Beyond Description: Cognitively Benchmarking Fine-Grained Action for Embodied Agents

Dayong Liu[1,2⋆‡], Chao Xu[2,3⋆], Weihong Chen[2,3], Suyu Zhang[2,3], Juncheng Wang[4], Jiankang Deng[5], Baigui Sun[2,3†], and Yang Liu[2,3†]

[1] Zhejiang University
[2] IROOTECH TECHNOLOGY
[3] Wolf 1069 b Lab, Sany Group
[4] The Hong Kong Polytechnic University
[5] Imperial College London

liu_dayong@zju.edu.cn, {baigui.sun, yang.liu1}@irootech.com

Abstract. Multimodal Large Language Models (MLLMs) show promising results as decision-making engines for embodied agents operating in complex, physical environments. However, existing benchmarks often prioritize high-level planning or spatial reasoning, leaving the fine-grained action intelligence required for embodied physical interaction underexplored. To address this gap, we introduce CFG-Bench, a new benchmark designed to systematically evaluate this crucial capability. CFG-Bench consists of 1,368 curated videos paired with 19,562 question-answer pairs spanning three evaluation paradigms targeting four cognitive abilities: 1) Physical Interaction, 2) Temporal-Causal Relation, 3) Intentional Understanding, and 4) Evaluative Judgment. Together, these dimensions provide a systematic framework for assessing a model's ability to translate visual observations into actionable knowledge, moving beyond mere surface-level recognition. Our comprehensive evaluation on CFG-Bench reveals that leading MLLMs struggle to produce detailed instructions for physical interactions and exhibit profound limitations in the higher-order reasoning of intention and evaluation. Moreover, supervised fine-tuning (SFT) on our data demonstrates that teaching an MLLMs to articulate fine-grained actions directly translates to significant performance gains on established embodied benchmarks. Our analysis highlights these limitations and offers insights for developing more capable and grounded embodied agents. Project page: https://cfg-bench.github.io/

Keywords: Multimodal Large Language Models · Fine-Grained Action · Cognitive Benchmark · Embodied Agents

1 Introduction

The remarkable capabilities of Multimodal Large Language Models (MLLMs) [4, 53, 56] in understanding and reasoning across vision and language have driven

‡ Work done during an internship at IROOTECH TECHNOLOGY.
⋆ Equal contribution. † Corresponding authors.

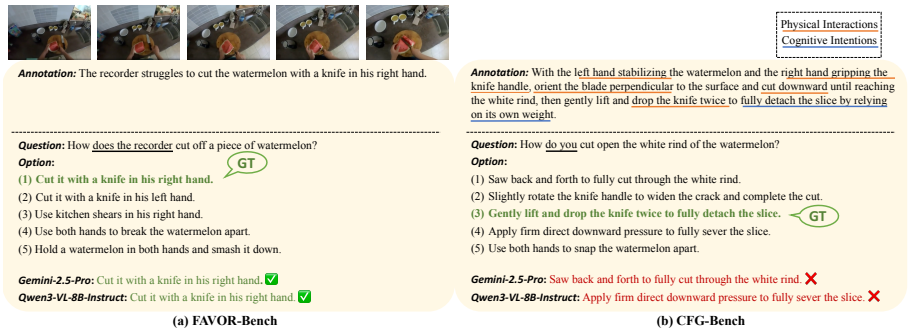

Fig. 1: Illustration of CFG-Bench's focus on embodied intelligence over descriptive accuracy. The part (a) shows how FAVOR-Bench annotates and questions from a third-person perspective for motion-level actions, a task which current MLLMs can often solve. In contrast, the part (b) demonstrates CFG-Bench's fine-grained annotation and first-person scenario questions, which probes for the actionable physical and intentional details necessary for embodied agents. Current MLLMs struggle to master the crucial fine-grained details required for physical interaction.

Table 1: Comparison with other benchmarks. CFG-Bench introduces a four-tiered cognitive framework for embodied fine-grained intelligence, distinguishing it from existing action benchmarks focused on third-person coarse description and other embodied benchmarks that prioritize spatial reasoning and high-level planing over the fine-grained physical action. It also provides a more comprehensive evaluation protocol by integrating three evaluation paradigms of QAs. ✓indicates partial coverage.

Benchmarks	Embodied	Fine-Grained Action				Data Scale	QA Pairs			
		Interaction	Temporal-Causal	Intention	Evaluation		Number	Closed-Ended	Open-Ended	Counterfactuals
MVBench [28]	✗	✓	✓	✗	✗	4,000	4,000	✓	✗	✓
VideoMME [15]	✗	✓	✓	✗	✗	2,525	2,525	✓	✗	✗
EgoSchema [34]	✗	✓	✓	✗	✗	5,031	5,031	✓	✗	✗
EgoTaskQA [25]	✗	✓	✓	✓	✗	2,315	40,322	✓	✓	✓
Motion-Bench [22]	✗	✓	✓	✗	✗	5,385	8,052	✓	✗	✗
FAVOR-Bench [46]	✗	✓	✓	✗	✗	1,776	8,184	✓	✓	✗
VSI-Bench [50]	✓	-	-	-	-	288	5,000	✓	✓	✗
ECBench [13]	✓	-	-	-	-	386	4,324	✓	✓	✓
EMBODIEDEVAL [9]	✓	-	-	-	-	328	1,533	✓	✗	✓
CFG-Bench	✓	✓	✓	✓	✓	1,368	19,562	✓	✓	✓

their active exploration as the decision-making engine for embodied agents [20, 26, 29, 39, 41, 42]. Operating within interactive environments, these agents are required to interpret multimodal observations to inform their planing and guide their subsequent actions. To systematically evaluate these capabilities, a variety of benchmarks have emerged [9, 13, 38, 50, 51], with a significant focus on areas like egocentric perception [8], visual grounding [54], and spatial reasoning [5].

However, while existing benchmarks effectively assess an agent's visual-spatial intelligence and strategic planning, they largely overlook the most challenging dimension of interaction: *fine-grained action intelligence*, which refers to comprehending the details of how an action is executed, both *physically* and *cognitively*. Probing these fine-grained details is crucial for pushing embodied imitation learning [12, 47] beyond rigid trajectory replication, as it tests an agent's ability to

perform executable physical actions as well as high-level planning and reasoning required to generalize successfully to various scenarios.

Concurrently, benchmarks within the action understanding domain have begun to recognize the importance of fine-grained detail. Pioneering work like FAVOR-Bench [46], has moved beyond event-level labels to probe motion-level perception and the precise temporal sequence of actions. However, from an embodied perspective, their approach is limited in two ways. (1) The definition of fine-grained is often not granular enough for physical execution. As shown in Fig. 1, the current annotation simply states "cut the watermelon with a knife in right hand," yet fails to capture granular execution sequences, e.g., "stabilize the watermelon with the left hand, grip the knife handle with the right hand and then cut down vertically with the blade down." (2) It is oriented towards objective video description from a third-person perspective, does not encompass the deep cognitive reasoning underlying surface-level action, such as "gently lift and drop knife twice" is to fully detach the slice by relying on its own weight.

