# Uni-Embodied: Towards Unified Evaluation for Embodied Planning, Perception, and Execution

Lingfeng Zhang*
Shenzhen International Graduate School, Tsinghua University
Shenzhen, Guangdong, China
lfzhang715@gmail.com

Yingbo Tang*
Institute of Automation, CAS
Beijing, China
tangyingbo2020@ia.ac.cn

Xinyu Zheng
NMTC, Tongji University
Shanghai, China
xzheng565@connect.hkust-gz.edu.cn

Qiang Zhang
HKUST (GZ)
X-Humaniod
Guangzhou, Guangdong, China
jony.zhang@x-humanoid.com

Yu Liu
Hefei University of Technology
Hefei, Anhui, China
yuliu@hfut.edu.cn

Renjing Xu
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, Guangdong, China
renjingxu@hkust-gz.edu.cn

Xiaoshuai Hao†
Beijing Academy of Artificial Intelligence (BAAI)
Beijing, China
xshao@baai.ac.cn

## Abstract

Embodied intelligence is a key challenge in artificial general intelligence (AGI), requiring the seamless integration of ***planning, perception, and execution*** capabilities for agents to perform physical tasks effectively. Although current vision-language models (VLMs) excel in individual capabilities, their ability to simultaneously exhibit all three necessary skills for embodied tasks is uncertain, hindering the development of unified embodied systems. In this paper, we introduce the novel ***Uni-Embodied***, the first comprehensive benchmark designed to evaluate the comprehensive capabilities of VLMs across various areas of embodied intelligence. Our benchmark encompasses three key dimensions—planning, perception, and execution—and includes nine specific tasks: complex and simple embodied task planning, navigation trajectory summarization, navigation map understanding, object affordance recognition, spatial pointing, manipulation trajectory analysis, and task execution for navigation and manipulation. Extensive evaluations of various state-of-the-art open-source and closed-source VLMs reveal that current models struggle to perform well in all three embodied capabilities. We find that integrating planning and perception weakens execution abilities, while focusing on execution significantly degrades planning and perception performance, highlighting critical limitations in existing approaches. Additionally, we identify effective strategies for enhancing different embodied capabilities, including ***chain-of-thought and hybrid training***. These insights pave the way for developing improved embodied intelligence systems, which are essential for advancing real-world robotics applications. The benchmark, along with its associated code, has been publicly released to support further research in this domain. Project website: *https://Uni-Embodied.github.io/*.

## CCS Concepts

• **Computing methodologies → Robotic planning**; **Vision for robotics**; **Cognitive robotics**.

## Keywords

Embodied Intelligence, Embodied Manipulation and Navigation

## 1 Introduction

Embodied intelligence empowers agents to interact with the physical world and perform complex tasks, serving as a pathway toward achieving artificial general intelligence (AGI) [7, 19, 35]. By enhancing their capabilities in perceiving and acting within dynamic environments, embodied agents can unlock new potentials in AI development, bringing us closer to a deeper understanding of intelligence. These tasks can be systematically decomposed into three interrelated components: planning [11, 21, 29, 32], perception [9, 18, 42], and execution [3, 8]. ***Planning*** involves high-level task decomposition and reasoning, such as breaking down the instruction to "prepare breakfast" into subtasks like "navigate to the kitchen," "open the refrigerator," "get ingredients," and "boil eggs." *This approach allows agents to manage complex tasks effectively.* ***Perception*** includes key functions such as trajectory summarization, scene understanding, affordance recognition, spatial pointing, and trajectory analysis. *These elements enable agents to navigate and interact with objects, ensuring appropriate responses to changing conditions.* ***Execution***

*Both authors contributed equally to this research.
†Corresponding Author.


