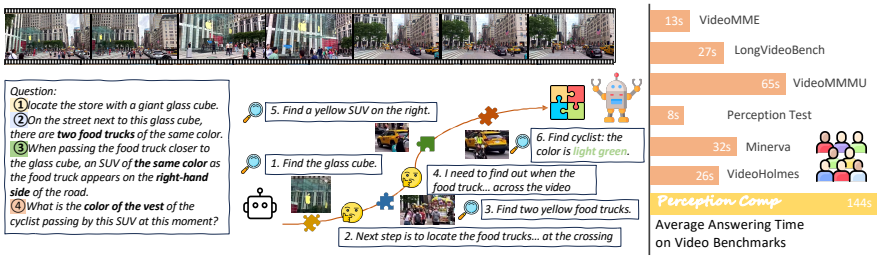
PerceptionComp: A Video Benchmark for Complex Perception-Centric Reasoning

Shaoxuan Li[*1] Zhixuan Zhao[*1] Hanze Deng[*1] Zirun Ma[*1]
Shulin Tian[3] Zuyan Liu[1] Yushi Hu[2] Haoning Wu[3]
Yuhao Dong[#3] Benlin Liu[#2] Ziwei Liu[†3] Ranjay Krishna[†2]

[1] Tsinghua University
[2] University of Washington
[3] Nanyang Technological University

Abstract. Deep video understanding requires long-horizon, perception-centric reasoning that repeatedly revisits a video to gather temporally distributed evidence. However, existing benchmarks are either relatively easy (perception-centric but often solvable after a single view) or logic-heavy with simplified visuals, and thus do not faithfully measure multimodal test-time thinking that depends on repeated perception. We introduce **PerceptionComp**, a fully manually annotated benchmark designed so that no single moment is sufficient: answering requires evidence from multiple temporally separated segments under compositional constraints. PerceptionComp contains **1,114** five-choice questions over **279** high-scene-complexity videos spanning diverse domains. Videos are selected using automatic proxies for scene complexity (SAM2 instance counts and optical-flow magnitude), and each question requires **10–20 minutes** of annotation. Human evaluation confirms the intended difficulty: PerceptionComp requires substantially longer response times than prior benchmarks, and under a single-view setting (no rewatching) human accuracy drops to near chance (**18.97%**), while experts can reach **100%** accuracy with unrestricted rewatching and sufficient time. State-of-the-art MLLMs perform notably worse: the best model in our evaluation (Gemini-3-Flash) reaches only **45.96%** accuracy, and open-source MLLMs remain below **40%**. Test-time reasoning helps but remains far from human-level (e.g., GPT-o3 exceeds GPT-4o by **11.04%**; Gemini-2.5-Pro exceeds Gemini-2.5-Flash by **6.19%**), and increasing test-time compute via larger thinking-token budgets or more input frames further improves performance. Finally, among the strongest frontier models we tested (Gemini-3 variants and GPT-o3), accuracies cluster in the mid-40s, suggesting a bottleneck in perception-centric long-horizon video reasoning. PerceptionComp provides a focused testbed for diagnosing these limitations and advancing multimodal visual thinking.

[*]Equal contribution. [#]Project co-lead. [†]Equal advising.



(a) Perception-Centric Reasoning (b) Human Study

Fig. 1: Overview of the PerceptionComp benchmark. (a) An example from PerceptionComp, where models are required to perform complex, perception-centric reasoning with various types of subconditions to arrive at the final answer. (b) Results from a human study measuring question-answering time, showing that PerceptionComp is more challenging for humans than previous perception and reasoning video benchmarks, largely due to its emphasis on perception-centric reasoning.

1 Introduction

Videos capture human activities and the physical world, and multimodal intelligence—from robots to AI glasses—must achieve *deep* video understanding to be broadly useful. Consider a seemingly simple query: *"On which floor did the person last appear in the video before dropping their* ***apartment*** *keys (not their office keys)?"* Answering it requires long-horizon, step-by-step repeated perception that composes multiple perceptual skills: *semantic recognition* (identify which object is a key), *correspondence* (track the apartment key rather than the office key), *temporal reasoning* (locate the drop event and trace back to the previous time the key appears), and *spatial reasoning* (infer the building layout and floor level). Recent breakthroughs in long-horizon reasoning for mathematics and coding suggest that *test-time scaling*—allocating more computation during inference—is a promising route for enabling multimodal language models (MLLMs) to perform such multi-step, perception-driven video reasoning [8, 12, 15, 21]. For deep video understanding, this should not mean only longer language-side thinking; it should also mean composing multiple perception skills and repeatedly revisiting the video to gather visual information across different dimensions.

However, existing video benchmarks do not adequately measure this capability. Many widely used benchmarks (e.g., VideoMME, Perception Test) [4,9,14,30] are perception-centric but relatively easy: humans can often answer after a single viewing with minimal deliberation (Fig. 1), leaving limited room to differentiate models' test-time thinking ability. In contrast, benchmarks that demand substantial reasoning, such as geometry or maze solving [32, 43], often derive difficulty primarily from logical structure rather than real-world perception, since their visual inputs are synthetic or overly simple. Even long-video understanding benchmarks [42] frequently stress memory more than evidence-seeking reasoning. As a result, there is still no benchmark that is simultaneously long-horizon, perception-centric, and truly forces repeated visual information gathering.

To fill this gap, we introduce **PerceptionComp**, a manually annotated benchmark for *complex, long-horizon, perception-centric* video reasoning. PerceptionComp is constructed so that no single moment is sufficient: solving a question requires multiple temporally separated pieces of visual evidence and compositional constraints. Concretely, each question combines several perceptual sub-conditions under two logics—*conjunctive* and *sequential*—so the model must satisfy multiple constraints and track entities/states across time. Each subcondition is itself a perceptual subtask that requires extracting diverse visually grounded information from the video, including objects, attributes, relations, locations, actions, and events. Completing these subtasks requires a range of perceptual skills, including semantic recognition, visual correspondence, temporal reasoning, and spatial reasoning; some questions additionally involve commonsense knowledge tightly linked to the visual content and simple near-future prediction from ongoing dynamics.

