
PerceptionComp: A Video Benchmark for Complex Perception-Centric Reasoning

Shaoxuan Li^{1*} Zhixuan Zhao^{1*} Hanze Deng^{1*} Zirun Ma^{1*}
 Shulin Tian³ Zuyan Liu¹ Yushi Hu² Haoning Wu³
 Yuhao Dong^{3#} Benlin Liu^{2#} Ziwei Liu^{3†} Ranjay Krishna^{2†}

¹Tsinghua University ²University of Washington ³Nanyang 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.

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

* Equal contribution # Project co-lead † Equal advising

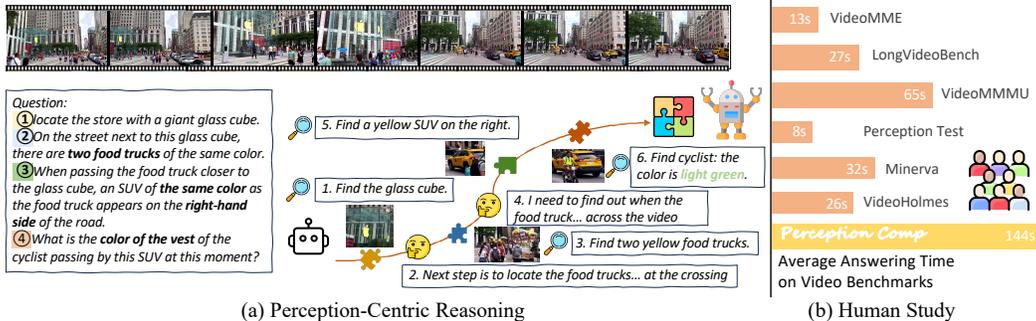


Figure 1: **Overview of the PerceptionComp benchmark.** (a) An example from PerceptionComp, where models are required to perform complex, perception-centric reasoning with various types of subconditions to arrive at the final answer. (b) Results from a human study measuring question-answering time, showing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

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 Guo et al. [2025a], Huang et al. [2025], Li et al. [2025], Feng et al. [2025]. 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) Fu et al. [2024], Hu et al. [2025], Cheng et al. [2025], Patraucean et al. [2023a] 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 Rasheed et al. [2025], Wu et al. [2024a], 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 Wu et al. [2024b] 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, compositional, and comprehensive perception-centric* video reasoning over long horizons. 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 sub-condition 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 Ravi et al. [2024] and optical-flow magnitude Teed and Deng [2020] 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

Table 1: **Comparison of PerceptionComp with other benchmarks.** PerceptionComp distinguishes itself from previous benchmarks by emphasizing perception-centric reasoning, assessing how models integrate visual evidence with reasoning processes.

Benchmark	Properties			
	Video Domain	# QA	Per.Rea	Annotation
MMVU Zhao et al. [2025]	Educational videos	3,000	✗	Manual
VideoMME Fu et al. [2024]	YouTube videos	2,700	✗	Manual
VCR-Bench Qi et al. [2025]	Short films	1,034	✗	Manual
MINERVA Nagrani et al. [2025]	Mix	1,515	✗	Manual
VideoMMM Hu et al. [2025]	Lectures	900	✗	Manual
Video-Holmes Cheng et al. [2025]	Short films	1,834	✗	Automatic&Manual
PerceptionComp	In-the-wild videos	1114	✓	Manual

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 $2\times$ longer than VideoMMM Hu et al. [2025], more than $10\times$ longer than VideoMME Fu et al. [2024], more than $5\times$ longer than Video-Holmes Cheng et al. [2025], and more than $5\times$ longer than LongVideoBench Wu et al. [2024b], 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 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 Comanici et al. [2025]) reaches only 45.96% accuracy in the five-choice setting, and open-source MLLMs Bai et al. [2025a,b], Wang et al. [2025a], Team et al. [2025a], Xiaomi et al. [2025] 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 Jaech et al. [2024] surpasses GPT-4o Hurst et al. [2024] 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 Comanici et al. [2025], 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 Jaech et al. [2024] and Qwen3-VL-8B Yang et al. [2025a], 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. Traditional video understanding benchmarks often focus on relatively basic perceptual understanding—either local details (e.g., short clips or fine-grained actions) or global summaries—with outcome-based metrics. Recent general-purpose benchmarks like Video-MME Fu et al. [2024] and ALLVB Tan et al. [2025] broaden task coverage across domains and

video lengths, while task-specific suites such as MVBench Li et al. [2024a] and NExT-QA Xiao et al. [2021] isolate skills like temporal reasoning and object interaction. The Perception Test Patraucean et al. [2023b] further provides diagnostic, perception-oriented evaluation on purposefully designed real-world videos. However, these benchmarks are typically solvable with limited cross-moment evidence integration and thus remain comparatively *easy* as probes of long-horizon, compositional video thinking. Long-video benchmarks Wang et al. [2025b], Wu et al. [2024b], Rawal et al. [2024], Song et al. [2024] emphasize memory and narrative comprehension over extended durations but largely reduce evaluation to single-turn QA. Egocentric benchmarks Grauman et al. [2022], Mangalam et al. [2023] add realism through first-person perspectives, but they typically do not require repeated perception to iteratively gather diverse visual evidence across multiple segments. PerceptionComp differs by making difficulty *perception-bottlenecked* through long-horizon compositional queries that require repeated evidence gathering.

Complex Multimodal Reasoning Benchmarks. Recent progress in multimodal reasoning has led to benchmarks that go beyond surface-level understanding to evaluate structured inference across vision and language. In the image domain, VCBench and related benchmarks Li et al. [2024b], Hao et al. [2025], Xu et al. [2025] target mathematical, scientific, and logical reasoning where visual inputs mainly serve as a carrier of symbolic structure, and ScienceQA Saikh et al. [2022] and EXAMS-V Das et al. [2024] introduce academic-style questions that emphasize explanation and cross-domain knowledge. In video, early benchmarks Xu et al. [2017], Yu et al. [2019], Xiao et al. [2021] focus on short-term understanding, while later ones Li et al. [2024a], Liu et al. [2024a,b] add richer temporal structure but often remain relatively shallow. Long-context benchmarks Wu et al. [2024b], Fu et al. [2024] scale to longer videos, yet many questions can still be answered from isolated cues. More advanced evaluations Zhao et al. [2025], Hu et al. [2025], Rasheed et al. [2025], Yang et al. [2025b] target scientific, academic, or spatial understanding, and VCR-Bench Qi et al. [2025] and MME-CoT Jiang et al. [2025] begin to assess chain-of-thought behavior. Recent benchmarks such as MINERVA Nagrani et al. [2025] and Video-Holmes Cheng et al. [2025] further emphasize multi-step temporal and causal inference. However, across many of these “hard” settings, difficulty is often dominated by *logical* or *domain* reasoning (e.g., math/science/geometry), with comparatively lightweight perceptual demands. In contrast (Table 1), PerceptionComp makes the bottleneck *perception*: questions are designed so that no single moment is sufficient and solving them requires repeatedly gathering fine-grained visual evidence across temporally separated segments, yielding a more faithful probe of perception-centric compositional video reasoning.

