EgoCoT-Bench: Benchmarking Grounded and Verifiable Operation-Centric Chain of Thought Reasoning for MLLMs

Yang Dai*
yangdai@zju.edu.cn
Zhejiang University
Hangzhou, Zhejiang, China

Dian Jiao*
jd_dcd@zju.edu.cn
Zhejiang University
Hangzhou, Zhejiang, China

Tianwei Lin
lintw@zju.edu.cn
Zhejiang University
Hangzhou, Zhejiang, China

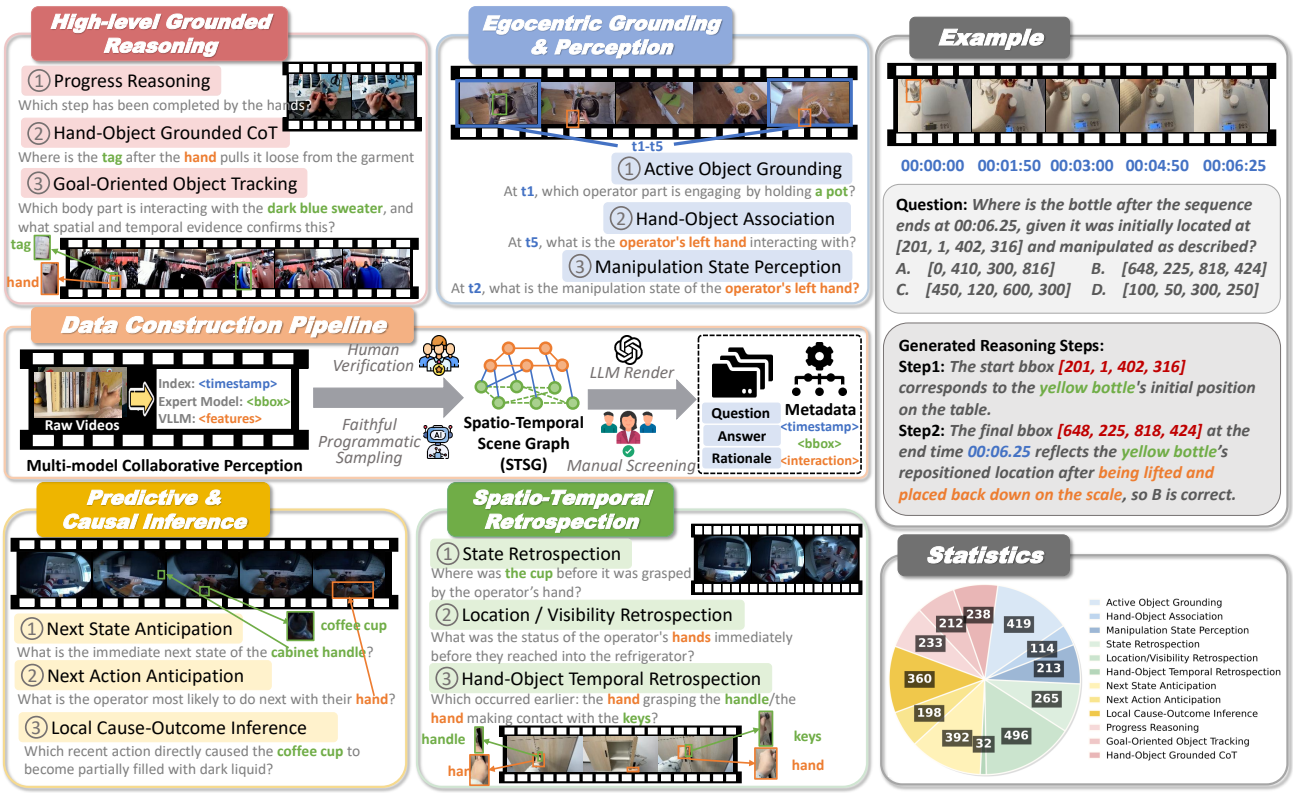


Figure 1: Overview of EgoCoT-Bench. EgoCoT-Bench is a fine-grained benchmark for grounded and verifiable operation-centric reasoning in egocentric videos, containing 3,172 QA pairs over 351 videos across four task groups and 12 subtasks. It is built through an STSG-guided human verification pipeline with explicit spatio-temporal evidence and rationale annotations.

Abstract

The rapid development of Multimodal Large Language Models (MLLMs) has led to growing interest in egocentric video understanding, specifically the ability for MLLMs to recognize fine-grained hand-object interactions, track object state changes over time, and reason about manipulative processes in dynamic environments from a first-person perspective. However, existing egocentric video benchmarks suffer from **limited grounded rationale evaluation**, offering limited support for fine-grained operation-centric reasoning and rarely examining whether model rationales are grounded in explicit spatio-temporal evidence. To address this gap, we introduce **EgoCoT-Bench**, a fine-grained egocentric benchmark for grounded and verifiable operation-centric reasoning with explicit step-by-step rationale annotations. Overall, EgoCoT-Bench comprises 3,172 verifiable QA pairs over 351 egocentric videos separated into four task groups for a total of 12 sub-task groups, encompassing perception and retrospection, anticipation, and high-level reasoning. The benchmark is constructed through a spatio-temporal scene graphs (STSG) guided generation framework and is further refined by human annotators to ensure correctness, egocentric relevance and fine-grained quality. Experimental results show continuing difficulties with egocentric fine-grained reasoning and further reveal that many multimodal models produce explanations that are answer-correct, but have evidence that is inconsistent with the answer. We hope EgoCoT-Bench can serve as a useful testbed for grounded and verifiable reasoning in egocentric video understanding. Project page and supplementary materials are available at: https://dstardust.github.io/EgoCoT/.