We manually annotate PerceptionComp following this question design on 279 videos drawn from diverse domains, such as city walk tours, large indoor villa tours, video games, and extreme outdoor sports, resulting in 1,114 highly complex questions. We quantify scene complexity using automatic signals: specifically, we use the number of instances detected by SAM2 [33] and optical-flow magnitude [38] as proxies for object density, motion intensity, and scene-change dynamics, and we select high-complexity videos with many objects, intense motion, and frequent transitions. Beyond video complexity, each question has high compositional complexity and requires multiple perceptual reasoning skills to extract different aspects of visual evidence. To ensure correctness under this extreme difficulty, we adopt 100% manual annotation: each question takes 10–20 minutes from video selection to final annotation. We use a five-choice format for reliable evaluation and report accuracy as the metric; answer options are designed to be plausible and closely confusable, so disambiguation requires video evidence rather than option priors.

We first evaluate PerceptionComp with humans and treat human performance as a gold standard for benchmark quality. In our human study, participants watch the video and answer; during answering, they may rewatch the video as needed. We measure response time and find that participants take substantially longer on PerceptionComp than on prior benchmarks (Fig. 1): more than 2× longer than VideoMMMU [14], more than 10× longer than VideoMME [9], more than 5× longer than Video-Holmes [4], and more than 5× longer than LongVideoBench [42], even though most of our videos are not longer than 10 minutes. This shows that video context length is not the only dimension of video thinking. We further evaluate a stricter setting where participants may watch the video only once and cannot rewatch while answering; accuracy in this setting is near chance (18.97%), even with extended thinking time. Together, these results show that PerceptionComp (i) requires substantial test-time thinking to solve and (ii) cannot be solved without repeated perception steps, ruling out shortcuts based on single-view memory or language priors. Given its coverage of

Table 1: Comparison of PerceptionComp with other benchmarks. PerceptionComp emphasizes perception-centric compositional reasoning, requiring models to integrate temporally distributed visual evidence.

Benchmark	Properties				
	Video Domain	# Videos	# QA	Per.Rea	Annotation
MMVU [50]	Specialized videos	1,529	3,000	✗	Manual
VideoMME [9]	YouTube videos	900	2,700	✗	Manual
VCR-Bench [31]	Short films	859	1,034	✗	Manual
MINERVA [25]	Mix	223	1,515	✗	Manual
VideoMMMU [14]	Lectures	300	900	✗	Manual
Video-Holmes [4]	Short films	270	1,837	✗	Automatic&Manual
PerceptionComp	**In-the-wild videos**	**279**	**1,114**	✓	**Manual**

diverse perceptual skills, PerceptionComp also serves as a testbed for perceptual competence.

We further evaluate PerceptionComp on state-of-the-art MLLMs. While these models achieve strong results on existing benchmarks, they perform notably worse on PerceptionComp. Even the best-performing model in our evaluation (Gemini-3-Flash [10]) reaches only 45.96% accuracy in the five-choice setting, and open-source MLLMs [1, 2, 36, 40, 45] remain below 40%. In contrast, human participants can reach 100% accuracy when given sufficient time with unrestricted rewatching. Moreover, test-time reasoning helps, but performance remains far from human-level. Thinking models outperform their non-thinking counterparts: GPT-o3 [29] surpasses GPT-4o [16] by 11.04%, and Gemini-2.5-Pro exceeds Gemini-2.5-Flash by 6.19%. By controlling the thinking-token budget for Gemini-2.5-Flash [6], we further show that allocating more test-time tokens improves performance. We also find that providing more input frames boosts accuracy for both GPT-o3 [29] and Qwen3-VL-8B [1], consistent with the benchmark's reliance on temporally distributed evidence. Finally, among the strongest frontier models we tested (Gemini-3 variants and GPT-o3), despite different architectures/interfaces, accuracies cluster in the mid-40s, suggesting a potential bottleneck in perception-centric long-horizon video reasoning. To better understand this regime, we analyze representative failure cases of frontier models and characterize common bottlenecks. We hope PerceptionComp will help the community recognize these limitations and drive progress in perceptual reasoning.

2 Related Work

General Video Understanding Benchmarks. Existing video benchmarks evaluate a wide range of abilities, from short-term action and object understanding [20, 44, 46, 49] to long-video memory and narrative comprehension [9, 34, 35, 39, 42]. Diagnostic benchmarks such as the Perception Test [30] and egocentric datasets [11, 24] further improve realism and perceptual coverage. However, many questions in these benchmarks can still be answered from a limited number of salient moments, or mainly test memory, narration, and single-turn QA.

PerceptionComp instead makes difficulty *perception-bottlenecked*: answering requires repeatedly locating, grounding, and integrating fine-grained evidence from temporally separated segments.

Complex Multimodal Reasoning Benchmarks. Recent multimodal reasoning benchmarks go beyond recognition and evaluate structured inference across vision and language. Image and multi-image benchmarks such as MME-CoT and related suites [13, 17, 47], ScienceQA [23], and EXAMS-V [7] emphasize mathematical, scientific, logical, or academic reasoning. In video, VideoCogQA [18], MMVU [50], VideoMMMU [14], VideoMathQA [32], VCR-Bench [31], MINERVA [25], Video-Holmes [4], and HumanPCR [19] evaluate increasingly complex temporal, causal, academic, narrative, or human-centric reasoning. Nevertheless, their difficulty is often dominated by logical, semantic, domain-specific, narrative, or human-centric reasoning. In contrast, PerceptionComp targets *perception-centric compositional reasoning*: each question composes perceptual primitives over objects, attributes, relations, events, and states under conjunctive or sequential constraints. Annotation verifies answer uniqueness and subcondition necessity, ensuring that no single cue or short segment is sufficient. As shown in Table 1, PerceptionComp is also comparable in video scale to recent complex video reasoning benchmarks while using diverse in-the-wild videos.