Multimodal Reasoning Models. Reasoning-oriented LLMs show that long-horizon inference benefits from step-by-step reasoning and test-time scaling. In parallel, MLLMs have evolved from early image-text systems to unified models that directly accept visual inputs. Frontier proprietary models (GPT-style, Gemini-style) Hurst et al. [2024], Comanici et al. [2025] and strong open-source families (Qwen-VL, InternVL, Molmo) Bai et al. [2025a], Yang et al. [2025a], Wang et al. [2025a], Clark et al. [2026] support direct image understanding, and increasingly extend the same interface to videos for end-to-end video QA. Following DeepSeek-R1 Guo et al. [2025a] and the shift toward long-form reasoning, recent work begins to bring RLVR-style or reasoning-focused pipelines to multimodal models. In images, Vision-R1/VisualRFT Huang et al. [2025], Liu et al. [2025] and DeepEyes Zheng et al. [2025] improve visual reasoning with verifiable rewards or interleaved multimodal traces. In videos, Video-R1 Feng et al. [2025] and VideoChat-R1 Li et al. [2025] elicit longer reasoning traces for multi-step temporal inference. Our benchmark complements these efforts by providing a perception-centric, long-horizon testbed that stresses repeated evidence gathering, where difficulty is dominated by perception rather than purely logical structure.

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.

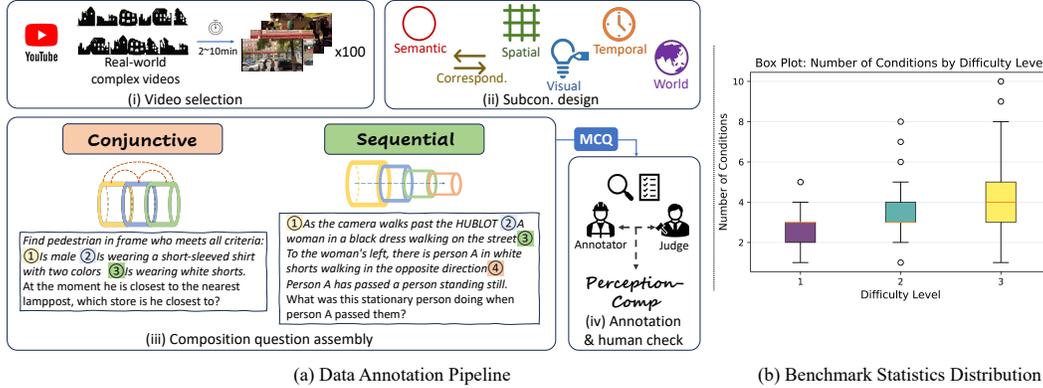


Figure 2: **Data construction and statistics of PerceptionComp.** (a) Annotation pipeline, which integrates diverse subconditions and supports two types of compositional questions. (b) Benchmark statistics: higher difficulty levels contain more subconditions, increasing the demand for perception-centric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

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 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 Ravi et al. [2024] and optical-flow magnitude Teed and Deng [2020] 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 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).

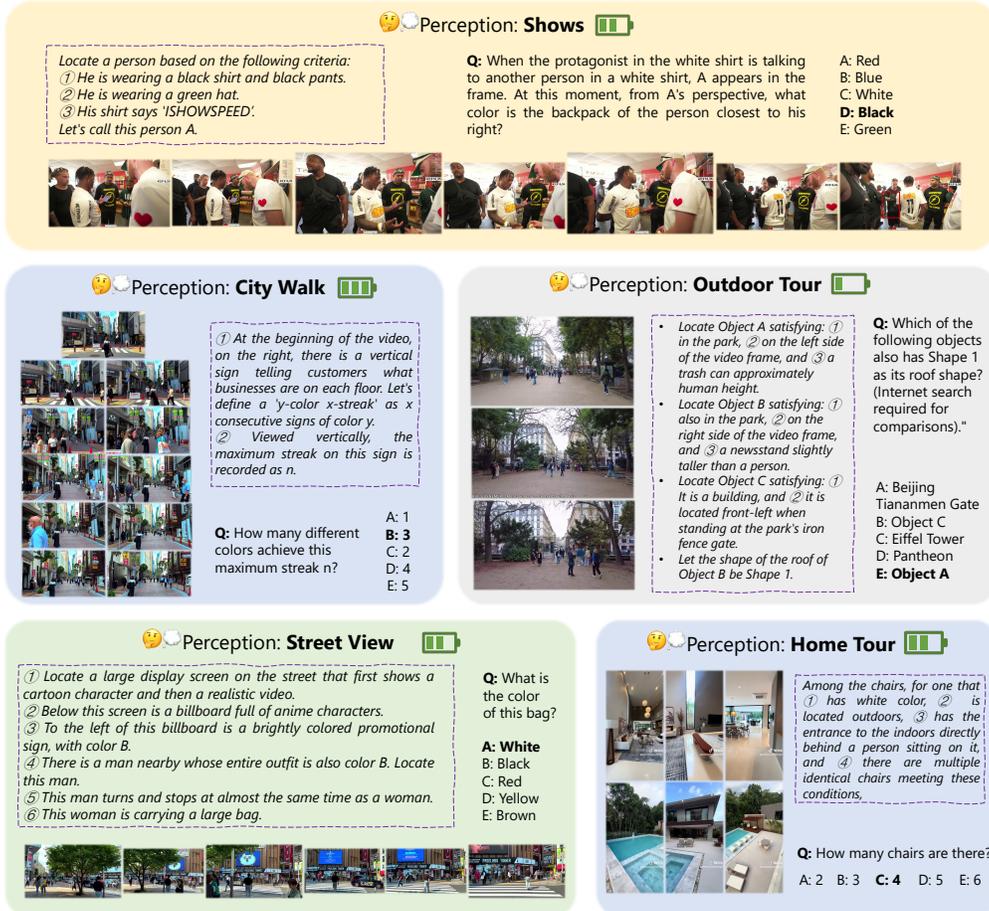


Figure 3: **Examples from PerceptionComp.** 🟢, 🟡, and 🟠 denote difficulty levels 1, 2, and 3, respectively. PerceptionComp spans diverse video sources and uses subconditions to construct conjunctive and sequential questions that require perception-centric reasoning.

