

PerceptionComp: A Video Benchmark for Complex Perception-Centric Reasoning

First Author Institution1 Institution1 address firstauthor@i1.org	Second Author Institution2 First line of institution2 address secondauthor@i2.org
--	--

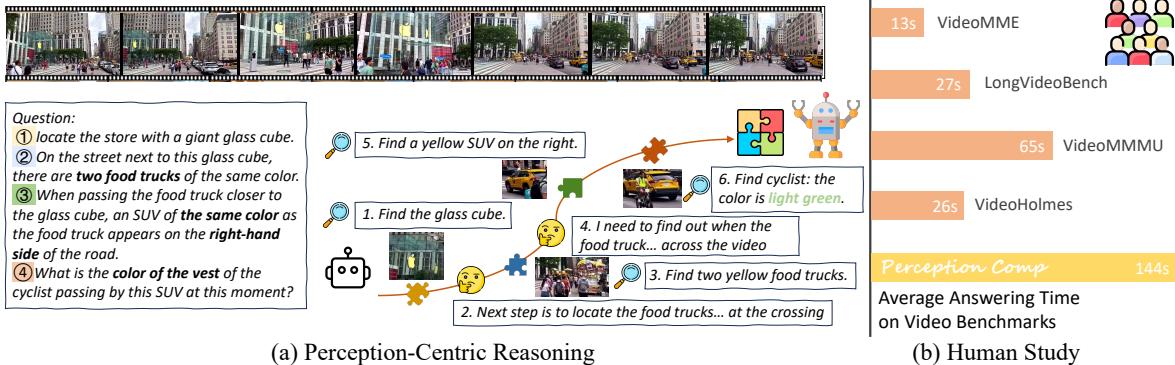


Figure 1. **Overview of the PerceptionComp benchmark.** (a) An example from PerceptionComp, where models are required to perform complex, perception-centric reasoning with various type of subconditions to arrive at the final answer. (b) Results from a human study measuring question-answering time, showing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

Abstract

Deep video understanding requires models to compose multiple perception skills: semantic recognition, spatial and temporal reasoning, cross-time correspondence, and to allocate additional inference-time computation when needed. Yet most existing video benchmarks admit shortcut solutions after a single viewing, or else emphasize symbolic/logical difficulty over perceptual complexity, making them ill-suited to measure multimodal test-time thinking. We introduce **PerceptionComp**, a perception-centric benchmark that enforces extended, stepwise inference by (i) composing questions from multiple perceptual subconditions under conjunctive or sequential logic, and (ii) sourcing real videos with high scene complexity (many objects, substantial motion, frequent transitions) across diverse domains. Each subcondition is itself a perceptual subtask, and answers are short, verifiable pieces of perceptual information (objects, attributes, relations, locations, events). A human study confirms that PerceptionComp requires longer reasoning and repeated perceptual checks to solve. Evaluating state-of-the-art MLLMs shows that models strong on prior benchmarks perform notably worse: Gemini-2.5-Pro reaches only **46.2%** under a 5-choice setting.

with most models below **40%**. Moreover, thinking variants achieve improved performance compared with non-thinking ones (e.g., GPT-o3 surpasses GPT-4o by **13** points), indicating that PerceptionComp is an effective testbed for assessing test-time scaling for video understanding.

1. Introduction

Videos record human life and the world we live in, and multi-modal intelligence, whether deployed on robots or AI glasses, must achieve deep understanding of videos in order to better serve human needs. For example, given a recorded clip, the system should be able to state on which floor of a building a person dropped their keys. Achieving this level of understanding requires a multi-modal agent to compose multiple perception skills, so that it can integrate diverse kinds of perceptual information. To find which floor the keys were dropped on, the agent must perform *spatial reasoning* to understand the building’s interior layout; *semantic recognition* to identify which object is a key; *temporal reasoning* to determine when the keys were dropped; and *correspondence* to disambiguate among multiple keys shown in the video which one is the user’s previously held item. Recent

breakthroughs in deep reasoning for largely symbolic tasks such as mathematics and coding show that the paradigm of test-time scaling [6, 10, 13, 19], which allocates more computation at inference time, is a highly promising route for enabling multi-modal language models (MLLMs) to attain such deep video understanding. In particular, MLLMs need to learn multi-modal test-time thinking that composes several perception skills so as to extract the necessary types of perceptual information from video.

