

RoboBench: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models as Embodied Brain

Yulin Luo^{1*}, Chun-Kai Fan^{1*}, Menghang Dong^{1*}, Jiayu Shi^{1*}, Xiangju Mi^{1*},
Mengdi Zhao^{3*†}, Bo-Wen Zhang^{4*}, Cheng Chi^{2*†}, Jiaming Liu¹, Gaole Dai¹,
Rongyu Zhang¹, Ruichuan An¹, Kun Wu⁵, Zhengping Che⁵, Shaoxuan Xie²,
Guocai Yao², Zhongxia Zhao^{1,2}, Pengwei Wang², Guang Liu²,
Zhongyuan Wang², Tiejun Huang^{1,2}, and Shanghang Zhang^{1,2✉}

¹State Key Laboratory of Multimedia Information Processing, School of Computer
Science, Peking University ²Beijing Academy of Artificial Intelligence
³Institute for Brain and Intelligence, Fudan University ⁴University of Science and
Technology Beijing ⁵Beijing Innovation Center of Humanoid Robotics
yulin@stu.pku.edu.cn, shanghang@pku.edu.cn

Abstract. Building robots that can perceive, reason, and act in dynamic, unstructured environments remains a core challenge. Recent embodied systems often adopt a dual-system paradigm, where System 2 handles high-level reasoning while System 1 executes low-level control. In this work, we refer to System 2 as the *embodied brain*, emphasizing its role as the cognitive core for reasoning and decision-making in manipulation tasks. Given this role, systematic evaluation of the embodied brain is essential for advancing robotic intelligence. Yet existing benchmarks emphasize execution success, or, when targeting high-level reasoning, suffer from incomplete dimensions and limited task realism, offering only a partial picture of cognitive capability. To bridge this gap, we introduce RoboBench, a benchmark that systematically evaluates multimodal large language models (MLLMs) as embodied brains. Motivated by the distinct cognitive roles required across the full manipulation pipeline, RoboBench defines five dimensions—Instruction Comprehension, Perception Reasoning, Generalized Planning, Affordance Prediction, and Failure Analysis—spanning 14 capabilities, 25 tasks, and 6092 QA pairs. To ensure realism, we curate datasets across diverse embodiments, attribute-rich objects, multi-view scenes, and memory-driven navigation, drawing from large-scale real robotic data and in-house collection. For planning, RoboBench introduces an evaluation framework that uses an MLLM as a world simulator. It moves beyond symbolic matching to evaluate embodied feasibility by simulating whether predicted plans can achieve critical object-state changes under physical and visual constraints, enabling faithful assessment of long-horizon reasoning. Experiments on 18 state-of-the-art MLLMs reveal fundamental limitations: difficulties with implicit instruction comprehension, spatiotemporal reasoning, cross-scenario planning, fine-grained affordance understanding, and execution failure diagnosis.

*Equal contribution. [†]Project leader. [✉]Corresponding author.

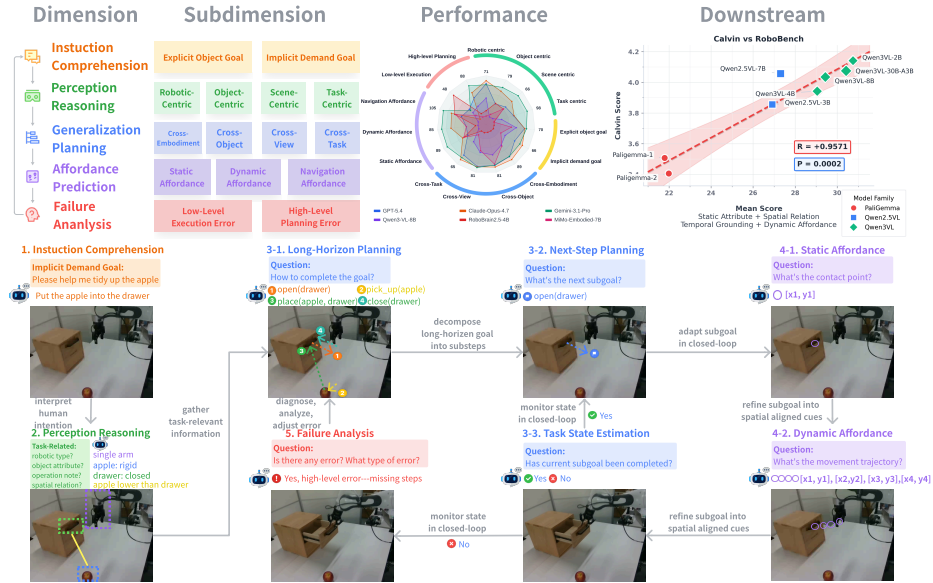


Fig. 1: Overview of RoboBench We evaluate MLLMs as embodied brains with 25 tasks color-coded by 5 dimensions (**top left**). These dimensions follow the embodied execution pipeline (**bottom**)—from intent understanding to failure diagnosis—capturing the core cognitive roles of System 2. Performance comparison (**top middle**) reveals significant gaps among state-of-the-art MLLMs (**top right**) RoboBench scores strongly correlate with downstream VLA performance on CALVIN.

sis. We further analyze how embodied cognitive abilities relate to downstream robotic control. RoboBench provides a comprehensive scaffold to quantify high-level cognition, clarify the role of the embodied brain, and guide the development of next-generation MLLMs for more robust robotic intelligence. Project page: <https://robo-bench.github.io>.

Keywords: Embodied AI · Multimodal Large Language Models · Vision-Language-Action Models · Benchmark

1 Introduction

Manipulation in dynamic, unstructured environments remains a core challenge for building generalizable embodied intelligence [14, 37, 48]. Such tasks demand not only precise motor control, but also high-level cognition: understanding instructions, perceiving the surroundings, formulating long-horizon plans, inferring affordances, and reflecting on failures [7, 57]. In this context, multimodal large language models (MLLMs) have shown strong potential for these roles, especially in instruction-following, commonsense reasoning, and general planning [27, 51]. To leverage these capabilities, recent embodied systems integrate MLLMs through

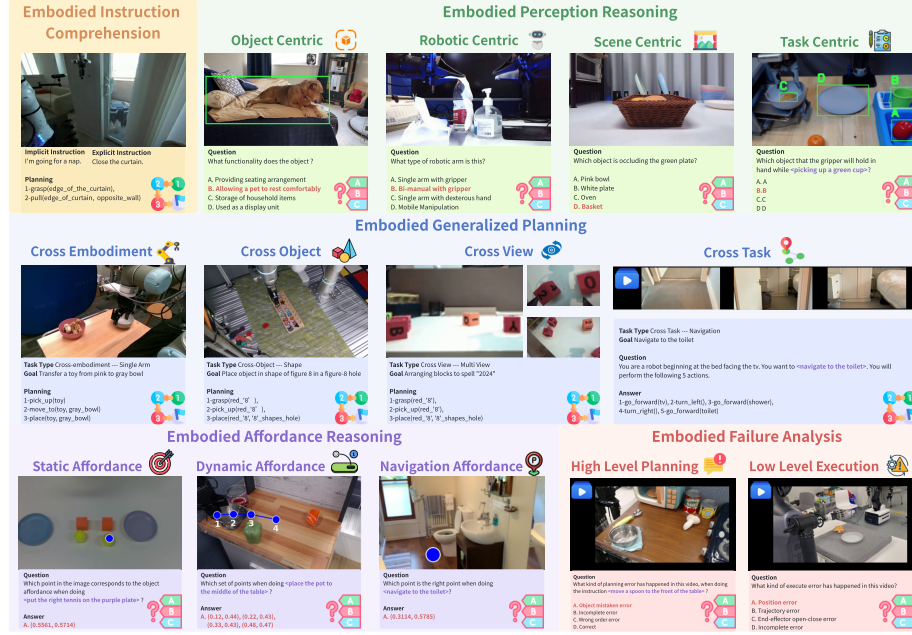


Fig. 2: Examples of RoboBench Our benchmark covers 5 dimensions, 14 capabilities, and 25 tasks. We design diverse question formats, including multiple-choice, planning, and point prediction.

a dual-system design [2, 3, 18], where System 2 performs high-level reasoning and System 1 handles low-level control [5, 35]. In vision-language-action (VLA) models, MLLMs are fine-tuned as backbones [3], while in multi-agent frameworks, they serve as high-level planners guiding specialized executors [18, 22, 23, 25]. In this work, we call System 2 the *embodied brain*, the cognitive core of reasoning and decision-making.