*Both authors contributed equally to this research.

CCS Concepts

• **Computing methodologies** → **Activity recognition and understanding**.

Keywords

egocentric video understanding, benchmark, multimodal large language models, grounded reasoning, fine-grained reasoning, verifiable rationales

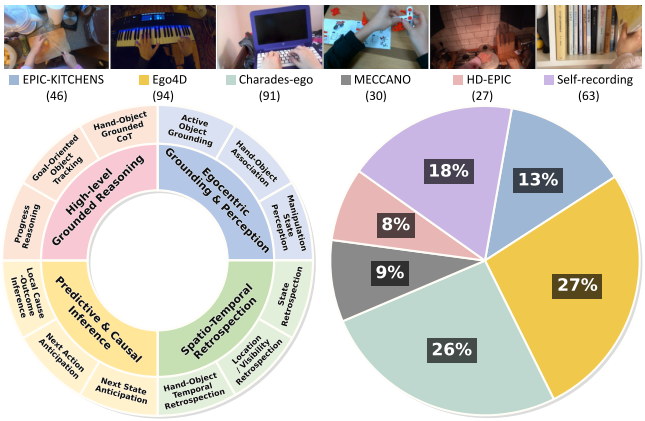
1 Introduction

The rapid progress of multimodal large language models (MLLMs) has greatly advanced video understanding, opening up new possibilities for question answering, temporal reasoning, and embodied perception [25, 34, 43]. Among these directions, egocentric video understanding is of particular importance for real-world assistive agents and embodied systems [10, 15, 26], since first-person observations directly capture how a user manipulates objects, shifts attention, and interacts with the surrounding environment during task execution. Compared with generic third-person videos, egocentric videos require models to reason about ongoing hand-object interactions, local state changes, and short-horizon action evolution from the operator's own viewpoint [33, 36].

However, understanding dynamic object interactions in egocentric videos remains particularly challenging. Owing to the first-person viewpoint, manipulated objects are often only partially visible, intermittently leave and re-enter the field of view, and are frequently occluded by the wearer's hands under rapid camera motion. The problem is further compounded by cluttered scenes and the presence of visually similar objects, which make the correct interaction target difficult to identify from instantaneous appearance alone. More fundamentally, answering egocentric questions requires reasoning over temporally evolving evidence rather than relying solely on the current frame, including prior contact history, earlier object states, and the immediate context of an ongoing manipulation sequence [12, 33, 36].

Despite growing interest in egocentric understanding, existing benchmarks still suffer from **limited grounded rationale evaluation**. As summarized in Table 1, general video benchmarks such as Video-MME [14], MMVU [39], and LongVideoBench [18] have substantially advanced video QA and temporal reasoning, but they are not designed for first-person interaction understanding and provide limited support for spatial grounding. Egocentric benchmarks such as EgoSchema [20], EgoThink [6], EgoTempo [8], and MultiHop-EgoQA [4] move evaluation closer to first-person settings, especially for temporal or open-ended reasoning, but they still provide limited support for explicit rationale supervision and fine-grained spatial grounding. More recent benchmarks such as EOC-Bench [42] and EASG-Bench [32] further incorporate egocentric temporal and spatial evaluation, but they still offer limited support for jointly assessing rationale faithfulness, temporal sensitivity, and evidence-aware evaluation.

To address this gap, we introduce **EgoCoT-Bench**, a fine-grained egocentric benchmark for grounded and verifiable operation-centric reasoning with explicit step-by-step rationale annotations and spatio-temporal grounding. EgoCoT-Bench contains 3,172 QA pairs over **351** egocentric videos and is organized into four task groups



(a) Dimensions of EgoCoT-Bench **(b) Video source distribution**

Figure 2: Overall statistics of EgoCoT-Bench. Top: representative video sources in the benchmark. (a) Dimensions of EgoCoT-Bench. (b) Distribution of EgoCoT-Bench samples.

with **12** fine-grained subtasks. These tasks cover egocentric grounding and perception, spatio-temporal retrospection, predictive and causal inference, and high-level grounded reasoning, targeting key capabilities required for first-person manipulation understanding beyond generic scene comprehension.

A central design goal of EgoCoT-Bench is to evaluate not only answer correctness but whether MLLMs reasoning are grounded in explicit first-person evidence. To this end, we construct the benchmark using a spatio-temporal scene graphs (STSG)-guided generation framework. Candidate QA samples are first derived from structured egocentric interaction traces, and subsequently refined through human annotation to ensure semantic correctness, first-person relevance, and fine-grained reasoning quality. Each accepted sample is further augmented with structured evidence annotations—including timestamps, object identities, interaction relations, action history, and localized bounding boxes-enabling evaluation at both the answer and the evidence grounding level.

Using EgoCoT-Bench, we benchmark a range of representative MLLMs such as GPT [27, 28], Qwen [2, 30] and LLaVA [1, 22] series, and observe that fine-grained egocentric reasoning remains highly challenging. While many models can produce correct answers, their underlying rationales are often temporally incomplete, weakly grounded, or inconsistent with the available object-level spatio-temporal evidence. This reveals a notable gap between answer correctness and reasoning faithfulness in current models, which may in turn limit performance gains and lead to error accumulation in more complex scenarios. Our findings highlight the importance of moving beyond final answer accuracy, advocating instead for evaluation protocols that explicitly assess whether model reasoning is consistent with the underlying spatio-temporal evidence in egocentric video understanding.