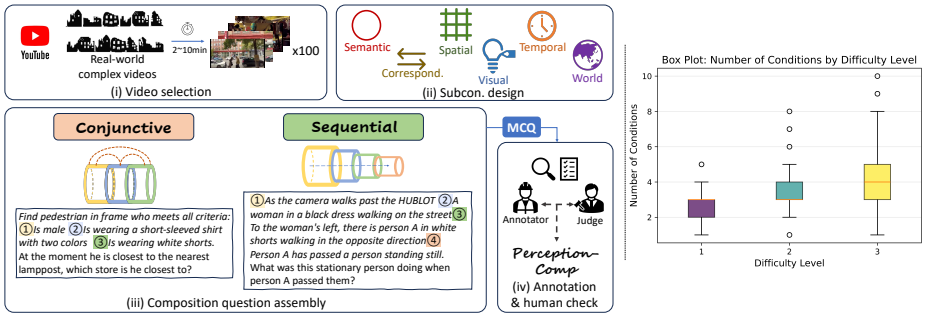
Multimodal Reasoning Models. Modern MLLMs, including proprietary GPT-style and Gemini-style systems [6,16] and open-source families such as Qwen-VL, InternVL, and Molmo [1, 2, 5, 40], increasingly support end-to-end video QA. Recent reasoning-oriented work further extends RLVR-style or long-form reasoning pipelines to multimodal settings, including Vision-R1/VisualRFT [15, 22], DeepEyes [51], Video-R1 [8], and VideoChat-R1 [21]. However, most evaluations still use a passive interface where models receive fixed frames or fixed video inputs. PerceptionComp naturally stresses *test-time perception scaling*: models must benefit not only from longer reasoning, but also from active evidence acquisition, segment revisiting, and compositional use of temporally distributed visual evidence.

3 PerceptionComp

We design PerceptionComp to evaluate video thinking by enforcing long-horizon, perception-centric reasoning. We do so in two ways: (i) selecting structurally complex videos, and (ii) composing questions from multiple subconditions that probe different perceptual skills, thereby increasing compositional complexity. Fig. 2(a) summarizes our end-to-end manual annotation pipeline, and Fig. 3 provides representative question examples across difficulty levels and composition types. Below, we describe our video selection process, the question–answer format, the annotation pipeline, and our difficulty annotation.

3.1 Video Selection

Many existing video benchmarks use clips that are visually simple: they often depict a single event or activity and contain only a small number of humans or



(a) Data Annotation Pipeline (b) Benchmark Statistics Distribution

Fig. 2: Data construction and statistics of PerceptionComp. (a) Annotation pipeline, which integrates diverse subconditions and supports two types of compositional questions. (b) Benchmark statistics: higher difficulty levels contain more subconditions, increasing the demand for perception-centric reasoning.

objects. As a result, many videos can be approximately replaced by a short textual caption without substantially affecting downstream performance, limiting their ability to diagnose perceptual competence.

To better probe perception-related abilities, we deliberately select videos with high scene and object complexity. Our videos span seven categories: city-walk tours (outdoor), shopping in malls, sports competitions, indoor home/villa tours, variety shows, movie clips, and game livestreams. These videos typically contain many objects, frequent scene transitions, and substantial camera motion, making them difficult to summarize with a single caption. The selected clips range from 2 to 10 minutes in length. Unlike benchmarks that increase difficulty primarily by extending duration, we also increase difficulty along an orthogonal axis: dynamic scene complexity.

We quantify complexity using automatic signals: we use the number of instances detected by SAM2 [33] and optical-flow magnitude [38] as proxies for object density, motion intensity, and scene-change dynamics, and prioritize clips with many objects, intense motion, and frequent transitions. All videos are sourced from real recordings rather than synthetic renderings; while some categories (e.g., game livestreams) are screen-captured, the videos still exhibit rich, naturally occurring dynamics and clutter that make the tasks challenging and practically relevant. This combination of dynamic scene complexity and rapidly changing content forces models to repeatedly gather and integrate visual evidence across temporally separated moments rather than relying on a coarse global summary. We defer more detailed video statistics to the supplementary.

3.2 Subconditions and Perceptual Skills

We explicitly increase compositional difficulty by combining multiple subconditions into a single query. Each subcondition targets a distinct perceptual–reasoning skill, so solving the full question requires coordinated use of multiple

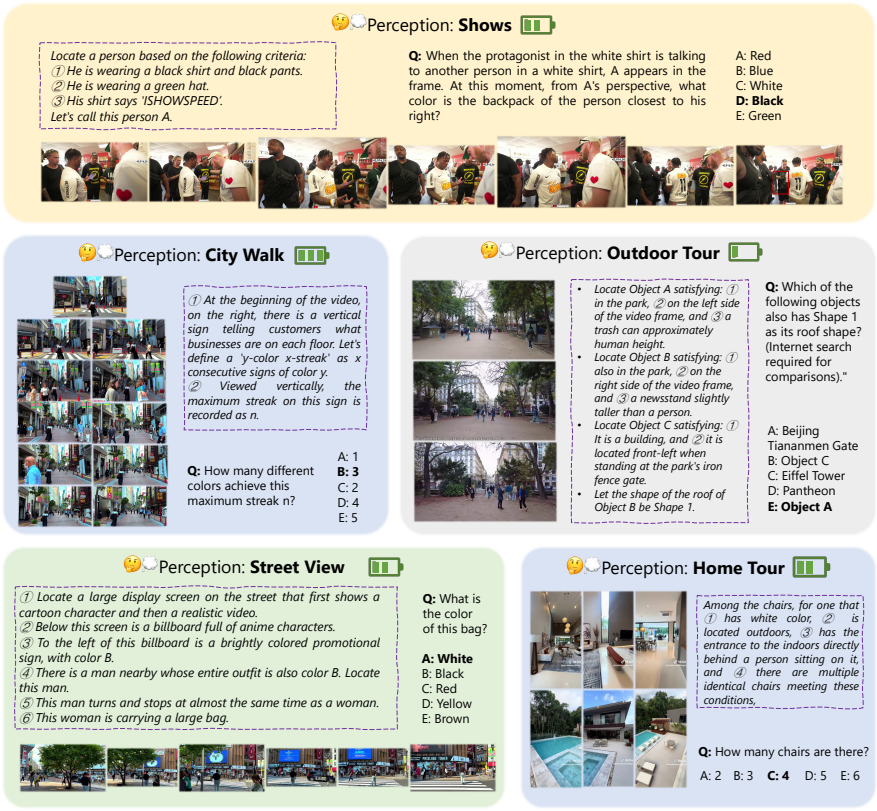


Fig. 3: Examples from PerceptionComp. , , and denote difficulty levels 1, 2, and 3, respectively. PerceptionComp spans diverse video sources and uses subconditions to construct conjunctive and sequential questions that require perception-centric reasoning.