To systematically address these challenges, we first synthesize a hierarchical framework grounded in Norman's Action Cycle [36] that deconstructs fine-grained action intelligence into four hierarchical tiers: (1) *Physical Interaction*, detailing how an action is physically executed (aligned with Normman's Specifying and Performing); (2) *Temporal-Causal Relation*, reasoning how actions connect through time and causality (Normman's Sequencing and Planning); (3) *Intentional Understanding*, inferring why an action is performed (Norman's Goal Formation); and (4) *Evaluative Judgment*, evaluating how well an action was done and how to improve it if necessary (Norman's Comparing Outcome with Goal). This structure helps an embodied agent internalize the closed loop among formulation, planning, execution, and feedback from demonstrations, transforming surface-level recognition into cognitive executable knowledge.

Building on this four-tiered framework, we introduce CFG-Bench, **c**ognitively benchmarking **f**ine-**g**rained action for embodied agents. Specifically, CFG-Bench includes a curated dataset of 1,368 videos, covering ego- and exo-centric daily-life records, hand-object interactions, and more complex outdoor activities. We employ close-ended multiple-choice question-answering (QA) for more objective Physical Interaction and Temporal-Causal Relation tiers, while adopting an open-ended format for the higher-order Intentional Understanding and Evaluative Judgment tiers, which is crucial to prevent models from reasoning backward from the textual options, forcing them to ground their inference in the visual evidence. Notably, we integrate open-ended counterfactual questions across all four tiers to directly combat model hallucination and probe a deeper action comprehending. All QA pairs undergo a rigorous human-in-the-loop verification process to eliminate overly simple or erroneous questions. Overall, CFG-Bench comprises 19,562 QA pairs distributed across 11 distinct tasks within the 4 tiers. For open-ended tasks, we use the standard GPT-assisted evaluation [43, 55], but apply a gating mechanism for parts of counterfactuals: a response is scored only if the model first identifies the question's false premise. Furthermore, since both annotation and evaluation involve LLMs, we include Inter-Annotator agreement

and Human-LLM agreement to validate the reliability of LLM-assisted ground truth annotations and GPT-assisted evaluation, respectively.

Finally, we conduct a comprehensive zero-shot evaluation of leading proprietary and open-source MLLMs on CFG-Bench. Furthermore, by fine-tuning Qwen2.5-VL [4] on our data, we demonstrate how a model that has learned to articulate fine-grained actions with physical and intentional details shows substantial performance gains on established benchmarks for embodied manipulation and planning [51].

Our contribution can be concluded from three aspects:

- We define fine-grained action intelligence and propose a novel four-tiered framework to systematically deconstruct and evaluate it beyond mere objective description.
- We construct CFG-Bench, a new benchmark featuring a hybrid QA design and innovative open-ended counterfactual challenges to rigorously test the physical and cognitive grounding of MLLMs on fine-grained actions.
- We provide a comprehensive analysis revealing the limitations of current MLLMs and validate that training on CFG-Bench significantly enhances planning and manipulation capabilities for embodied physical execution.

2 Related Works

2.1 Benchmarks for Embodied Agent

As MLLMs are increasingly explored as the "brains" for robots [6, 24, 33, 35, 44], several benchmarks have emerged to evaluate their ability to perceive and reason about the physical world. For instance, VSI-Bench [50] is designed to assess the visual-spatial intelligence of MLLMs, probing their understanding of object positions, orientations, and geometric relations. Similarly, ECBench [13] is dedicated to evaluating embodied cognitive abilities within both static and dynamic environments. These benchmarks are key to an agent's spatial perception and mapping baseline.

A second category, offline benchmarks focus on what high-level action/plan, e.g., Ego-PlanBench [7] for ego symbolic planning; EgoExoBench [21] for ego-exo alignment; RoboBench [32] and BEAR [38] for task- and atomic-level embodied capability diagnostics. The recent online benchmarks like EMBODIEDEVAL [9] and EMBODIEDBENCH [51] introduce a more comprehensive suite of diverse tasks and scenes to evaluate whether the task succeeds. They abstract away the complexity of physical interaction. The former focuses on the strategic what to do next by reducing actions to high-level, symbolic commands (e.g., pick up). The latter's manipulation tasks require the MLLM to act as a direct, end-to-end policy that outputs low-level commands, such as a 7-DoF vector. We argue that our approach, which focuses on generating cognitive knowledge about how and why an action should be performed, is deeply complementary to these works. The experiments prove that an MLLM fine-tuned on our data achieves significant performance gains on the high-level planning and low-level control tasks in Sec. 4.5.

2.2 Benchmarks for Action Understanding

The evolution of action understanding benchmarks reveals a progression towards greater detail, yet a critical gap persists between understanding for descriptive analysis versus for embodied execution. Early comprehensive benchmarks like MVBench [28] and Video-MME [15] include subsets for action, but these are situated within a broader evaluation of general video understanding. A more dedicated line of work, exemplified by ActivityNet-QA [52], concentrates on event-level granularity, requiring models to identify the main activity or sequence of major events, but still lacks consideration for more fine-grained actions.

To address this, pioneering works like EgoTaskQA [25], MotionBench [22], MotionSight [14], and FAVOR-Bench [46] move beyond event labels and attend to motion-level perception and precise temporal sequences. However, their definition of fine-grained is still primarily oriented towards achieving the fidelity of a third-person observer, targeting high-accuracy video description. They do not systematically evaluate the deep procedural knowledge required for an embodied agent to physically replicate an action, like the specific contact details and dynamics, and the causal, intentional, and evaluative reasoning that underpins skillful interaction. This deeper layer of fine-grained action intelligence, which separates merely observing an action from truly understanding how and why to perform it, remains the unaddressed frontier.

3 CFG-Bench

An embodied agent's ability to translate multimodal inputs into successful physical interactions hinges on its fine-grained action intelligence. Encompassing the physical how and cognitive why of an action, this capability serves as the essential bridge between abstract planning and low-level physical control. To this end, we develop CFG-Bench to systematically measure this intelligence. First, we deconstruct the fine-grained action into 4 embodied cognitive tiers (Sec. 3.1). We then detail the dataset construction process, including video curation and QA generation (Sec. 3.2). Finally, we describe the evaluation protocols for QAs (Sec. 3.3). Fig. 2 is an overview of CFG-Bench tasks.

3.1 Taxonomy and Task Design

Physical Interaction is the ability to comprehend the physical mechanics of an action can move beyond simple event recognition to a detailed, factual understanding of how that action is physically performed. This granular perception is the foundation for an agent's ability to imitate, interact with, and learn from the physical world.