*ACM MM, Dublin, Ireland*

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

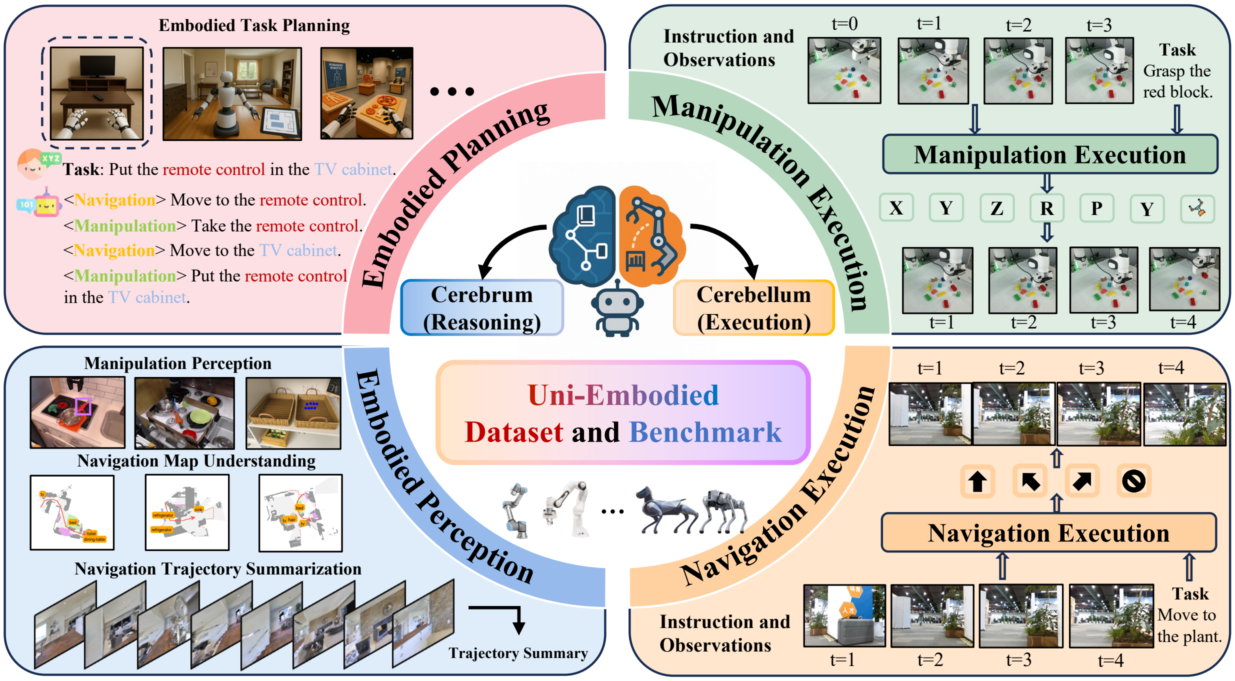


**Figure 1: Overview of the Uni-Embodied Benchmark. It encompasses three key dimensions: planning, perception, and execution.**

refers to the precise implementation of low-level actions and motion control, including generating robotic actions to grasp objects and navigate to designated locations based on instructions. *This execution is essential for translating plans and perceptions into tangible actions in the real world.*

Recent advancements in vision-language models (VLMs) [2, 4, 13, 27, 30] show significant promise for embodied intelligence applications [12, 31, 38–40, 43]. By leveraging their robust pre-trained representations and fine-tuning on domain-specific datasets, VLMs excel in various embodied tasks, including planning complex sequences, perceiving dynamic scenes, and executing precise robotic actions. However, existing benchmarks typically focus on individual domains—such as planning, perception, or execution—failing to provide a comprehensive evaluation of VLM capabilities in embodied intelligence. Therefore, a unified benchmark that integrates these essential tasks is needed for more thorough assessment.

To address this fundamental gap, we propose the ***Uni-Embodied***, the first comprehensive benchmark designed to uniformly evaluate embodied planning, perception, and execution. This benchmark aims to assess VLM performance across these three critical capabilities. Specifically, as shown in Fig. 1, ***Uni-Embodied Benchmark*** includes three components: ***Planning***, which evaluates VLMs' planning capabilities in navigation-manipulation integrated tasks and pure desktop manipulation scenarios; ***Perception***, which encompasses navigation map understanding, trajectory summarization, object affordance recognition, spatial pointing, and manipulation trajectory prediction to thoroughly assess VLMs' perception capabilities; and ***Execution***, which provides evaluation snippets for navigation and manipulation tasks along with an interactive platform to directly test VLMs' execution capabilities. Extensive experiments conducted on both open-source and closed-source models reveal that while existing VLMs perform well in one or two capabilities, none manage to excel across all tasks. Our findings indicate that integrating planning and perception can weaken execution, while a focus on execution significantly degrades planning and perception performance, highlighting the key limitations of current methods. Additionally, ablation experiments show that ***chain-of-thought and hybrid training*** can enhance VLM performance across three capabilities.