Given this design, systematic evaluation of the embodied brain is essential, yet current benchmarks remain inadequate. Most existing efforts focus narrowly on overall task success, offering little insight into underlying reasoning [8, 24, 33].

Even benchmarks that explicitly target embodied cognition exhibit three major shortcomings: (1) fragmented coverage of cognitive abilities, typically isolating perception [40], planning [9, 30, 44], or error reflection [19, 39] rather than assessing them as an integrated whole; (2) limited task realism and complexity, relying heavily on simulations [12, 45, 46, 54, 57], or ignoring practical challenges such as diverse embodiment, object properties, and occlusion; and (3) simplistic planning evaluation, often reduced to multiple-choice [1, 9, 50] or text similarity metrics such as BLEU [44] and LLM scoring [13], failing to capture embodied priors such as skill-object dependencies, execution order flexibility, and feasibility.

To overcome these gaps, we introduce **RoboBench**, a benchmark systematically designed to evaluate MLLMs as the cognitive core for robotic manipulation. Specifically, we highlight three key features of RoboBench: (1) **Comprehen-**

sive evaluation dimensions. RoboBench defines five dimensions—Instruction Comprehension, Perception Reasoning, Generalized Planning, Affordance Prediction, and Failure Analysis—to capture the key capabilities required for embodied cognition. These dimensions follow the manipulation execution pipeline (Fig. 1): interpreting human intent, perceiving task-relevant context, decomposing long-horizon goals, grounding subgoals into spatial affordances, and diagnosing failures during execution [19, 25, 34, 49, 57]. (2) **Realistic and diverse tasks.** RoboBench covers diverse embodied scenarios including single-arm, dual-arm, and mobile manipulation, objects with rich physical attributes, and multi-view scenes with occlusion and navigation memory [53]. The dataset combines large-scale real-world robotic data [4, 15, 20, 37, 52] with curated in-house collection to better reflect real-world embodied interaction. (3) **World-simulation rollout for planning evaluation.** We propose an *MLLM-as-world-simulator* framework for evaluating long-horizon planning. Given the initial scene, reference plan, and a human-annotated DAG capturing task dependencies, the evaluator simulates plan execution step-by-step under visual and physical constraints, assessing whether predicted plans can achieve key state milestones and remain physically executable.

We evaluate 18 state-of-the-art MLLMs and uncover fundamental limitations: poor grounding of implicit instructions, fragile embodiment-specific and spatiotemporal perception, difficulty in long-horizon and cross-scenario planning, shallow fine-grained affordance understanding, and weak execution failure diagnosis. These findings show that while MLLMs hold promise as embodied brains, their reasoning remains superficial. We further correlate RoboBench scores with downstream VLA benchmark performance and find that abilities such as perception reasoning and affordance prediction strongly predict manipulation success. This establishes RoboBench as a principled framework for selecting VLM backbones and guiding the development of stronger embodied reasoning.

In summary, our contributions are as below:

- We introduce RoboBench, a comprehensive benchmark for evaluating MLLMs as embodied brains across 5 key dimensions, spanning 14 capabilities, 25 tasks, and 6092 QA pairs.
- We design realistic datasets across embodiments, objects, and viewpoints that reflect real-world embodied interaction.
- We propose an MLLM-as-world-simulator evaluation framework that faithfully assesses long-horizon planning by simulating whether plans can achieve critical object-state milestones.
- We conduct a large-scale evaluation of SOTA MLLMs, revealing key embodied cognitive limitations and demonstrating that RoboBench abilities predict downstream VLA performance and guide VLM selection.

2 Related Work

With the rapid development of embodied AI, the number of benchmarks has also grown [9, 10, 17, 29, 43, 44, 54]. However, current benchmarks still have significant

Table 1: Comparison of various benchmarks for embodied AI, evaluating their features across comprehensiveness, reality, complexity, and planning evaluation methodology.

Benchmark	Comprehensiveness					Reality		Complexity		Planning Evaluation method	Size
	Language Instruction	Perception Reasoning	Generalized Planning	Affordance Prediction	Failure Analysis	Real Scene	Real Robot	Cross Embodiment	Multi View		
RoboVQA [44]	✓	×	✓	×	×	✓	✓	✓	×	BLEU	18,248
EgoPlanBench [9]	×	✓	✓	×	×	✓	×	×	×	Multi-Choice	3,355
MMRo [29]	×	✓	✓	×	×	✓	✓	×	×	LLM Score	26,175
EAI [30]	×	✓	✓	×	×	×	×	×	×	Simulator Task SR	438
OpenEQA [40]	×	✓	✓	×	×	✓	×	×	×	LLM Score	1,636
EgoThink [11]	×	✓	✓	×	×	✓	×	×	×	LLM Score	700
VideoEgoThink [10]	×	✓	✓	×	×	✓	×	×	×	LLM Score	4,993
EmbodiedEval [12]	×	✓	✓	×	✓	×	×	×	×	Multi-Choice	327
EmbodiedBench [54]	×	✓	✓	×	✓	×	×	✓	×	Simulator Task SR	1,128
VLABench [57]	✓	✓	✓	×	×	×	✓	✓	✓	Simulator Task SR	2,427
BEAR [43]	×	✓	✓	✓	×	✓	×	✓	✓	Multi-Choice	4,469
ECBench [17]	×	✓	×	×	×	✓	×	×	×	LLM Score	4,324
RoboBench (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	MLLM World Simulator	6,092

limitations in evaluating the capabilities of embodied intelligence. We analyze these limitations from four aspects: comprehensiveness, realism, complexity, and evaluation of planning tasks.

Existing benchmarks related to embodied cognition tend to assess individual abilities separately — for example, perception [11, 40], planning [9, 30, 44], and affordance prediction [55, 58] or failure analysis [19, 39, 59]. None of them comprehensively evaluate Multimodal Large Language Models (MLLMs) as integrated embodied brains. In terms of realism, some benchmarks rely on simulators, incurring a sim-to-real gap [12, 54, 57], while others build on human daily-activity or egocentric video [9, 10, 17] that still differs from real robotic manipulation. Regarding complexity, many benchmarks fail to account for cross-embodiment and cross-viewpoint scenarios [9, 29, 40]. For planning evaluation, existing metrics are also limited. Text-similarity metrics such as BLEU [44] capture only surface-level wording, ignoring the structural and physical feasibility of plans. Multiple-choice evaluations [9] deviate from real-world tasks whose solutions are open-ended. LLM-based judgments [13] introduce subjectivity, often failing to credit logically valid plans phrased differently from the reference.

Closest to our work, ECBench [17] and BEAR [43] pursue broad-coverage embodied evaluation, yet neither covers the full embodied-brain loop: ECBench builds on egocentric daily-activity video and omits generalized planning, affordance, and failure analysis, while BEAR—the most comprehensive prior effort—still diagnoses atomic skills in isolation, omits execution failure analysis, and adopts multiple-choice scoring that constrains open-ended planning. In contrast, RoboBench unifies all five embodied-cognition dimensions on real-robot data with an MLLM world-simulator judge for open-ended planning. A comprehensive comparison between our benchmark and prior works is presented in Table 1.