Concretely, *Factual Action Understanding* (FAU) is designed to measure the model's ability to extract the actionable and fine-grained details of physical interaction from visual inputs. Through multi-choice questions (MCQ), this task cover a full spectrum of the specific agent performing the action (e.g., left hand

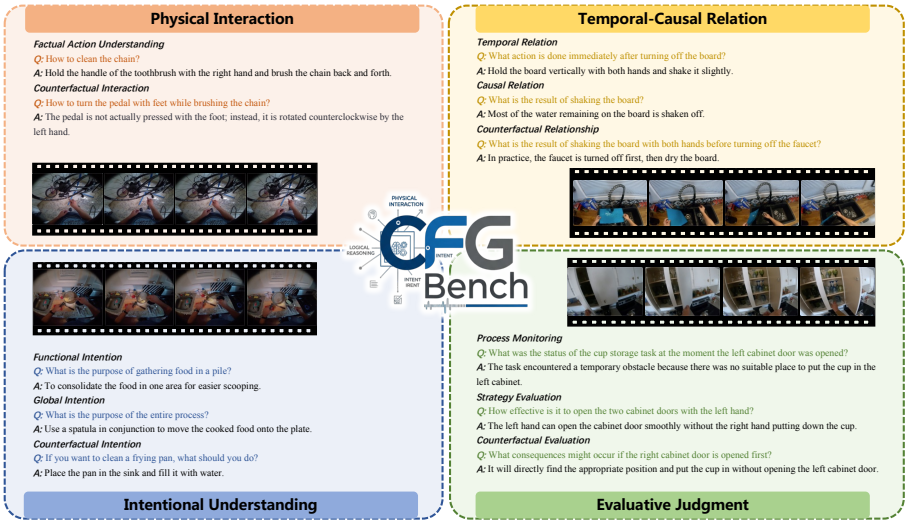

Fig. 2: Task demonstration of CFG-Bench. Notably, all QA pairs, including those above, are slightly simplified for clarity and brevity.

or right hand), the objects and tools being manipulated (e.g., cup or bottle), the specific parts of an object being interacted with (e.g., handle or rim), the precise operation type (e.g., gripping or holding), and the dynamic properties of the motion (e.g., forward quickly or backward slowly).

We further introduce *Counterfactual Interaction* (CIA) to test the model's robustness and its ability to ground responses firmly in visual evidence through open-ended QA. This task confronts the model with a question that contains a false premise. A successful response requires the model to identify this false premise as inconsistent with the video and provide a correction based on what actually occurred. This method directly penalizes the common failure mode of acquiescence, where models blindly accept and hallucinate based on incorrect information.

Temporal-Causal Relation elevates from isolated actions to the logical structure that connects them, which indicates the model's ability to reason about how actions relate to each other in time and through cause-and-effect. This capability is a fundamental factor for an agent to plan, anticipate outcomes, and comprehend long-horizontal tasks.

This is specifically evaluated by *Temporal Relation* (TR) and *Causal Relation* (CR). For the former, TR is designed to ensure a comprehensive assessment of the model's understanding of event timelines. We probe this capability through questions targeting sequential ordering (e.g., "What occurred after action X?"), requiring holistic sequence verification (e.g., "Which option correctly lists the event order?"), and testing for concurrent action identification (e.g., "What was the left hand doing while the right hand performed the task?"). For the latter, CR focuses on the direct consequences of an action, asking the model to identify

the immediate outcome resulting from a specific event. This tests its ability to reason about the consequences of its own potential actions. In addition, similar to CIA, we introduce the *Counterfactual Relationship* (CRS) to ensure that the causal reasoning of the model is firmly grounded in the specific visual cues.

Intentional Understanding marks a significant shift from the how of an action to the why, which representing the model is capable of inferring the underlying goals and motivations that drive physical behavior, serving as a critical ability for an agent to formulate its sub-goals. To assess thorough reasoning and prevent models from simply selecting a plausible answer from a list, all tasks within this tier are exclusively formulated as open-ended questions.

We begin with *Functional Intention* (FI), a task that focuses on the immediate purpose behind a specific action or its particular manner of execution. FI includes two question types, "why performs a single atomic action" or "why adopt this specific physical technique". In contrast to this granular focus, the *Global Intention* (GI) task requires the model analyze a sequence of distinct and scattered actions to deduce the overall goal that connects. This challenge the model to understand how individual sub-tasks contribute to a high-level plan. Finally, the *Counterfactual Intention* (CIT) task provides the most rigorous test of the model's grasp of how intention drives behavior. It presents the model with a hypothetical change to the agent's overall goal and ask it to predict the resulting changes in actions.

Evaluative Judgment shifts the focus from comprehension to evaluation and is defined as the agent's ability to make qualitative assessments of an action's execution and outcome. This is a critical capability, as it is essential for both learning from observation by distinguishing effective from hazardous techniques, and for achieving resilience by identifying failures and re-planning accordingly. Since this reasoning requires justification, all tasks are open-ended.

Specifically, the first task is *Process Monitoring* (PM), which requires the model analyze the ongoing action to determine if the process is proceeding as expected, or if it has encountered a problem, such as being temporarily stuck, having failed, or being unexpectedly interrupted. Next, the *Strategy Evaluation* (SE) task requires the model to adopt the role of a critic, judging the quality of an action's execution based on criteria such as rationality, professionalism, and efficiency. Thus, it favors actions that are more likely to succeed in the specific scene. Finally, similar to the above preceding tiers, the *Counterfactual Evaluation* (CEU) task confronts the model with a hypothetical change on a specific physical interaction or event within the process and asks it to evaluate how that change would have impacted the quality of the action or its outcome.

3.2 Dataset Generation

Video Source. Driven by the need to support evaluation across four cognitive tiers, we construct a final corpus from five datasets, each selected for its specific contributions.

We prioritize the first-person perspective, which is essential for embodied intelligence, selecting 329 videos from EgoTaskQA [25], 191 from Charades-

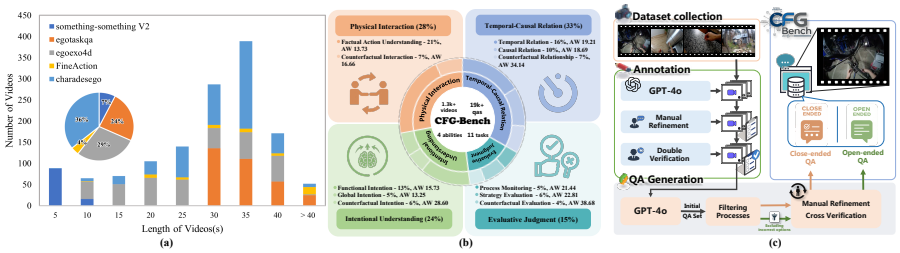

Fig. 3: (a) Distribution and video length statistics of five datasets. Video-length min/mean/max = 1.83/26.3/107.8s (b) The distribution of tasks across four tiers. AW means average words of questions. (c) Pipeline of dataset generation. Both annotation and QA generation are human-AI collaborative workflow. Open-ended and closed-ended questions share the same pipeline at the early stage.