abilities rather than a single narrow competence. Concretely, our subconditions cover:

- **Semantic understanding**: Recognize object categories, attributes (e.g., shape, color, material), and higher-level semantic relations (e.g., roles or interactions).
- **Spatial understanding**: Reason about scene layout and relative geometry (e.g., left/right, front/behind, near/far) and occlusion.
- **Temporal understanding**: Follow motion patterns and localize events in time (e.g., what happens before/after a reference event).
- **Correspondence**: Match instances or parts across time/views (e.g., tracking the same object across shots, or part–whole matching).
- **Visual knowledge**: Some questions require commonsense knowledge tightly coupled to visual content.

– **World modeling**: Some questions involve simple near-future prediction from ongoing dynamics.

By sampling and composing subconditions from these categories, each question assesses perceptual competence in complex video reasoning more comprehensively than single-skill probes. Example subconditions and their compositions are illustrated in Fig. 3.

3.3 Compositional Question Design

We combine subconditions into full questions using two composition logics. **Conjunctive composition:** All subconditions refer to the *same* target, forming an "and"-style conjunction. To ensure every subcondition matters, we verify that no proper subset uniquely determines the answer: each condition only removes part of the candidate set, and only the full conjunction yields a single solution. This prevents shortcuts where the model can ignore part of the query.
Sequential composition: Subconditions must be resolved *in order*, where later subconditions depend on intermediate entities or states established earlier. For example, a first subcondition identifies an object, the second constrains its behavior at a later time, and a third asks about a relation involving that same object after another event. The model must carry the referent forward across steps, inducing a multi-hop perceptual reasoning process in which early errors propagate.

3.4 Answer Space

Each question is formed by composing subconditions, and the final answer is a piece of perceptual information extracted from the video. Answers fall into six categories:

– **Objects**: Category names (e.g., "car", "sofa").
– **Attributes**: Properties such as color, count, or shape.
– **Relationships**: Semantic/spatial/social relations between entities.
– **Location**: Place descriptors (e.g., room type, country, or region).
– **Action**: The name of an action performed by an agent.
– **Event**: A higher-level event or composite situation that occurs in the video.

We cast every question as a five-way multiple-choice problem. To discourage reliance on language priors, all distractors are drawn from the *same* answer category as the correct option (e.g., all colors or all object categories). Each option is constrained to a single word or a very short phrase, minimizing extra linguistic cues.

3.5 Difficulty Annotation

To enable difficulty-aware evaluation, we ask expert annotators to assign each question to one of three difficulty levels (Level 1/2/3) based on both (i) the

Table 2: Comprehensive evaluation results of MLLMs on PerceptionComp. We report both category-wise accuracies and accuracies across different difficulty levels.

Model	Size	Frame	Accuracy by Category							Accuracy by Difficulty			Overall
			Outdoor	Shopping	Sport	Home	Show	Movie	Game	Level 1	Level 2	Level 3	
Human Performance													
Expert (unrestricted rewatch)	-	-	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	**100.00**
Human	-	-	81.33	86.80	87.56	86.72	87.25	88.00	87.10	90.18	86.00	72.25	**85.10**
Single-view Human (no rewatch)	-	-	19.40	17.80	20.60	18.20	21.10	17.00	18.90	20.30	19.10	17.60	**18.97**
Proprietary Models													
Gemini-3-Flash [10]	-	-	45.27	45.18	38.34	42.97	59.73	52.00	48.39	43.75	47.26	47.85	45.96
Gemini-3-Pro [10]	-	-	42.20	42.13	40.41	37.50	60.40	48.00	61.29	45.09	45.51	40.67	44.43
Gemini-2.5-Pro [6]	-	-	45.78	45.18	32.64	42.19	52.35	52.00	58.06	44.64	44.20	44.02	44.34
Seed2.0 Pro [3]	-	64	43.73	47.72	30.57	47.66	51.68	40.00	70.97	48.21	45.51	33.49	44.34
Gemini-3.1-Pro [10]	-	-	39.90	44.16	39.90	42.97	56.38	48.00	51.61	45.76	45.08	36.36	43.72
GPT-o3 [29]	-	50	43.22	46.70	32.64	44.53	50.34	56.00	48.39	43.08	43.98	43.54	43.54
GPT-5.2 [28]	-	64	42.97	44.16	27.46	38.28	48.99	48.00	38.71	44.42	38.07	38.76	40.75
Gemini-2.5-Flash [6]	-	-	39.39	41.12	24.35	39.84	47.65	44.00	32.26	41.96	35.89	34.93	38.15
GPT-5 [26]	-	64	26.60	47.21	29.53	40.62	49.66	56.00	38.71	40.18	36.32	28.71	36.45
GPT-4.1 [27]	-	50	26.09	46.70	27.46	28.91	44.97	52.00	48.39	37.28	34.79	24.40	34.02
GPT-4o-latest [16]	-	50	30.18	36.04	25.39	32.81	40.94	48.00	29.03	35.04	30.63	29.67	32.50
Open-Source Instruct Models													
Qwen2.5-VL [2]	7B	64	26.34	24.87	13.54	17.97	26.85	20.00	22.58	24.11	21.44	22.60	22.73
InternVL-3.5 [40]	8B	64	31.20	35.03	27.98	25.78	41.61	48.00	38.71	34.82	31.51	30.62	32.32
Qwen3-VL [1]	8B	64	34.53	33.50	30.05	29.69	45.64	52.00	38.71	34.60	34.57	33.01	34.06
Qwen3-VL [1]	30B	64	32.48	42.64	20.21	31.25	48.32	27.00	45.16	38.03	34.87	25.48	34.38
Qwen2.5-VL [2]	72B	64	32.99	35.53	19.69	23.44	44.97	20.00	41.94	34.82	29.10	28.71	31.33
GLM-4.5V [37]	106B	64	36.57	37.56	30.77	28.12	51.01	52.00	22.58	39.10	34.37	36.59	36.69
Qwen3-VL [1]	235B	64	39.64	36.04	23.83	32.03	40.27	23.00	0.00	35.57	33.99	30.77	34.02
Open-Source Thinking Models													
Video-R1 [8]	7B	64	28.31	27.16	16.43	20.09	30.22	23.00	24.87	26.63	24.38	25.42	26.27
VideoChat-R1 [21]	7B	64	31.42	29.68	19.17	22.94	33.11	26.00	27.46	29.39	27.02	28.21	28.63
Qwen3-VL-Thinking [1]	8B	64	33.26	58.41	38.62	26.18	64.77	29.00	38.49	36.14	32.79	32.11	33.82
Qwen3-VL-Thinking [1]	30B	64	39.13	36.04	26.94	25.78	46.31	38.00	32.26	35.65	37.06	32.69	35.68
Qwen3-VL-Thinking [1]	235B	64	38.87	41.12	30.57	33.59	48.99	43.00	22.58	39.46	38.16	35.58	38.20