The contributions of this paper are mainly three-fold:

- We propose the ***Uni-Embodied Benchmark***, the first comprehensive benchmark for evaluating three embodied capabilities essential for developing general embodied intelligence systems.
- Our benchmark encompasses three key dimensions: ***planning***, ***perception***, and ***execution***. It includes nine specific tasks: complex and simple embodied task planning, navigation trajectory summarization, navigation map understanding, object affordance recognition, spatial pointing, manipulation trajectory analysis, and task execution for both navigation and manipulation, enabling a thorough evaluation of the model's embodied capabilities.
- We conduct an extensive evaluation of existing open-source and closed-source models based on our benchmark. We find that all current models cannot perform well in all three capabilities. At the same time, we propose three strategies to enhance the embodied capabilities of VLMs: ***chain-of-thought and hybrid training***.

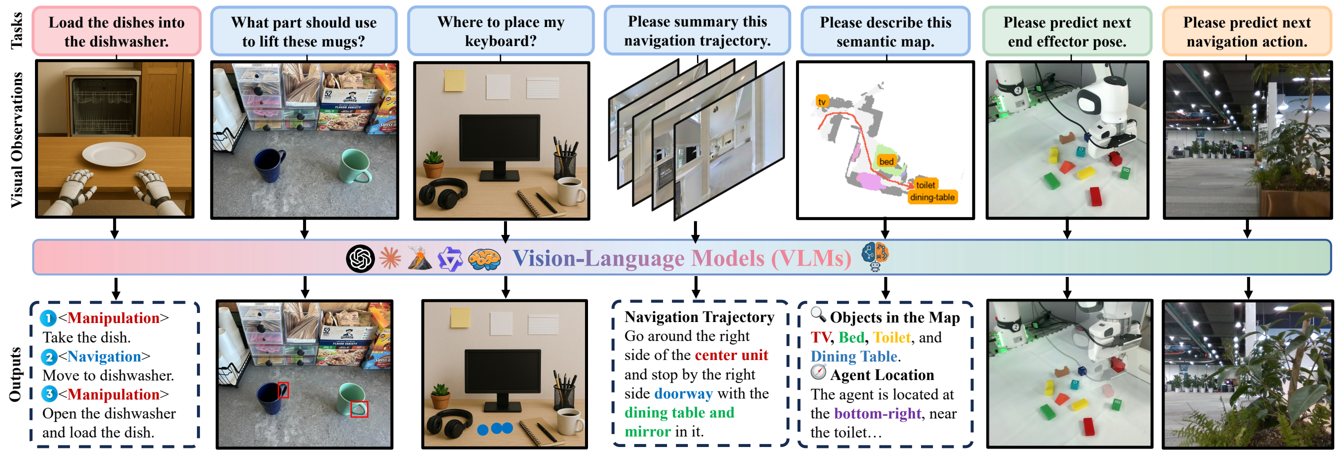


**Figure 2: Tasks in Uni-Embodied Benchmark including planning, perception and execution.**

## 2 Related Work

**Embodied Datasets and Benchmarks** As embodied intelligence evolves, numerous large-scale datasets and benchmarks have emerged to support its core components: perception, planning, and execution. For planning, the Embodied-Reasoner [41] introduces a three-stage process—observation, thinking, and action—for spatial reasoning and task planning. EgoCOT [24] is a chained reasoning dataset designed for complex task execution. In perception, the ShareRobot dataset [13] supports affordance detection and semantic understanding, while Where2Place [36] uses large-scale scene annotations to assist agents in identifying object placements. For execution, Open X-Embodiment [26] consolidates millions of robot trajectories across 22 platforms for cross-platform learning, and RH20T [10] provides human demonstration data for household tasks to aid multi-task imitation learning. Despite these advancements, many datasets remain domain-specific and lack integrative coverage. To address this, we propose the ***Uni-Embodied Benchmark***, the first unified benchmark designed to evaluate all three core capabilities, supporting the development of general-purpose embodied intelligence models.