3 RoboBench

3.1 Core Capabilities

Through a detailed analysis, we identify five critical dimensions of the MLLM-based embodied brain, each aligned with the task execution pipeline: understand-

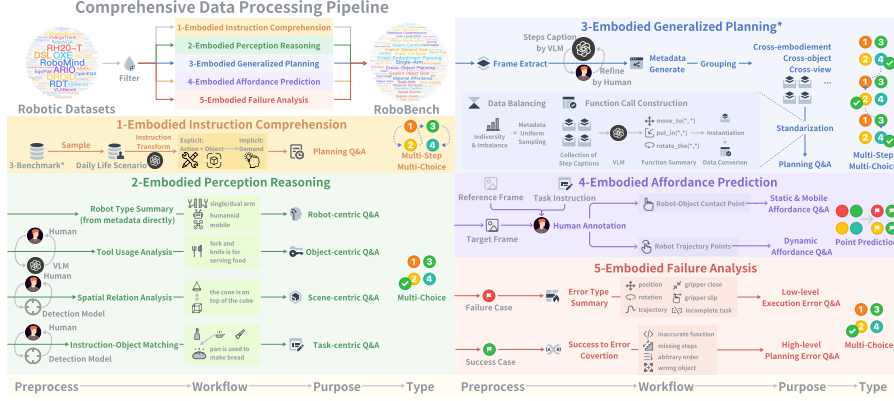


Fig. 3: Dataset Construction Pipeline. RoboBench customizes data workflows for each dimension: orange for instruction, green for perception, blue for planning, purple for affordance, and red for reflection. Each workflow generally follows three stages—preprocessing; annotation with VLMs, detection models, or human experts; and Q&A format generation.

ing human intent, perceiving the environment, formulating and adapting plans, refining actions via affordance prediction, and diagnosing failures. RoboBench evaluates these capabilities to uncover bottlenecks in embodied cognition.

Embodied Instruction Comprehension: *Can the embodied brain understand human intention?* Most embodied tasks rely on explicit instructions where actions and targets are clearly specified [31, 44, 54]. In contrast, real-world instructions are often implicit (e.g., “I’m thirsty” instead of “Retrieve a drink”) [57]. This dimension evaluates whether models can translate both explicit and implicit instructions into actionable plans.

Embodied Perception Reasoning: *Can the embodied brain perceive the environment to gather task-relevant information?* Reliable planning and execution rely on accurate perception [6, 29, 40, 47], RoboBench evaluates this through four aspects: Robotic-centric considers embodiment type and viewpoint. Object-centric examines static and functional attributes. Scene-centric evaluates spatial relations, temporal grounding, and causality analysis. Task-centric assesses the identification of instruction-relevant objects.

Embodied Generalized Planning: *Can the embodied brain generalize planning across embodiments, objects, views, and tasks?* Planning begins with decomposing long-horizon goals into subgoals, and continues during execution through predicting the next subtask, monitoring completion, and adapting subsequent steps [44, 54]. We evaluate generalized planning across four aspects: embodiments (single-arm, dual-arm, mobile manipulator, and humanoid robot), objects (material affordance, physical attribute, and world knowledge), views (multi-view integration under occlusion), and tasks (navigation planning using spatial cues from videos [53]).

Embodied Affordance Prediction: *Can the embodied brain refine sub-task plans through spatial affordances?* Beyond high-level planning, each subgoal

should be transformed into spatial cues to guide low-level execution [26, 28, 34, 50, 55, 58]. Affordance prediction links subgoals with the embodiment, objects, and environment, enabling System 2 to instruct System 1 more effectively [28]. RoboBench evaluates three types: Static affordance identifies contact points (e.g., grasp an apple); Dynamic affordance predicts motion trajectories (e.g., open a drawer); Navigation affordance determines robot base position (e.g., approach a microwave on a distant table).

Embodied Failure Analysis: *Can the embodied brain detect and correct failures?* Open-world manipulation inevitably introduces errors, requiring the brain not only to identify them but also to diagnose causes and suggest corrections [19, 39]. RoboBench evaluates low-level execution errors (e.g., position misalignment, trajectory deviations, gripper failures, and incomplete actions), and high-level planning errors (e.g., wrong object, missing steps, incorrect ordering), providing insights for more robust and generalizable execution.

3.2 Benchmark Construction

Dataset Collection and Processing Pipeline To evaluate the five cognitive dimensions, we build datasets from open-source and in-house robotic data, enriched with MLLM- and human-provided annotations, aligning each sub-benchmark with the embodied task pipeline for realism. The examples and construction pipeline are shown in Figures 2 and 3.

Instruction Comprehension: Evaluated through planning tasks, this dimension adopts a paired explicit-implicit design. Explicit instructions are drawn from daily-life scenarios with clearly specified actions and targets, whereas implicit instructions are generated by an MLLM conditioned on both the scene image and the explicit goal, then manually checked. A sampled audit confirms the generated implicit instructions are natural and inferable from the scene (scores in the supplementary), testing a model’s ability to interpret human intentions.

Perception Reasoning: Accurate perception is essential for reliable planning and execution. We construct datasets across four aspects: robotic-centric, using real robot data with type and view metadata for template-based QA; object-centric, combining curated static attributes [21] with GPT-generated functional properties and distractors; scene-centric, leveraging Gemini-segmented video steps for temporal grounding, together with manual annotations of relative positions and keypoint changes for spatial and causal reasoning; and task-centric, with human-labeled bounding boxes linking long-horizon instructions to target objects. All data are standardized into multiple-choice QA.

Generalized Planning: We build a planning pool from high-quality robotic videos [15, 20, 32, 36, 37, 42, 52, 53], as standardized inputs. Gemini [16] generates structured annotations—task summaries, step-wise instructions with timestamps, and metadata (objects, actions, scenes, embodiments)—which are refined by human annotators. Each step is then mapped into function templates (e.g., `pick_up(object)`, `move_to(object, target)`) grouped into manipulation or navigation skill lists [10] for structured plan generation. Evaluation covers three types [44]: (1) Q1: long-horizon planning, predicting the full action sequence

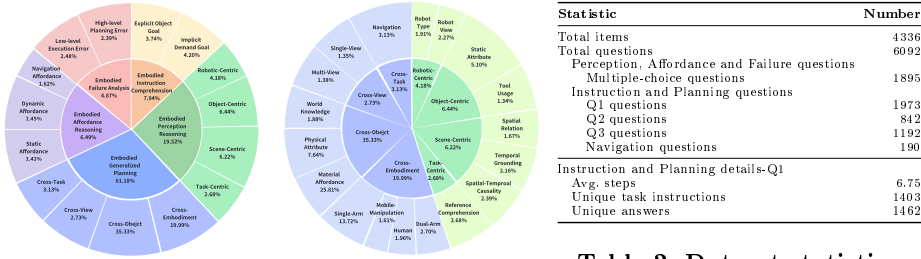


Fig. 4: Dimension Distribution of RoboBench.

Table 2: Dataset statistics.

from the first frame; (2) Q2: next-step planning, forecasting the (n+1)-th step; and (3) Q3: task state estimation, deciding whether a subtask is completed.

Affordance Prediction: Affordances refine high-level subgoals into spatial cues for low-level execution [34]. From the planning pool, representative frames are sampled and annotated with three types of affordance: static (contact points), dynamic (motion trajectories), and mobile (base positions). Each annotated affordance is then cast into a multiple-choice question in which the model selects, given the task instruction, the correct contact point, motion trajectory, or base position from scene-grounded candidate options.

Failure Analysis: It evaluates whether models can detect and reason about errors during execution [19, 39]. Execution-level failures (e.g., position misalignment, trajectory deviations, gripper errors, premature release, and trajectory overshoot) are collected from real VLA rollouts in RoboMIND [52] and labeled by experts, so they reflect actual low-level execution failures rather than perturbations of successful trajectories. Because real high-level planning-failure data is scarce, planning-level failures follow the REFLECT-style protocol [39]: we perturb correct plans with canonical classes including wrong object, missing step, wrong order, and wrong condition, then manually verify that the injected failure is identifiable and task-relevant.

Quality Control We adopt a two-stage quality control process to ensure benchmark quality: data filtering during construction and post-construction validation. In the construction stage, we apply both general and task-specific filters. General criteria address image quality and task validity. Sub-benchmarks also have tailored rules, such as excluding single-arm data from dual-arm tasks. After construction, professional researchers validate linguistic clarity, answerability, and correctness [38]. Every MLLM-assisted annotation receives at least one human cross-pass, and validation disagreements are tracked per dimension before adjudication; residual disagreement on the validation subset is below 4%. We further adopt a majority-vote strategy [38]: items all models solve are removed, and items all models fail undergo manual review.