number of composed subconditions and (ii) the intrinsic difficulty of the subconditions. This avoids treating difficulty as a function of subcondition count alone. As shown in Fig. 2(b), higher difficulty levels typically contain more subconditions, reflecting increased compositional depth and stronger requirements for long-horizon, perception-centric reasoning.

3.6 Annotation Pipeline

Following the procedure above, we select 279 videos with high scene complexity and annotate 1,114 questions. Because each question is highly compositional, we adopt fully manual annotation to ensure correctness. Representative annotated examples are shown in Fig. 3. Annotators first create the subconditions and final answer, then verify that (i) the answer is uniquely determined by the video and (ii) every subcondition is necessary.

Each question is subsequently checked by at least one additional annotator who did not create it. During verification, we confirm again that there is a single correct answer and that no proper subset of subconditions suffices to uniquely identify it. Items that fail either requirement are revised or discarded. This protocol ensures each question admits a unique solution and genuinely requires the full set of composed perceptual subconditions.

4 Experiments

This section evaluates state-of-the-art multimodal LLMs (MLLMs) on PerceptionComp to quantify its difficulty and diagnose current model limitations. We first report comprehensive benchmark results across a broad set of open-source and proprietary models (Sec. 4.1). We then analyze how perception and reasoning budgets affect performance by varying the number of input frames and the allocated thinking-token budget (Sec. 4.3). Finally, we present qualitative case studies and error patterns to highlight common failure modes in perception-centric long-horizon video reasoning (Sec. 4.4). Throughout, our goal is not only to rank models, but also to identify which aspects of perception-centric long-horizon reasoning remain brittle under clutter, scene changes, and compositional constraints.

4.1 Evaluation Setup

Models. We evaluate a broad set of video MLLMs reported in Table 2, spanning (i) proprietary frontier models (Gemini-2.5/3 series, GPT-4o/4.1, GPT-5/5.2, GPT-o3), (ii) open-source instruction-tuned MLLMs (Qwen2.5-VL, Qwen3-VL at multiple scales, InternVL-3.5, GLM-4.5V), and (iii) open-source "thinking" and video-reasoning models (Video-R1, VideoChat-R1, and Qwen3-VL-Thinking variants). This coverage enables comparisons across proprietary vs. open-source systems, instruction-following vs. thinking-style variants, and different backbone scales under the same benchmark.

Input format and prompting. For models with native video inputs (e.g., Gemini), we directly feed raw videos without frame extraction. For models without native video support, we uniformly sample 64 frames per video as input; for certain GPT APIs we use 50 frames due to input-length constraints (Table 2). Proprietary models are evaluated with Chain-of-Thought prompting. For open-source models, instruction-tuned variants are prompted to output the answer choice directly, while thinking-style variants are prompted with Chain-of-Thought [41] (temperature 0.7, max generation length 16,384 tokens).

Human baselines. We report three human baselines. **Expert (unrestricted rewatch)** corresponds to a careful setting where an annotator is given sufficient time and can repeatedly rewatch and cross-check the video until confident, yielding 100% accuracy. **Human** corresponds to ordinary participants, who may lose patience as questions become highly compositional, leading to occasional mistakes even when rewatching is allowed. Finally, **Single-view Human** evaluates a stricter setting where participants may watch the video only once (no rewatch) and must answer from a single pass; performance drops to near random guess (overall 18.97%), highlighting that PerceptionComp cannot be solved from a single viewing or language priors alone and instead requires sustained, long-horizon evidence gathering and reasoning.

4.2 Overall Results

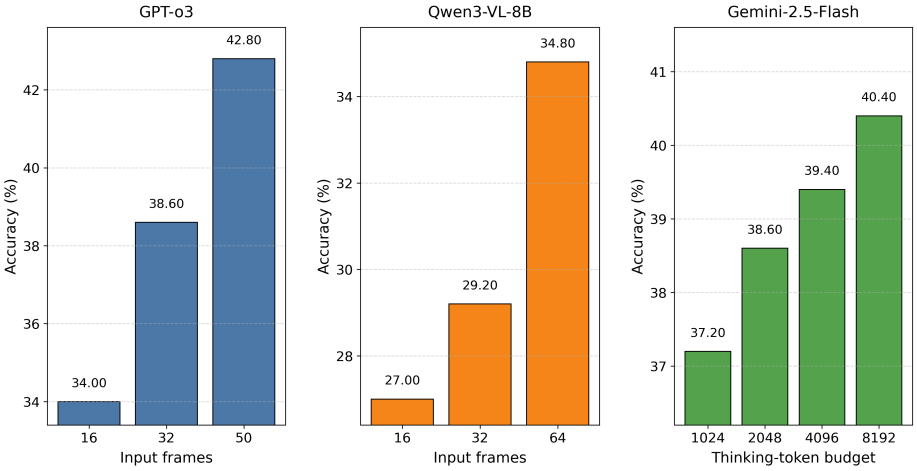
Benchmark performance. We report comprehensive results in Table 2. Most models achieve accuracy below 40%, indicating that PerceptionComp is challenging for current video MLLMs. The best-performing model in our evaluation is Gemini-3-Flash (45.96%). In contrast, strong open-source instruction-tuned models remain substantially lower (e.g., Qwen3-VL-8B: 34.80%, Qwen3-VL-235B: 34.02%). Scaling model size does not consistently improve performance, suggesting that PerceptionComp is bottlenecked less by generic capacity and more by reliably extracting fine-grained evidence under clutter and temporal discontinuities and integrating that evidence across multiple steps.