However, existing video benchmarks are still insufficient for evaluating an MLLM’s multimodal test-time thinking ability. On the one hand, many commonly used video benchmarks [3, 7, 12] are relatively simple: as shown in Figure 1, for humans, questions in these benchmarks can often be solved very quickly after merely watching the video and reading the question. This implies there typically exists a shortcut that does *not* require additional inference-time computation, and thus these benchmarks may fail to reveal differences in multimodal test-time thinking among models. Even for some long-video understanding benchmarks [35], success more often hinges on whether the model has sufficient memory, rather than on its ability to perform multimodal test-time thinking after watching the video. On the other hand, multi-modal tasks that do demand test-time thinking, such as geometry or maze solving [25, 36], tend to derive their difficulty primarily from logical reasoning rather than perception, because their visual inputs are synthetic or drawn from overly simple instructional videos. The structural complexity of such inputs is far lower than that of general real-world videos, so these tasks are poor probes of perceptual competence. As a result, there is a lack of a video benchmark that can faithfully reflect an MLLM’s capacity for deep video understanding, as such a benchmark must be both *complex* and *perception-centric*.

To address these gaps, we introduce PerceptionComp, a highly complex video reasoning benchmark that increases compositional complexity in order to force models to engage in extended inference-time reasoning to arrive at correct answers. Specifically, we design questions by combining multiple sub-conditions through two distinct logics—conjunctive and sequential—and require the model to extract perceptual information from the video that satisfies the composed conditions. This design ensures that the model must reason step by step during inference to identify answers satisfying each condition. To guarantee that perception dominates the reasoning process, each sub-condition is itself formulated as a perceptual subtask, such that the overall problem can be viewed as requiring recursive perception over the video. To comprehensively evaluate perceptual capacity, different sub-conditions are designed to target different aspects of perception, such as semantic, spatial, and temporal understanding, while the extracted perceptual information spans diverse types, including objects, attributes, relationships, lo-

cations, and events. Consequently, correctly answering a full question requires models to employ a broad set of perceptual primitives and to integrate heterogeneous visual information, while also knowing how to compose them to complete complex tasks. Furthermore, to increase task difficulty and pose greater challenges for video-based reasoning, we deliberately source videos with higher scene complexity than in many existing benchmarks, including videos with numerous objects, significant motion, and frequent scene transitions. These videos span diverse domains such as city walk tours, large villa indoor tours, and extreme outdoor sports. With these carefully designed questions and videos, PerceptionComp demands long reasoning processes that include repeated perceptual steps in order to arrive at correct answers. As shown in Figure 1, this is validated in our human study: participants required substantially longer response times than on other benchmarks, and could only answer correctly after repeatedly watching the videos—otherwise, accuracy after a single viewing was close to random guess.

We evaluate PerceptionComp on state-of-the-art MLLMs. While these models achieve impressive results on existing video perception benchmarks, they perform notably worse on ours. The best-performing model, Gemini-2.5-Pro [4], achieves only 46.2% accuracy under a 5-choice setting, and most other MLLMs [2, 30, 32, 38, 41] fall below 40%. Moreover, we find that thinking models demonstrate improved performance compared with non-thinking models; for instance, GPT-o3 [15] surpasses GPT-4o [14] by 13% in accuracy. This clearly demonstrates that our benchmark serves as an effective testbed for evaluating video thinking ability.

2. Related Work

General Video Understanding Benchmarks. Traditional video understanding benchmarks primarily evaluate either local details (e.g., short clips or fine-grained actions) or global summaries, focusing on predefined tasks with outcome-based metrics. This setup often overlooks how models compose local observations into coherent global understanding. Recent general-purpose benchmarks like Video-MME [7] and ALLVB [29] broaden task coverage across domains and video lengths, while task-specific suites such as MVBench [18] and NExT-QA [37] isolate skills like temporal reasoning and object interaction. Long-video benchmarks [26, 28, 33, 35] emphasize memory and narrative comprehension over extended durations but largely reduce evaluation to single-turn QA. Egocentric benchmarks [8, 22] add realism through first-person perspectives, yet remain outcome-focused. While these benchmarks offer broad perceptual coverage, they largely omit the reasoning dimension, overlooking the compositional processes needed to link perceptual cues into extended inferences.

Multimodal Reasoning Benchmarks. Recent progress in multimodal reasoning has led to a rise in benchmarks that

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

Benchmark	Attributes			
	Video Domain	# QA	Per.Rea	Anno.
MMVU [44]	Educational videos	3,000	✗	M
VideoMME [7]	YouTube videos	2,700	✗	M
VCR-Bench [24]	Short films	1034	✗	M
MINERVA [23]	Mix	1,515	✗	M
VideoMMMU [12]	Lectures	900	✗	M
Video-Holmes [3]	Short films	1,834	✗	A&M
PerceptionComp	Instructional videos	500	✓	A&M

