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 Liu[1] Gaole Dai[1] Rongyu Zhang[1]
Ruichuan An[1] Kun Wu[5] Zhengping Che[5] Shaoxuan Xie[2] Guocai Yao[2] Zhongxia Zhao[1,2] Pengwei Wang[2] Guang Liu[2] Zhongyuan Wang[2] Tiejun Huang[1,2] Shanghang Zhang[1,2#]

[1]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University [2]Beijing Academy of Artificial Intelligence
[3]Institute for Brain and Intelligence, Fudan University [4]University of Science and Technology Beijing [5]Beijing Innovation Center of Humanoid Robotics
*Equal contribution †Project leader #Corresponding author

ECCV 2026 | MALMÖ | SEPT 8-12

Project Page

Dataset (HF)

1. Motivation & Highlights

What is the embodied brain?

In dual-system robots an MLLM is **System 2** — the cognitive core that reasons and decides — while System 1 handles low-level control. Completing a task takes a full cognitive chain:

understand → perceive → plan → ground affordances → diagnose

Existing benchmarks fall short

- **Incomplete** — perception, planning or reflection tested in isolation
- **Not grounded in real robots** — built on simulation or generic web imagery, a domain gap away from real deployment
- **Shallow metrics** — success rate conflates cognition with execution; MC / BLEU / LLM scores miss embodied feasibility

RoboBench highlights

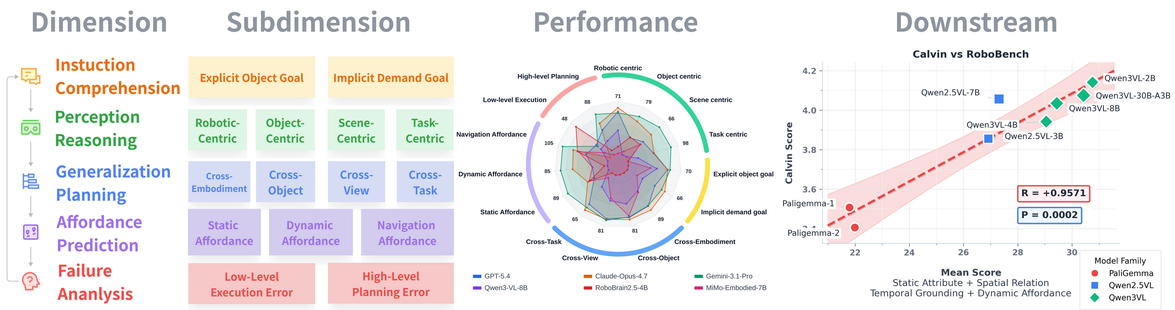
- **Full-pipeline cognition:** 5 dimensions / 14 capabilities / 25 tasks, from instruction to failure analysis
- **Realistic & diverse:** large-scale real robot data — single/dual-arm, mobile & humanoid, attribute-rich objects, occluded multi-view scenes; no domain gap, no simulation shortcut
- **Faithful planning eval:** MLLM-as-world-simulator rolls plans out against a DAG to test embodied feasibility, not text overlap
- **Largest study of its kind:** 18 SOTA MLLMs × 3 runs; abilities predict downstream VLA success

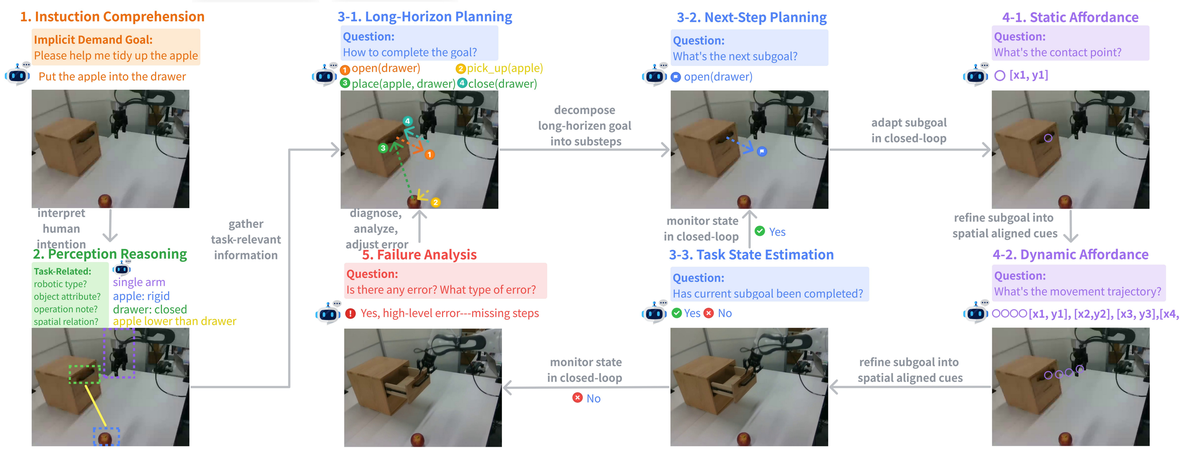
Already adopted as the cognition-evaluation suite of **RoboBrain 2.0** and **Tencent HY-Embodied-0.5**