Dataset Statistics RoboBench consists of 6092 samples and 4336 unique items, providing a balanced mix of diversity and complexity to evaluate embodied

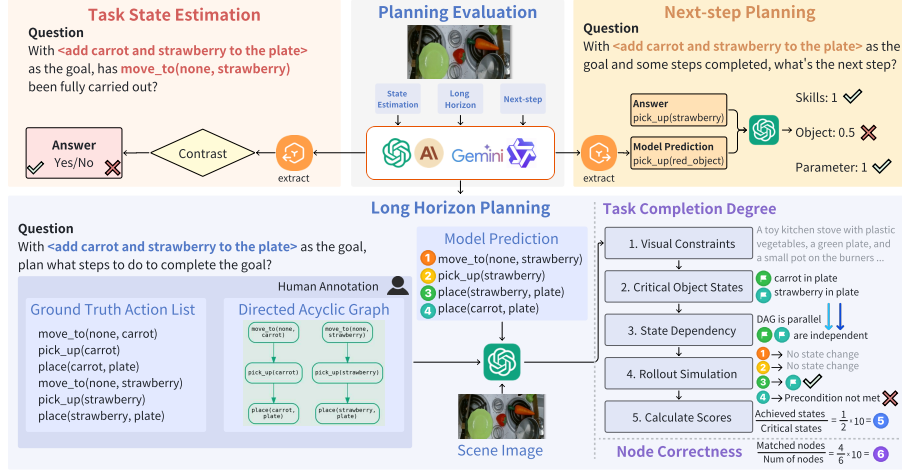


Fig. 5: Planning Evaluation Pipeline. The planning benchmark includes three task types: long-horizon planning, next-step planning, and task state estimation. They are evaluated respectively with our proposed MLLM-as-world-simulator framework, LLM scoring, and binary accuracy.

brain capabilities. It covers 5 dimensions, 14 capabilities, and 25 tasks, ensuring comprehensive and challenging assessments. The detailed distribution and per-dimension question statistics are shown in Figure 4 and Table 2.

4 Evaluation Metrics

To capture the diverse cognitive demands of embodied manipulation, we tailor metrics to each dimension. Perception reasoning, affordance prediction, and failure analysis are evaluated by multiple-choice accuracy [1, 50]. Planning is assessed across three tasks (Q1–Q3). Q1 serves as the core, leveraging an MLLM world simulator to evaluate both structural correctness and embodied feasibility.

Q1 – Long-horizon planning. We model a task as a partially ordered set of *atomic actions* parameterized by $\langle \text{skill}, \text{object}, \text{args} \rangle$. Inter-step constraints are encoded by a Directed Acyclic Graph (DAG) $G = (V, E)$, where each $v \in V$ is an action node and each edge $(u \rightarrow v) \in E$ enforces precedence. The DAG exposes milestones (e.g., **drawer=open**) while allowing valid permutations. Manually annotated G serves as the reference for evaluation. The pipeline is in Figure 5.

We calculate two metrics—*NodeCorrectness* and *TaskCompletion*—both computed via an MLLM. The final score is $\text{LongHorizon} = (\text{NodeCorrectness} + \text{TaskCompletion})/20 \in [0, 1]$.

Node correctness. Let V^* denote ground-truth nodes and $\hat{V}$ predicted nodes. An MLLM performs one-to-one matching under exact alignment on $\{\text{skill}, \text{object}, \text{parameter}\}$. The calculation is: $\text{NodeCorrectness} = \left\lfloor \frac{|V^* \cap \hat{V}|}{|V^*|} \times 10 \right\rfloor$.

Task completion. An MLLM conducts a lightweight world-simulation rollout using: (i) first-frame image I_0 for spatial/physical priors, (ii) reference action list A^* , and (iii) DAG G . We define *critical states* as object-level state changes (e.g., `drawer=open`, `apple on plate`). Let S^* be the set of such milestones and $\hat{S}$ those achieved during rollout. The MLLM proceeds as follows:

1. **Visual constraint analysis:** from I_0 , identify objects, spatial relations, and physical constraints to form the initial world W_0 (e.g., `drawer=closed`, `apple on table`).
2. **Critical object-state detection:** parse A^* and $G = (V, E)$ to extract state-transition predicates from action effects, then aggregate into S^* .
3. **State order & concurrency validation:** use G to enforce precedence and allowable parallelism among states in S^* , and establish causal links.
4. **Rollout simulation:** execute predicted actions step-by-step, checking preconditions against W_t , updating to W_{t+1} , and marking $s \in S^*$ as *achieved* $\mathcal{E}$ *protected* when it becomes true and remains valid to its last consumer, accumulating $\hat{S} \subseteq S^*$.

The score is computed as $\text{TaskCompletion} = \left\lfloor \frac{|\hat{S}|}{|S^*|} \times 10 \right\rfloor$. Across four vendor judges and three prompt variants on 200 stratified Q1 items, this protocol attains high inter-rater agreement (NodeCorrectness ICC(2,k)=0.948), confirming the evaluator is not tied to a single grader or prompt; the full robustness analysis is provided in the supplementary material.

Q2 – Next-step planning. Given an observation I_t , the model predicts the next action $\hat{a}_{t+1}$. An MLLM evaluates the prediction against the ground truth a_{t+1}^* : skill must match exactly ($s_{\text{skill}} \in \{0, 1\}$), while object and parameters are scored by visual reasonableness ($s_{\text{obj}}, s_{\text{param}} \in \{0, 0.5, 1\}$). The per-sample score is $\text{NextStep} = (s_{\text{skill}} + s_{\text{obj}} + s_{\text{param}})/3$.

Q3 – Task state estimation. Given I_{t-1} or I_t , the model predicts whether the subtask $\hat{a}_t$ has been achieved. Performance is measured using binary accuracy against the ground-truth task state.

5 Experiment

We evaluate several MLLMs on RoboBench, including closed-source MLLMs, open-source multi-image MLLMs, and open-source embodied MLLMs. Detailed descriptions of the models are provided in the supplementary materials. As additional references, we also report a text-only LLM baseline and a human upper bound; the full human-evaluation protocol is detailed in the supplementary materials, and we validate that our automated evaluator aligns with these human judgments in the In-Depth Analysis (Sec. 5.3).

5.1 Overall Results

Frontier MLLMs already answer embodied questions with reasonable quality: Among all evaluated models, Gemini-3.1-Pro shows the most consistent

Table 3: Results on Perception Reasoning(%) Attr. stands for Attribute; Temp. stands for Temporal; Refer. Compreh. stands for Reference Comprehension.

Model	Perception Reasoning								Avg
	Robotic-centric		Object-centric		Scene-centric		Task-centric		
	Robot-type	Robot-view	Static Attr.	Functional Attr.	Spatial Relation	Temp. Grounding	Causality	Refer. Compreh.	
Basic Reference									
Human Evaluation	80.67	79.08	43.77	83.89	70.91	51.61	91.22	93.22	74.30
GPT-5.4-ext-only	25.86	28.26	8.81	45.57	32.67	22.90	34.48	18.40	27.12
Closed-Source MLLMs									
GPT-5.4	73.28	50.00	42.86	73.42	54.46	38.93	45.52	71.17	56.20
GPT-5.2	68.10	39.86	38.60	77.22	47.52	30.53	54.48	71.17	53.44
GPT-5	64.66	47.10	49.24	69.62	54.46	48.09	74.48	78.53	60.77
GPT-4.1	66.38	50.00	40.43	68.35	47.52	22.14	56.55	73.01	53.05
GPT-4o	75.00	39.13	18.24	60.76	49.50	22.14	43.45	55.21	45.43
Claude-Opus-4.7	76.72	53.62	57.14	81.01	48.51	46.56	65.52	71.78	62.61
Claude-Sonnet-4.6	53.45	47.83	53.80	69.62	52.48	29.01	57.24	69.33	54.09
Claude-Sonnet-4.5	46.55	33.33	37.08	72.15	48.51	33.59	51.72	36.81	44.97
Claude-Haiku-4.5	44.83	33.33	30.70	56.96	25.74	22.14	45.52	27.61	35.85
Gemini-3.1-Pro	71.55	49.28	66.26	78.48	61.90	31.93	88.97	90.18	67.32
Gemini-2.5-Pro	67.24	43.48	57.14	82.28	57.43	52.07	73.10	80.37	64.14
Gemini-2.5-Flash	66.38	34.78	57.75	74.68	55.45	34.92	75.17	76.69	59.48
Open-Source Multi-Image MLLMs									
Qwen-3-VL-8B	52.59	36.96	27.66	65.82	36.63	25.95	31.72	54.60	41.49
Qwen-2.5-VL-7B-Ins	37.07	23.19	24.32	56.96	26.73	22.14	33.10	34.36	32.23
LLaVA-OneVision-7B	31.03	26.81	39.21	68.35	42.57	18.32	33.79	50.92	38.88
Embodied MLLMs									
RoboBrain-2.0-7B	31.90	19.57	28.57	44.30	34.65	21.37	24.83	33.13	29.79
RoboBrain-2.5-4B	35.34	24.64	39.82	77.22	53.47	18.32	59.31	44.17	44.04
MiMo-Embodied-7B	25.86	21.74	32.22	65.82	49.50	24.43	57.93	43.56	40.13

advantages across the perception, affordance, and failure dimensions, clearly outperforming both other closed-source and open-source MLLMs (e.g., 67.3 perception average vs. the next-best 64.1). Notably, the strongest MLLMs already answer many embodied questions with reasonable quality—reaching 85.7 on affordance prediction and 67.3 on perception reasoning—indicating a solid foundation of embodied cognition, with only a modest residual gap to the human reference. In contrast, most other MLLMs remain highly uneven or generally weak.