go beyond surface-level understanding to evaluate models’ ability to perform structured inference across visual and language modalities. In the image domain, benchmarks such as VCBench and others [11, 17, 40] target mathematical, scientific, and logical reasoning tasks where visual and textual modalities must be jointly interpreted, and ScienceQA [27] and EXAMS-V [5] extend this by introducing academic-style multimodal questions that require explanation generation and cross-domain understanding. For video reasoning, early benchmarks [37, 39, 43] focused on short-term video understanding, primarily testing action recognition and basic factual retrieval, where recent ones [18, 20, 21] incorporate richer temporal structure but still lean on relatively shallow reasoning. Some works [7, 35] attempt to scale reasoning to longer temporal contexts, though many of their tasks remain solvable through isolated visual cues. Advanced evaluations [12, 25, 42, 44] target scientific, academic, or spatial understanding, while VCR-Bench [24] and MME-CoT [16] begin to assess chain-of-thought reasoning performance. More recently, MINERVA [23] and Video-Holmes [3] step further to emphasize multi-step temporal and causal inference in temporal reasoning. In contrast to these existing benchmarks, as shown in Table 1, PerceptionComp presents a novel setting by requiring extended inference-time reasoning through recursive visual perception, which better evaluates models capabilities in fine-grained perceptual inference and compositional reasoning.

3. PerceptionComp

We increase task difficulty to force models to engage in longer, perception-centric reasoning in order to more faithfully evaluate their video thinking ability. Concretely, we achieve this in two ways: (i) select structurally more complex videos, and (ii) compose questions from multiple subconditions that each probe a different perceptual skill, thereby raising the compositional complexity of the task. Below, we describe in detail our video selection process, the format of questions and answers, and the full annotation pipeline.

3.1. Video Selection

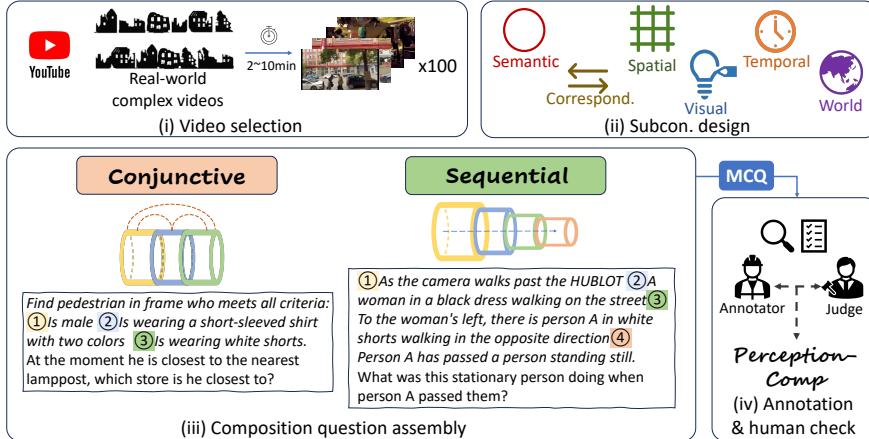
Many existing video benchmarks use clips that are visually simple: they often depict a single event or activity, and contain only a small number of humans or objects. As a result, many videos can be approximately replaced by a short textual caption without substantially affecting downstream performance. This limits their ability to diagnose a model’s perceptual competence, since solving the tasks does not truly require rich visual understanding.

To better probe perception-related abilities, we deliberately select videos with high scene and object complexity, including city-walk tour vlogs, indoor tours of large houses or villas, and highlight reels of sports competitions. These videos typically contain many objects, frequent scene transitions, and substantial camera motion, making them far more challenging to summarize with a single caption. The selected clips range from 2 to 10 minutes in length. Unlike many benchmarks that increase difficulty primarily by extending video duration, we additionally increase the difficulty along an orthogonal axis: the underlying dynamic scene complexity. All videos are recorded in real-world environments rather than synthetically generated, ensuring that the resulting tasks are both genuinely challenging and closely aligned with practical applications. This dual design—longer videos with complex, rapidly changing scenes—forces models to repeatedly query and integrate visual evidence rather than relying on a coarse global summary.

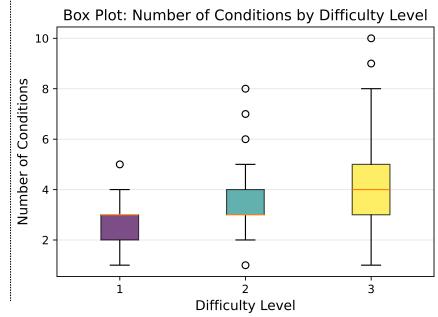
3.2. Subconditions and Perceptual Skills

We explicitly increase the compositional complexity of each question by combining multiple subconditions into a single query. Each subcondition requires distinct perceptual–reasoning skills, so that solving the full question demands coordinated use of several abilities rather than a single narrow competence. Concretely, our subconditions cover:

- **Semantic understanding:** the model must recognize object categories, attributes (e.g., shape, color, material), and high-level semantic relations (e.g., social roles or interactions between people).
- **Spatial understanding:** the model must reason about spatial layout in the depicted scene, including relative position (left/right, front/behind), distance (near/far), and occlusion relationships.
- **Temporal understanding:** the model must follow motion patterns and localize activities or events in time, such as determining what happens before or after a given reference event.
- **Correspondence:** the model must establish correspondences between instances or parts across time or views (e.g., tracking the same object across shots, or matching a part to its whole).
- **Visual knowledge:** the model must recall external knowledge that is tightly coupled to visual content.