In summary, our contributions are three-fold: **(1)** we introduce **EgoCoT-Bench**, a fine-grained egocentric benchmark for operation-centric reasoning, comprising 3,172 QA pairs over 351 videos across 12 subtasks; **(2)** we construct the benchmark via an STSG-guided generation and human refinement pipeline, with temporal and

Table 1: Comparison with representative video and egocentric benchmarks.

Benchmark	#Clips	#Samples	Question Type	Annotation	Egocentric	CoT / Rationale	Temporality	Spatial Grounding	Metric
Video-MME [14]	900	2,700	Close	Human	✗	✗	✗	✗	Accuracy
MMVU [39]	1,529	3,000	Open/Close	Human	✗	✓	✓	✗	Accuracy
LongVideoBench [18]	3,763	6,678	Close	Human	✗	✗	✓	✗	Accuracy
EgoSchema [20]	250 hours+	5,000+	Close	Human	✓	✗	✓	✗	Accuracy
EgoThink [6]	595	700	Open	Human	✓	✗	✗	✗	LLM-Judge
EgoTempo [8]	365	500	Open	Auto&Human	✓	✗	✓	✗	LLM-Judge
MultiHop-EgoQA [4]	360	1,080	Open	Auto&Human	✓	✗	✓	✗	Accuracy/LLM-Judge
EOC-Bench [42]	656	3,277	Open/Close	Human	✓	✗	✓	✓	Accuracy
EASG-Bench [32]	221	1,807	Open	Auto	✓	✗	✓	✓	LLM-Judge
EgoCoT-Bench (Ours)	**351**	**3,172**	**Open/Close**	**Auto&Human**	✓	✓	✓	✓	**Accuracy/LLM-Judge**

spatial evidence attached to each accepted samples for grounded first-person reasoning; and **(3)** we propose an evaluation protocol that jointly measures answer correctness, reasoning quality, and spurious correctness for a more faithful assessment.

2 Related Work

2.1 Egocentric Video Understanding

Egocentric video understanding has received increasing attention in recent years, driven by its importance for embodied AI, assistive systems, and first-person human activity analysis [10, 15, 17, 23, 24]. Large-scale datasets such as Ego4D and EPIC-KITCHENS have advanced research on egocentric perception, activity recognition, narration, forecasting, and long-form video understanding [9, 11, 13, 16, 31, 33, 36]. Beyond these foundational resources, more recent benchmarks have extended evaluation toward egocentric question answering, scene-text understanding, object-centric cognition, and cross-view reasoning [5, 19, 26, 41, 42, 45]. These benchmarks have played an important role in promoting first-person video understanding, but many of them primarily emphasize broad scene understanding, long-context comprehension, or object-centric reasoning at a relatively coarse level [12, 20]. As a result, they are less suited for systematically evaluating fine-grained operation-centric reasoning in dynamic first-person scenarios, such as identifying active manipulation targets, tracking short-horizon state changes, or recovering temporally localized hand-object interaction evidence.

2.2 Video Reasoning Benchmarks

A large body of work has also studied reasoning-oriented evaluation for general video understanding [18, 21, 40]. Existing benchmarks have explored temporal perception, event ordering, motion understanding, long-context reasoning, and multi-step video question answering [7, 18, 37, 40, 44]. These resources have significantly improved the diagnosis of multimodal reasoning ability in video-based settings, especially for temporal comprehension and general event-level inference [3, 38]. However, compared with egocentric manipulation scenarios, generic video reasoning benchmarks are typically less sensitive to the distinctive challenges of first-person interaction, where reasoning often depends on local object contact, operator viewpoint, immediate action history, and subtle state transitions. Consequently, they provide only limited support for evaluating whether a model can reason over object-centered manipulation processes in a temporally and spatially grounded manner.

3 EgoCoT-Bench

3.1 Overview

We introduce **EgoCoT-Bench**, a fine-grained benchmark for egocentric video understanding that focuses on operation-centric reasoning in dynamic first-person environments. EgoCoT-Bench contains 3,172 QA pairs collected from **351** egocentric video clips. It is organized into four task groups, covering a total of 12 fine-grained subtasks. These tasks are designed to systematically assess egocentric grounding and perception, temporal retrospection, predictive and causal inference, and high-level grounded reasoning. Together, they target a core challenge of first-person understanding: whether a model can reason about dynamic object-centered interactions in a temporally and spatially grounded manner.

3.2 Benchmark Construction

3.2.1 Video Collection. To ensure both diversity and task relevance, the video collection of EgoCoT-Bench is curated from a wide range of egocentric sources. Specifically, we integrate public first-person datasets including **Ego4D** [15], **EPIC-KITCHENS** [10], **MECCANO** [13, 31], **Charades-Ego** [17], and **HD-EPIC** [29], together with a supplementary set of **self-recorded videos**. These sources provide complementary coverage of interaction scenarios, ranging from daily object use and kitchen activities to more structured manipulation and assembly processes, thereby supporting a comprehensive evaluation of egocentric reasoning tasks.

3.2.2 Task Taxonomy. To systematically characterize first-person operation-centric understanding, we organize EgoCoT-Bench into four task groups with twelve fine-grained subtasks.

(i) Egocentric Grounding & Perception. This group evaluates current interaction grounding in first-person videos: **Active Object Grounding (AOG)** identifies the object currently attended to, touched, or manipulated by the operator; **Hand-Object Association (HOA)** determines which hand is interacting with which object; and **Manipulation State Perception (MSP)** recognizes the current manipulation-related state of the object.