Closed-source models outperform open-source models: As summarized in Tabs. 3 and 4, closed-source models maintain a clear lead in four of the five dimensions, with performance margins often reaching 10–15%. Only in perception reasoning do open-source models approach comparable results, while gaps in instruction comprehension, planning, affordance reasoning, and failure analysis remain substantial. Additionally, we find that, within the same model family, performance improves consistently as model size increases.

5.2 Fine-grained Results across Tasks

We further break down results across tasks, highlighting distinct weaknesses in instruction comprehension, perception, planning, and failure analysis.

Implicit vs. Explicit Instruction Comprehension: Models perform significantly worse on implicit instructions compared to explicit ones, even when images and ground truth remain the same: the strongest model drops from 79.9 on explicit goals to 61.4 on implicit ones, and the gap widens further for weaker MLLMs. This reveals a clear weakness in grounding indirect human demands into actionable goals, likely because current MLLMs fail to fully integrate scene context when inferring task intent. Future models must jointly reason over language, perception, and context to infer and execute human intent.

Table 4: Results on Instruction Comprehension and Generalized Planning Q1.

Model	Instruction Comprehension			Generalized Planning												
	Explicit	Implicit	Avg	Cross-Embodiment Planning			Cross-Object Planning				Cross-View Planning		Cross-Task Planning		Avg	
				Single-arm	Dual-arm	Mobile-manip.	Human	Material Afford.	Physical Attr.	World Knowl.	Multi	Single	Navigation Plan.			
Basic Reference																
Human Evaluation	59.94	81.33	69.54	72.50	41.93	41.33	62.28	56.74	58.98	49.36	52.82	51.59	45.23	54.38		
GPT-4o Creativity	74.88	39.24	56.71	83.53	66.87	74.21	62.95	76.93	89.36	66.95	71.33	72.41	47.86	73.74		
Cloud-Source MLLMs																
GPT-4o	74.58	59.88	62.69	85.56	70.28	75.59	56.26	78.64	64.83	72.44	73.33	72.58	53.92	70.91		
GPT-4o	75.90	48.85	62.38	86.92	70.80	88.28	60.88	81.60	64.82	75.49	75.84	71.22	55.93	72.31		
GPT-4o	77.63	54.71	66.17	84.25	69.48	81.27	70.13	81.58	58.88	71.95	70.95	72.76	58.29	71.84		
GPT-4o	76.23	57.30	66.77	88.98	68.26	78.37	55.38	81.37	65.26	64.88	73.70	71.83	52.94	71.79		
GPT-4o	74.22	54.90	64.56	86.92	66.40	76.44	62.21	80.86	72.83	66.24	66.50	65.63	57.62	73.30		
Claude-3.5 Sonnet	73.92	61.52	67.72	89.43	71.98	84.30	68.53	84.56	67.70	70.80	75.86	72.86	58.53	75.54		
Claude-3.5 Sonnet	79.94	61.28	70.66	88.89	73.93	84.42	66.39	84.76	79.10	73.24	80.55	77.90	67.06	79.38		
Claude-3.5 Sonnet	76.65	53.62	65.13	89.11	75.06	81.70	64.27	85.30	69.78	76.22	82.17	73.86	59.52	76.62		
Claude-3.5 Haiku	73.78	42.88	58.33	86.13	74.87	76.63	60.38	81.37	58.93	71.75	78.48	71.73	58.87	71.61		
Gemini-2.0 Flash	73.25	59.90	66.58	89.64	69.25	79.80	59.13	73.11	71.86	69.63	76.64	74.39	57.14	70.71		
Gemini-2.0 Flash	76.28	68.96	72.62	83.23	69.88	84.21	59.36	76.72	66.93	77.68	73.57	75.24	58.14	72.93		
Gemini-2.0 Flash	71.45	49.80	60.67	83.58	69.41	81.86	58.43	75.72	79.76	74.88	72.14	72.65	55.88	70.98		
Open-Source Multi-Image MLLMs																
Qwen2.5-72B	59.46	38.80	45.13	71.49	44.54	57.38	52.25	61.88	54.66	55.49	49.15	58.75	37.22	56.71		
Qwen2.5-72B Instruct	56.84	23.80	39.97	73.89	30.19	56.86	53.85	58.90	57.78	53.90	25.83	37.58	11.95	49.92		
LLaVA-3.1-Janus-7B	39.25	18.61	24.43	54.87	31.85	35.88	43.99	57.59	51.07	39.80	31.61	36.60	25.11	41.82		
Embodied MLLMs																
RoboBrain-2.0	43.54	21.30	32.32	62.49	30.36	44.42	42.90	46.92	52.87	45.24	31.25	32.68	25.98	45.12		
RoboBrain-2.0 Instruct	36.20	16.65	26.47	39.32	23.89	45.87	54.53	31.69	29.36	24.29	28.39	25.75	23.97	31.85		
M3-Embodied	66.87	37.30	52.08	82.20	37.11	61.76	63.93	73.05	66.85	70.88	58.95	63.88	28.71	62.72		

Commonsense vs. Embodied Grounding: The text-only baseline helps separate language-prior reasoning from embodied grounding. It remains close to random on many multiple-choice perception and affordance tasks—a 27.1 perception and 25.6 affordance average, far below the 67.3 and 85.7 of the best vision-conditioned MLLM—and is consistently below vision-conditioned MLLMs on task-state and planning questions that require scene evidence. Moreover, during benchmark filtering we remove items that all evaluated models solve correctly, reducing the chance that the final benchmark is dominated by questions answerable from commonsense alone. The remaining gaps therefore reflect failures to connect language priors with visual state, embodiment, and physical feasibility rather than generic commonsense weakness alone.

Perception Challenges: While models perform reasonably on object property analysis—the strongest model reaches 82.3 on functional-attribute reasoning—they struggle with basic robotic perception and spatiotemporal reasoning: even the best model attains only 53.6 on robot-view understanding and 52.1 on temporal grounding, the two weakest perception tasks. Common failures include misidentifying robot type and viewpoint, or failing to reason under different reference frames. These gaps suggest that future models should incorporate stronger embodiment-aware perception modules and explicit spatiotemporal reasoning.

Planning Limitations: As shown in Tab. 5, models exhibit limited capacity in complex planning scenarios. **Cross-Embodiment:** Current models, largely trained on single-arm settings, often fail to coordinate dual-arm actions or perform mobile manipulation, leading to poor spatial search and movement decisions. **Cross-Object:** Performance drops sharply when tasks involve uncommon objects, symbolic reasoning, or world knowledge, suggesting weak integration of heterogeneous information, even though results on common objects remain acceptable. **Cross-View:** When the front-view camera is occluded, multi-view inputs effectively improve model performance. This underscores the need to strengthen multi-view reasoning capabilities, thereby enhancing planning robustness in real-world settings.

Table 5: Results on Planning Q2, Q3. Instr. Compre. stands for Instruction Comprehension. str. Compre. stands for Instruction Comprehension. Naviga. stands for Navigation.