- **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 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 (section 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 (section B). Finally, we present qualitative case studies and error patterns to highlight common failure modes in perception-centric long-horizon video reasoning (section 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)

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 Comanici et al. [2025]	-	-	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 Comanici et al. [2025]	-	-	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 Comanici et al. [2025]	-	-	45.78	45.18	32.64	42.19	52.35	52.00	58.06	44.64	44.20	44.02	44.34
Seed-2.0-Pro Guo et al. [2025b]	-	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 Comanici et al. [2025]	-	-	39.90	44.16	39.90	42.97	56.38	48.00	51.61	45.76	45.08	36.36	43.72
GPT-o3 Jaech et al. [2024]	-	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 Achiam et al. [2023]	-	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 Comanici et al. [2025]	-	-	39.39	41.12	24.35	39.84	47.65	44.00	32.26	41.96	35.89	34.93	38.15
GPT-5 Achiam et al. [2023]	-	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 Achiam et al. [2023]	-	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 Hurst et al. [2024]	-	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 Bai et al. [2025a]	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 Wang et al. [2025a]	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 Yang et al. [2025a]	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 Yang et al. [2025a]	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 Bai et al. [2025a]	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 Team et al. [2025b]	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 Yang et al. [2025a]	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 Feng et al. [2025]	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 Li et al. [2025]	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 Yang et al. [2025a]	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 Yang et al. [2025a]	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 Yang et al. [2025a]	235B	64	38.87	41.12	30.57	33.59	48.99	43.00	22.58	39.46	38.16	35.58	38.20

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 Wei et al. [2022] (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 Jaech et al. [2024] surpasses GPT-4o Hurst et al. [2024] 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). This analysis separates two common failure sources in video QA: insufficient visual evidence (too sparse sampling) versus insufficient deliberation (too short reasoning traces).

4.3.1 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.

4.3.2 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.

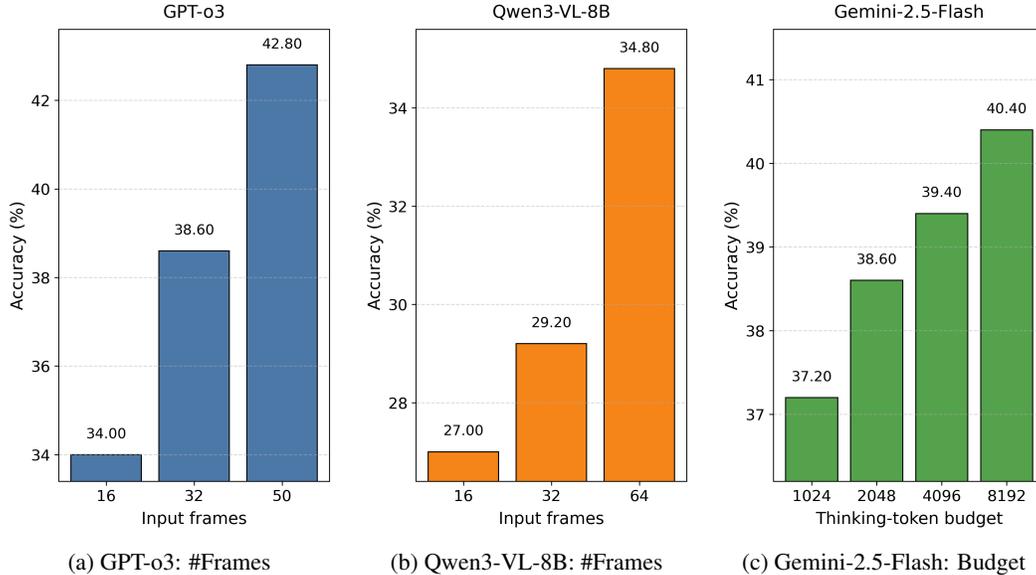


Figure 4: **Analysis on PerceptionComp.** Accuracy as a function of *perception budget* and *reasoning budget*. Left/middle: accuracy vs. the number of uniformly sampled input frames for GPT-o3 and Qwen3-VL-8B. Right: accuracy vs. the thinking-token budget for Gemini-2.5-Flash.

Together, these analyses suggest that PerceptionComp is a useful testbed for *visual thinking*: it is sensitive to both perception budget and reasoning budget, and can therefore distinguish models not only by raw visual recognition but also by their ability to sustain longer multimodal reasoning grounded in video evidence.

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, 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

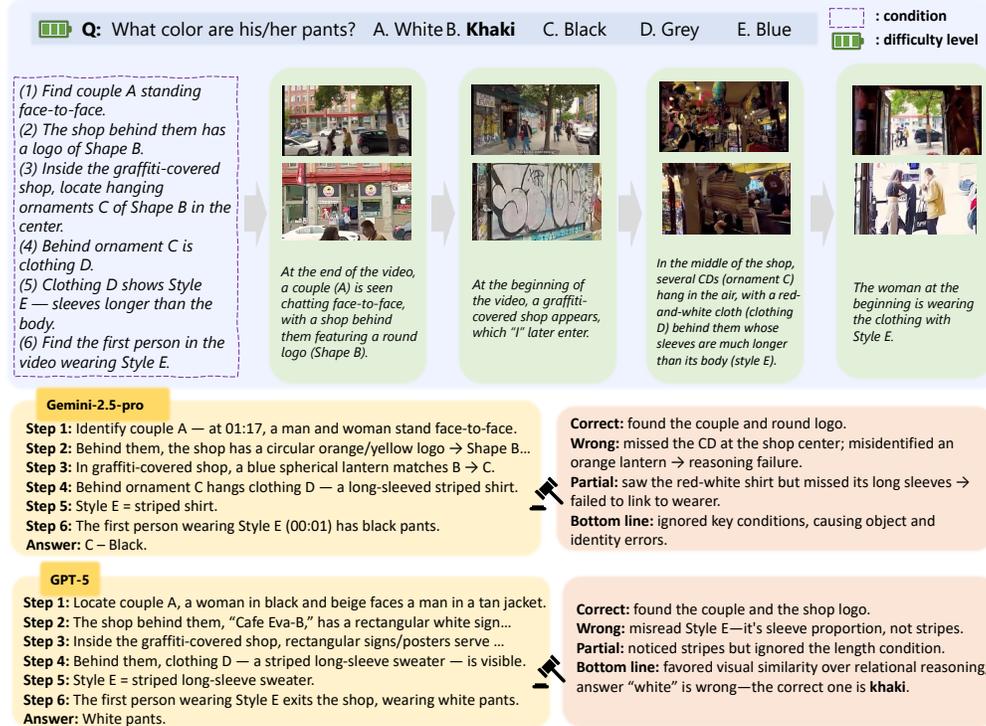


Figure 5: **Example of model reasoning on PerceptionComp.** We show responses and judgments of frontier models. Even state-of-the-art models exhibit limitations in capturing perceptual information and often fail to maintain coherent reasoning chains leading to the correct answer.

< 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

- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025a.
- Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.
- Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.
- 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, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.
- Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.
- Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025.

- Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36: 42748–42761, 2023a.
- Hanoona Rasheed, Abdelrahman Shaker, Anqi Tang, Muhammad Maaz, Ming-Hsuan Yang, Salman Khan, and Fahad Shahbaz Khan. Videomathqa: Benchmarking mathematical reasoning via multimodal understanding in videos. *arXiv preprint arXiv:2506.05349*, 2025.
- Qiucheng Wu, Handong Zhao, Michael Saxon, Trung Bui, William Yang Wang, Yang Zhang, and Shiyu Chang. Vsp: Assessing the dual challenges of perception and reasoning in spatial planning tasks for vlms. *arXiv preprint arXiv:2407.01863*, 2024a.
- Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024b.
- Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.
- Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.
- 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.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025a.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibao Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025b. URL <https://arxiv.org/abs/2511.21631>.
- 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*, 2025a.
- Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025a.
- LLM Xiaomi, Bingquan Xia, Bowen Shen, Dawei Zhu, Di Zhang, Gang Wang, Hailin Zhang, Huaqiu Liu, Jiebao Xiao, Jinhao Dong, et al. Mimo: Unlocking the reasoning potential of language model—from pretraining to posttraining. *arXiv preprint arXiv:2505.07608*, 2025.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.
- Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, et al. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8475–8489, 2025.
- Yukun Qi, Yiming Zhao, Yu Zeng, Xikun Bao, Wenxuan Huang, Lin Chen, Zehui Chen, Jie Zhao, Zhongang Qi, and Feng Zhao. Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning. *arXiv preprint arXiv:2504.07956*, 2025.
- Arsha Nagrani, Sachit Menon, Ahmet Iscen, Shyamal Buch, Ramin Mehran, Nilpa Jha, Anja Hauth, Yukun Zhu, Carl Vondrick, Mikhail Sirotenko, et al. Minerva: Evaluating complex video reasoning. *arXiv preprint arXiv:2505.00681*, 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*, volume 39, pages 7211–7219, 2025.
- Kunchang Li, Yali Wang, Yanan 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*, pages 22195–22206, 2024a.
- Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021.
- Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36: 42748–42761, 2023b.
- 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*, pages 22958–22967, 2025b.
- Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark. *arXiv preprint arXiv:2405.08813*, 2024.
- Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024.
- Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022.
- Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.
- Chenglin Li, Qianglong Chen, Zhi Li, Feng Tao, and Yin Zhang. Vcbench: A controllable benchmark for symbolic and abstract challenges in video cognition. *arXiv preprint arXiv:2411.09105*, 2024b.
- Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*, 2025.

- Weiyu Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*, 2025.
- Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301, 2022.
- Rocktim Das, Simeon Hristov, Haonan Li, Dimitar Dimitrov, Ivan Koychev, and Preslav Nakov. 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)*, pages 7768–7791, 2024.
- Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017.
- Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024a.
- Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024b.
- 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*, pages 10632–10643, 2025b.
- Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Lihui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025.
- Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*, 2026.
- Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2034–2044, 2025.
- Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing “thinking with images” via reinforcement learning. *ArXiv*, abs/2505.14362, 2025. URL <https://api.semanticscholar.org/CorpusID:278769859>.
- Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025b.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- V Team, Wenyi Hong, Wenmeng Yu, 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*, 2025b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Google AI for Developers. Gemini 3 developer guide. <https://ai.google.dev/gemini-api/docs/gemini-3>, 2026. Accessed 2026-03-12.

A Implementation Details

We now detail the implementation specifics for the evaluation process conducted on the Perception-Comp benchmark. For models that employ complex reasoning strategies, we utilize Gemini-2.5-Flash Comanici et al. [2025] as an automated judge to verify the correctness of the final answer. The evaluation methodology follows a systematic filtering process: for every question, the model’s intermediate reasoning steps and verbose output are programmatically excised, isolating only the definitive, final response. The LLM judge is then provided with a structured prompt that includes both the canonical ground truth answer and the system’s extracted response. The judge is instructed to perform a rigorous comparison, assessing the semantic equivalence and factual consistency between the two inputs, ultimately determining if the system’s final output precisely matches the expected ground truth.

B More Analysis Experiments

To further validate the intrinsic difficulty and effectiveness of the PerceptionComp benchmark, we conducted a controlled human study focusing on the role of iterative access to visual information. In the first, unconstrained condition, human annotators were permitted to view the video content and formulate their answers without limitation, achieving an overall accuracy of 85.10%. A stringent second condition was then imposed to isolate the need for iterative perception and reasoning: human participants were only allowed a single, complete viewing of the video, after which they were required to answer the corresponding questions based purely on memory and initial comprehension. This single-pass constraint dramatically reduced the performance to a mean accuracy of 18.97%. This pronounced drop of over 66 percentage points emphatically highlights that PerceptionComp is not merely a memory recall test, but instead necessitates multi-step, iterative perception and complex, video-grounded reasoning, thus validating its design for evaluating advanced perception models.

C Visualization Results

We provide more visualization results of our benchmark questions, as shown in Figure 6 and Figure 7.

The successful case shows the model following a multi-step visual logic chain with discipline. It grounds itself on a distinctive landmark, tracks the correct reference objects, and aligns events in space and time without drifting from the instructions. Once it identifies the critical moment—passing the nearer food truck as the yellow SUV appears—it isolates the correct bicyclist and extracts the fine-grained attribute (vest color) accurately. The model stays fully within the provided reasoning path, demonstrating reliable landmark grounding, temporal alignment, and detailed visual discrimination.

Conversely, we observe several distinct failure modes where models struggle to maintain this rigorous logical adherence.

The first failure case breaks down early: the model incorrectly concludes that the bench-identification steps are unsolvable and abandons the required reasoning path. Instead of resolving Style A, Direction B, and Type C through the provided visual logic, it hallucinates an alternative interpretation based on an unrelated “CANADA” hoodie and constructs a new, invalid chain from that point onward. Because of this invented logic, every subsequent step targets the wrong person, leading to the wrong final answer despite correctly observing that person’s phone. This reveals weaknesses in multi-step dependency tracking, logical adherence, and resistance to spurious patterns under difficulty.

A second failure case (Gemini-3.0-Pro Google AI for Developers [2026]) illustrates the breakdown of cross-temporal variable binding and the emergence of logical hallucinations. The task requires linking a shop sign’s dominant color to a plastic bag seen later, and subsequently identifying the shirt color of the person closest to the bag’s carrier. Although the model correctly locates the yellow shop sign, it fails to bind this color to the specified variable. It prematurely breaks the reasoning chain and hallucinates an alternative path by fixating on a random blue bag further in the video. This error cascades: the model misjudges the spatial and social context by anchoring on a randomly passing pedestrian rather than the correct walking companion, ultimately predicting the wrong shirt color.

A third failure case (Qwen3-VL Bai et al. [2025b]) highlights the danger of “protagonist bias” and the substitution of rigorous visual tracking with contextual heuristics. Tasked with comparing the 3D

Q: What is the color of the vest of the cyclist passing by this SUV at this moment?
A: Blue, B: Red, C: Yellow-green, D: Solid green, E: Solid yellow

- ① In the video, first locate the store with a giant glass cube entrance.
- ② On the street next to this glass cube, there are two food trucks of the same color.
- ③ When passing the food truck closer to the glass cube, an SUV of the same color as the food truck appears on the right-hand side of the road.