(a) Data Annotation Pipeline



(b) Benchmark Statistics Distribution

Figure 2. **Data construction and statistics of PerceptionComp benchmark.** (a) Data annotation pipeline of PerceptionComp, which integrates various subconditions and designs two types of compositional questions. (b) Benchmark statistics, showing that questions with higher difficulty levels contain more subconditions, emphasizing the growing importance of perception-centric reasoning.

- **World modelling:** the model must form expectations about plausible future outcomes (e.g., predicting what is likely to happen next given ongoing dynamics).

By sampling and composing subconditions from these categories, each question requires the model to deploy a *combination* of perceptual skills even at the level of individual sub-tasks, rather than relying on any single capability. This design enables a more comprehensive assessment of a model’s perceptual competence in complex video reasoning.

3.3. Compositional Question Design

We design two distinct composition logics to combine subconditions into a full question:

- **Conjunctive.** All subconditions refer to the *same* underlying target, and the conditions stand in a parallel, “and”-style relationship. To ensure that every subcondition is necessary, we verify that no proper subset of the subconditions suffices to uniquely determine the answer: each individual condition can only eliminate a subset of candidate targets, and only the conjunction of all conditions yields a single correct solution. This prevents shortcuts where the model could ignore part of the question and still guess the answer.
- **Sequential.** Each subcondition resolves an intermediate entity or state, and later subconditions explicitly depend on the output of earlier ones. For example, a first subcondition may identify an object, the second may constrain its behavior at a later time, and a third may ask about a relation involving that same object after another event occurs. In this setting, the model must first solve the earlier subconditions to locate the correct referent, and then carry that referent forward when interpreting subsequent conditions. This induces a multi-hop perceptual reasoning

process in which errors in early steps propagate, making it impossible to answer the final question correctly without following the intended sequence of perceptual inferences.

3.4. Answer Space

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

- **Objects:** object category names (e.g., “car”, “sofa”).
- **Attributes:** properties such as color, count, or shape.
- **Relationships:** semantic, spatial, or social relations between entities.
- **Location:** the place where an object is situated (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 models from exploiting language priors instead of relying on visual evidence, all distractors are drawn from the *same* answer category as the correct option (e.g., all are color words, or all are object categories). Furthermore, each option is constrained to be a single word or a very short phrase, minimizing additional linguistic cues that might allow the model to guess the answer without truly understanding the video.

3.5. Annotation Pipeline

Following the procedure above, we select 100 videos with high scene complexity and annotate 500 questions. Because each question is highly compositional, we adopt a fully manual annotation process to ensure correctness. In particular,

💡 Perception: Shows

Locate a person based on the following criteria:
 ① He is wearing a black shirt and black pants.
 ② He is wearing a green hat.
 ③ His shirt says 'ISHOWSPEED'.
 Let's call this person A.



💡 Perception: Shows

Q: When the protagonist in the white shirt is talking to another person in a white shirt, A appears in the frame. At this moment, from A's perspective, what color is the backpack of the person closest to his right?

A: Red
B: Blue
C: White
D: Black
E: Green

💡 Perception: City Walk



① At the beginning of the video, on the right, there is a vertical sign telling customers what businesses are on each floor. Let's define a 'y-color x-streak' as x consecutive signs of color y.
 ② Viewed vertically, the maximum streak on this sign is recorded as n.

Q: How many different colors achieve this maximum streak n?
 A: 1
B: 3
 C: 2
 D: 4
 E: 5

💡 Perception: Outdoor Tour



- Locate Object A satisfying: ① in the park, ② on the left side of the video frame, and ③ a trash can approximately human height.
- Locate Object B satisfying: ① also in the park, ② on the right side of the video frame, and ③ a newsstand slightly taller than a person.
- Locate Object C satisfying: ① It is a building, and ② it is located front-left when standing at the park's iron fence gate.
- Let the shape of the roof of Object B be Shape 1.