**Vision-Language Models for Embodied Intelligence** Recent advancements in vision-language models (VLMs) have significantly enhanced robots' planning, perception, and execution capabilities. In planning, EmbodiedGPT [24] connects natural language instructions to executable plans using thought chain reasoning, while Embodied-Reasoner [41] improves planning for interactive tasks through deep reasoning. For perception, RoboBrain [13] uses large-scale datasets and three-stage training to enhance embodied perception, and RoboPoint [36] improves spatial key point prediction with synthetic instruction tuning data. In execution, NaVid [37] employs video frames for end-to-end navigation predictions, while MapNav [38] utilizes structured annotated semantic maps to optimize memory usage. RT-2 [5] excels in manipulation by predicting action sequences through autoregressive poses, and OpenVLA [14] enhances task generalization via large-scale pre-training on the Open-X-Embodiment [26] dataset. However, these models primarily focus on individual capabilities. To address this limitation, we propose the ***Uni-Embodied Benchmark*** for comprehensive evaluation of embodied capabilities.

## 3 Uni-Embodied Benchmark

As shown in Fig. 2, our ***Uni-Embodied Benchmark*** includes 156.5k evaluation samples organized into three components: planning (11k samples), perception (125k samples), and execution (20.5k samples). This benchmark comprehensively evaluates VLM performance across these three core capabilities, ensuring their ability to complete full embodied tasks.

### 3.1 Embodied Planning

The embodied planning component features both complex and simple planning tasks. Complex tasks integrate navigation and manipulation across various robot instances and environments, focusing on multi-step coordination challenges. Simple tasks emphasize desktop-level manipulation with different object categories. Each task uses natural language instructions and visual scenes as input, expecting the model to output a structured sequence of subtasks.

**Complex Planning Tasks** We generated 1,000 complex planning scenarios using Claude for task generation and scene description. These tasks require multi-step coordination between navigation and manipulation, such as "Install the light bulb at the door to the ceiling" or "Retrieve the book from the shelf and place it on the dining table." Our process systematically specifies contextual dimensions, including perspective (first-person or third-person), environment type (simulator, real indoor, or outdoor), and robot embodiment (single-arm wheeled, dual-arm wheeled, or humanoid). This approach ensures diverse task complexity and realistic scenarios. The generated scene description prompt is input into GPT-4o [25] to create corresponding visual scenes, accurately representing all necessary objects and the robot. To ensure high data quality, all scene-task-answer triplets underwent rigorous human expert validation, confirming task feasibility, logical consistency, and appropriate difficulty levels.

**Simple Planning Tasks** We manually curated 10,000 high-quality desktop manipulation samples from the RoboBrain dataset, covering a wide range of basic manipulation scenarios. Our curation focused on tasks with clear planning sequences and diverse object interactions, selecting samples that demonstrate various manipulation primitives, such as grasping, placing, pushing, and rearranging objects from categories like household items, tools, and geometric shapes. Each sample presents a unique manipulation challenge, ranging from simple single-object tasks to complex multi-object coordination. Our selection criteria emphasized task diversity, clear visual observations, and well-defined subtask sequences, enabling comprehensive evaluation of VLMs' embodied planning capabilities. This benchmark effectively assesses the transferability of basic planning skills across different object types and task complexities.

## 3.2 Embodied Perception

The embodied perception component includes five key aspects: navigation trajectory summarization, navigation semantic map understanding, object affordance recognition, spatial pointing, and manipulation trajectory analysis. Sample selection accounts for diverse scene configurations, object types, and task complexities, ensuring a comprehensive evaluation of capabilities.