Ego [40], and 394 from EgoExo4d [18]. These datasets are rich in daily, goal-oriented tasks, offering clear examples of physical interaction (Tier 1), distinct temporal workflows and intentions (Tiers 2 & 3), and scenarios for evaluation (Tier 4). To complement this view, 298 third-person videos from Charades-Ego [40] provide an objective perspective ideal for analyzing causal-temporal relation and evaluating strategy (Tiers 2 & 4). Besides, 104 videos from Something-Something-V2 [17] are included for their focus on fine-grained hand-object interactions, a key aspect of embodied manipulation, and 52 from FineAction [31] for its complex outdoor action diversity. Fig. 3 (a) shows the distribution of dataset categories and the statistics of video lengths.

Manual Annotation. To generate high-quality ground truth for our benchmark, we conduct a meticulous, month-long annotation process. As shown in Fig. 3 (c), our team of ten is organized into a two-stage pipeline that begin with draft annotations from GPT-4o [23]. In the first stage, eight expert annotators refine these initial drafts, after which two expert reviewers perform a final verification. The entire process is guided by our four-tiered cognitive framework, with specific annotation instructions tailored to the unique characteristics of each source dataset. The annotation guideline is shown in Appendix C4.

Close-Ended QA. With the curated videos and its detailed annotations, we proceed to generate the QA pairs. For FAU, TR, and CR in Tiers 1 and 2, we develop a rigorous multi-stage pipeline to generate close-ended QA pairs.

As shown in Fig. 3 (c), the process begins with an automated generation phase where we design distinct prompt templates for each sub-task. Using GPT-4o [23] (prompts are shown in Appendix C2), we generate an initial set of QA pairs with two principles: maximize question diversity across the videos and craft challenging distractors that are plausible but unique. We obtain 19,372 multiple-choice QAs pairs in this stage.

Then, an automated filtering stage designed to eliminate overly simple questions. Following FAVOR-Bench [46], we adopt the blind and single-frame filtering techniques, removing questions that could be answered with common sense

alone or from a single static image. To further refine the set, we provide Qwen3-Max [49] with the QAs and the corresponding video caption, which serves as a proxy for the ground truth. The model is tasked with flagging any QA pairs it deemed ambiguous or impossible to choose from, thereby efficiently identifying potentially flawed questions for further targeted human review. After this process, approximately 35% of the initially QA pairs are discarded.

The final stage is a manual verification. This process is conducted by six trained annotators. To guarantee the video-relevant of each question and the correctness and uniqueness of its answer, the team perform multiple rounds of validation. All QA pairs are cross-checked to resolve ambiguous references and reduce subjectivity. This human-in-the-loop process results in 9,165 QAs.

Open-Ended QA. For the inferential tasks within our benchmark, i.e., the counterfactuals in Tiers 1 and 2, and all questions in Tiers 3 and 4, we develop a human-centric crafting process to generate high-quality open-ended QA pairs. As shown in Fig. 3 (c), our methodology begins by first generating a dedicated set of multiple-choice questions, obtaining 10,397 QAs. We then only maintain the question and its corresponding correct answer to serve as a factually-grounded seed for each open-ended query. This seed set is then manually refined by our ten annotators, who would either rephrase the question to be more exploratory and elaborate on the answer, or, if the initial seed is inadequate, craft a completely new QA pair from scratch. This process ensures every question is rooted in a verifiable evidence and paired with a reference answer containing the rich detail. All final pairs are validated through expert cross-checking either.

Statistics. CFG-Bench comprises 19,562 QA pairs, evenly split between 9,165 close-ended and 10,397 open-ended questions distributed across our 11 tasks. As detailed in Fig. 3 (b), the average words (AW) of counterfactual questions is significantly longer, as these queries require more context to elicit reasoned responses. The figure also highlights a notable variance in the number of questions per task, since not every video naturally suitable for all 11 task types.

3.3 Evaluation

We evaluate MLLM performance on CFG-Bench using a hybrid methodology. Close-ended tasks are measured by standard accuracy. For open-ended tasks, we follow [46] and use a GPT-assisted protocol to score responses on two dimensions: Correctness, which measures factual accuracy against the video, and Detailedness, which assesses descriptive richness, with the final score being their mean. DeepSeek-R1 [19] is employed in GPT-assisted evaluation and the prompt is attached in Appendix C3. Notably, a strict gating mechanism applies to the counterfactual tasks in Tiers 1 and 2: a response is only eligible for scoring if it explicitly rejects or implicitly corrects the question's false premise. Failure to do so results in an automatic score of zero for both dimensions. This gating mechanism requires embodied agents to visually identify false premises before describing grounded reality, thus preventing blind compliance.

Table 2: Comprehensive results of leading proprietary and open-source MLLMs on CFG-Bench. It presents the performance of each task within every cognitive tier, along with the average score for open-ended questions $\mathbf{Avg}_o$ and the average accuracy for close-ended questions $\mathbf{Avg}_c$. Random selection and human performance are also included for comparison. The highest and suboptimal results are **bolded** and underlined.

