exceed 20.
3. Each segment must contain:
   - start_time (seconds, one decimal)
   - end_time (seconds, one decimal)
   - summary (1–2 short objective sentences in Chinese)

Output JSON format:

```
{
  "overall_summary": "...",
  "segments": [
    {
      "start_time": ...,
      "end_time": ...,
      "summary": "..."
    }
  ]
}
```

System Prompt (Question Insight Generation)

You are a video understanding and benchmark construction assistant. Your goal is to help humans design challenging spatio-temporal reasoning questions for this video. Each hint must include approximate evidence timestamps (mm:ss). Each hint should reference relevant atomic abilities. Output must be valid JSON with no extra explanations. Use Chinese.

User Prompt (Question Insight Generation)

Below is the dense captioning result of the video. It may contain errors and should be used only as reference.

[dense_caption_json]

[Atomic ability list]

Please propose around 10 challenging question hints.

Each hint must contain:
- hint_id
- angle (what capability the question tests)
- evidence_windows (1–3 windows, format mm:ss-mm:ss)
- why_hard (one sentence explanation)

Output JSON format:

```
{
  "hints": [
    {
      "hint_id": 1,
      "angle": "...",
      "evidence_windows": ["...-..."],
      "why_hard": "..."
    }
  ]
}
```

Figure 9: Prompt used to generate dense captions and question insights for annotation reference.

A.5 Human Performance Study

The human study provides a reference for answering and grounding on a 50-question subset. Table 2 reports Levels 3–5 and Table 11 gives the Level-3 comparison on the same subset.

We further conduct a human performance study on a randomly sampled 50-question subset of VideoZeroBench. The subset covers all 13 video categories and includes 27 Chinese questions and 23 English questions. We recruit 20 participants whose birth years range from the 1970s to the 2000s, and all participants are independent of the dataset annotators. Each participant answers at most 10 questions. Each question is answered by multiple human participants. Human scores are averaged across participants within each question, then averaged across questions.

The Level-4/5 evaluation follows the same procedure, with participants providing temporal intervals and bounding boxes at designated key frames. We apply the same IoU thresholds as in model evaluation. Participants leave answers blank if they cannot find them in the video, and blank answers count as incorrect. Human accuracy is 63.5% at Level-4 and 55.8% at Level-5. Some errors involve imprecise intervals or boxes when questions require multiple evidence segments or targets.

The 50-question study provides a reference estimate under a practical evaluation budget. Its scope is limited to this subset, where answering requires reviewing long videos, searching for subtle evidence, and replaying relevant segments.

Table 11 shows that humans outperform frontier proprietary models on this subset. The human accuracy reaches 67.6%, while Gemini-3-Pro and Gemini-2.5-Pro achieve only 22.0% and 20.0%, respectively. This large gap indicates that current video MLLMs remain far from human-level performance on this benchmark.

Humans consistently outperform the models on fine-grained perception tasks. In particular, humans achieve much higher accuracy on *Counting*, *Small-Object*, *Spatial Orientation*, and *Object Tracking*. These capabilities require identifying subtle visual cues and tracking objects across frames, which remain challenging for current video MLLMs. Many questions require fine-grained inspection, so participants typically need around one hour to complete ten questions.

In contrast, the Gemini models perform relatively better on *World Knowledge*. This suggests that the models possess substantial common knowledge about everyday scenarios. For *Audio Perception*, human performance is relatively low. One possible reason is that participants may overlook audio cues while answering, or find it difficult to convert auditory information into precise textual responses. Overall, the results suggest that the primary bottleneck of current video MLLMs lies in fine-grained visual perception in long videos.

A.6 Evaluation Protocol Design

Level-4/5 require both correct answers and accurate evidence grounding. Our protocol uses separate prompts for answer generation, temporal localization, and spatial grounding. An alternative, joint prediction, requests the answer and required evidence in a single response. We explain why we use separate prompts and compare the protocols to assess their impact on evaluation results.