RoboBench at a glance

5	14	25	6,092
Dimensions	Capabilities	Tasks	QA pairs

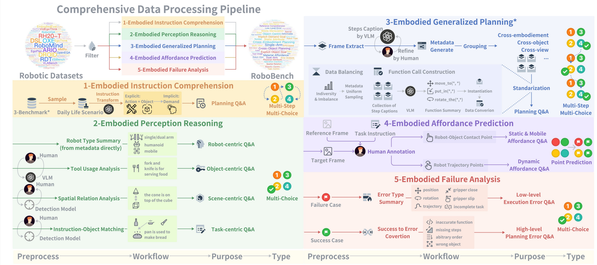
2. The RoboBench Benchmark — Taxonomy, Performance, Downstream, and Task Walkthrough



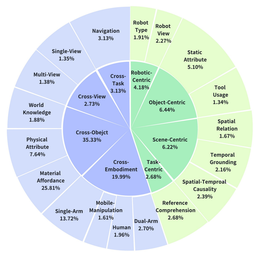


Top: 5 dimensions / 14 capabilities; per-capability performance; RoboBench score predicts downstream control. Bottom: one episode exercises all five cognitive roles — from intent to failure diagnosis.

3. Construction & Statistics

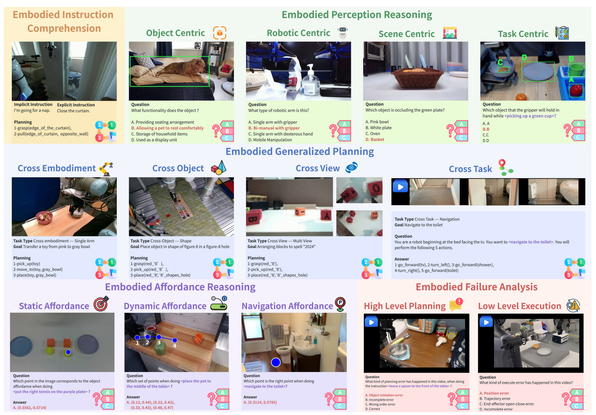


- **Sources:** large-scale real robot corpora (OXE, RH20T, RoboMIND, AgiBot, …) + in-house collection across single-arm, dual-arm, mobile and humanoid embodiments
- **Annotation:** per-dimension workflows — VLM / detector / expert annotation → QA generation with scene-grounded distractors
- **Real failures:** execution errors come from real VLA rollouts (RoboMIND), expert-labeled — not perturbed successes
- **Quality control:** two-stage filtering + majority vote, residual disagreement < 4%



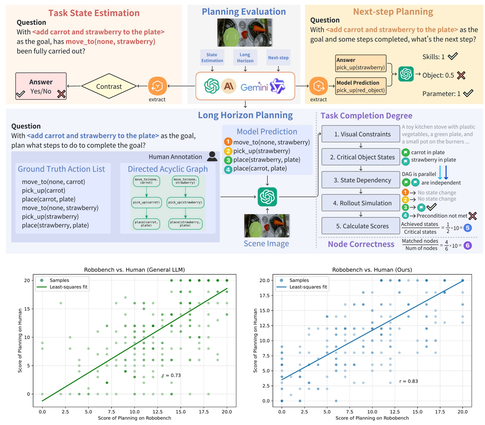
MC 1,895 · Planning Q1 1,973 / Q2 842 / Q3 1,192 · Navigation 190 · avg 6.75 steps (max 55)

4. Demo Cases — One per Capability



- **Scene-grounded:** the model must read the actual image / video, not recall priors
- **Formats:** every task is multiple-choice — except planning, which requires structured plan generation
- **Human-verified:** every MLLM-assisted annotation gets a human cross-pass; majority-vote filtering drops items all models solve or all fail
- **One framework, any brain:** pure-cognition QA — VLA backbones and agentic planners are comparable on equal footing
- **Open release:** all 6,092 items on Hugging Face — scan the Dataset QR above

5. World-Simulator Evaluation



How a plan is judged — not by wording, but by whether it runs:

- **Annotate:** experts encode each task as a dependency DAG over atomic actions (skill, object, args)
- **Ground:** the judge MLLM reads first frame + reference plan + DAG → derives critical object-state milestones (e.g. drawer = open)
- **Roll out:** the candidate plan executes step-by-step; each action's preconditions are checked and the world state updated
- **Score:** (NodeCorrectness + TaskCompletion) / 20 — matched action nodes + milestones truly achieved; Q2 next-step & Q3 state estimation complete the suite
- **Trustworthy: r = 0.83** vs human (baseline 0.73) · judge/prompt-robust **ICC(2,k) = 0.948**