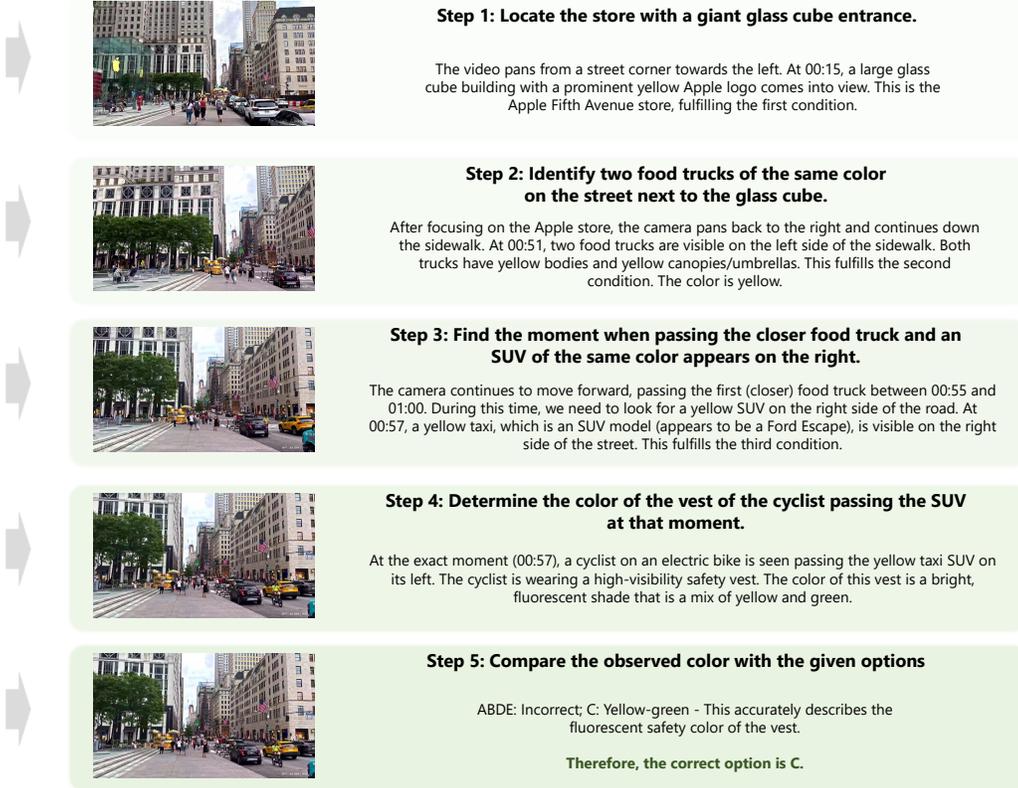


Figure 6: **Illustration of the model’s successful multi-step visual reasoning.** The system first anchors on the iconic glass-cube Apple Store, then tracks two yellow food trucks as spatial reference points. At the moment the camera passes the nearer truck, a yellow SUV appears on the right, enabling the model to localize the bicyclist adjacent to it and accurately identify the yellow-green vest color. This showcases robust landmark grounding, spatio-temporal alignment, and fine-grained visual detail recognition across a dynamic street scene.

flip directions of four parkour runners across two separate timestamps, the model is overly primed by the prompt’s use of a “grey-clothed runner” as an initial temporal anchor. It incorrectly elevates this runner to the central “logical lead” of the sequence and applies a “behavioral consistency” heuristic, assuming the athlete will naturally repeat the same movement pattern. By relying on this narrative guesswork rather than performing the necessary fine-grained spatio-temporal modeling of all four individuals, the model entirely bypasses the actual visual verification, leading to an incorrect, elimination-based conclusion.

Together, these failure cases demonstrate that while current models excel at specific visual recognitions, they remain highly vulnerable in long-horizon reasoning tasks where success strictly depends on robust variable retention, suppression of contextual biases, and unwavering adherence to a prescribed logic chain.

D Gemini-3-Pro & Gemini-3-Flash

Interestingly, our evaluation reveals a counter-intuitive phenomenon: the seemingly lightweight Gemini-3.0-Flash Google AI for Developers [2026] model achieves higher overall accuracy on our multi-step visual reasoning benchmark compared to its more capable counterpart, Gemini-3.0-



Q: What is D holding in their hand(s)?

A: Holding hands with someone, B: Mobile phone, C: Nothing, D: Document bag, E: Coffee cup

(1) In the last 10 seconds of the video, (2) chairs appear on both the left and right sides, with the more numerous style being Style A. (3) The third Style A chair being used by a person that passes the camera faces Direction B. (4) The first time the camera faces Direction B, identify the closest person's type of bag as Type C. (5) Identify D, the second person carrying a Type C bag who passes the camera