Design rationale. The protocol evaluates temporal search and spatial inspection as separate tasks alongside answering. For example, identifying the counseling-room scene and locating its empty chairs test different aspects of the evidence needed to answer the counting question (Figure 4).

Manual box annotations are available at a few fixed key frames. Because objects move or leave view, predictions must be compared with annotations at matching timestamps. We provide the annotated timestamps and include these frames in the visual input, allowing spatial localization to be evaluated without requiring frame selection. Evaluating arbitrary model-selected timestamps requires additional manual annotations, as in the joint-prediction experiment below.

Joint prediction requires stricter formats, such as JSON with separate fields for the answer, temporal intervals, and bounding boxes. In our trials, some post-trained video reasoning specialists fail to follow this multi-field format. We use separate prompts to limit each prediction's formatting demands. Standalone tIoU and vIoU measure grounding regardless of answer correctness.

Table 11: **Human performance on a 50-question subset of VideoZeroBench.** The subset covers all 13 categories and includes 27 Chinese and 23 English questions. We report the overall Level-3 accuracy and results on several representative atomic capabilities.

Method	Acc.	Count.	Small-Obj.	Spatial Ori.	Obj. Track.	World Know.	Audio Perc.
Gemini-3-Pro	22.0	12.0	0.0	20.0	42.9	50.0	**50.0**
Gemini-2.5-Pro	20.0	12.0	11.8	26.7	28.6	50.0	0.0
Human	**67.6**	**64.4**	**70.4**	**64.4**	**63.5**	**75.0**	42.0

Table 12: Separate versus joint prediction on a shared random 100-question subset (accuracy in %).

Model	Level-3	Level-4		Level-5	
		Separate	Joint	Separate	Joint
Gemini-3-Pro	23.0	10.0	7.0	2.0	1.0
Gemini-2.5-Pro	21.0	9.0	7.0	2.0	0.0
Qwen3-VL-4B	9.0	3.0	2.0	0.0	0.0

Relation to joint prediction. In joint prediction, the model receives only the video and question. At Level-4, it predicts the answer and temporal intervals in a single response. At Level-5, it predicts the answer, temporal intervals, and corresponding object bounding boxes with model-selected key-frame timestamps in a single response.

Our protocol reduces joint-generation demands through task-specific prompts and provides annotated key-frame timestamps for spatial grounding. These relaxations motivate interpreting its scores as an optimistic estimate of joint answer-and-grounding performance. Since MLLM predictions depend on the prompting conditions, this interpretation is heuristic and does not establish a theoretical upper bound. The shared-subset results in Table 12 are consistent with this interpretation for the evaluated models.

Experimental verification. We compare the protocols on the same randomly sampled 100-question subset, with additional ground-truth boxes annotated at model-selected timestamps. This annotation cost limits the comparison to a subset. Joint prediction yields no accuracy improvement for any of the three models at either level (Table 12). Level-5 accuracy is 0–2% with our protocol and 0–1% with joint prediction. Both settings show low answer-and-grounding accuracy for the three tested models on this subset.

A.7 Robustness of Evaluation Results

We examine whether variation across runs affects our main findings. We evaluate five representative models five times at temperature 0.7, keeping the questions and other evaluation settings fixed.

Table 13 reports the mean accuracy and standard deviation for each model and level. All reported standard deviations are below 0.7 percentage points. These results characterize decoding variability on the fixed question set. They do not estimate uncertainty from sampling different questions or videos.

A.8 More Visualization and Error Cases

To better illustrate typical reasoning failures of current video MLLMs, we present several representative examples from VideoZeroBench in Figure 10, 11, and 12. These cases show errors in temporal and spatial grounding, including incorrect evidence predictions accompanying correct answers. They also illustrate weaknesses in counting, small-object perception, and spatial orientation.