Model	Instr. Compre.		Generalized Planning					
	Explicit	Goal	Single Arm		Mater. Afford.		World Knowl.	
	Q2	Q3	Q2	Q3	Q2	Q3	Q2	Q3
Basic Reference								
Human Evaluation	45.28	74.32	27.52	71.35	43.62	71.20	43.89	69.83
GPT-5.4-text-only	36.98	46.25	40.98	52.86	40.43	52.33	43.01	41.46
Closed-Source MLLMs								
GPT-5.4	49.48	62.50	48.20	67.85	44.38	64.67	42.19	51.22
GPT-5.2	39.32	75.00	42.86	73.02	41.07	66.67	37.50	36.10
GPT-5	44.09	72.97	47.26	75.75	44.38	62.83	39.58	63.41
GPT-4.1	45.31	70.00	48.62	63.76	44.32	63.67	39.58	38.54
GPT-4o	42.86	65.00	43.94	59.95	41.23	55.33	42.19	46.34
Claude-Opus-4.7	44.53	67.50	51.08	65.12	45.97	63.00	45.31	63.41
Claude-Sonnet-4.6	43.75	70.00	43.22	66.21	43.90	64.33	39.58	60.98
Claude-Sonnet-4.5	41.67	61.25	44.73	56.68	41.99	54.00	41.15	48.78
Claude-Haiku-4.5	34.13	62.50	43.07	60.38	38.14	63.18	31.77	70.73
Gemini-3.1-Pro	44.79	71.25	50.65	64.58	46.97	67.67	26.56	65.85
Gemini-2.5-Pro	37.76	72.50	52.02	71.66	49.20	70.00	35.42	68.29
Gemini-2.5-Flash	47.66	77.50	49.85	54.22	44.81	68.83	45.31	68.29
Open-Source Multi-Image MLLMs								
Qwen3-VL-8B	49.74	62.50	50.94	61.58	44.80	55.50	36.98	36.10
Qwen2.5-VL-7B-Ins	29.43	55.00	31.88	52.59	30.17	49.33	20.31	51.22
LLaVA-OneVision-7B	33.59	41.25	35.06	46.05	35.97	40.50	34.90	43.90
Embodied MLLMs								
RoboBrain-2.0-7B	34.90	52.50	33.84	53.95	34.29	49.00	27.60	48.78
RoboBrain-2.5-4B	32.81	58.75	37.37	54.50	35.25	55.17	32.29	36.10
MMo-Embodied-7B	38.17	55.00	43.87	55.31	42.90	51.30	46.88	63.41

Model	Affordance Prediction				Failure Analysis			
	Static	Dynamic	Naviga.	Avg	Execution	Planning	Avg	
	Q2	Q3	Q2	Q3	Q2	Q3	Q2	Q3
Basic Reference								
Human Evaluation	86.08	80.02	81.85	82.63	47.30	80.67	63.99	
GPT-5.4-text-only	23.81	27.52	25.51	25.61	11.92	32.64	22.28	
Closed-Source MLLMs								
GPT-5.4	44.22	36.91	58.16	46.43	26.49	65.97	46.23	
GPT-5.2	43.54	39.60	47.96	43.70	26.49	68.06	47.27	
GPT-5	62.59	49.66	62.24	58.16	19.87	80.56	50.21	
GPT-4.1	29.93	42.95	68.37	47.08	20.53	70.83	45.68	
GPT-4o	40.82	42.28	50.00	44.37	31.79	57.64	44.71	
Claude-Opus-4.7	53.74	62.42	79.59	65.25	14.57	72.22	43.40	
Claude-Sonnet-4.6	37.41	52.35	41.84	43.87	17.88	77.78	47.83	
Claude-Sonnet-4.5	34.69	38.93	53.06	42.23	14.57	63.19	38.88	
Claude-Haiku-4.5	27.89	26.17	21.43	25.16	17.22	45.83	31.53	
Gemini-3.1-Pro	82.31	77.85	96.94	85.70	25.17	80.74	52.95	
Gemini-2.5-Pro	65.99	61.07	93.88	73.65	18.54	74.29	46.41	
Gemini-2.5-Flash	61.22	69.80	36.73	55.92	25.83	65.49	45.66	
Open-Source Multi-Image MLLMs								
Qwen3-VL-8B	23.81	17.45	22.45	21.24	22.52	55.56	39.04	
Qwen2.5-VL-7B-Ins	18.37	31.54	26.53	25.48	13.91	35.42	24.66	
LLaVA-OneVision-7B	38.78	33.56	66.33	46.22	20.53	31.25	25.89	
Embodied MLLMs								
RoboBrain-2.0-7B	31.97	27.52	31.63	30.37	15.23	40.28	27.75	
RoboBrain-2.5-4B	30.34	21.48	72.45	48.09	43.71	46.53	45.12	
MMo-Embodied-7B	51.70	36.91	70.41	53.01	19.21	42.56	30.78	

Failure Analysis: As reported in Tab. 6, analyzing execution errors proves harder than diagnosing planning errors, with scores consistently lower on the former: even the strongest model reaches only 43.7 on execution-error diagnosis while most models fall between 15 and 27, whereas planning-error diagnosis reaches 80.7; the human upper bound exhibits the same asymmetry (47.3 versus 80.7). Execution-level failures often require fine-grained distinctions—such as separating position errors from rotation errors (correct position but wrong gripper angle)—which demand expert-level embodied knowledge. This suggests that fine-grained perception is crucial for improving error diagnosis.

5.3 In Depth Analysis

Planning Error Analysis: To understand the limitations of current multi-modal models as embodied planners, we analyze representative failure cases across RoboBench tasks, and find that most failures stem not from misunderstanding instructions but from difficulty grounding semantic reasoning into executable actions. We categorize common failures into four types (Fig. 6). **Execution Errors** account for the majority of failures (45%), arising from incomplete or incorrect action sequences such as missing steps or wrong functions. **Identification Errors** (24%) occur when models confuse objects or assign incorrect parameters. **Common Sense Errors** (25%) involve violations of physical or spatial constraints. **Mode-Specific Errors** (6%) result from failing to follow structured output conventions. These results highlight a persistent perception-action gap: models can generate plausible reasoning but struggle to consistently translate it into executable embodied actions.

Human Alignment of the Evaluation Framework: To validate reliability, we compare our framework with a baseline LLM-based scoring method [13] that

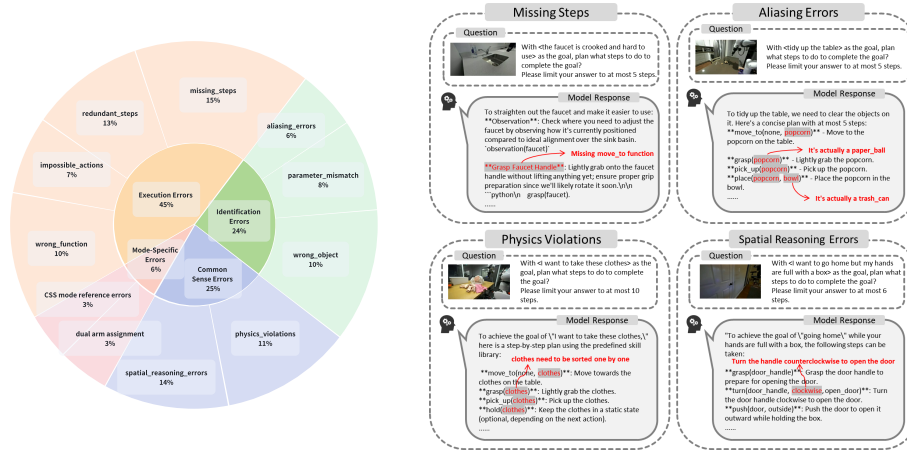
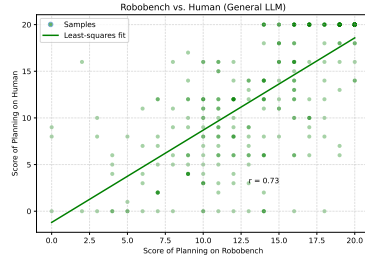


Fig. 6: Planning error analysis. Left: distribution of error types in RoboBench planning tasks. Right: representative failure cases illustrating common planning errors.