Q: Which of the following objects also has Shape 1 as its roof shape? (Internet search required for comparisons)."
 A: Beijing Tiananmen Gate
 B: Object C
 C: Eiffel Tower
 D: Pantheon
E: Object A

💡 Perception: Street View

① Locate a large display screen on the street that first shows a cartoon character and then a realistic video.
 ② Below this screen is a billboard full of anime characters.
 ③ To the left of this billboard is a brightly colored promotional sign, with color B.
 ④ There is a man nearby whose entire outfit is also color B. Locate this man.
 ⑤ This man turns and stops at almost the same time as a woman.
 ⑥ This woman is carrying a large bag.



💡 Perception: Home Tour



Among the chairs, for one that ① has white color, ② is located outdoors, ③ has the entrance to the indoors directly behind a person sitting on it, and ④ there are multiple identical chairs meeting these conditions.

Q: How many chairs are there?
 A: 2
 B: 3
C: 4
 D: 5
 E: 6

Figure 3. Examples of questions from PerceptionComp. **💡**, **💡**, and **💡** denote questions with difficulty levels 1, 2, and 3, respectively. PerceptionComp includes a diverse range of video sources and employs subconditions to construct conjunctive and sequential questions that require perception-centric reasoning.

annotators first create the subconditions and the final answer, and then carefully 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 this verification stage, we confirm that there is a single correct answer in the video and that no proper subset of the subconditions suffices to uniquely identify it. Items that fail either requirement are revised or discarded. This protocol guarantees that all questions admit a unique solution and genuinely require the full set of composed perceptual subconditions.

4. Experiments

In this section, we first conduct comprehensive evaluation experiments on a set of widely used MLLMs to demonstrate the challenging nature of the PerceptionComp benchmark. Our evaluation covers a broad range of state-of-the-art models, and the results are presented in Section 4.1. We then perform detailed analysis experiments in Section 4.2 to further investigate why the questions in the benchmark pose significant challenges to existing models. Finally, Section 4.3 presents representative case studies illustrating common failure patterns of current video MLLMs, providing deeper insights into their limitations in perception and reasoning.

Table 2. **Comprehensive evaluation results of MLLMs on PerceptionComp.** We report both category-wise accuracies and accuracies across different difficulty levels. For open-source thinking models, performance improvements over instruction-tuned models are highlighted in green, while performance drops are highlighted in red.

Model	Size	Frame	Accuracy by Category					Accuracy by Difficulty			Overall
			Outdoor Tour	Shopping	Sports	Home Tour	Shows	Level 1	Level 2	Level 3	
Human Performance											
Human	-	-	82.10	80.00	81.08	93.33	100.00	91.22	85.12	68.61	82.40
Proprietary Models											
Gemini-2.5-Pro[4]	-	-	45.78	80.00	33.33	47.30	46.67	50.68	44.65	43.07	46.00
GPT-03[15]	-	50	43.22	80.00	40.00	43.24	20.00	37.84	45.58	43.80	42.80
Gemini-2.5-Flash[4]	-	-	39.39	80.00	20.00	45.95	46.67	50.00	37.21	35.04	40.40
Seed-1.6-VL[9]	200B	64	39.39	40.00	35.71	39.19	53.33	37.16	43.72	36.03	39.60
GPT-4o-latest[14]	-	50	30.18	40.00	33.33	31.08	20.00	33.02	29.05	27.00	30.20
GPT-5[1]	-	64	26.60	40.00	26.67	35.14	46.67	28.38	32.56	22.63	28.60
GPT-4.1[1]	-	50	26.09	40.00	13.33	28.38	20.00	29.77	26.39	19.71	26.00
Open-Source Instruct Models											
Qwen2.5-VL[2]	7B	64	26.34	40.00	14.29	16.22	33.33	24.32	26.98	22.06	25.40
InternVL-3.5 [32]	8B	64	31.20	20.00	13.33	27.03	53.33	31.76	30.23	29.93	30.60
Qwen3-VL[41]	8B	64	34.53	60.00	40.00	27.03	66.67	37.16	33.95	33.57	34.80
Kimi-VL[30]	16B	64	27.88	40.00	26.67	35.14	40.00	33.11	29.77	24.82	29.40
Qwen3-VL[41]	30B	64	34.02	80.00	20.00	36.49	53.33	36.74	35.81	32.85	35.40
Qwen2.5-VL[2]	72B	64	33.07	20.00	26.67	20.27	53.33	33.11	31.16	27.74	30.80
GLM-4.5V[31]	106B	64	36.57	60.00	35.71	31.08	46.67	37.84	34.42	37.50	36.20
Qwen3-VL[41]	235B	64	36.06	40.00	33.33	22.97	40.00	35.77	34.46	33.49	34.40
Open-Source Thinking Models											
Mimo-VL-SFT[38]	7B	64	31.46	60.00	20.00	21.62	53.33	32.56	30.41	27.74	30.60
Mimo-VL-RL[38]	7B	64	31.20	60.00	20.00	37.84	26.67	35.14	32.56	29.20	32.40
Video-R1[6]	7B	64	26.34	40.00	33.33	25.68	26.67	22.97	26.97	29.93	26.60
VideoChat-R1[19]	7B	64	28.13	20.00	26.67	25.68	60.00	32.09	26.35	25.55	28.60
Qwen3-VL-Thinking[41]	8B	64	34.02	40.00	40.00	27.03	53.33	35.14	34.42	31.39	33.80
Kimi-VL-Thinking[30]	16B	64	30.70	40.00	13.33	27.03	33.33	32.43	29.77	26.28	29.60
Qwen3-VL-Thinking[41]	30B	64	32.99	60.00	20.00	25.68	60.00	33.58	28.38	34.88	32.60
Qwen3-VL-Thinking[41]	235B	64	32.48	40.00	26.67	20.27	53.33	31.76	30.70	31.39	31.20