A.9 Related Works

Video MLLM Benchmarks. Most video MLLM benchmarks (Fu et al., 2024a; Hu et al., 2025; Wu et al., 2024; Wang et al., 2025e; Li et al., 2024; Tan et al., 2025; Zhao et al., 2025; Fang et al., 2024; Liu et al., 2024) evaluate models' ability to understand dynamic scenes, integrate multimodal signals,

Table 13: Accuracy (%, mean ± standard deviation) over five runs at temperature 0.7.

Model	Level-1	Level-2	Level-3	Level-4	Level-5
Gemini-3-Pro	23.6 ± 0.36	20.2 ± 0.50	16.9 ± 0.30	7.8 ± 0.35	0.9 ± 0.09
Gemini-2.5-Pro	25.3 ± 0.48	21.0 ± 0.35	16.4 ± 0.32	7.3 ± 0.22	0.6 ± 0.09
Qwen3-VL-4B	22.4 ± 0.52	16.5 ± 0.58	7.7 ± 0.31	1.2 ± 0.15	0.0 ± 0.00
Qwen2.5-VL-7B	16.1 ± 0.44	12.4 ± 0.50	5.8 ± 0.33	0.0 ± 0.09	0.0 ± 0.00
VideoChat-R1.5-7B	17.6 ± 0.50	12.2 ± 0.68	8.2 ± 0.42	0.0 ± 0.09	0.0 ± 0.00

and reason over long videos. Video-MME (Fu et al., 2024a) provides a comprehensive evaluation suite across short, medium, and long videos. MVBench (Li et al., 2024) focuses on temporally sensitive tasks, such as action understanding, object interaction, and event reasoning. MMBench-Video (Fang et al., 2024) extends multimodal benchmark design to video scenarios with broad coverage of perception and reasoning skills. LVBench (Wang et al., 2025e) and LongVideoBench (Wu et al., 2024) stress long-form video comprehension, where key information may be distributed across distant temporal positions. Video-MMMU (Hu et al., 2025) targets knowledge-intensive instructional videos and evaluates whether models can connect visual evidence with domain concepts. CG-Bench (Chen et al., 2024) introduces clue-grounded question answering and tests whether models can localize supporting temporal clues. ALLVB (Tan et al., 2025) and MMVU (Zhao et al., 2025) broaden evaluation toward diverse video categories and university-level multimodal understanding. TempCompass (Liu et al., 2024) focuses on temporal reasoning, including event order, duration, frequency, and temporal relation understanding. These benchmarks have greatly advanced video MLLM evaluation. However, many of them rely on multiple-choice protocols or coarse answer accuracy, which may overestimate model capability. They also rarely verify whether a correct answer is supported by precise spatio-temporal evidence. In contrast, VideoZeroBench is designed to probe the performance boundary of current video MLLMs through substantially harder open-ended questions and a hierarchical evaluation protocol that exposes unresolved challenges in answering, temporal grounding, and spatial grounding.

Detailed Spatial Benchmarks. Another line of work studies fine-grained visual perception and deeper reasoning grounded in visual evidence. In the image domain, ZeroBench (Roberts et al., 2025) constructs extremely difficult visual questions that require careful inspection of subtle image details. BLINK (Fu et al., 2024b) evaluates perception abilities that are easy for humans but difficult for MLLMs, such as relative depth, counting, visual correspondence, and spatial relation. TreeBench (Wang et al., 2025a) emphasizes traceable visual reasoning, where models must follow structured evidence paths rather than relying on shortcut recognition. These image benchmarks reveal that strong global semantics do not imply reliable detailed perception. In videos, related efforts mainly fall into two directions: spatio-temporal localization and spatial intelligence. V-STaR (Cheng et al., 2025), Know-Show (Sugandhika et al., 2025), and ToG-Bench (Xu et al., 2025) evaluate object and event grounding through spatio-temporal reasoning, action-conditioned evidence, and embodied tasks. Dr. V (Luo et al., 2025) uses spatio-temporal grounding to diagnose video hallucination on examples derived from existing public datasets. In contrast, VideoZeroBench uses newly collected long videos with fully human-annotated questions, answers, temporal intervals, and spatial boxes. VidSTG (Zhang et al., 2020) and HC-STVG (Tang et al., 2021) provide classic settings for spatio-temporal video grounding, where models localize queried objects or events in space and time. However, they focus on grounding given queries rather than solving hard QA problems. VSI-Bench (Yang et al., 2025a), SpatialTree (Xiao et al., 2025), and MMSI-Video-Bench (Lin et al., 2025) instead examine spatial relations, motion direction, and structural understanding in dynamic scenes. However, they mainly evaluate spatial intelligence itself, without requiring joint QA and evidence localization.