**Navigation Trajectory Summarization** The navigation trajectory summarization task requires VLMs to generate a natural language description from a series of navigation frames. We collected 10,000 samples from expert trajectories in the R2R-CE [15] dataset to evaluate the model's ability to summarize navigation across various scenarios, task lengths, and instruction complexities. Each sample includes 8-10 carefully selected frames from individual navigation episodes, representing key decision points and environmental transitions. These frames capture critical moments like path initiation, direction changes, landmark recognition, and goal achievement, providing essential context for understanding the trajectory. The corresponding ground truth summary is derived from the original human instruction, detailing the intended navigation in natural language.

**Navigation Semantic Map Understanding** The navigation semantic map understanding task evaluates models' spatial reasoning using annotated semantic maps created from RGB images, depth maps, and odometry data. We collected 10,000 samples from diverse indoor navigation episodes to ensure comprehensive coverage of room types, layouts, and object distributions. Using Claude and GPT-4o [25], we generated annotations covering key aspects of spatial intelligence, including visible object inventories, spatial relationships, room properties, agent positioning, and historical path analysis. Each annotation was rigorously reviewed by human experts for accuracy and consistency.

**Object Affordance Recognition** The object affordance recognition task uses natural language instructions and RGB images as input, requiring the model to output bounding box coordinates indicating operable areas. We created 50,000 samples to help the model identify actionable parts of objects in complex scenes. Each sample includes an instruction describing the desired interaction (*e.g.*, "Which part should I use to lift these cups?"), an RGB image of the target object, and the corresponding affordance bounding box (x, y, width, height). Our data construction relies on the PACO dataset [28], which provides detailed part-based annotations. We extract object part masks and convert them to bounding box coordinates to indicate specific interaction areas. To enhance instruction diversity and quality, we used GPT-4o [25] to generate natural language instructions for various affordance types, such as graspable surfaces and pushable areas. This comprehensive approach enables the model to effectively assess manipulable parts of objects.

**Spatial Pointing** The spatial pointing task requires the model to output pixel coordinate predictions based on a spatial command and an image. We constructed a benchmark from RoboPoint [36], featuring two datasets: RoboRefIt [22] and Where2Place [36]. RoboRefIt [22] focuses on object reference tasks in cluttered images where objects are distinguished by relational references, while Where2Place addresses free space recognition with 100 images from real home and office environments. Each sample includes an RGB image, a natural language command specifying a spatial reference (*e.g.*, "point to the red square on the left," "find the cup near the window"), and the corresponding pixel coordinates of the target location. This task effectively evaluates the spatial intelligence of visual language models in embodied scenes, essential for agents to accurately interpret and respond to location commands.

**Manipulation Trajectory Analysis** The manipulation trajectory analysis task takes a scene image with manipulation instructions and requires the model to predict a sequence of 2D coordinates representing the manipulation path. We curated 5,000 high-quality samples from the RoboBrain dataset [13], each including an RGB image of the manipulation scene, a natural language instruction (*e.g.*, "move the red block to the top left corner"), and the corresponding trajectory coordinates as a sequence of x, y pixel positions. Our selection process emphasized trajectories with clear visual progression and sufficient coordinate density, ensuring each sample contains at least three coordinate pairs and covers various manipulation scenarios. These 2D coordinates represent key points along the manipulation path from the camera perspective, allowing us to evaluate the model's ability to analyze the spatial relationship between instructions and trajectories in pixel space.

## 3.3 Embodied Execution

The execution evaluation focuses on models' action generation capabilities, encompassing three complementary categories: navigation execution, manipulation execution, and real-world execution.

**Navigation Execution** For navigation execution, we selected 100 episodes from the val-unseen split of R2R-CE [15]. At each timestep $t$, the agent obtains the visual observation $o_t$, the robot state $s_t$, and the natural language instruction $I$. The navigation model then outputs the next action $a_t \in \{\text{move\_forward}, \text{turn\_left}, \text{turn\_right}, \text{stop}\}$. This process can be expressed as:

$$a_t = \pi(o_t, s_t, I, h_{t-1}), \tag{1}$$

where $\pi$ is the navigation policy function, $s_t$ is the current robot state, $o_t$ is the visual observation, $I$ is the instruction, and $h_{t-1}$ is the historical state information. The agent performs navigation tasks based on language instructions in an unknown environment through this temporal decision-making process.