4.1. PerceptionComp Benchmark

Evaluation Models. We conduct a comprehensive evaluation of representative video MLLMs, including both open-source systems and closed-source API-based models, to rigorously assess the current state of complex perception-driven video reasoning capabilities. Our evaluation encompasses a diverse range of state-of-the-art architectures, such as InternVL-3.5 [32], the Qwen2.5-VL and Qwen3-VL series [2], as well as advanced open-source variants like Seed-VL-1.6 [9], Kimi-VL [30], and Mimo-VL [38]. To approximate the current performance upper bound, we further include leading closed-source models, namely Gemini-2.5 Pro, Gemini-2.5 Flash [4], the latest GPT-5 [1], and other top-performing commercial systems such as GLM-4.5V [31]. In addition, we assess video reasoning models, including Video-R1 [6] and VideoChat-R1 [19], to explore the role of explicit reasoning capabilities in our benchmark. This diverse selection ensures a balanced and representative comparison across different model types and training paradigms.

Implementation Details. For MLLMs with native video-processing capabilities, such as Gemini-2.5 Pro, we directly utilize the raw video streams without performing any frame extraction or temporal preprocessing. In contrast, for models lacking native video-handling functionality, we uniformly sample 64 frames per video sequence as input. For certain GPT-based APIs, the number of sampled frames is reduced to 50 due to input length constraints. All proprietary models are evaluated using a Chain-of-Thought reasoning approach. For open-source models, instruction-following models are prompted to directly provide the correct answer choice, while thinking models are evaluated using a Chain-of-Thought [34] prompting strategy with the temperature set to 0.7 and a maximum generation length of 16,384 tokens.

Results. The overall performance of different MLLMs on the PerceptionComp benchmark is presented in Table 2. Most models achieved accuracy below 40%, highlighting the difficulty of the benchmark. Gemini-2.5 Pro achieved the best performance with 46.0% accuracy, while open-source

 Q: What color are his/her pants? A. White B. **Khaki** C. Black D. Grey E. Blue

 : condition
 : difficulty level

- (1) Find couple A standing face-to-face.
- (2) The shop behind them has a logo of Shape B.
- (3) Inside the graffiti-covered shop, locate hanging ornaments C of Shape B in the center.
- (4) Behind ornament C is clothing D.
- (5) Clothing D shows Style E — sleeves longer than the body.
- (6) Find the first person in the video wearing Style E.



At the end of the video, a couple (A) is seen chatting face-to-face, with a shop behind them featuring a round logo (Shape B).



At the beginning of the video, a graffiti-covered shop appears, which "I" later enter.



In the middle of the shop, several CDs (ornament C) hang in the air, with a red-and-white cloth (clothing D) behind them whose sleeves are much longer than its body (style E).



The woman at the beginning is wearing the clothing with Style E.

Gemini-2.5-pro

- Step 1:** Identify couple A — at 01:17, a man and woman stand face-to-face.
Step 2: Behind them, the shop has a circular orange/yellow logo → Shape B...
Step 3: In graffiti-covered shop, a blue spherical lantern matches B → C.
Step 4: Behind ornament C hangs clothing D — a long-sleeved striped shirt.
Step 5: Style E = striped shirt.
Step 6: The first person wearing Style E (00:01) has black pants.

Answer: C – Black.

Correct: found the couple and round logo.

Wrong: missed the CD at the shop center; misidentified an orange lantern → reasoning failure.

Partial: saw the red-white shirt but missed its long sleeves → failed to link to wearer.

Bottom line: ignored key conditions, causing object and identity errors.

GPT-5

- Step 1:** Locate couple A, a woman in black and beige faces a man in a tan jacket.
Step 2: The shop behind them, "Cafe Eva-B," has a rectangular white sign...
Step 3: Inside the graffiti-covered shop, rectangular signs/posters serve ...
Step 4: Behind them, clothing D — a striped long-sleeve sweater — is visible.
Step 5: Style E = striped long-sleeve sweater.
Step 6: The first person wearing Style E exits the shop, wearing white pants.