Table 2: Main results on EgoCoT-Bench. Results are reported as accuracy (%).

Method	Mean	Egocentric Grounding & Perception				Spatio-Temporal Retrospection				Predictive & Causal Inference				High-level Grounded Reasoning			
		AOG	HOA	MSP	Mean	SR	LVR	HOTR	Mean	NSA	NAA	LCOI	Mean	PR	HGC	GOT	Mean
Human	95.93	96.18	93.86	97.18	96.11	93.96	95.36	96.88	94.96	98.47	94.44	95.00	96.32	98.71	97.90	91.98	96.34
Proprietary Multimodal Foundation Models																	
GPT-5.1 [27]	66.71	64.20	**63.16**	77.00	67.69	66.04	49.40	56.25	55.23	**86.99**	67.17	70.56	76.63	68.24	79.83	**45.28**	65.15
GPT-5.2 [28]	67.91	64.92	62.28	72.30	66.62	67.55	59.88	59.38	62.42	84.69	64.14	72.22	75.68	65.24	**84.03**	42.92	64.86
Qwen3-VL-Plus [2]	67.12	69.21	61.40	77.00	70.24	69.70	54.66	56.25	59.75	84.69	66.16	71.94	76.00	68.53	77.54	32.70	60.53
Qwen3.5-Plus [30]	70.68	68.26	62.28	85.92	**72.39**	67.55	60.69	53.12	62.67	85.20	**70.71**	**74.72**	**78.21**	**81.12**	82.77	35.85	67.64
Open-Source Multimodal Foundation Models																	
InternVL3.5-1B [35]	53.91	41.77	55.26	69.95	51.88	44.91	51.21	**75.00**	50.06	64.29	55.56	56.94	59.68	55.36	67.65	32.55	52.56
InternVL3.5-2B [35]	61.79	50.60	**63.16**	79.81	60.86	58.11	66.94	71.88	64.18	77.30	59.09	58.33	66.32	60.94	71.01	26.42	53.73
InternVL3.5-4B [35]	61.95	58.71	59.65	73.24	63.00	48.30	54.44	62.50	52.71	81.63	59.60	63.06	70.00	64.81	75.63	38.21	60.32
LLaVA-OneVision-1.5-4B [1]	60.78	55.74	54.39	72.30	60.27	61.51	55.04	40.62	56.62	81.38	61.11	61.67	69.68	63.95	73.11	21.23	53.88
LLaVA-NeXT-Video-7B [22]	44.26	35.08	55.26	46.95	41.55	46.42	40.93	56.25	43.38	62.24	48.99	48.61	54.32	33.91	49.58	17.45	34.26
InternVL3.5-8B [35]	64.06	56.32	61.40	70.89	61.26	54.34	**67.74**	68.75	63.30	80.36	66.16	67.22	72.42	66.95	73.11	25.94	56.37
LLaVA-OneVision-1.5-8B [1]	60.81	53.94	58.77	74.65	60.59	57.36	51.61	40.62	53.09	82.14	69.19	61.94	71.79	68.67	71.01	21.23	54.76
Qwen3-VL-8B [2]	65.42	**69.54**	59.29	81.60	71.43	64.02	60.69	59.38	61.75	83.03	63.13	64.72	71.91	59.91	78.15	25.00	55.43
InternVL3.5-14B [35]	64.09	56.32	56.14	75.12	61.66	56.98	63.71	**75.00**	61.92	78.32	65.66	70.00	72.53	61.37	72.27	36.79	57.54
Qwen3.5-27B [30]	**71.28**	68.26	61.40	84.51	71.85	**72.83**	**67.74**	59.38	**69.10**	84.65	68.69	73.33	77.03	77.68	80.67	34.43	65.30
Qwen3-VL-30B-A3B [2]	64.63	62.44	61.06	82.16	67.88	65.15	65.86	62.50	65.49	81.89	59.09	66.94	71.47	66.52	73.11	9.05	51.10
Qwen3-VL-32B [2]	67.09	67.78	62.28	79.81	70.38	64.53	55.04	68.75	58.76	84.18	70.20	71.67	76.53	69.10	79.83	27.83	60.03
Qwen3.5-122B-A10B [30]	69.96	68.26	61.40	86.38	**72.39**	70.72	62.70	56.25	65.11	81.63	**70.71**	73.61	76.32	79.83	79.41	30.19	64.28
Qwen3-VL-235B-A22B [2]	65.86	67.54	57.02	77.00	68.63	71.32	52.82	56.25	59.14	85.97	68.18	62.50	73.37	70.82	78.99	27.36	60.18
Qwen3.5-397B-A17B [30]o	70.11	68.26	58.77	**87.79**	**72.39**	69.81	59.07	56.25	62.55	84.95	68.18	70.83	76.11	78.97	83.61	38.68	**68.08**

(ii) Spatio-Temporal Retrospection. This group measures whether a model can recover object-centric evidence from preceding moments: **State Retrospection (SR)** recalls an object's earlier state; **Location / Visibility Retrospection (LVR)** recovers its previous location or status; and **Hand-Object Temporal Retrospection (HOTR)** infers the temporal order of hand-object interactions.

(iii) Predictive & Causal Inference. This group evaluates short-horizon anticipation and local causal reasoning grounded in the current manipulation context: **Next State Anticipation (NSA)** predicts an object's most likely next state; **Next Action Anticipation (NAA)** predicts the operator's most likely next action; and **Local Cause-Outcome Inference (LCOI)** identifies the recent action directly responsible for the observed outcome or state change.