Thinking-with-Videos. Recent multimodal models such as Gemini-3 (Google, 2025), GPT-5.2 (OpenAI, 2025a), Qwen3-VL (Bai et al., 2025a), and post-trained variants like Video-R1 (Feng et al., 2025) already demonstrate structured reasoning and improved temporal understanding. Inspired by OpenAI-o3 (OpenAI, 2025b), a growing line of work (Yan et al., 2025; Ding et al., 2025; Zhang et al., 2025a; Ouyang et al., 2025; Meng et al., 2025; Wang et al., 2025b; Zeng et al., 2026; Zhang et al., 2025b;a; Wang et al., 2025d; Fu et al., 2025; Yuan et al., 2025; Ge et al., 2025; Yang et al., 2025b; Tang et al., 2025; Gao et al., 2025) studies thinking-with-videos, where reasoning is interleaved with explicit visual grounding. VideoChat-R1.5 (Yan et al., 2025), VideoZoomer (Ding et al., 2025)



Figure 10: Examples from **VideoZeroBench** organized by video category, with predictions from several models. The left side shows the annotated spatio-temporal evidence. In the text on the right, green indicates correct answers or IoU > 0.3, while red indicates incorrect answers or IoU ≤ 0.3. Spatial coordinates are normalized to the $[0, 1]$ range for key timestamps. **(1)** The model answers correctly but fails to locate the correct temporal evidence. **(2)(3)** Gemini and GPT show relatively accurate localization, but still make mistakes in counting or spatial orientation. Qwen3-VL-235B achieves Level-5 in (3).

and Video-o3 (Zeng et al., 2026) focus on temporal evidence seeking via iterative zoom-in to refine relevant intervals. Open-o3-Video (Meng et al., 2025) directly emits timestamps and bounding boxes within a reasoning trace without external tool calls. These approaches make evidence more explicit and motivate evaluation on difficult long-video questions with fine-grained manual evidence