6. Leaderboard & Key Findings

Model	Instruction	Perception	Planning (Q1)	Affordance	Failure
Reference					
Human Evaluation	60.5	74.3	54.5	82.6	64.0
GPT-5.4-text-only	56.7	27.1	73.7	25.6	22.3
Closed-Source MLLMs					
GPT-5.4	62.7	56.2	70.9	46.4	46.2
GPT-5.2	62.4	53.4	72.3	43.7	47.3
GPT-5	66.2	60.8	71.8	58.2	50.2
GPT-4.1	66.8	53.0	71.8	47.1	45.7
GPT-4o	64.6	45.4	73.5	44.4	44.7
Claude-Opus-4.7	67.7	62.6	75.5	65.2	43.4
Claude-Sonnet-4.6	70.7	54.1	79.4	43.9	47.8
Claude-Sonnet-4.5	65.1	45.0	76.6	42.2	38.9
Claude-Haiku-4.5	58.3	35.9	71.0	25.2	31.5
Gemini-3.1-Pro	66.6	67.3	70.7	85.7	53.0
Gemini-2.5-Pro	68.6	63.9	71.5	73.7	45.4
Gemini-2.5-Flash	60.7	59.5	71.0	55.9	45.7
Open-Source MLLMs					
Qwen3-VL-8B	45.1	41.5	56.7	21.2	39.0
Qwen2.5-VL-7B-Ins	40.0	32.2	49.9	25.5	24.7
LLaVA-OneVision-7B	24.4	38.9	41.0	46.2	25.9
Embodied MLLMs					
RoboBrain-2.0-7B	32.3	29.8	45.1	30.4	27.8
RoboBrain-2.5-4B	26.5	44.0	31.9	48.1	45.1
MiMo-Embodied-7B	52.1	40.1	62.7	53.0	30.8

Best = red cell, runner-up = shaded; per-dimension averages (planning = Q1), 3-run mean.

- **Closed > open source:** ~20 points on average (~50% relative) — widest in instruction (~28) and planning (~25), narrowest in failure analysis (~13); within a family, scores rise with size & generation.
- **Implicit instructions are hard:** best model drops 79.9 → 61.4 when the same goal is stated indirectly; CoT rewriting does not close it.
- **Weakest perception:** robotic-centric view (53.6) & temporal grounding (50.4) vs 82.3 on object attributes.
- **Planning degrades** across embodiments, uncommon objects & views; multi-view input helps under occlusion.
- **Execution-failure diagnosis is hardest:** best 43.7 vs 80.7 on planning errors — fine-grained embodied perception is the bottleneck.
- **Vision is required:** text-only GPT-5.4 is near random on perception & affordance — language priors cannot solve RoboBench.

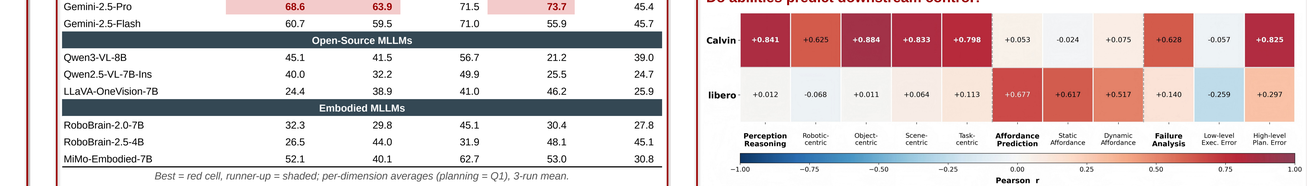
7. In-depth Analysis & Takeaway

Why do plans fail? A breakdown of 4 planning-error types:

- **Execution 45%** — missing steps, wrong functions
- **Common sense 25%** — physical / spatial violations
- **Identification 24%** — wrong objects or parameters
- **Mode 6%** — output-format failures

Takeaway: *errors are dominated by grounding, not language — models reason well but act poorly.*

Do abilities predict downstream control?



Object-centric perception predicts CALVIN (r ≈ 0.88); affordance (static+dynamic) predicts LIBERO-10 (r ≈ 0.68) — a guide for VLA backbone selection.

Takeaway

- **Success rate tells whether a robot finished — RoboBench tells whether it understood:** today's MLLMs imitate tasks more than they understand them; plans read fluently yet break under physical rollout.
- **Not all VLM abilities matter equally for embodiment —** perception & affordance are what transfer downstream: diagnose and improve those, rather than only scaling end-to-end data.