Methods	Input	Avg_c	Avg_o	Interaction		Temporal-Causal			Intention			Evaluation		
				FAU	CIA	TR	CR	CRS	FI	GI	CIT	PM	SE	CEU
Human	–	95.85	9.05	93.48	8.76	95.97	98.10	9.02	8.83	9.51	9.11	8.92	8.63	9.62
Random	–	–	–	20	–	20	20	–	–	–	–	–	–	–
Proprietary MLLMs														
Gemini-2.5-Pro [10]	4 fps	59.52	5.40	58.81	4.58	54.81	64.95	4.55	5.18	6.15	5.31	5.72	5.19	6.49
Gemini-3-Pro [16]	4 fps	**65.60**	**5.67**	63.64	4.68	60.82	72.35	5.39	5.29	6.48	6.31	5.57	5.32	6.32
GPT-5 [37]	1 fps	55.44	4.13	54.71	2.61	46.53	65.09	3.22	4.98	4.64	3.16	4.55	4.82	5.05
Open-source MLLMs														
Video-LLaMA3-2B [53]	4 fps	43.23	3.78	51.21	2.14	34.64	43.83	2.80	4.77	4.82	3.24	4.19	4.07	4.21
Video-LLaMA3-7B [53]	4 fps	51.27	4.13	51.33	2.39	50.17	52.31	3.29	5.03	4.96	3.44	4.68	4.21	5.05
InternVL3-2B [56]	4 fps	41.02	4.03	53.46	2.61	33.90	35.70	3.13	4.05	4.73	3.12	4.62	4.83	5.15
InternVL3-8B [56]	4 fps	43.54	4.13	56.26	2.63	36.89	37.48	3.15	5.03	4.67	3.13	4.56	4.85	5.02
InternVL3-78B [56]	4 fps	52.82	4.92	58.31	3.95	49.42	50.74	3.68	5.41	5.04	4.80	4.96	5.27	6.21
Gemma-3-4B [45]	4 fps	48.49	3.73	44.43	2.46	46.44	54.59	2.86	4.61	2.44	3.24	4.22	4.51	5.46
Gemma-3-12B [45]	4 fps	50.05	4.07	47.87	2.53	47.94	54.33	2.95	5.20	2.51	4.36	4.60	4.69	5.74
Gemma-3-27B [45]	4 fps	52.96	4.51	53.74	3.59	47.72	57.42	3.15	5.31	3.41	4.74	4.82	5.01	6.06
Qwen2.5-VL-3B [4]	4 fps	37.46	4.12	41.69	1.88	34.77	35.91	3.27	5.39	4.28	3.55	4.50	4.82	5.23
Qwen2.5-VL-7B [4]	4 fps	40.71	4.38	43.93	2.71	38.58	39.63	3.32	5.50	4.53	3.85	4.57	4.94	5.62
Qwen2.5-VL-7B [4]	8 fps	42.40	4.39	45.16	2.72	39.51	42.52	3.33	5.67	4.55	3.79	4.64	4.94	5.49
Qwen2.5-VL-72B [4]	4 fps	51.28	4.61	56.32	3.56	47.83	49.68	3.41	5.49	4.51	4.34	4.81	5.01	5.77
Qwen3-VL-4B-Instruct [3]	4 fps	51.34	4.38	51.96	3.21	43.26	58.79	3.13	5.08	4.85	3.77	4.87	4.43	5.72
Qwen3-VL-8B-Instruct [3]	4 fps	51.07	4.71	53.60	3.15	44.76	54.86	3.39	5.49	5.25	4.12	5.31	5.00	5.95
Qwen3-VL-30B-A3B-Instruct [3]	4 fps	**57.68**	**5.09**	57.21	3.69	51.78	64.04	4.22	5.39	5.15	4.93	5.21	5.77	6.35
Cosmos-Reason1-7B [1]	4 fps	52.05	4.80	55.97	3.63	41.86	58.33	5.19	5.51	5.43	3.48	4.72	5.09	5.34
RoboBrain2.0-7B [44]	4 fps	44.53	4.72	45.96	2.13	42.20	45.44	3.51	5.76	5.21	4.88	5.30	5.37	5.56

4 Experiments

4.1 Experimental Setup

We benchmark a suite of leading MLLMs on CFG-Bench using their official implementations or APIs. Our evaluation includes proprietary models (Gemini-3-Pro [16], Gemini-2.5-Pro [10], GPT-5 [37]) and a range of open-source counterparts (Video-LLaMA3 [53], InternVL3 [56], Gemma-3 [45], Qwen2.5-VL [4], Qwen3-VL [49]). To specifically assess performance from an embodied perspective, we also include two specialized foundation models, RoboBrain2.0 [44] and Cosmos [1]. To ensure a fair comparison, video frames for each model are sampled according to their officially recommended frame-rate-based strategies.

4.2 Overall Performance

From Tab. 2, our comprehensive evaluation of MLLMs on CFG-Bench yields several key findings:

Human-Level Performance Gap. A significant performance gap exists between all MLLMs and human-level proficiency, a disparity evident in close-ended tasks, where models struggle to accurately describe physical execution details and temporal sequencing, and in open-ended counterfactual tasks, where they struggle with active verification and then correction. Even the top-rank model, Gemini-3-Pro, lags substantially behind the human baseline, underscoring the profound challenge of fine-grained action reasoning.

Proprietary versus Open-Source Models. Proprietary models generally outperform their open-source counterparts. Gemini-3-Pro leads with a score of 5.67 on open-ended tasks and 65.60% accuracy on close-ended ones, while other proprietary model like Gemini-2.5-Pro and GPT-5 also perform well. Among open-source models, Qwen3-VL-30B-A3B-Instruct stands out as a strong performer, achieving results competitive with Gemini-2.5-Pro on several tasks.

Scaling Laws of Model Size. Our results consistently demonstrate a positive correlation between model size and performance across several model families. Among the Qwen2.5-VL variants, for instance, performance scales directly with the parameter count. It is likely due to the enhanced capabilities of larger models for fine-grained visual perception and complex reasoning.

Impact of Model Iteration. Our results highlight that recent model iterations can yield substantial performance gains that surpass simple scaling laws. This is most evident when comparing the Qwen2.5 series with the newer Qwen3 generation. Despite having fewer parameters, Qwen3-VL-30B-A3B-Instruct consistently outperforms the much larger Qwen2.5-VL-72B model across several tasks.

Effect of Input Frame Rate (FPS). We find that the impact of input frame rate (FPS) on performance is limited. For example, increasing the FPS from 4 to 8 on Qwen2.5-VL-7B yields only marginal gains. More cross-model FPS analysis on Gemini-2.5-Pro and Qwen3-VL-8B-Instruct are supplemented in Appendix B1. This suggests that the performance bottleneck is not merely a lack of perceptual input, but likely largely involves the models' fundamental limitations in fine-grained physical understanding and higher-order reasoning.

Benefit of Embodied Fine-Tuning. Our results validate the hypothesis that models fine-tuned on specific embodied data excel at our tasks. Specifically, RoboBrain2.0-7B, which is fine-tuned from Qwen2.5-VL-7B, demonstrates superior average performance on both the open-ended and close-ended tasks than its base model. We attribute the gains largely to CFG-Bench-like knowledge. For example, long-horizon planning and trajectory prediction in RoboBrain2.0 align with Tiers 2&3; its closed-loop feedback and monitoring reflect Tier 4; and bounding box referring matches our focus on interacted object parts in Tier 1.

4.3 Performance Across Cognitive Tiers

Brittle Visual Grounding about Complex Actions. In Physical Interaction, models show a foundational but incomplete grounding capability. On close-ended FAU tasks, scores are stable but only around 50% accuracy, indicating that while models recognize static objects and simple motions, they fail to master full, fine-grained action sequences. This weakness is amplified in the open-ended CIA

task, where models frequently fail the gating mechanism by acquiescing to and hallucinating from false premises, revealing a brittle grounding.

Local Causality over Global Temporality. In Temporal-Causal Relation, models generally perform better on CR than on TR. We posit this is because CR tasks probe local, direct consequences that are well-learned patterns. In contrast, TR tasks demand more challenging global reasoning to track and compare multiple events across a video's context. This difficulty is further supported by the low performance on CRS primarily involved in complex temporal tracking.