(iv) High-level Grounded Reasoning. This group focuses on compositional reasoning over progress, evidence chains, and goal-oriented tracking: **Progress Reasoning (PR)** infers the current operation step or whether a step has been completed; **Hand-Object Grounded CoT (HGC)** generates interpretable reasoning chains that combine hand-object cues, temporal evidence, and visual grounding; and **Goal-Oriented Object Tracking (GOT)** tracks an object over time according to its functional role in the ongoing manipulation goal.

3.2.3 *Construction Pipeline.*

STSG-Guided Candidate Generation. To ensure the quality and verifiability of EgoCoT-Bench, we build the benchmark through a structured human-in-the-loop pipeline in which candidate generation is grounded in verified spatio-temporal scene graph (STSG) rather than free-form video description as illustrated in Figure 1. In the first stage, each video clip is converted into an ego-adapted STSG, which serves as an intermediate representation for candidate construction. Before candidate generation, the STSG is manually inspected and refined to correct unreliable object identities, temporally inconsistent interaction links, ambiguous state transitions, and misaligned spatial grounding. The STSG organizes object instances, operator body parts, action traces, interaction relations, temporal states, and bounding boxes across time.

Based on the refined STSG, we derive candidate samples by traversing task-specific evidence paths that connect each target answer to concrete first-person interaction cues. The LLM is then used to render these verified structural facts into natural-language questions, answer options, and rationales, rather than to invent the underlying evidence. For every candidate sample, we preserve the associated structural metadata, including timestamps, object identities, action history, interaction relations, and bounding boxes, so that the sample remains traceable and can be explicitly checked during downstream evidence-aware evaluation.

Human Refinement and Quality Control. All generated candidates are then subjected to careful manual screening under a multi-round review protocol. First, four human annotators independently perform an initial screening pass to remove obviously invalid or weak candidates, such as those with ambiguous targets, weak first-person relevance, inconsistent reasoning, or low-quality distractors. Next, the retained candidates are cross-checked by different reviewers, who verify the consistency among the question, answer, and reasoning. Finally, a lead reviewer performs the last-round inspection and adjudication, resolving disagreements, rejecting low-confidence cases, and confirming the final accepted version. We keep a sample

Table 3: Reasoning Score (R) and Spurious Correct Rate (SCR) evaluation on EgoCoT-Bench. Results are reported with R on a strict 0-5 scale and SCR in percentage (%). Higher R is better, while higher SCR indicates worse answer-reasoning consistency.

Method	Mean		Egocentric Grounding & Perception		Spatio-Temporal Retrospection		Predictive & Causal Inference		High-level Grounded Reasoning	
	R ↑	SCR ↓	R ↑	SCR ↓	R ↑	SCR ↓	R ↑	SCR ↓	R ↑	SCR ↓
Proprietary Multimodal Foundation Models										
GPT-5.1	2.77	4.91	2.65	5.35	2.16	5.25	3.43	**1.79**	2.67	9.21
GPT-5.2	2.85	**4.27**	2.39	**3.23**	**2.73**	**4.02**	3.40	2.36	2.76	8.80
Qwen3-VL-Plus	**3.08**	7.84	**3.04**	7.44	2.61	5.29	**3.64**	6.37	2.87	13.87
Qwen3.5-Plus	2.92	9.10	2.88	7.41	2.31	10.87	3.50	7.67	2.87	11.47
Open-Source Multimodal Foundation Models										
InternVL3.5-1B	2.21	13.33	2.11	8.53	1.77	17.63	2.70	8.99	2.14	20.61
InternVL3.5-2B	2.53	7.86	2.43	9.69	2.29	7.86	2.98	2.70	2.29	14.44
InternVL3.5-4B	2.45	9.57	2.39	8.72	1.86	9.57	3.09	4.96	2.31	17.96
LLaVA-OneVision-1.5-4B	2.50	9.57	2.32	10.47	2.03	14.92	3.17	3.77	2.32	13.59
LLaVA-NeXT-Video-7B	1.85	22.93	1.57	25.16	1.51	35.46	2.59	8.53	1.53	33.33
InternVL3.5-8B	2.56	5.61	2.39	5.47	2.25	4.98	3.15	3.92	2.30	9.61
LLaVA-OneVision-1.5-8B	2.21	24.73	2.08	28.31	1.76	27.31	2.77	21.11	2.08	24.06
Qwen3-VL-8B	2.73	10.07	2.74	9.25	2.29	14.40	3.27	6.46	2.47	12.17
InternVL3.5-14B	2.60	5.36	2.50	5.43	2.27	4.48	3.17	2.46	2.30	11.45
Qwen3.5-27B	2.96	7.25	2.87	8.39	2.49	8.94	3.56	3.15	2.78	10.54
Qwen3-VL-30B-A3B	2.79	7.25	2.73	8.91	2.43	10.42	3.36	5.89	2.46	**7.76**
Qwen3-VL-32B	2.96	7.99	2.88	9.90	2.40	7.51	3.63	5.36	2.77	10.73
Qwen3.5-122B-A10B	2.94	9.73	2.83	11.29	2.43	11.26	3.48	7.45	**2.88**	9.79
Qwen3-VL-235B-A22B	2.78	11.01	2.70	11.32	2.32	9.38	3.42	8.90	2.53	16.05
Qwen3.5-397B-A17B	2.87	10.93	2.86	9.81	2.29	12.70	3.37	9.82	2.87	12.04

only when its question, answer, rationale, and supporting evidence are mutually consistent and clearly grounded in the video. Through this process, EgoCoT-Bench retains only samples that satisfy semantic correctness, egocentric relevance, and evidence consistency, yielding a human-refined benchmark with structured temporal and spatial evidence support.