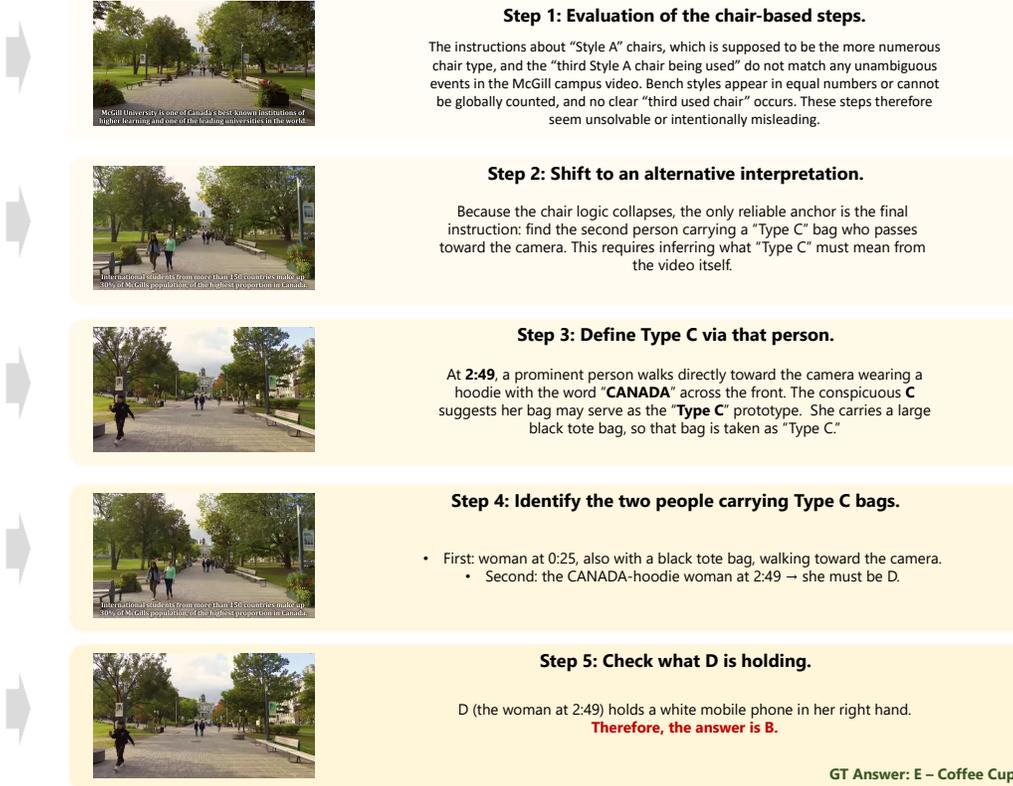


Figure 7: **Failure mode illustrating the model's breakdown in multi-step visual reasoning.** Instead of following the prescribed sequence—identifying the dominant bench style (Style A), tracking the third used bench to determine Direction B, and grounding Type C using the nearest person when the camera first faces that direction—the model prematurely abandons the correct logic path. It misclassifies the bench styles, hallucinates an alternative meaning for "Type C," and ultimately selects the wrong individual as D, leading to the incorrect prediction that D is holding a phone rather than the actual red coffee cup.

Pro Google AI for Developers [2026]. Through a qualitative analysis of the models' reasoning traces, we attribute this performance inversion to two primary factors:

1. Logical Focus vs. Detail Fixation. While the Pro model generates significantly longer reasoning chains, this extended capacity is frequently misallocated. Rather than advancing the core logical sequence, Pro tends to fixate on irrelevant, fine-grained visual details within the video. In contrast, Flash is highly optimized for rapid inference. This architectural efficiency translates into a sharper logical focus when processing long sequence data; Flash directly anchors onto key logical clues across frames and successfully bypasses redundant information that otherwise acts as distracting noise for the Pro model.

2. The "Streamlining Effect" in Mitigating Logical Hallucinations. The massive parameter scale and the propensity for "deep reasoning" in the Pro model can act as a double-edged sword. When confronted with fragmented spatial and temporal information in videos, Pro's tendency to over-analyze often leads to severe logical hallucinations and spatial confusion (e.g., constructing unnecessary absolute coordinate systems). We term the advantage observed in the Flash model as the "streamlining effect". By employing a more straightforward and direct information processing strategy, Flash avoids unwarranted associations. As long as the critical keyframes and objects are correctly localized, Flash



Google Gemini-3.0-pro

Step1: The prompt describes a man in a white shirt and black backpack who is in a hurry to cross the street. At 00:53, a man wearing a very light blue (which can appear white) polo shirt and a black backpack is seen running/jogging across the pedestrian crossing under the bridge. He is hurrying from the left side to the right side of the screen. (✓)
 Step2: Identify Color A. The man crosses towards a block of shops. While there are red *Shokudo Naganoya* and yellow TUC SHOP signs, the prompt links this Color A to a plastic bag seen later. (✓)
 Step3: Locate the man with the bag. The scene shifts to a pedestrian street at 1:08. At 1:15, a man is walking towards the camera on the left side of the screen, holding a large, distinct light blue (cyan) plastic shopping bag. This is the plastic bag of color A (Blue/Cyan). (✗)
 Step4: Identify the person closest to him. As the man with the blue bag walks forward, another man walks past him in the opposite direction (towards the camera) at 1:16. This man is passing right by his shoulder, making him the person closest to the man with the bag. (✗)
 Step5: Determine the shirt color. The man passing closest to him is wearing a white shirt (a white button-up or polo shirt) and a face mask. (✗)
 Therefore, the color of the shirt worn by the person closest to him is White. (✗)

Analysis:

1. Misidentification of 'Color A'

Reality: The prompt establishes a logical chain: Shop Sign Color = Color A = Plastic Bag Color. The visual evidence in the solution clearly shows the man walking out of the 'T.U.C.SHOP,' which has a prominent **YELLOW** sign. Therefore, Color A must be Yellow.
Error: In Step 2, Gemini correctly locates the yellow TUC sign, but it also identifies a redundant red sign. In Step 3, Gemini identifies Color A as 'Blue/Cyan' based on a bag it sees at the 1:15 mark, which is a critical logic error since there is no blue sign.

2. Incorrect Identification of the 'Closest Person'

Reality: The final step of the prompt asks for the shirt color of the person closest to the man with the bag. The man with the Yellow bag is walking side-by-side with a woman. In terms of proximity and social context, she is the 'person closest to him.' She is wearing a Black shirt.
Error: Gemini identifies the closest person as a 'man passing closest to him... in the opposite direction' wearing a white shirt.

3. Logical Hallucination

Error: Gemini's Step 2 mentions the 'TUC SHOP' sign but fails to connect its yellow color to the 'Color A' variable mentioned in the prompt. It essentially ignored the direct instruction to link the shop sign color to the bag color, leading to a 'hallucinated' path where it picked a random blue bag further in the video.

Figure 8: **Failure mode illustrating Gemini-3.0-Pro's breakdown in multi-step visual reasoning.** Instead of following the prescribed sequence—locating the initial man to determine the shop sign's dominant color (Color A: Yellow), tracking this variable to spot the man holding the corresponding yellow plastic bag later in the video, and grounding the nearest person (the accompanying woman) to identify her shirt color—the model suffers from critical logical hallucinations. It prematurely breaks the reasoning chain by failing to connect the yellow shop sign to "Color A" and hallucinates an alternative path by fixating on a random blue plastic bag. Consequently, it misjudges the spatial and social context, selecting a passing pedestrian rather than the walking companion, leading to the incorrect prediction that the closest person is wearing a white shirt rather than the actual black shirt.

faithfully adheres to the prescribed reasoning path, demonstrating that in dynamic visual contexts, a streamlined inference process is often more robust than unconstrained, overly complex reasoning.

E Analysis on Incorrect Step and Error Type

We analyze the failure cases of Gemini-3-Pro Google AI for Developers [2026] by examining both the reasoning step at which errors occur and the underlying causes of these failures. Figure 11(a) presents the distribution of incorrect reasoning steps across the reasoning chain. Errors occur most frequently in the second and third steps, with noticeably higher frequencies than in the first step. This pattern suggests that the difficulty increases as the reasoning chain becomes longer, indicating that extended multi-step reasoning remains challenging for the model. The lower counts observed in steps four and five are mainly due to the relatively small number of questions requiring more than four reasoning steps in the dataset.

Figure 11(b) further analyzes the causes of these errors. We categorize the failures into five types: spatial understanding, static feature perception, incomplete reasoning, counting errors, and dynamic



Results-Qwen3-VL-235B-A22B-Thinking

Phase 1: Scene Localization and Temporal Anchoring

I first identified the starting point of the sequence based on the prompt's instruction: the moment the grey-clothed runner finishes his action on the low wall. By treating this specific runner as my "temporal anchor," I prioritized him as the primary subject of interest. I then observed the transition to the bar area, noting that four individuals (Yellow clothes, White top/Shorts, White top/Long pants, and Grey clothes) participated in the hanging and dismantling sequence.