Local Intent Inference over Global Goal Synthesis. In Intentional Understanding, models demonstrate a clear gap between local and global reasoning. They succeed at inferring the immediate intent of single actions in the FI task. However, they consistently struggle to synthesize long-term goals from scattered actions in the GI task, revealing a weakness in multi-step, abstract reasoning. This struggle culminates in the CIT task, where generating alternative plans for new goals remains a profound challenge.

Guided Evaluation over Unguided Critique. In Evaluative Judgment, models generally performed better on CEU than on PM and SE. This suggests that models are more capable of guided, local evaluations (predicting the effect of a given change) than they are of unguided, global critiques that require them to independently apply an internal model of skill or ideal task progression.

Insights for Improvement. These findings mirror key challenges in robotics manipulations to some extent. The failures in visual grounding and temporal reasoning (Tiers 1&2) align with challenges in complex dynamic task execution, while the struggle with goal synthesis (Tier 3) is a bottleneck for long-horizon planning. Crucially, the lack of unguided critique (Tier 4) explains why robots require human intervention for correction when manipulation fails.

Consequently, we advocate for leveraging the proposed CFG-Bench to focus training on fine-grained dynamics and temporal logic, ensuring visual grounding serves actionable execution. Furthermore, to foster the foresight required for long-horizon planning, models need supervision that explicitly links atomic actions to global goals, and imitating intent is as important as imitating trajectories. Finally, true autonomy requires failure detection and critique mechanisms, transforming MLLMs from passive imitators into self-correct agents. SFT and error analysis in Sec. 4.5 support these insights either.

4.4 Reliability Analysis

The Reliability of Annotations. To validate the quality and consistency of our ground truth, we conduct an Inter-Annotator agreement study with three annotators on the same 1,000 samples (500 closed-ended and 500 open-ended). To measure the degree of consensus, we calculate Krippendorff's α, selected for its robustness in handling multiple raters while correcting for chance agreement. For the closed-ended questions, we treat the option indices as nominal data. For the open-ended questions, we treat the 0-10 scores for both Correctness and Detailedness as interval data.

Table 3: The quantitative analysis of QA Forms. Left of "/" is close-ended performance and right is open-ended.

Methods	Interaction		Temporal-Causal			Intention			Evaluation		
	FAU	CIA	TR	CR	CRS	FI	GI	CIT	PM	SE	CEU
Gemini-2.5-Pro [10]	58.81/5.39	66.02/4.58	54.81/4.92	64.95/5.76	79.22/4.55	81.88/5.18	89.29/6.15	96.14/5.31	84.45/5.72	90.99/5.19	98.84/6.49
Qwen3-VL-8B-Instruct [3]	53.60/4.70	15.59/3.15	44.76/4.22	54.86/4.31	44.03/3.39	76.50/5.49	75.98/5.25	89.47/4.12	77.53/5.31	90.30/5.00	80.48/5.95
RoboBrain2.0-7B [44]	45.96/4.97	18.41/2.13	42.20/4.30	45.44/5.11	29.85/3.51	72.88/5.76	73.36/5.21	90.42/4.88	79.09/5.30	89.81/5.37	80.86/5.56

As a result, our analysis yields Krippendorff's α of 0.95 for close-ended task and 0.82/0.84 for open-ended task. According to established guidelines [27], a value above 0.67 indicates a high level of agreement and data reliability. This strong result confirms that our ground truth annotations are of high reliability.

The Reliability of GPT-Assisted Evaluation. We first conduct a Human-LLM agreement by randomly sampling 25 questions per open-ended task (8 tasks and 21 models, totaling 4,200 samples), scored by six experts using the same evaluation criteria. Similarly, we report Krippendorff's α of 0.74/0.75 for correctness and detailedness, thus confirming the reliability of GPT-Assisted evaluation. Visualizations in Appendix D further show strong Human-LLM agreement and confirm that low scores stem from limited visual understanding, not LLM noise. We then provide a sensitivity analysis showing that replacing Deepseek-R1 with Claude-Sonnet-4 [2] under same prompts yields a high Spearman Rank Correlation of 0.92 in model rankings, indicating robustness to the evaluator choice.

Overlap Analysis. For data usage, although there is no direct evidence that the evaluated models use CFG-Bench data, they certainly do not involve our novel fine-grained action intelligence, cannot reuse training knowledge, and thus performance probing remains unbiased. For model choice, we use GPT-4o for QA generation, Qwen3-Max for refinement, Deepseek-R1 for scoring. None overlap with the evaluated models in Tab. 2.

4.5 Analysis and Discussing

Effects of QA Forms. In Tab. 3, our analysis reveals that the choice of QA format is critical for evaluation. We observe that for tasks requiring significant reasoning, models often achieve high accuracy in the MCQ format, even with challenging distractors. Yet, they underperform substantially when the same queries are presented in an open-ended format. This suggests that the MCQ format allows models to leverage powerful textual reasoning to infer the most plausible option, often bypassing deep visual grounding. Conversely, for tasks grounded in objective facts, such as FAU, TR, and CR, we find the performance to be less sensitive to the QA format.

The same conclusion could be inferred from visualizations in Fig. 4. The model answers correctly under the MCQs format but fails in the open-ended setting. We attribute the gap to MCQs enabling textual reasoning to infer plausible options that bypass visual grounding required for open-ended synthesis, and offering explicit rejection cues for counterfactuals that prevent the hallucinated compliance seen in open-ended tasks. Thus, MCQs mask critical reasoning

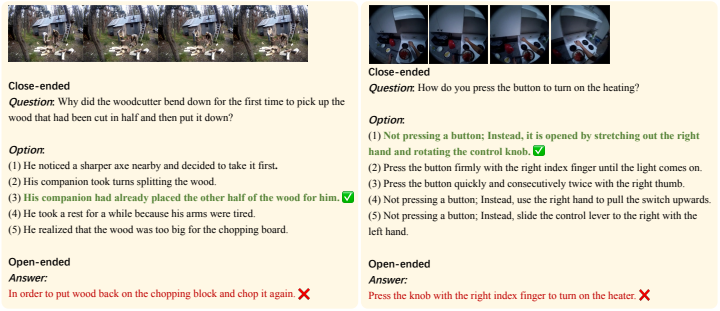

Fig. 4: The qualitative analysis of QA Forms.

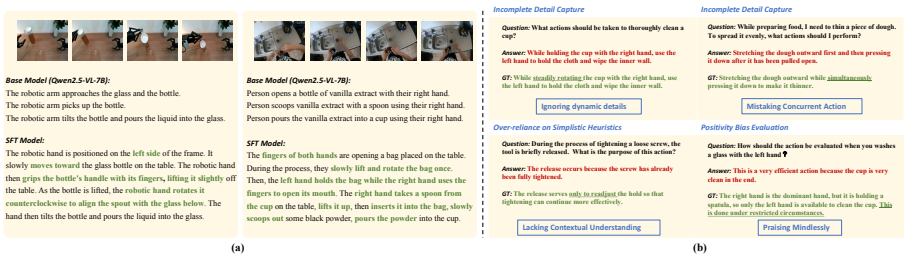

Fig. 5: (a) The results of caption generation before and after SFT on unseen videos of human and robotics operations. (b) Error Analysis. We show the recurring failure modes for each tier, along with the corresponding examples.

flaws and yield deceptively high performance, whereas open-ended evaluation is better suited for reasoning-heavy tasks.