3.2.4 Evaluation Metrics. To provide a comprehensive assessment of multimodal large language models (MLLMs) in egocentric environments, EgoCoT-Bench evaluates not only the final answer correctness but also the quality of the reasoning process and the consistency between them. Conventional benchmarks often rely solely on answer accuracy, which may overestimate model capability when the correct answer is obtained without sound reasoning. To address this issue, we adopt a three-metric evaluation protocol consisting of Answer Accuracy (Acc), Reasoning Score (R), and Spurious Correct Rate (SCR).

Answer Accuracy (Acc). All tasks in EgoCoT-Bench are formulated as four-way multiple-choice questions. We adopt a strict exact-match criterion to evaluate the final prediction.

Reasoning Score (R). In egocentric video understanding, predictions may be unsupported by grounded and coherent reasoning. We therefore evaluate the model's reasoning quality by scoring its generated reasoning steps against the annotated reference reasoning by employing a strong LLM (Qwen-Max) as a judge to assess each prediction from the perspectives of logical coherence, factual consistency, and alignment with the visual evidence on a 0-5 scale.

Spurious Correct Rate (SCR). To quantify how often a model arrives at the correct answer with weak reasoning, we introduce Spurious Correct Rate (SCR), which measures the proportion of answer-correct cases whose reasoning remains weak. Specifically, a prediction is considered *spurious correct* if it satisfies both $\hat{y}_i = y_i$ and $S_{\text{judge}}(\hat{c}_i, c_i) \leq 2$. The SCR is defined as:

$$\text{SCR} = \frac{\sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = y_i)\, \mathbb{I}\big(S_{\text{judge}}(\hat{c}_i, c_i) \leq 2\big)}{\sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = y_i)}. \tag{1}$$

SCR is reported as a percentage where a higher value indicates worse answer-reasoning consistency.

4 Experiment

4.1 Models and Human Evaluation.

We evaluate a broad set of multimodal large language models (MLLMs) on **EgoCoT-Bench**, including 4 proprietary MLLMs and 15 open-source MLLMs spanning different parameter scales and architectural families. Among proprietary models, we evaluate GPT-5.1 [27], GPT-5.2 [28], Qwen3-VL-Plus [2] and Qwen3.5-Plus [30]. For open-source models, we test InternVL3.5 [35], LLaVA-OneVision-1.5 [1], LLaVA-NeXT-Video [22], Qwen3-VL [2] and Qwen3.5 [30]. In addition to model evaluation, we also measure human performance on EgoCoT-Bench with three volunteers.

4.2 Main Results Analysis.

The overall results in Table 2 show that EgoCoT-Bench remains highly **challenging for current MLLMs**. The best overall accuracy is achieved by Qwen3.5-27B with Benchmark Models and Human Evaluation, followed by Qwen3.5-Plus with 70.68% and Qwen3.5-397B-A17B with 70.11%. However, even the strongest model still trails human performance (95.93%) by a large margin, indicating

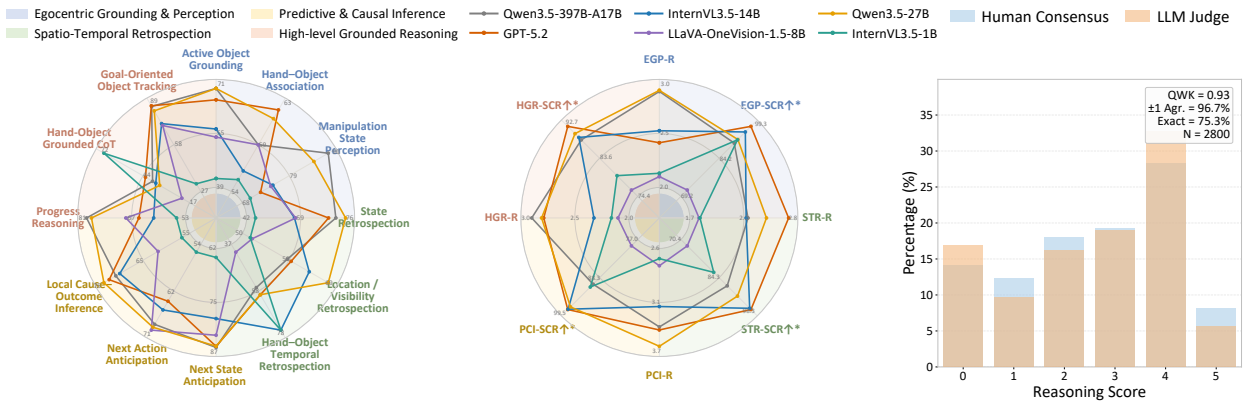


Figure 3: Fine-grained radar analysis on EgoCoT-Bench. Left: answer accuracy (%) across 12 subtasks. Middle: reasoning quality across four task groups using Reasoning Score (R) and inverted Spurious Correct Rate (SCR↑*, i.e., 100-SCR), where larger radii indicate better performance. Right: comparison between human ratings and LLM-judge scores on a randomly sampled subset of model responses.

that fine-grained egocentric reasoning is far from solved. This gap is consistently observed across all four task groups.