Thinking vs. instruct. We further compare instruction-tuned models with thinking-style or reasoning-trained variants. Overall, stronger test-time reasoning can help: GPT-o3 [29] surpasses GPT-4o [16] by 11.04%, and Gemini-2.5-Pro exceeds Gemini-2.5-Flash by 6.19%, suggesting that additional deliberation improves performance on PerceptionComp. However, the effect is not uniform. Some video-reasoning models (e.g., VideoChat-R1) benefit from explicit reasoning, while Qwen3-VL thinking variants can underperform their instruction-tuned counterparts (e.g., Qwen3-VL-Thinking-8B vs. Qwen3-VL-8B). This indicates that stronger language-side reasoning does not automatically translate into better perception-driven video reasoning: when perceptual evidence is misread or underspecified, longer reasoning may amplify errors rather than correct them. PerceptionComp makes this failure mode visible because it requires models to repeatedly ground intermediate steps in temporally separated visual evidence, where uncertainty compounds over multi-hop chains.

Category-wise and difficulty-wise trends. Table 2 further reports accuracy by video category and by difficulty level, where difficulty is annotated by experts based on both the number of subconditions and subcondition difficulty. Across models, accuracy drops substantially on Level 3 questions, indicating that current MLLMs struggle as compositional complexity increases. Notably, Level 3 questions require maintaining consistent intermediate hypotheses while repeatedly gathering evidence from different moments, stressing both temporal integration and error correction. These trends align with our human baselines: even humans require sustained effort and repeated verification to achieve perfect accuracy, underscoring PerceptionComp's strong demand for long-horizon, perception-centric reasoning.

4.3 Analysis

We conduct controlled analysis experiments to understand why PerceptionComp is difficult for current models. Specifically, we vary (i) the number of input frames (perception budget) and (ii) the thinking-token budget (reasoning budget), and measure the resulting accuracy changes. Due to evaluation budget constraints, all three analyses are conducted on a fixed subset of 100 videos (500 samples).



(a) GPT-o3: #Frames **(b)** Qwen3-VL-8B: #Frames **(c)** Gemini-2.5-Flash: Budget

Fig. 4: Analysis on PerceptionComp. Accuracy as a function of *perception budget* and *reasoning budget*. Left/middle: accuracy vs. the number of uniformly sampled input frames for GPT-o3 and Qwen3-VL-8B. Right: accuracy vs. the thinking-token budget for Gemini-2.5-Flash.

This analysis separates two common failure sources in video QA: insufficient visual evidence (too sparse sampling) versus insufficient deliberation (too short reasoning traces).

We further examine an active video-access setting where models can revisit selected video segments during reasoning. This analysis tests whether models benefit from actively acquiring missing visual evidence, rather than only passively consuming a fixed video input.

Effect of Input Frames To study how the density of temporal visual information influences perception-centric reasoning, we vary the number of input frames $\mathcal{F}$ for two representative models: GPT-o3 with $\mathcal{F} \in \{16, 32, 50\}$ and Qwen3-VL-8B with $\mathcal{F} \in \{16, 32, 64\}$. As shown in Fig. 4, both models improve monotonically as $\mathcal{F}$ increases: GPT-o3 rises from 34.0% (16 frames) to 43.54% (50 frames), and Qwen3-VL-8B gains 7.8 points from 27.0% (16 frames) to 34.80% (64 frames).

We attribute these gains to PerceptionComp being highly perception-centric: denser sampling increases the chance of capturing the relevant moments and provides richer visual evidence, which improves perceptual accuracy (e.g., identifying objects/attributes and tracking entities through motion and transitions), and in turn improves end-task accuracy. Importantly, the strong dependence on $\mathcal{F}$ also supports the intended benchmark property: PerceptionComp typically requires aggregating information from many frames and multiple temporally separated moments, rather than relying on a small number of salient frames.

Effect of Thinking Budget We further examine how reasoning effort affects performance by varying the thinking-token budget for Gemini-2.5-Flash (1,024 / 2,048 / 4,096 / 8,192 tokens). As shown in Fig. 4, larger thinking budgets consistently improve accuracy. This indicates that longer test-time thinking is beneficial for PerceptionComp: with more tokens, the model can better maintain intermediate hypotheses (e.g., entity identity, timing, and relational constraints), avoid premature commitment, and more reliably follow sequential subconditions to connect intermediate observations to the final answer.

Agentic Rewatching Method Increasing input frames or thinking tokens improves performance, but still assumes a fixed video input. To approximate human-like revisiting of uncertain moments, we evaluate two tool-based protocols: LongVT [48] for Qwen2.5-VL and a lightweight agentic-rewatch wrapper for Gemini-3-Flash, where models can select video segments to revisit before answering.

Active rewatching brings clear gains: LongVT improves Qwen2.5-VL from 22.73% to 28.80%, while agentic rewatch improves Gemini-3-Flash from 45.96% to 50.60%. This shows that active evidence acquisition helps beyond passive video input. However, PerceptionComp remains far from solved, indicating that success requires not only revisiting relevant segments, but also reliably grounding and composing temporally distributed evidence across semantic, spatial, temporal, correspondence, visual-knowledge, and world-modeling dimensions.

These results suggest that PerceptionComp is a useful testbed for *visual thinking*: it is sensitive to both perception and reasoning budgets, and exposes whether models can actively acquire, ground, and compose video evidence during long-horizon multimodal reasoning.