Answer: White pants.

Correct: found the couple and the shop logo.

Wrong: misread Style E—it's sleeve proportion, not stripes.

Partial: noticed stripes but ignored the length condition.

Bottom line: favored visual similarity over relational reasoning; answer "white" is wrong—the correct one is **khaki**.

Figure 4. Example of model reasoning on PerceptionComp. Responses and judgments of frontier models on our benchmark are shown. 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.

models such as the Qwen3-VL family performed notably worse, with the 8B and 235B MoE variants reaching only 34.80% and 34.40%. Interestingly, scaling up the model size did not consistently improve results, suggesting that current MLLMs still struggle with perception-centric reasoning tasks that require fine-grained visual comprehension and temporal integration. These findings indicate that PerceptionComp poses distinctive challenges for existing MLLMs, particularly in integrating video information and performing multi-step, long-horizon reasoning beyond static visual understanding.

To further analyze why current models fall short on the proposed perception-centric reasoning tasks, we extend our evaluation to thinking MLLMs. The results show a clear polarization. While several open-source reasoning models achieve modest gains over the baseline systems, the Qwen3-VL series exhibits the opposite trend. All “thinking” variants of Qwen3-VL perform worse than their instruct models, suggesting that these reasoning models are predominantly optimized for abstract reasoning benchmarks while under-emphasizing perceptual understanding. This observation

highlights a critical limitation of current thinking MLLMs: despite stronger reasoning capabilities, they often lack robust perception skills, leading to suboptimal performance on PerceptionComp benchmark, which requires the integration of fine-grained perception and multi-step reasoning.

To provide deeper insights into the performance of current MLLMs on the PerceptionComp benchmark, we report accuracy scores across different categories and difficulty levels. The difficulty levels are defined and annotated by human experts based not only on the number of subconditions but also on the intrinsic difficulty of each subcondition, offering a more reliable estimation of overall task complexity. As shown in the table, most models achieve satisfactory performance on Level 1 and Level 2 tasks. However, their accuracy drops sharply at Level 3, suggesting that current MLLMs still lack the essential capability to effectively integrate perception and reasoning for perception-centric reasoning tasks. Interestingly, the results further reveal model-specific trends among the thinking models. For example, Mimo-VL-RL and VideoChat-R1 consistently benefit from performing a thinking step before answering across all difficulty levels.

Table 3. Analysis experiment results on PerceptionComp, focusing on input configurations and reasoning strategies.

(a) Number of input frames			(b) Thinking budget		
Model	Frames	Acc	Model	Thinking budget	Acc
GPT-o3	16	34.00	Gemini-2.5-Flash	1024	37.20
GPT-o3	32	38.60	Gemini-2.5-Flash	2048	38.60
GPT-o3	50	42.80	Gemini-2.5-Flash	4096	39.40
Qwen3-VL-8B	16	27.00	Gemini-2.5-Flash	8192	40.40
Qwen3-VL-8B	32	29.20			
Qwen3-VL-8B	64	34.80			

In contrast, the Qwen3-VL series encounters performance degradation when incorporating thinking capabilities, even on Level 1 tasks that primarily require straightforward perceptual understanding. This further confirms the previous hypothesis that current thinking MLLMs focus more on complex abstract reasoning while neglecting the need to jointly enhance perceptual and reasoning abilities, a gap that underscores the significance of the PerceptionComp benchmark.

4.2. Analysis

In this section, we present a series of analysis experiments to analyze the challenging nature of the PerceptionComp benchmark. Models are evaluated under various experimental settings to explore the factors that contribute to the difficulty of solving the benchmark questions. Specifically, we vary both the number of input frames and the allocated thinking budget to provide deeper insight into how perception and reasoning jointly affect model performance.

4.2.1. Input Frames

To investigate how the density of temporal visual information influences perception-centric video reasoning, we conduct a sensitivity analysis by varying the number of input frames (\mathcal{F}) provided to two representative MLLMs: *GPT-o3* and *Qwen3-VL-8B*. Specifically, *GPT-o3* is evaluated with $\mathcal{F} \in \{16, 32, 50\}$ and *Qwen3-VL-8B* with $\mathcal{F} \in \{16, 32, 64\}$. As shown in Table 3, both models exhibit a clear positive correlation between \mathcal{F} and overall accuracy (Acc): *GPT-o3* improves from 34.0% at 16 frames to 42.8% at 50 frames, while *Qwen3-VL-8B* gains 7.8% when \mathcal{F} increases from 16 to 64. We attribute these gains to two factors intrinsic to PerceptionComp: (1) larger frame budgets enhance perceptual grounding by offering richer and more diverse visual evidence, enabling more accurate identification and localization of objects and subtle attributes; and (2) denser temporal sampling improves continuity, allowing the models to better integrate visual cues over time, track objects across scene transitions, and capture fine-grained actions crucial for multi-step temporal and compositional reasoning.