From a group-level perspective, Predictive & Causal Inference is comparatively more tractable than the other dimensions, where the best group accuracy reaches 78.21% by Qwen3.5-Plus. In contrast, Spatio-Temporal Retrospection and High-level Grounded Reasoning remain notably harder, with the best group results being only 69.10% and 68.08%, respectively. This pattern suggests that **current MLLMs are relatively better at short-horizon anticipation and local cause-outcome reasoning than at recovering earlier interaction evidence or performing compositional object-centered reasoning over longer temporal context.**

Fine-grained Task Analysis. Fig. 3 further reveals a highly uneven capability profile across the 12 subtasks. Among individual subtasks, models are relatively strong on Manipulation State Perception (MSP), Next State Anticipation (NSA), and Hand-Object Grounded CoT (HGC), where the best accuracies reach 87.79%, 86.99%, and 84.03%, respectively. These results suggest that **current MLLMs can often capture immediate object state cues and some short-range action consequences when the visual evidence is sufficiently explicit.**

By contrast, Goal-Oriented Object Tracking (GOT) is by far the most difficult subtask. The best model achieves only 45.28%, which is lower than the human score of 91.98%. This large gap indicates that tracking an object according to its functional role in an evolving manipulation process is still beyond the capability of current systems. In addition, tasks such as Hand-Object Association (HOA) and Location / Visibility Retrospection (LVR) also remain challenging, suggesting **persistent weaknesses in local interaction grounding and in recalling object-centric evidence from earlier moments.**

Reasoning Quality and Answer-Reasoning Consistency. Table 3 shows that answer correctness and reasoning quality do not fully align. Although Qwen3.5-27B achieves the best overall accuracy, the highest mean reasoning score is obtained by Qwen3-VL-Plus with 3.08/5. Meanwhile, GPT-5.2 yields the lowest SCR at only 4.27%, indicating the strongest consistency between correct answers and acceptable reasoning among the evaluated models. These results confirm that **a model may obtain the right answer while still relying on weak, incomplete, or weakly grounded rationales.** The right panel of Fig. 3 further shows that the LLM-judge is well aligned with human evaluation, as evidenced by a high quadratic weighted kappa (QWK = 0.93), 96.7% ±1 agreement, and 75.3% exact agreement on 2,800 randomly selected responses.

This inconsistency becomes more evident in the group-wise reasoning analysis. On Predictive & Causal Inference, several models achieve relatively high reasoning scores together with low SCR, suggesting that short-horizon causal judgments are easier to verbalize coherently. In contrast, High-level Grounded Reasoning exhibits substantially worse SCR for many models, despite moderate answer accuracy. This suggests that **models can sometimes guess the correct option, yet fail to provide reasoning that is faithfully aligned with the relevant hand-object interactions, temporal evidence, or functional object roles.**

Overall, these results highlight the importance of evaluating egocentric reasoning beyond answer accuracy alone. EgoCoT-Bench exposes a non-trivial amount of *spurious correctness*, where answer-level success can mask insufficiently grounded reasoning. We believe this is an important property for future benchmark design, especially for embodied or assistive systems that must justify their decisions using temporally and spatially verifiable evidence.

5 Conclusion

We present **EgoCoT-Bench**, a fine-grained benchmark for grounded, verifiable operation-centric reasoning in egocentric videos, featuring explicit spatio-temporal evidence and rationale annotations. Extensive evaluations of state-of-the-art MLLMs reveal that, despite strong answer accuracy on certain subtasks, models still struggle with evidence grounding and rationale consistency. These findings underscore the need for more reliable benchmarks and models for egocentric reasoning. We hope EgoCoT-Bench serves as a robust testbed for advancing grounded, verifiable, and temporally coherent reasoning in egocentric video understanding.

References

[1] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. 2025. LLaVA-OneVision-1.5: Fully Open Framework for Democratized Multimodal Training. In *arXiv*.

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631* (2025).

[3] Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Yu-Chiang Frank Wang. 2024. ReXTime: A Benchmark Suite for Reasoning-Across-Time in Videos. *arXiv preprint arXiv:2406.19392* (2024).

[4] Qirui Chen, Shangzhe Di, and Weidi Xie. 2025. Grounded multi-hop videoqa in long-form egocentric videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 2159–2167.

[5] Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. 2024. EgoPlan-Bench: Benchmarking Multimodal Large Language Models for Human-Level Planning. arXiv:2312.06722 [cs.CV] https://arxiv.org/abs/2312.06722

[6] Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. 2024. EgoThink: Evaluating First-Person Perspective Thinking Capability of Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14291–14302.

[7] Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. 2025. V-STaR: Benchmarking Video-LLMs on Video Spatio-Temporal Reasoning. arXiv:2503.11495 [cs.CV] https://arxiv.org/abs/2503.11495

[8] Plizzari Chiara, Tonioni Alessio, Yongqin Xian, Ace Kulshrestha, and Tombari Federico. 2025. Omnia de EgoTempo: Benchmarking Temporal Understanding of Multi-Modal LLMs in Egocentric Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2022. Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision (IJCV)* 130 (2022), 33–55. https://doi.org/10.1007/s11263-021-01531-2

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2018. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[11] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. 2022. EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*.

[12] Shangzhe Di and Weidi Xie. 2024. Grounded Question-Answering in Long Egocentric Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12934–12943.

[13] Ragusa Francesco, Furnari Antonino, and Farinella Giovanni, Maria. 2022. MECCANO: A Multimodal Egocentric Dataset for Humans Behavior Understanding in the Industrial-like Domain. arXiv:2209.08691 [cs.CV]

[14] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *CVPR*.