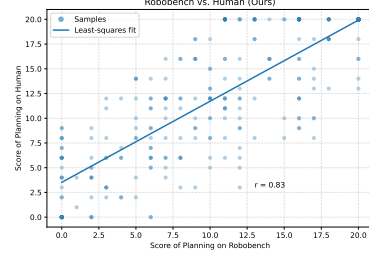
judges criteria such as planning rationality and object correctness, whereas we judge whether the plan, when executed, completes the task. As shown in Fig. 7, our MLLM-as-world-simulator achieves stronger alignment with human judgments ($r = 0.83$) than the baseline ($r = 0.73$). By scoring plans at structured skill-object nodes guided by the underlying DAG, it improves consistency with humans while reducing the bias of free-form language scoring. The protocol is also robust to grader and prompt: across four judge models and three prompt variants, NodeCorrectness remains stable ($\text{ICC}(2,k) = 0.948$), confirming that the protocol—not a particular model or phrasing—drives the evaluation (full analysis in the supplementary).

Correlation between RoboBench and downstream VLA performance: Following VLM4VLA [56], we convert several open-source VLMs into VLA policies via minimal fine-tuning and evaluate them on two manipulation benchmarks, CALVIN [41] and LIBERO-10 [33]. This ensures that downstream VLA performance largely reflects the intrinsic capabilities of the VLM backbone, with implementation details in the supplementary material.

We correlate RoboBench scores with downstream VLA success rates across three dimensions: *Perception Reasoning*, *Affordance Prediction*, and *Failure Analysis*. As shown in Fig. 8, perception reasoning strongly correlates with CALVIN (object-centric $r = 0.884$, scene-centric $r = 0.833$), reflecting the role of object and scene understanding in long-horizon manipulation, whereas LIBERO correlates more with affordance prediction (static $r = 0.617$, dynamic $r = 0.517$), indicating greater reliance on fine-grained interaction dynamics. Thus different benchmarks emphasize different abilities, and RoboBench helps identify VLM capabilities predictive of VLA performance.



(a) Baseline evaluation framework.



(b) MLLM-as-world-simulator evaluation.

Fig. 7: Alignment between automated evaluation scores and human judgments. Pearson correlation coefficient (r) measures the consistency between each evaluation method and human evaluation. Our MLLM-as-world-simulator evaluation shows stronger alignment with human judgments than the baseline LLM-based scoring method.

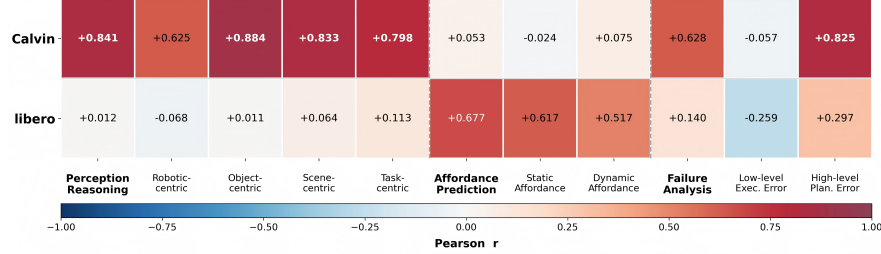


Fig. 8: Correlation between RoboBench cognitive abilities and downstream VLA performance. Results show that perception reasoning strongly correlates with long-horizon manipulation in CALVIN, while affordance prediction is more related to fine-grained manipulation tasks in LIBERO-10.

6 Conclusion

We presented RoboBench, a benchmark to systematically evaluate MLLMs as the embodied brain for robotic manipulation. It traces the full execution pipeline across five dimensions—Instruction Comprehension, Perception Reasoning, Generalized Planning, Affordance Prediction, and Failure Analysis—covering 25 tasks with 6092 QA pairs, and contributes a world-simulator framework that evaluates not only symbolic plan fidelity but also embodied feasibility. Experiments expose major gaps across all five dimensions and reveal the correlation between embodied cognitive abilities and downstream VLA performance. RoboBench thus provides a unified scaffold for measuring embodied cognition and guiding the development of more robust, generalizable embodied intelligence.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62476011), the Beijing Natural Science Foundation (L252060), and the Beijing Natural Science Foundation (4264125).

References

1. Azzolini, A., Bai, J., Brandon, H., Cao, J., Chattopadhyay, P., Chen, H., Chu, J., Cui, Y., Diamond, J., Ding, Y., et al.: Cosmos-reason1: From physical common sense to embodied reasoning. arXiv preprint arXiv:2503.15558 (2025)
2. Bjorck, J., Castañeda, F., Cherniadev, N., Da, X., Ding, R., Fan, L., Fang, Y., Fox, D., Hu, F., Huang, S., et al.: Gr00t n1: An open foundation model for generalist humanoid robots. arXiv preprint arXiv:2503.14734 (2025)
3. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., Jakubczak, S., Jones, T., Ke, L., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Shi, L.X., Tanner, J., Vuong, Q., Walling, A., Wang, H., Zhilinsky, U.: π_0 : A vision-language-action flow model for general robot control (2024), <https://arxiv.org/abs/2410.24164>
4. Bu, Q., Cai, J., Chen, L., Cui, X., Ding, Y., Feng, S., Gao, S., He, X., Hu, X., Huang, X., et al.: Agibot world colosseum: A large-scale manipulation platform for scalable and intelligent embodied systems. arXiv preprint arXiv:2503.06669 (2025)
5. Chen, H., Liu, J., Gu, C., Liu, Z., Zhang, R., Li, X., He, X., Guo, Y., Fu, C.W., Zhang, S., et al.: Fast-in-slow: A dual-system foundation model unifying fast manipulation within slow reasoning. arXiv preprint arXiv:2506.01953 (2025)
6. Chen, K., Xie, S., Ma, Z., Sanketi, P.R., Goldberg, K.: Robo2vlm: Visual question answering from large-scale in-the-wild robot manipulation datasets. arXiv preprint arXiv:2505.15517 (2025)
7. Chen, S., Wu, Z., Zhang, K., Li, C., Zhang, B., Ma, F., Yu, F.R., Li, Q.: Exploring embodied multimodal large models: Development, datasets, and future directions (2025), <https://arxiv.org/abs/2502.15336>
8. Chen, T., Chen, Z., Chen, B., Cai, Z., Liu, Y., Liang, Q., Li, Z., Lin, X., Ge, Y., Gu, Z., et al.: Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. arXiv preprint arXiv:2506.18088 (2025)
9. 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 (2024), <https://arxiv.org/abs/2312.06722>
10. Cheng, S., Fang, K., Yu, Y., Zhou, S., Li, B., Tian, Y., Li, T., Han, L., Liu, Y.: Videogothink: Assessing egocentric video understanding capabilities for embodied ai (2024), <https://arxiv.org/abs/2410.11623>
11. 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 (CVPR). pp. 14291–14302 (June 2024)
12. Cheng, Z., Tu, Y., Li, R., Dai, S., Hu, J., Hu, S., Li, J., Shi, Y., Yu, T., Chen, W., Shi, L., Sun, M.: Embodiedeval: Evaluate multimodal llms as embodied agents (2025), <https://arxiv.org/abs/2501.11858>
13. Chi, X., Zhang, H., Fan, C.K., Qi, X., Zhang, R., Chen, A., min Chan, C., Xue, W., Luo, W., Zhang, S., Guo, Y.: Eva: An embodied world model for future video anticipation (2024), <https://arxiv.org/abs/2410.15461>
14. Clark, A.: Being there: Putting brain, body, and world together again. MIT press (1998)
15. Collaboration, O.X.E., et al.: Open X-Embodiment: Robotic learning datasets and RT-X models. <https://arxiv.org/abs/2310.08864> (2023)