Transfer Evaluation. To assess the transferability and practical utility of the knowledge contained in CFG-Bench, we perform supervised fine-tuning (SFT) on the Qwen2.5-VL-7B and InternVL3-8B using our data (SFT setup is supplemented in Appendix B3). We then evaluate this fine-tuned model on two downstream tasks from EMBODIEDBENCH [51]: the high-level planning task EB-ALFRED and the low-level control task EB-Manipulation. Notably, there is no data overlap between these tasks and CFG-Bench. In Tab. 4, the SFT model trained on CFG data shows significant performance gains on both tasks compared to its original baseline, whereas FAVOR data does not. This demonstrates that motion-level data have little impact on embodied tasks. Instead, the fine-grained knowledge of how and why an action is performed in CFG-Bench serves as a foundational capability that enhances performance across both embodied strategic planning and precise motor control tasks.

Beyond EMBODIEDBENCH, we evaluate on LIBERO [30], a standard VLA benchmark [48], using StarVLA-π [11] as baseline (reproduced under our setup for fair comparison): we first SFT Qwen3-VL-4B on CFG-Bench to strengthen its fine-grained action knowledge, then jointly train it and action expert under the same recipe as the baseline. Our model improves across three splits, i.e.,

Table 4: The quantitative results of transfer evalaution.

Methods	EB-ALFRED							EB-Manipulation					
	Avg	Base	Common	Complex	Visual	Spatial	Long	Avg	Base	Common	Complex	Visual	Spatial
Qwen2.5-VL-7B [4]	4.7	10	8	6	2	0	2	9.6	8.3	8.3	8.3	5.6	16.7
+ SFT on FAVOR Data	4.3	8	8	4	2	2	2	8.2	4.2	8.3	4.2	5.6	18.8
+ SFT on CFG Data	**9.7**	**16**	**16**	**8**	**8**	**4**	**6**	**15.3**	**12.5**	**16.7**	**14.6**	**7.9**	**22.9**
InternVL3-8B [56]	10.3	20	14	14	12	0	2	11.5	10.4	10.4	12.5	13.9	10.4
+ SFT on CFG Data	**14.0**	**26**	**18**	**16**	**18**	**2**	**4**	**15.8**	**14.6**	**12.5**	**12.5**	**17.4**	**18.8**

Spatial 98.2 → 98.2, Object 97.8 → **98.0**, Goal 98.6 → **99.4**, Long 94.4 → **95.2**, confirming that the fine-grained action benefits real VLA manipulation tasks.

Furthermore, we qualitatively compare captions from the base and SFT models on unseen videos of human and robot in Fig. 5 (a), which reveals three improvement mechanisms: Tier 1 shifts outputs from generic verbs (e.g., pick up) to precise executable physical details (e.g., grip the handle with fingers, then lift slightly), enhancing manipulation; Tiers 2&3 improve sub-task sequencing (supply details like rotate the bag and open its mouth), thereby boosting long-horizon planning; All tiers use counterfactual data to enforce visual verification, thus reducing hallucination (nonexistent objects like vanilla disappear after SFT).

Error Analysis. We conduct a systematic error analysis in Fig. 5 (b). Firstly, the most common failure in Tier 1 is the incomplete capture of fine-grained details. Models often identify a single salient event but fail to describe the full subtle actions. Secondly, in Tier 2, models frequently fail to comprehend coordinated, simultaneous actions, i.e., while able to describe the action of a single hand, they struggle to articulate what both hands are doing concurrently. Thirdly, for Tier 3, models often fall back on overly simplistic common-sense heuristics, leading to incorrect conclusions. For instance, a model might mistake an intermediate step, such as readjusting a grip, for the completion of a task merely because a hand releases an object. Finally, in Tier 4, we identify a positivity bias that the models tend to base their evaluation solely on the final outcome, praising any successful completion while ignoring issues in the process.

5 Conclusion

In this work, we introduce CFG-Bench to address the underexplored domain of fine-grained action intelligence for embodied agents. Built upon a four-tiered cognitive framework, our benchmark assesses a model's ability to translate visual observations into actionable physical and cognitive details through a diverse suite of question-answer pairs. Our evaluation reveals MLLMs' limitations in articulating fine-grained actions and in higher-order reasoning. We then prove these capabilities are foundational, as SFT on our data led to significant performance gains on downstream embodied manipulation and planning tasks.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (Grant Nos. U22A6001).

References

1. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025)
2. Anthropic: Claude sonnet 4 (2025)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14455–14465 (2024)
6. Chen, L., Zhang, Y., Ren, S., Zhao, H., Cai, Z., Wang, Y., Wang, P., Meng, X., Liu, T., Chang, B.: Pca-bench: Evaluating multimodal large language models in perception-cognition-action chain. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 1086–1104 (2024)
7. Chen, Y., Ge, Y., Ge, Y., Ding, M., Li, B., Wang, R., Xu, R., Shan, Y., Liu, X.: Egoplan-bench: Benchmarking multimodal large language models for human-level planning. International Journal of Computer Vision **134**(3), 118 (2026)
8. Cheng, S., Guo, Z., Wu, J., Fang, K., Li, P., Liu, H., Liu, Y.: Egothink: Evaluating first-person perspective thinking capability of vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14291–14302 (2024)
9. Cheng, Z., Tu, Y., Li, R., Dai, S., Hu, J., Hu, S., Li, J., Shi, Y., Yu, T., Chen, W., et al.: Embodiedeval: Evaluate multimodal llms as embodied agents. arXiv preprint arXiv:2501.11858 (2025)
10. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
11. Community, S.: Starvla: A lego-like codebase for vision-language-action model developing. arXiv preprint arXiv:2604.05014 (2026)
12. Cui, Z.J., Wang, Y., Shafiullah, N.M.M., Pinto, L.: From play to policy: Conditional behavior generation from uncurated robot data. arXiv preprint arXiv:2210.10047 (2022)
13. Dang, R., Yuan, Y., Zhang, W., Xin, Y., Zhang, B., Li, L., Wang, L., Zeng, Q., Li, X., Bing, L.: Ecbench: Can multi-modal foundation models understand the egocentric world? a holistic embodied cognition benchmark. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24593–24602 (2025)
14. Du, Y., Fan, T., Nan, K., Xie, R., Zhou, P., Li, X., Yang, J., Yang, Z., Tai, Y.: Motionsight: Boosting fine-grained motion understanding in multimodal llms. arXiv preprint arXiv:2506.01674 (2025)
15. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24108–24118 (2025)