[15] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18995–19012.

[16] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. 2024. Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19383–19400.

[17] Sigurdsson Gunnar, A., Gupta Abhinav, Schmid Cordelia, Farhadi Ali, and Alahari Karteek. 2018. Actor and Observer: Joint Modeling of First and Third-Person Videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[18] Wu Haoning, Li Dongxu, Chen Bei, and Li Junnan. 2024. LongVideoBench: A Benchmark for Long-context Interleaved Video-Language Understanding. arXiv:2407.15754 [cs.CV] https://arxiv.org/abs/2407.15754

[19] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. 2022. EgoTaskQA: Understanding Human Tasks in Egocentric Videos. In *The 36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks*.

[20] Mangalam Karttikeya, Akshulakov Raiymbek, and Malik Jitendra. 2023. EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding. arXiv:2308.09126 [cs.CV] https://arxiv.org/abs/2308.09126

[21] Li Kunchang, Wang Yali, He Yinan, Li Yizhuo, Wang Yi, Liu Yi, Wang Zun, Xu Jilan, Chen Guo, Luo Ping, Wang Limin, and Qiao Yu. 2023. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. *arXiv* (2023). https://arxiv.org/abs/2311.17005

[22] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024. LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models. *arXiv preprint arXiv:2407.07895* (2024).

[23] Yin Li, Miao Liu, and James M. Rehg. 2018. In the Eye of Beholder: Joint Learning of Gaze and Actions in First Person Video. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[24] Yin Li, Miao Liu, and James M Rehg. 2018. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*. 619–635.

[25] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

[26] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. 2024. OpenEQA: Embodied Question Answering in the Era of Foundation Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16488–16498.

[27] OpenAI. 2025. GPT-5.1 Model. https://developers.openai.com/api/docs/models/gpt-5.1. Official OpenAI API documentation; accessed 2026-03-27.

[28] OpenAI. 2025. GPT-5.2 Model. https://developers.openai.com/api/docs/models/gpt-5.2. Official OpenAI API documentation; accessed 2026-03-27.

[29] Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti, Michael Wray, Hazel Doughty, and Dima Damen. 2025. HD-EPIC: A Highly-Detailed Egocentric Video Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[30] Qwen Team. 2026. Qwen3.5: Towards Native Multimodal Agents. https://qwen.ai/blog?id=qwen3.5

[31] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. 2021. The MECCANO Dataset: Understanding Human-Object Interactions from Egocentric Videos in an Industrial-like Domain. In *IEEE Winter Conference on Application of Computer Vision (WACV)*. arXiv:2010.05654

[32] Ivan Rodin, Tz-Ying Wu, Kyle Min, Sharath Nittur Sridhar, Antonino Furnari, Subarna Tripathi, and Giovanni Maria Farinella. 2025. EASG-Bench: Video Q&A Benchmark with Egocentric Action Scene Graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 2732–2737.

[33] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. 2022. Assembly101: A Large-Scale Multi-View Video Dataset for Understanding Procedural Activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21096–21106.

[34] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. 2025. Video Understanding with Large Language Models: A Survey. *IEEE Transactions on Circuits and Systems for Video Technology* (2025). doi:10.1109/TCSVT.2025.3566695

[35] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. 2025. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265* (2025).

[36] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. 2023. HoloAssist: an Egocentric Human Interaction Dataset for Interactive AI Assistants in the Real World. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 20270–20281.

[37] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 2021. STAR: A Benchmark for Situated Reasoning in Real-World Videos. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS)*.

[38] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9777–9786.

[39] Zhao Yilun, Xie Lujing, Zhang Haowei, Gan Guo, Long Yitao, Hu Zhiyuan, Hu Tongyan, Chen Weiyuan, Li Chuhan, Song Junyang, et al. 2025. MMVU: Measuring Expert-Level Multi-Discipline Video Understanding. arXiv:2501.12380 [cs.CV] https://arxiv.org/abs/2501.12380

[40] Liu Yuanxin, Li Shicheng, Liu Yi, Wang Yuxiang, Ren Shuhuai, Li Lei, Chen Sishuo, Sun Xu, and Hou Lu. 2024. TempCompass: Do Video LLMs Really Understand Videos? *arXiv preprint arXiv: 2403.00476* (2024).

[41] He Yuping, Huang Yifei, Chen Guo, Pei Baoqi, Xu Jilan, Lu Tong, and Pang Jiangmiao. 2025. EgoExoBench: A Benchmark for First- and Third-person View Video Understanding in MLLMs. *arXiv* (2025). https://arxiv.org/abs/2507.18342

[42] Yuan Yuqian, Dang Ronghao, Li Long, Li Wentong, Jiao Dian, Li Xin, Zhao Deli, Wang Fan, Zhang Wenqiao, Xiao Jun, and Zhuang Yueting. 2025. EOC-Bench: Can MLLMs Identify, Recall, and Forecast Objects in an Egocentric World? *arXiv* (2025). https://arxiv.org/abs/2506.05287

[43] Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Yansong Feng and Els Lefever (Eds.). Association for Computational Linguistics, Singapore, 543–553. doi:10.18653/v1/2023.emnlp-demo.49

[44] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264* (2024).

[45] Sheng Zhou, Junbin Xiao, Qingyun Li, Yicong Li, Xun Yang, Dan Guo, Meng Wang, Tat-Seng Chua, and Angela Yao. 2025. EgoTextVQA: Towards Egocentric Scene-Text Aware Video Question Answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*. 3363–3373.