**Manipulation Execution** For manipulation execution, we selected 100 episodes from the D part of the CALVIN dataset [23], covering various tasks such as sliding doors, grasping objects, pressing buttons, and opening drawers. At each timestep $t$, the agent receives the task instruction $T$, the current visual observation $o_t$, and the robot state $s_t$ from the CALVIN simulator. Based on these inputs, the model outputs the target pose $p_t = [x, y, z, r, p, y, g]$ for the 7-DOF end effector, where the first three dimensions are position coordinates, the next three represent Euler angles (roll, pitch, yaw), and the last dimension indicates the gripper state. This process can

**Table 1: Performance Comparison of VLMs on Embodied Perception Tasks.**

| VLMs | | Navigation Trajectory Summarization | Navigation Map Understanding | Object Affordance | Spatial Pointing | Manipulation Trajectory Analysis | | |
|---|---|---|---|---|---|---|---|---|
| Type | Models | Similarity Score↑ | Similarity Score↑ | AP↑ | Accuracy↑ | HD↓ | RMSE↓ | DFD↓ |
| Closed-source | GPT-4o [25] | 68.5±2.4 | **85.7±1.8** | **37.2** | **22.17** | 0.158 | 0.145 | 0.112 |
| | Claude-3.7-Sonnet [1] | **72.3±2.1** | 82.1±2.2 | 35.8 | 20.85 | **0.142** | **0.128** | **0.095** |
| | Qwen-VL-Max [33] | 54.1±3.5 | 72.8±3.1 | 28.9 | 17.29 | 0.172 | 0.156 | 0.124 |
| Open-source | Qwen2-VL-72B [33] | **45.2±3.8** | **63.4±3.6** | **24.1** | 15.18 | **0.189** | **0.174** | **0.139** |
| | Qwen2.5-VL-7B [33] | 35.7±4.1 | 55.2±3.8 | 22.1 | **16.22** | 0.188 | 0.178 | 0.142 |
| | Qwen2-VL-7B [33] | 33.4±4.3 | 48.7±4.0 | 19.6 | 15.86 | 0.203 | 0.185 | 0.155 |
| | LLaVA-NeXT-7B [20] | 30.2±4.6 | 52.6±4.2 | 18.2 | 14.73 | 0.215 | 0.194 | 0.168 |

be expressed as:

$$p_t = \phi(o_t, s_t, T, h_{t-1}), \tag{2}$$

where $\phi$ is the operation policy function, $o_t$ is the visual observation, $T$ is the task instruction, $s_t$ is the robot state, and $h_{t-1}$ is historical information.

**Real-World Execution** We conducted 100 real-world experiments, consisting of 50 navigation and 50 manipulation episodes, to evaluate the model's execution performance in physical environments. For navigation, we deployed the Unitree Go2 quadruped robot in five scenarios: office, conference room, pantry, appliance room, and corridor, testing the model's zero-shot capabilities. For manipulation, we utilized the Franka Research 3 robotic arm for tasks such as object grasping, movement, drawer manipulation, and container placement. Baseline models for manipulation were fine-tuned on 200 collected trajectories.

## 3.4 Evaluation Metrics

Our evaluation employs task-specific metrics tailored to each embodied capability.

**Planning, Navigation Trajectory Summarization, and Map Understanding** For these tasks, we use GPT-4o [25] to compute similarity scores between model predictions and ground truth. Following RoboBrain [13], we categorize simple planning tasks into planning, context-aware planning, remaining steps, future predictions, generative affordance, past descriptions, and success (positive/negative), including discriminative affordance (positive/negative).

**Object Affordance Recognition** We use Average Precision (AP) as the primary metric, measuring model performance by calculating the area under the precision-recall curve. AP values are computed at multiple intersection over union (IoU) thresholds and averaged for a comprehensive assessment.

**Spatial Pointing** For spatial pointing tasks, we adopt the RoboPoint evaluation method, focusing on the hit rate, which measures the percentage of predicted points within the true target mask area.