16. 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)
17. 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)
18. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., et al.: Palm-e: An embodied multimodal language model. In: ICML (2023)
19. Duan, J., Pumacay, W., Kumar, N., Wang, Y.R., Tian, S., Yuan, W., Krishna, R., Fox, D., Mandlekar, A., Guo, Y.: Aha: A vision-language-model for detecting and reasoning over failures in robotic manipulation. arXiv preprint arXiv:2410.00371 (2024)
20. Fang, H.S., Fang, H., Tang, Z., Liu, J., Wang, C., Wang, J., Zhu, H., Lu, C.: Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. arXiv preprint arXiv:2307.00595 (2023)
21. Gao, J., Sarkar, B., Xia, F., Xiao, T., Wu, J., Ichter, B., Majumdar, A., Sadigh, D.: Physically grounded vision-language models for robotic manipulation. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 12462–12469. IEEE (2024)
22. Huang, W., Wang, C., Li, Y., Zhang, R., Fei-Fei, L.: Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation (2024), <https://arxiv.org/abs/2409.01652>
23. Huang, W., Wang, C., Zhang, R., Li, Y., Wu, J., Fei-Fei, L.: Voxposer: Composable 3d value maps for robotic manipulation with language models (2023), <https://arxiv.org/abs/2307.05973>
24. James, S., Ma, Z., Arrojo, D.R., Davison, A.J.: Rlbench: The robot learning benchmark & learning environment. IEEE Robotics and Automation Letters **5**(2), 3019–3026 (2020)
25. 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)
26. Lee, J., Duan, J., Fang, H., Deng, Y., Liu, S., Li, B., Fang, B., Zhang, J., Wang, Y.R., Lee, S., et al.: Molmoact: Action reasoning models that can reason in space. arXiv preprint arXiv:2508.07917 (2025)
27. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer (2024), <https://arxiv.org/abs/2408.03326>
28. Li, J., Zhu, Y., Tang, Z., Wen, J., Zhu, M., Liu, X., Li, C., Cheng, R., Peng, Y., Feng, F.: Improving vision-language-action models via chain-of-affordance. arXiv preprint arXiv:2412.20451 (2024)
29. Li, J., Zhu, Y., Xu, Z., Gu, J., Zhu, M., Liu, X., Liu, N., Peng, Y., Feng, F., Tang, J.: Mmro: Are multimodal llms eligible as the brain for in-home robotics? (2024), <https://arxiv.org/abs/2406.19693>
30. Li, M., Zhao, S., Wang, Q., Wang, K., Zhou, Y., Srivastava, S., Gokmen, C., Lee, T., Li, E.L., Zhang, R., Liu, W., Liang, P.S., Li, F.F., Mao, J., Wu, J.: Embodied agent interface: Benchmarking llms for embodied decision making.

- In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) *Advances in Neural Information Processing Systems*. vol. 37, pp. 100428–100534. Curran Associates, Inc. (2024), https://proceedings.neurips.cc/paper_files/paper/2024/file/b631da756d1573c24c9ba9c702fde5a9-Paper-Datasets_and_Benchmarks_Track.pdf
31. Li, M., Zhao, S., Wang, Q., Wang, K., Zhou, Y., Srivastava, S., Gokmen, C., Lee, T., Li, E.L., Zhang, R., et al.: Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems* **37**, 100428–100534 (2025)
 32. Lin, F., Hu, Y., Sheng, P., Wen, C., You, J., Gao, Y.: Data scaling laws in imitation learning for robotic manipulation. *arXiv preprint arXiv:2410.18647* (2024)
 33. 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)
 34. Liu, F., Fang, K., Abbeel, P., Levine, S.: Moka: Open-vocabulary robotic manipulation through mark-based visual prompting. In: *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024* (2024)
 35. Liu, J., Chen, H., An, P., Liu, Z., Zhang, R., Gu, C., Li, X., Guo, Z., Chen, S., Liu, M., et al.: Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model. *arXiv preprint arXiv:2503.10631* (2025)
 36. Liu, S., Wu, L., Li, B., Tan, H., Chen, H., Wang, Z., Xu, K., Su, H., Zhu, J.: Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864* (2024)
 37. Liu, Y., Chen, W., Bai, Y., Liang, X., Li, G., Gao, W., Lin, L.: Aligning cyber space with physical world: A comprehensive survey on embodied ai (2024), <https://arxiv.org/abs/2407.06886>
 38. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., Lin, D.: Mmbench: Is your multi-modal model an all-around player? (2024), <https://arxiv.org/abs/2307.06281>
 39. Liu, Z., Bahety, A., Song, S.: Reflect: Summarizing robot experiences for failure explanation and correction. *arXiv preprint arXiv:2306.15724* (2023)
 40. Majumdar, A., Ajay, A., Zhang, X., Putta, P., Yenamandra, S., Henaff, M., Silwal, S., Mcvay, P., Maksymets, O., Arnaud, S., Yadav, K., Li, Q., Newman, B., Sharma, M., Berges, V., Zhang, S., Agrawal, P., Bisk, Y., Batra, D., Kalakrishnan, M., Meier, F., Paxton, C., Sax, A., Rajeswaran, A.: Openeqa: Embodied question answering in the era of foundation models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 16488–16498 (June 2024)
 41. Mees, O., Hermann, L., Rosete-Beas, E., Burgard, W.: Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters* **7**(3), 7327–7334 (2022)
 42. 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)
 43. 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)
 44. Sermanet, P., Ding, T., Zhao, J., Xia, F., Dwibedi, D., Gopalakrishnan, K., Chan, C., Dulac-Arnold, G., Maddingeni, S., Joshi, N.J., et al.: Robovqa: Multimodal long-horizon reasoning for robotics. In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. pp. 645–652. IEEE (2024)

45. Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., Fox, D.: Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
46. Shridhar, M., Yuan, X., Côté, M.A., Bisk, Y., Trischler, A., Hausknecht, M.: Alf-world: Aligning text and embodied environments for interactive learning. arXiv preprint arXiv:2010.03768 (2020)
47. Song, C.H., Blukis, V., Tremblay, J., Tyree, S., Su, Y., Birchfield, S.: Robospacial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15768–15780 (2025)
48. Sun, F., Chen, R., Ji, T., Luo, Y., Zhou, H., Liu, H.: A comprehensive survey on embodied intelligence: Advancements, challenges, and future perspectives. CAAI Artificial Intelligence Research **3**, 9150042 (2024). <https://doi.org/10.26599/AIR.2024.9150042>, <https://www.sciopen.com/article/10.26599/AIR.2024.9150042>
49. Team, B.R.: Robobrain 2.0 technical report. arXiv preprint arXiv:2507.02029 (2025)
50. Team, G.R., Abeyruwan, S., Ainslie, J., Alayrac, J.B., Arenas, M.G., Armstrong, T., Balakrishna, A., Baruch, R., Bauza, M., Blokzijl, M., et al.: Gemini robotics: Bringing ai into the physical world. arXiv preprint arXiv:2503.20020 (2025)
51. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution (2024), <https://arxiv.org/abs/2409.12191>
52. Wu, K., Hou, C., Liu, J., Che, Z., Ju, X., Yang, Z., Li, M., Zhao, Y., Xu, Z., Yang, G., et al.: Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. arXiv preprint arXiv:2412.13877 (2024)
53. 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 (2024), <https://arxiv.org/abs/2412.14171>
54. Yang, R., Chen, H., Zhang, J., Zhao, M., Qian, C., Wang, K., Wang, Q., Koripella, T.V., Movahedi, M., Li, M., Ji, H., Zhang, H., Zhang, T.: Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents (2025), <https://arxiv.org/abs/2502.09560>
55. Yuan, W., Duan, J., Blukis, V., Pumacay, W., Krishna, R., Murali, A., Mousavian, A., Fox, D.: Robopoint: A vision-language model for spatial affordance prediction for robotics. arXiv preprint arXiv:2406.10721 (2024)
56. Zhang, J., Chen, X., Wang, Q., Li, M., Guo, Y., Hu, Y., Zhang, J., Bai, S., Lin, J., Chen, J.: Vlm4vla: Revisiting vision-language-models in vision-language-action models. arXiv preprint arXiv:2601.03309 (2026)
57. Zhang, S., Xu, Z., Liu, P., Yu, X., Li, Y., Gao, Q., Fei, Z., Yin, Z., Wu, Z., Jiang, Y.G., Qiu, X.: Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks (2024), <https://arxiv.org/abs/2412.18194>
58. Zhou, E., An, J., Chi, C., Han, Y., Rong, S., Zhang, C., Wang, P., Wang, Z., Huang, T., Sheng, L., et al.: Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. arXiv preprint arXiv:2506.04308 (2025)
59. Zhou, E., Su, Q., Chi, C., Zhang, Z., Wang, Z., Huang, T., Sheng, L., Wang, H.: Code-as-monitor: Constraint-aware visual programming for reactive and proactive

robotic failure detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6919–6929 (2025)