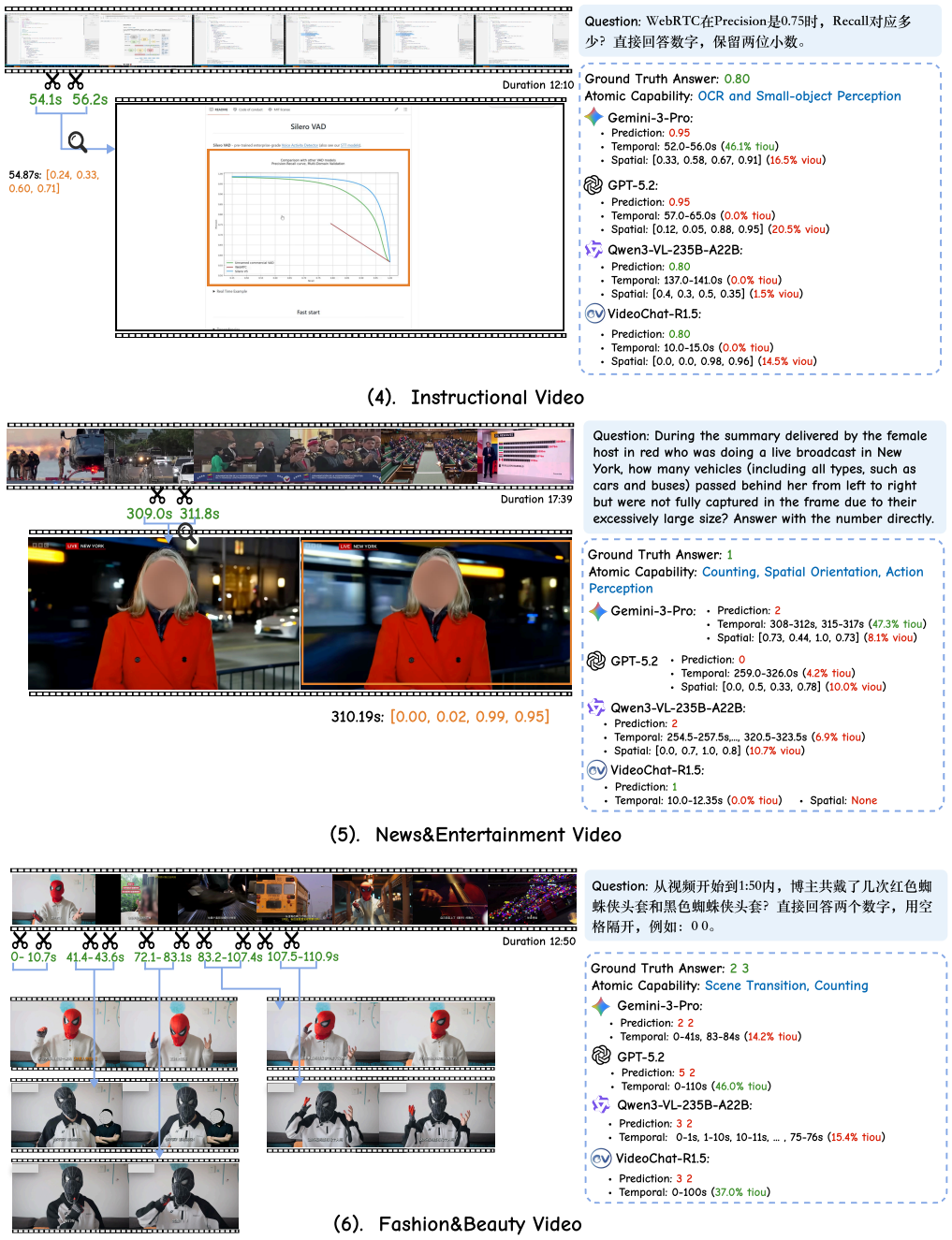

Figure 11: Examples from **VideoZeroBench** by video category, with model predictions. **(4)** Models struggle to localize and interpret chart information, even when the final answer is correct. **(5)(6)** Models perform poorly on counting tasks under complex constraints with easily confused objects.

annotations. Our benchmark is designed to fill this gap by evaluating challenging questions and explicitly verifying temporal and spatial evidence.

A.10 Limitations and Future Works

VideoZeroBench focuses on deliberately challenging questions from a limited collection of videos. Its results characterize this diagnostic set. Broader domains, languages, and video sources are needed to