4.2.2. Thinking Budget

We further investigate the effect of reasoning effort on solving PerceptionComp questions. Specifically, we experiment with Gemini-2.5-Flash using different thinking budgets of

1,024, 2,048, 4,096, and 8,192 tokens. As shown in Table 3, allocating more extensive thinking resources allows the MLLM to better handle the long dependency chains featured in PerceptionComp. This enables the model to more reliably follow the sequential logic of subconditions and perform the step-by-step reasoning required to connect initial visual observations to the final answer.

4.3. Case Study

We present qualitative results of state-of-the-art MLLMs and analyze their response behaviors to reveal two prevailing error patterns: perceptual errors and reasoning chain failures. These error patterns underscore the importance of perception-centric reasoning. As shown in Fig. 4, we examine Gemini-2.5-Pro and GPT-5 by comparing their generated explanations with human-annotated references, focusing on how perception and reasoning interact within each system.

- **Perceptual errors.** Both models exhibit generally stable visual grounding but often misinterpret key visual cues. They may correctly localize or detect an object yet misidentify its defining properties or relational roles. In some cases, visual similarity leads to false recognition, indicating insufficient perceptual discrimination.
- **Reasoning chain failures.** Even when perceptual grounding is accurate, both models frequently struggle to maintain coherent reasoning chains. The reasoning process tends to break when multiple spatial or semantic conditions must be integrated, leading to inconsistent or over-confident yet conceptually invalid conclusions.

These error modes suggest that current MLLMs tend to prioritize surface-level matching rather than jointly optimizing perceptual understanding and structured reasoning. PerceptionComp explicitly exposes this gap by requiring models to reason through perception rather than beyond it, thus highlighting the necessity of developing perception-grounded reasoning capabilities for MLLMs.

5. Conclusion

We introduced PerceptionComp, a video reasoning benchmark designed to evaluate MLLMs’ ability to perform extended, perception-centric reasoning tasks. Its difficulty arises from visually complex videos and highly compositional questions that require integrating multiple perceptual cues through sequential and conjunctive logic. Experiments on state-of-the-art MLLMs show significant challenges, revealing the benchmark’s effectiveness in exposing current limitations. Our analysis further indicates a gap between abstract reasoning and perceptual grounding for current open-source MLLMs. These findings highlight the need for architectures and training paradigms that explicitly couple perception and reasoning. We hope PerceptionComp will serve as a valuable testbed to advance the development of more deeply integrated and perceptually grounded MLLMs.

References

- [1] 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. 6
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 6
- [3] 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. 2, 3
- [4] 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. 2, 6
- [5] 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. 3
- [6] 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. 2, 6
- [7] 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. 2, 3
- [8] 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. 2
- [9] 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*, 2025. 6
- [10] 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*, 2025. 2
- [11] 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. 3
- [12] 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. 2, 3
- [13] Wenzuan 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. 2
- [14] 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. 2, 6
- [15] 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. 2, 6
- [16] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui 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. 3
- [17] 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*, 2024. 3
- [18] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 2, 3
- [19] 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. 2, 6
- [20] 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, 2024. 3
- [21] 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*, 2024. 3
- [22] 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. 2
- [23] 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. 3
- [24] Yukun Qi, Yiming Zhao, Yu Zeng, Xikun Bao, Wenzuan 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. 3
- [25] 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. 2, 3
- [26] 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. 2
- [27] 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. 3
- [28] 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. 2
- [29] 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*, pages 7211–7219, 2025. 2
- [30] 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*, 2025. 2, 6
- [31] 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*, 2025. 6
- [32] 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*, 2025. 2, 6
- [33] 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, 2025. 2
- [34] 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. 6
- [35] 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, 2024. 2, 3
- [36] Quicheng 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*, 2024. 2
- [37] 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. 2, 3
- [38] LLM Xiaomi, Bingquan Xia, Bowen Shen, Dawei Zhu, Di Zhang, Gang Wang, Hailin Zhang, Huaiqiu Liu, Jiebao Xiao, Jinbiao Dong, et al. Mimo: Unlocking the reasoning potential of language model—from pretraining to posttraining. *arXiv preprint arXiv:2505.07608*, 2025. 2, 6
- [39] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueling 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. 3
- [40] Weiye 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. 3
- [41] 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*, 2025. 2, 6
- [42] 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, 2025. 3
- [43] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueling 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*, pages 9127–9134, 2019. 3
- [44] 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 multidiscipline video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8475–8489, 2025. 3