4.4 Case Study and Error Patterns

We present qualitative case studies to illustrate common failure modes on PerceptionComp. Fig. 5 analyzes Gemini-2.5-Pro and GPT-5: both models often localize the relevant moment or object category, but fail on fine-grained attributes or relations (e.g., confusing similar objects, missing cues under occlusion/viewpoint shifts, or misreading spatial relations). Such perceptual errors often cascade into reasoning failures: once an intermediate entity/state is wrong, the remaining chain stays internally coherent but drifts from the ground truth, especially for sequential questions where later conditions depend on earlier results. Overall, PerceptionComp highlights the tight coupling between perception and reasoning in long-horizon, evidence-integrating video QA.

We further analyze the Gemini-3 family and find that many remaining errors stem from mid-chain collapses driven by spatial misunderstanding. Specifically, we decompose each question into a fixed sequence of subcondition-solving steps (following the annotated subconditions) and identify the first step where the model's intermediate conclusion diverges from the expert trace (protocol in supp.). Failures peak mid-chain (5% at Step 1, 20% at Step 2, 40% at Step 3,

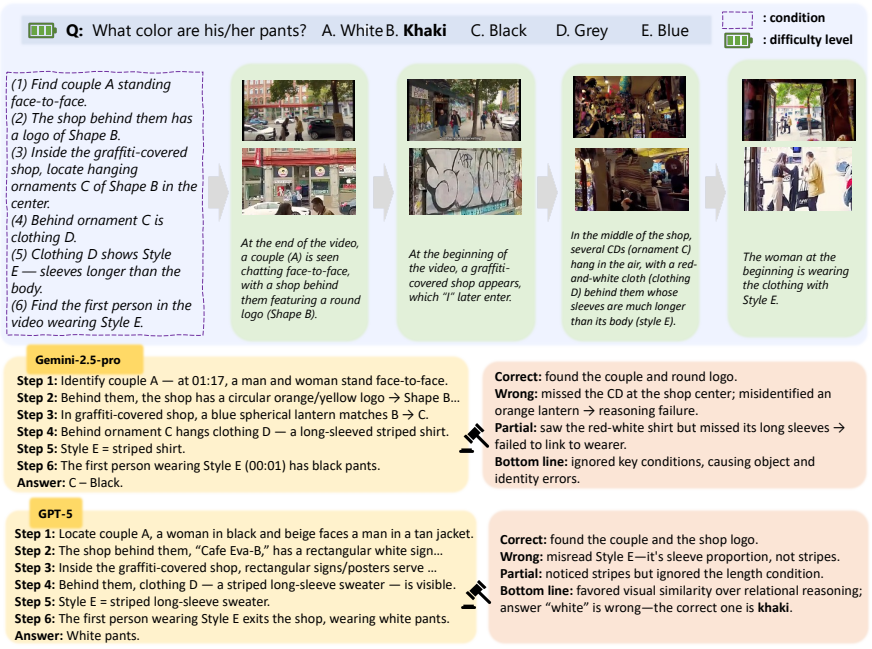


Fig. 5: Example of model reasoning on PerceptionComp. We show responses and judgments of frontier models. Even state-of-the-art models exhibit limitations in capturing perceptual information and often fail to maintain coherent reasoning chains leading to the correct answer.

25% at Step 4, and 10% thereafter), and expert review attributes 60% of these mid-stage failures to violated *spatial* subconditions (e.g., incorrect 3D spatial relationships). In such cases, the model anchors on an object that partially matches identity keywords but violates critical spatial or temporal constraints, causing the rest of the chain to drift. Due to space limits, we defer additional Gemini-3 case studies and full annotation details to the supplementary.

5 Conclusion

We introduce **PerceptionComp**, a fully manually annotated benchmark for complex, long-horizon, perception-centric video reasoning that requires repeated evidence gathering across temporally separated segments. PerceptionComp contains 1,114 five-choice questions over 279 high-complexity videos. Human studies confirm the intended difficulty: response times are much longer than prior benchmarks, single-view performance is near chance, while unrestricted rewatching achieves 100% accuracy. In contrast, state-of-the-art MLLMs reach at most 45.96% accuracy (open-source models $< 40\%$); increasing test-time reasoning or perceptual compute (more tokens or frames) helps but leaves a large gap. We additionally analyze Gemini-3 failure cases to characterize current bottlenecks. We hope PerceptionComp will serve as a reliable testbed for measuring and driving progress in perception-centric long-horizon video understanding.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631, accessed 2026-07-17
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Bytedance Seed: Seed2.0 model card: Towards intelligence frontier for real-world complexity (2026). https://doi.org/10.48550/arXiv.2607.00248, https://arxiv.org/abs/2607.00248, accessed 2026-07-17
4. Cheng, J., Ge, Y., Wang, T., Ge, Y., Liao, J., Shan, Y.: Video-holmes: Can mllm think like holmes for complex video reasoning? arXiv preprint arXiv:2505.21374 (2025)
5. Clark, C., Zhang, J., Ma, Z., Park, J.S., Salehi, M., Tripathi, R., Lee, S., Ren, Z., Kim, C.D., Yang, Y., et al.: Molmo2: Open weights and data for vision-language models with video understanding and grounding. arXiv preprint arXiv:2601.10611 (2026)
6. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., 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)
7. Das, R., Hristov, S., Li, H., Dimitrov, D., Koychev, I., Nakov, P.: Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 7768–7791 (2024). https://doi.org/10.18653/v1/2024.acl-long.420
8. Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wu, J., Zhang, X., Wang, B., Yue, X.: Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776 (2025)
9. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24108–24118 (2025)
10. Google AI for Developers: Gemini 3 developer guide. https://ai.google.dev/gemini-api/docs/gemini-3 (2026), accessed 2026-07-17
11. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., 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)
12. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)