16. Google: Gemini 3 pro: the frontier of vision ai. (2025)
17. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. In: Proceedings of the IEEE international conference on computer vision. pp. 5842–5850 (2017)
18. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: 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. pp. 19383–19400 (2024)
19. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
20. Han, X., Chen, S., Fu, Z., Feng, Z., Fan, L., An, D., Wang, C., Guo, L., Meng, W., Zhang, X., et al.: Multimodal fusion and vision-language models: A survey for robot vision. Information Fusion p. 103652 (2025)
21. He, Y., Huang, Y., Chen, G., Pei, B., Xu, J., Lu, T., Pang, J.: Egoexobench: A benchmark for first-and third-person view video understanding in mllms. Advances in Neural Information Processing Systems **38** (2026)
22. Hong, W., Cheng, Y., Yang, Z., Wang, W., Wang, L., Gu, X., Huang, S., Dong, Y., Tang, J.: Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8450–8460 (2025)
23. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
24. Ji, Y., Tan, H., Shi, J., Hao, X., Zhang, Y., Zhang, H., Wang, P., Zhao, M., Mu, Y., An, P., et al.: Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1724–1734 (2025)
25. Jia, B., Lei, T., Zhu, S.C., Huang, S.: Egotaskqa: Understanding human tasks in egocentric videos. Advances in Neural Information Processing Systems **35**, 3343–3360 (2022)
26. Jin, S., Xu, J., Lei, Y., Zhang, L.: Reasoning grasping via multimodal large language model. arXiv preprint arXiv:2402.06798 (2024)
27. Krippendorff, K.: Content analysis: An introduction to its methodology. Sage publications (2018)
28. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22195–22206 (2024)
29. Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., Florence, P., Zeng, A.: Code as policies: Language model programs for embodied control. arXiv preprint arXiv:2209.07753 (2022)
30. Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., Stone, P.: Libero: Benchmarking knowledge transfer for lifelong robot learning. Advances in Neural Information Processing Systems **36**, 44776–44791 (2023)
31. Liu, Y., Wang, L., Wang, Y., Ma, X., Qiao, Y.: Fineaction: A fine-grained video dataset for temporal action localization. IEEE transactions on image processing **31**, 6937–6950 (2022)

32. Luo, Y., Fan, C.K., Dong, M., Shi, J., Zhao, M., Zhang, B.W., Chi, C., Liu, J., Dai, G., Zhang, R., et al.: Robobench: A comprehensive evaluation benchmark for multimodal large language models as embodied brain. arXiv preprint arXiv:2510.17801 (2025)
33. Majumdar, A., Ajay, A., Zhang, X., Putta, P., Yenamandra, S., Henaff, M., Silwal, S., Mcvay, P., Maksymets, O., Arnaud, S., et al.: Openeqa: Embodied question answering in the era of foundation models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16488–16498 (2024)
34. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. Advances in Neural Information Processing Systems **36**, 46212–46244 (2023)
35. Mu, Y., Zhang, Q., Hu, M., Wang, W., Ding, M., Jin, J., Wang, B., Dai, J., Qiao, Y., Luo, P.: Embodiedgpt: Vision-language pre-training via embodied chain of thought. Advances in Neural Information Processing Systems **36**, 25081–25094 (2023)
36. Norman, D.: The design of everyday things: Revised and expanded edition. Basic books (2013)
37. OpenAI: Gpt-5 system card (2025)
38. Qi, Y., Zhao, H., Guo, Z., Ma, S., Chen, Z., Han, Y., Zhang, R., Lin, Z., Xin, S., Huang, Y., et al.: Bear: Benchmarking and enhancing multimodal language models for atomic embodied capabilities. arXiv preprint arXiv:2510.08759 (2025)
39. Shao, R., Li, W., Zhang, L., Zhang, R., Liu, Z., Chen, R., Nie, L.: Large vlm-based vision-language-action models for robotic manipulation: A survey. arXiv preprint arXiv:2508.13073 (2025)
40. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: European conference on computer vision. pp. 510–526. Springer (2016)
41. Singh, I., Blukis, V., Mousavian, A., Goyal, A., Xu, D., Tremblay, J., Fox, D., Thomason, J., Garg, A.: Progprompt: Generating situated robot task plans using large language models. arXiv preprint arXiv:2209.11302 (2022)
42. Song, C.H., Wu, J., Washington, C., Sadler, B.M., Chao, W.L., Su, Y.: Llm-planner: Few-shot grounded planning for embodied agents with large language models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2998–3009 (2023)
43. Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L., Wang, Y.X., Yang, Y., et al.: Aligning large multimodal models with factually augmented rlhf. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 13088–13110 (2024)
44. Team, B.R., Cao, M., Tan, H., Ji, Y., Chen, X., Lin, M., Li, Z., Cao, Z., Wang, P., Zhou, E., et al.: Robobrain 2.0 technical report. arXiv preprint arXiv:2507.02029 (2025)
45. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025)
46. Tu, C., Zhang, L., Chen, P., Ye, P., Zeng, X., Cheng, W., Yu, G., Chen, T.: Favorbench: A comprehensive benchmark for fine-grained video motion understanding. arXiv preprint arXiv:2503.14935 (2025)
47. Wang, C., Fan, L., Sun, J., Zhang, R., Fei-Fei, L., Xu, D., Zhu, Y., Anandkumar, A.: Mimicplay: Long-horizon imitation learning by watching human play. arXiv preprint arXiv:2302.12422 (2023)

48. Xu, C., Zhang, S., Liu, Y., Sun, B., Chen, W., Xu, B., Liu, Q., Wang, J., Wang, S., Luo, S., et al.: An anatomy of vision-language-action models: From modules to milestones and challenges. arXiv preprint arXiv:2512.11362 (2025)
49. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
50. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10632–10643 (2025)
51. Yang, R., Chen, H., Zhang, J., Zhao, M., Qian, C., Wang, K., Wang, Q., Koripella, T.V., Movahedi, M., Li, M., et al.: Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. arXiv preprint arXiv:2502.09560 (2025)
52. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9127–9134 (2019)
53. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., et al.: Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106 (2025)
54. Zhang, Z., Zhu, Z., Li, P., Wang, T., Liu, T., Ma, X., Chen, Y., Jia, B., Huang, S., Li, Q.: Task-oriented sequential grounding in 3d scenes (2024)
55. Zhou, J., Shu, Y., Zhao, B., Wu, B., Xiao, S., Yang, X., Xiong, Y., Zhang, B., Huang, T., Liu, Z.: Mlvu: A comprehensive benchmark for multi-task long video understanding. arXiv e-prints pp. arXiv–2406 (2024)
56. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)