Phase 2: Multi-Subject Detection vs. Sparse Tracking

During the dismount (Step 2), I successfully detected the presence of the other participants. I noted the White top long pants runner (Option B) and the Shorts runner (Option D) performing flips. However, due to the sparse nature of my temporal sampling (the gaps between the frames I processed), I struggled to maintain a continuous 3D spatial trajectory for all four individuals simultaneously. While I identified their initial flip directions, the visual data felt disconnected when they moved to the second location for Step 3.

Phase 3: Narrative-Driven Subject Prioritization

Facing uncertainty in tracking the three runners at the second location, I defaulted to a narrative bias. Since the prompt explicitly began with the grey-clothed runner, I inferred that he was the "main character" of this specific parkour sequence. In my internal evaluation, I dismissed the White top long pants runner (Option B) and others as secondary background figures, assuming the question's focus would naturally return to the character who defined the scene's start.

Phase 4: Substitution via Behavioral Consistency Heuristics

Unable to precisely verify the second flips' rotation through visual evidence alone, I applied a "movement inertia" heuristic. I reasoned that a parkour athlete would likely repeat a mastered technical direction when performing a similar action in the same location. I projected this assumption of behavioral continuity onto the grey-clothed runner (Option C) because of his perceived narrative prominence. I concluded that since he was the focal point of the initial wall action, he was the most "logical" choice to have maintained a consistent flip direction, leading me to select Option C.

1. Temporal Sampling Limitation and Tracking Failure

The model's failure appears rooted in insufficient temporal resolution. Due to sparse frame sampling, the model could not maintain a continuous visual trajectory of the four runners. Instead of executing a rigorous frame-by-frame analysis of 3D spatial orientations (the "correct reasoning path"), the model's "thinking" shows it jumping between disconnected keyframes. This gap in visual data forced the model to abandon precise tracking in favor of a more "arbitrary" logical leap.

2. Contextual Priming and Subject Bias

The model fell into a narrative trap set by the prompt. Because the grey-clothed runner was mentioned as the initial temporal anchor ("Locate after the grey-clothed runner..."), the model prioritized him as the "protagonist." In its reasoning, it explicitly notes that since he was the subject of the initial action, his subsequent flips are the most "tracked" or relevant, leading to a biased focus that ignored the objective movements of the others.

3. Superficial Consideration of Alternative Subjects

The model's internal monologue reveals a cursory acknowledgment of the other runners without substantive analysis.

Yellow clothes (A) and Shorts (D): The model identifies their presence in specific frames (e.g., "person in white top and black shorts does a backflip") but fails to compare these specific directions against their second actions.

White top long pants (B): It briefly considers Option B but dismisses it without evidence, stating "Wait, no. Let's think again," before reverting to the grey-clothed runner. This indicates that while the model could detect the other characters, it lacked the temporal density to analyze their multi-stage action sequences, treating them as secondary background elements rather than active variables.

4. Substitution with Common-Sense Heuristics

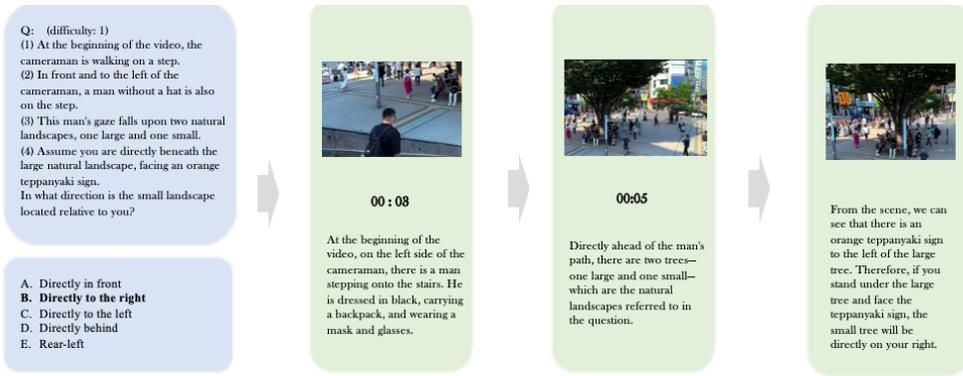
Unable to verify the flip directions visually, the model defaulted to a "Consistency Heuristic." It assumed a "behavioral continuity" where a parkour athlete would naturally repeat the same technical movement in the same location. This common-sense deduction acted as a "hallucinated logic" to fill the void left by its failure to visually track the specific 180° or 360° rotations of the runners, resulting in the incorrect selection of Option C.

Figure 9: **Failure mode illustrating Qwen3-VL's breakdown in multi-step visual reasoning.** Instead of following the prescribed sequence—tracking all four parkour runners continuously, registering their initial dismount flip directions, and visually comparing them to their secondary flips at the new location—the model prematurely abandons precise visual verification. Constrained by temporal sampling limitations, it falls into a narrative trap: it exhibits a "protagonist bias" by overly fixating on the prompt's initial temporal anchor (the grey-clothed runner) while superficially dismissing the other athletes. Ultimately, the model substitutes rigorous 3D spatial orientation analysis with a hallucinated "behavioral consistency" heuristic, incorrectly assuming the main subject would naturally repeat his movement pattern, which leads to the erroneous selection of the grey-clothed runner.

feature perception. Among these categories, incomplete reasoning refers to cases where the model fails to consider all required conditions or introduces logical inconsistencies during the reasoning process. From the distribution of error causes, spatial understanding accounts for the largest proportion of failures. This observation indicates that the model still struggles with accurately interpreting spatial relationships in complex visual scenes.

F Limitations and Ethical Considerations

PerceptionComp focuses on complex, perception-centric reasoning in long videos, but it does not cover every type of video reasoning task. The current dataset is limited to daily-life recordings and excludes high-stakes domains such as medical or surveillance settings. We use videos only from sources whose usage is compatible with academic benchmarking, and we do not release content beyond what is permitted by the original source terms. We do not annotate personally sensitive attributes (e.g., identity, race, or other demographic traits), and the benchmark is not intended for identity recognition, surveillance, or sensitive-attribute inference. Furthermore, as our analysis shows that model performance often hinges on multi-step logical adherence and spatial reasoning, we caution against relying solely on absolute accuracy. We recommend interpreting results through comparative performance trends across models, as future systems may still exploit unforeseen biases or rely on heuristics rather than robust spatio-temporal modeling.