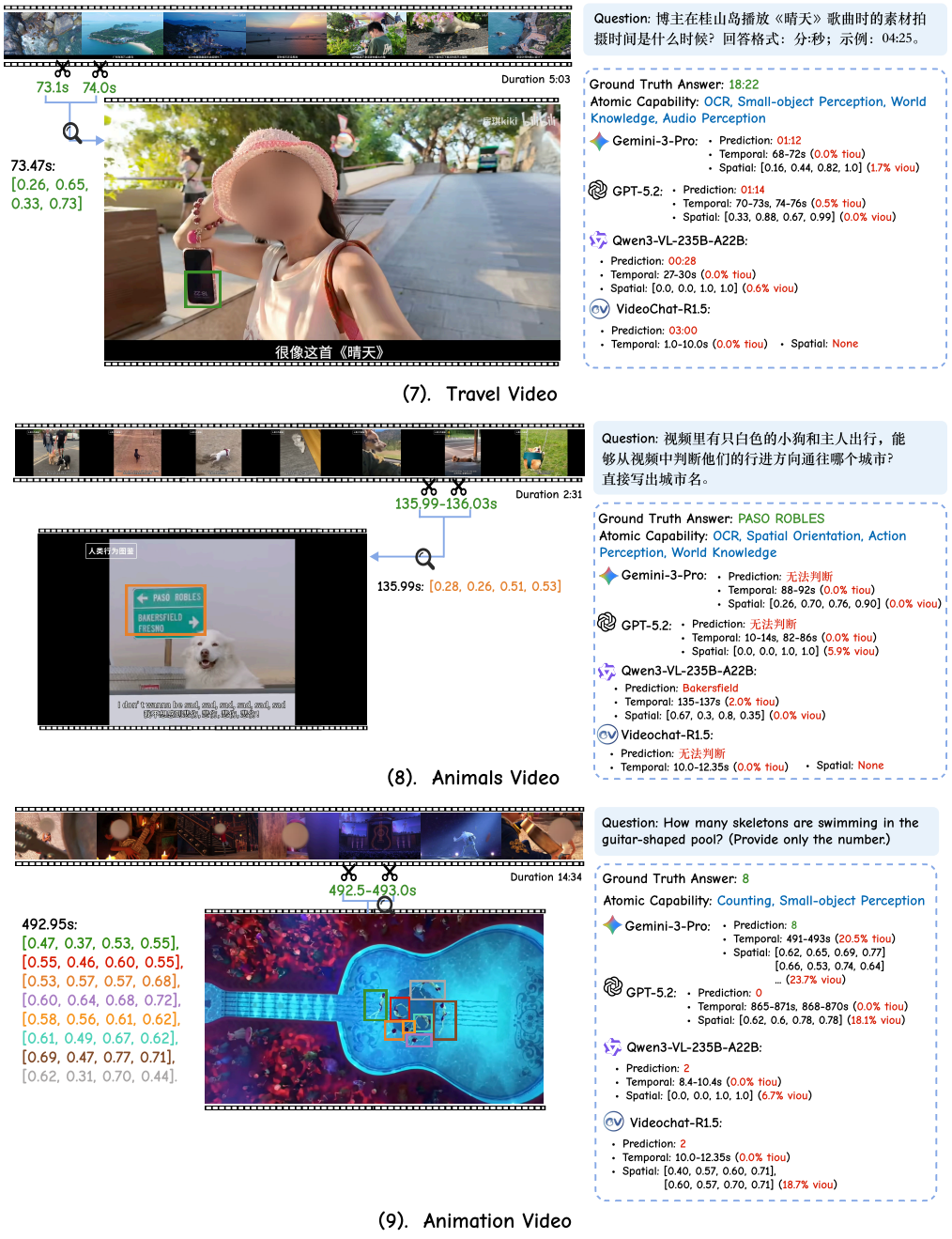

Figure 12: Examples from **VideoZeroBench** with model predictions. **(7)** Models struggle with a small visual cue in an audio-conditioned question. **(8)** The annotated event lasts 0.04s. A 135–137s prediction covers it but receives only 2.0% tIoU, illustrating sensitivity to temporal precision. **(9)** Gemini answers correctly, but its wider temporal window and spatial predictions fall below the grounding thresholds.

assess generalization. Evidence annotations consist of temporal intervals and sparse key-frame boxes. Extending coverage and adding trajectories or dense masks would support finer-grained evaluation. These extensions can also help assess thinking-with-videos systems that actively search, zoom, and verify evidence before answering.