13. Hao, Y., Gu, J., Wang, H.W., Li, L., Yang, Z., Wang, L., Cheng, Y.: Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. arXiv preprint arXiv:2501.05444 (2025)
14. Hu, K., Wu, P., Pu, F., Xiao, W., Zhang, Y., Yue, X., Li, B., Liu, Z.: Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. arXiv preprint arXiv:2501.13826 (2025)
15. Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y., Lin, S.: Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749 (2025)
16. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
17. Jiang, D., Zhang, R., Guo, Z., Li, Y., Qi, Y., Chen, X., Wang, L., Jin, J., Guo, C., Yan, S., et al.: Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. arXiv preprint arXiv:2502.09621 (2025)
18. Li, C., Chen, Q., Li, Z., Tao, F., Zhang, Y.: Videocogqa: A controllable benchmark for evaluating cognitive abilities in video-language models. arXiv preprint arXiv:2411.09105 (2024)
19. Li, K., Shen, H., Shi, H., Hou, R., Chang, H., Huang, J., Jia, C., Wang, W., Wu, Y., Jiang, D., Shan, S., Chen, X.: Humanpcr: Probing mllm capabilities in diverse human-centric scenes. arXiv preprint arXiv:2508.13692 (2025)
20. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., 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)
21. Li, X., Yan, Z., Meng, D., Dong, L., Zeng, X., He, Y., Wang, Y., Qiao, Y., Wang, Y., Wang, L.: Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. arXiv preprint arXiv:2504.06958 (2025)
22. Liu, Z., Sun, Z., Zang, Y., Dong, X., Cao, Y., Duan, H., Lin, D., Wang, J.: Visual-rft: Visual reinforcement fine-tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2034–2044 (2025)
23. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 2507–2521. Curran Associates, Inc. (2022). https://doi.org/10.52202/068431-0182, https://proceedings.neurips.cc/paper_files/paper/2022/file/11332b6b6cf4485b84afadb1352d3a9a-Paper-Conference.pdf, accessed 2026-07-17
24. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. Advances in Neural Information Processing Systems **36**, 46212–46244 (2023)
25. Nagrani, A., Menon, S., Iscen, A., Buch, S., Mehran, R., Jha, N., Hauth, A., Zhu, Y., Vondrick, C., Sirotenko, M., et al.: Minerva: Evaluating complex video reasoning. arXiv preprint arXiv:2505.00681 (2025)
26. OpenAI: Gpt-5 system card. https://openai.com/index/gpt-5-system-card/ (2025), accessed 2026-07-17
27. OpenAI: Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/ (2025), accessed 2026-07-17

28. OpenAI: Introducing gpt-5.2. https://openai.com/index/introducing-gpt-5-2/ (2025), accessed 2026-07-17
29. OpenAI: Openai o3 and o4-mini system card. https://openai.com/index/o3-o4-mini-system-card/ (2025), accessed 2026-07-17
30. Patraucean, V., Smaira, L., Gupta, A., Recasens, A., Markeeva, L., Banarse, D., Koppula, S., Malinowski, M., Yang, Y., Doersch, C., et al.: Perception test: A diagnostic benchmark for multimodal video models. Advances in Neural Information Processing Systems **36**, 42748–42761 (2023)
31. Qi, Y., Zhao, Y., Zeng, Y., Bao, X., Huang, W., Chen, L., Chen, Z., Zhao, J., Qi, Z., Zhao, F.: Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning. arXiv preprint arXiv:2504.07956 (2025)
32. Rasheed, H., Shaker, A., Tang, A., Maaz, M., Yang, M.H., Khan, S., Khan, F.S.: Videomathqa: Benchmarking mathematical reasoning via multimodal understanding in videos. arXiv preprint arXiv:2506.05349 (2025)
33. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
34. Rawal, R., Saifullah, K., Farré, M., Basri, R., Jacobs, D., Somepalli, G., Goldstein, T.: Cinepile: A long video question answering dataset and benchmark. arXiv preprint arXiv:2405.08813 (2024)
35. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Chi, H., Guo, X., Ye, T., Zhang, Y., 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)
36. Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., et al.: Kimi-vl technical report. arXiv preprint arXiv:2504.07491 (2025)
37. Team, V., Hong, W., Yu, W., et al.: Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv preprint arXiv:2507.01006 (2025)
38. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision. pp. 402–419. Springer (2020). https://doi.org/10.1007/978-3-030-58536-5_24
39. Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Ding, M., Gu, X., Huang, S., Xu, B., et al.: Lvbench: An extreme long video understanding benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22958–22967 (2025)
40. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
41. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
42. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. Advances in Neural Information Processing Systems **37**, 28828–28857 (2024)
43. Wu, Q., Zhao, H., Saxon, M., Bui, T., Wang, W.Y., Zhang, Y., Chang, S.: Vsp: Assessing the dual challenges of perception and reasoning in spatial planning tasks for vlms. arXiv preprint arXiv:2407.01863 (2024)
44. Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9777–9786 (2021)

45. Xiaomi LLM-Core Team: Mimo-vl technical report (2025), https://arxiv.org/abs/2506.03569, accessed 2026-07-17
46. Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y.: Video question answering via gradually refined attention over appearance and motion. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1645–1653 (2017). https://doi.org/10.1145/3123266.3123427
47. Xu, W., Wang, J., Wang, W., Chen, Z., Zhou, W., Yang, A., Lu, L., Li, H., Wang, X., Zhu, X., et al.: Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. arXiv preprint arXiv:2504.15279 (2025)
48. Yang, Z., Wang, S., Zhang, K., Wu, K., Leng, S., Zhang, Y., Qin, C., Lu, S., Li, X., Bing, L.: Longvt: Incentivizing "thinking with long videos" via native tool calling. arXiv preprint arXiv:2511.20785 (2025)
49. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9127–9134 (2019). https://doi.org/10.1609/aaai.v33i01.33019127
50. Zhao, Y., Zhang, H., Xie, L., Hu, T., Gan, G., Long, Y., Hu, Z., Chen, W., Li, C., Xu, Z., et al.: Mmvu: Measuring expert-level multi-discipline video understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8475–8489 (2025)
51. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing "thinking with images" via reinforcement learning. arXiv preprint arXiv:2505.14362 (2025), https://arxiv.org/abs/2505.14362, accessed 2026-07-17