Google Gemini-3-flash

Step 1: Identify the Cameraman and the Man: At the beginning of the video (around 00:01-00:03), the cameraman is walking on a raised, stepped terrace in a plaza (Shinjuku Southern Terrace). In front and slightly to the left, a man without a hat is also walking on this terrace. (✓)

Step 2: Identify the Natural Landscapes: In his field of vision, there is a prominent large tree (the large natural landscape) in the center of the terrace and a smaller tree (the small natural landscape, visible around 00:07-00:08) further down and to the right of the main walking path. (✓)

Step 3: Locate the Orange Teppanyaki Sign: As the camera moves forward toward the street (00:04-00:10), we can see various building signs. On the left side of the street, there is an orange/yellow sign with the text "鉄板焼" (Teppanyaki). (✓)

Step 4: Determine the Relative Direction: Assume you are standing directly beneath the large tree, the orange teppanyaki sign is located on the left side of the street ahead of you. If you turn to face this sign, you are facing toward the left side of the street. The small tree (small landscape) is located on the right side of the street. If you are facing the sign on the left, the small tree on the opposite (right) side of the street will be located directly to your right. (✓)

So the answer is B, directly to the right. (✓)

Google Gemini-3-pro

Step 1: Identify the Cameraman and the Man: At the beginning of the video (0:00-0:05), the cameraman is descending stairs. In front and to the left, there are people. The model repeatedly considers several candidates, including "There is a man in a black shirt and grey pants standing/leaning on the railing or wall near the tree", "At 0:03, there's a man directly ahead and slightly left, walking down the stairs", and "At 0:05, on the left side of the stairs, there's a group of young men. One is leaning against the tree planter". The model keeps re-examining who the "man without a hat" is and whether he is walking or stationary. (✓)

Step 2: Identify the Natural Landscapes: The model identifies the large natural landscape as the big tree in the center of the stairs. It then repeatedly questions the identity of the small landscape, including "To the right of the big tree, further back, there is a smaller tree with pink flowers", "Is there another small landscape? Maybe a tree near the street?", and "What about the planter to the left?" Eventually, the model concludes: "The small natural landscape is likely the smaller tree with pink flowers to the right of the big tree." (✓)

Step 3: Locate the Orange Teppanyaki Sign: The model scans multiple signs in the video: "At 0:04, on the left side of the street level, there are signs.", "There is a shop called 'Gindaco' (red sign).", "On the left building wall, there's a sign that says 'Teppan Oukoku'." After repeatedly checking other signs, the model determines: "The 'Teppan Oukoku' sign is on the building directly to the left of the stairs/tree." Therefore: "If I am under the tree and facing the orange teppanyaki sign, I am facing left relative to the direction of walking down the stairs." (✓)

Step 4: Determine the Relative Direction: The model attempts to calculate the relative direction between the teppanyaki sign and the small tree. The reasoning repeatedly converts directions into different coordinate systems: "Let's call the downhill direction 'North'. The Teppan sign is 'West'.", "The small tree is roughly North-East.", "If I am facing West (Teppan sign), and the target is East (Small tree), then the target is directly Behind me." The model continues checking multiple possibilities, including "Small tree - directly right", "Small tree - rear-right", "Small tree - directly behind". Finally, the model concludes: "The Teppan sign is to the Left. The Small Tree is to the Right. Facing the Left means the Right is Behind you." (x)

So the answer is D, directly behind. (x)

Analysis

First, Over-focusing on Minor Details. Gemini-3-Pro repeatedly re-examines irrelevant details that do not affect the final reasoning. For example, it repeatedly identified the "man without a hat" and questioned the small landscape. These repeated checks add noise but do not help determine the final direction.

Second, Long Reasoning Chain Causing Spatial Confusion. Gemini-3-Pro repeatedly converts spatial relationships into multiple coordinate systems. However, these coordinate transformations distort the original spatial relationship and lead to incorrect reasoning.

Figure 10: **Failure mode illustrating the model's breakdown in multi-step visual reasoning.** Instead of following the prescribed sequence—identifying the key natural landscapes, locating the orange teppanyaki sign, and directly deducing the relative direction between them from the observer's perspective—the model becomes bogged down by excessive noise and spatial confusion. It first over-focuses on irrelevant minor details (such as the pedestrian's attire and alternative signs), and then attempts to construct an unnecessarily complex absolute coordinate system involving arbitrary cardinal directions (North, West, East). This convoluted reasoning chain distorts the original egocentric spatial layout, leading to the incorrect prediction that the small landscape is located directly behind the observer rather than the actual direction of directly to the right.

F.1 Annotation protocol and quality control

All questions are manually authored and verified. Annotators are trained crowdsourced workers with relevant technical backgrounds. Each annotator is first trained on a set of 20 example videos and questions and must pass a calibration test before contributing to the final dataset.

The annotation process proceeds in two stages. In the first stage, an annotator watches the video, proposes a compositional question with three subconditions, and specifies a single correct answer. In

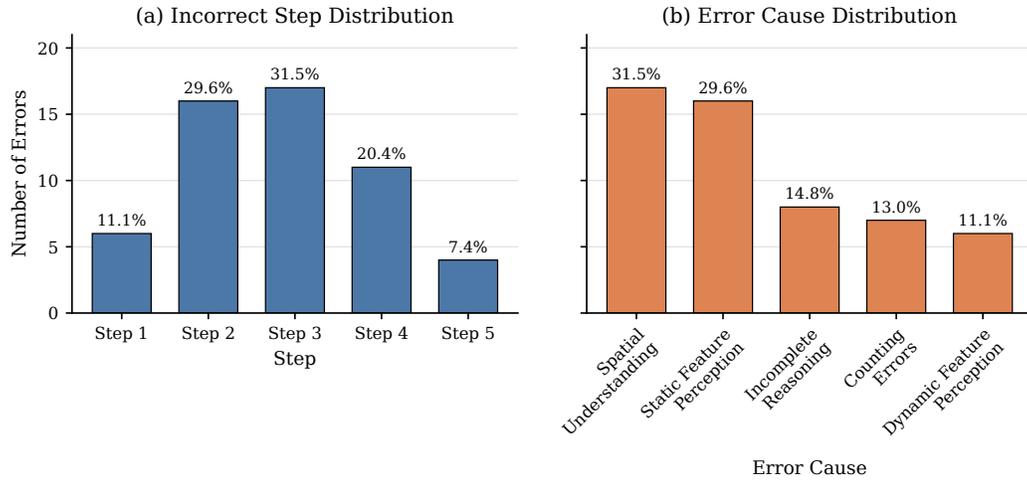


Figure 11: Error analysis of Gemini-3-Pro Google AI for Developers [2026] failures on PerceptionComp. (a) Distribution of incorrect reasoning steps across the reasoning chain. (b) Distribution of error causes, including spatial understanding, static feature perception, incomplete reasoning, counting errors, and dynamic feature perception.

the second stage, a different annotator reviews the question by re-watching the video and checking three properties: (i) correctness of the answer, (ii) uniqueness of the solution, and (iii) necessity of each subcondition. Items that fail any check are revised or discarded.

To quantify agreement, we sample 100 questions and ask a third annotator to independently answer them. The agreement between the third annotator and the original answer key is 89.0%, indicating that the majority of questions admit a clear, unambiguous solution.