**Manipulation Trajectory Analysis** We compare real and model-predicted trajectories represented as two-dimensional sequences of waypoints. Three distance metrics are employed: Discrete Fréchet Distance (DFD) for shape similarity, Hausdorff Distance (HD) for maximum local deviation, and Root Mean Square Error (RMSE) for average point deviation. Together, these metrics assess trajectory prediction accuracy and consistency.

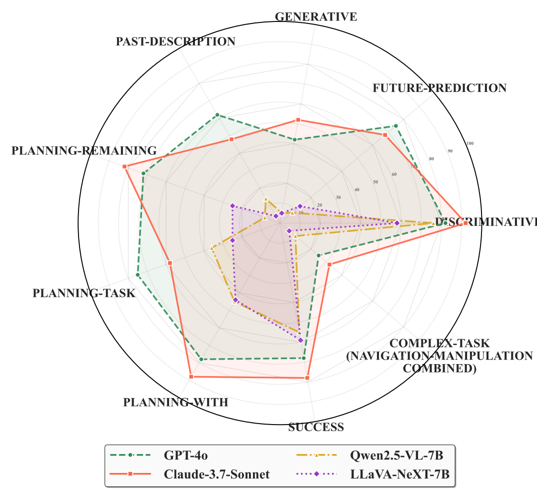


**Figure 3: Performance Comparison of VLMs on Planning Tasks.**

**Navigation and Manipulation Execution** In simulation, navigation tasks utilize VLN-CE metrics: Success Rate (SR), Success weighted by Path Length (SPL), and Navigation Error (NE). For manipulation, we apply CALVIN [23] metrics, including the success rate of consecutive tasks and average task completion (Avg. Len.). In real-world deployment, we use the Mean Success Rate (Mean SR) to measure the proportion of successfully completed tasks, providing a clear performance assessment.

# 4 Experiments

## 4.1 Experimental Details

**Environment Setup**

We evaluate state-of-the-art visual-language models (VLMs) on the ***Uni-Embodied* Benchmark**. Planning and perception tasks use a visual question answering (VQA) approach, while execution tasks employ the Habitat simulator for navigation and the CALVIN [23] simulator for manipulation.

**Implementation Details** We evaluate both closed-source and open-source VLMs for a thorough performance analysis. Closed-source models are accessed via API, while open-source models are deployed on A800 GPUs with optimized inference configurations to ensure efficiency and reliability. All models are assessed using the same prompt format for fair comparison.

**Table 2: Performance Comparison of Models on Navigation.**

| Models | Params | Fine-tuned | Navigation (R2R-CE) | | | |
|---|---|---|---|---|---|---|
| | | | NE ↓ | OSR↑ | SR↑ | SPL↑ |
| GPT-4o [25] | - | ✗ | 12.5 | 10.2 | 6.5 | 3.9 |
| Claude-3.7-Sonnet [1] | - | ✗ | 11.6 | 9.8 | 7.2 | 4.6 |
| Qwen2-VL [33] | 7B | ✗ | 15.8 | 6.0 | 3.2 | 1.3 |
| Navid [37] | 7B | ✓ | 5.12 | 52.3 | **40.8** | **38.7** |
| MapNav [38] | 7B | ✓ | **4.89** | **53.6** | 39.4 | 37.1 |

**Table 3: Performance Comparison of Models on Manipulation.**

| Models | Params | Manipulation (CALVIN) | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1↑ | 2↑ | 3↑ | 4↑ | 5↑ | Avg. Len.↑ |
| RT-1 [6] | 35M | 51.8 | 20.7 | 8.9 | 3.2 | 1.1 | 0.85 |
| RoboFlamingo [17] | 3B | 79.8 | 58.4 | 43.2 | 30.7 | 21.8 | 2.34 |
| GR-1 [34] | 195M | 82.7 | 67.9 | 56.2 | 46.4 | 37.6 | 2.91 |
| OpenVLA [14] | 7B | **88.9** | 74.1 | 58.6 | 49.3 | 41.2 | 3.12 |
| RoboVLMs [16] | 1.7B | 86.4 | **77.5** | **70.8** | **64.9** | **57.8** | **3.28** |

## 4.2 Experimental Results

**Planning** We evaluated four vision-language models (VLMs) on eight simple planning tasks and one complex navigation-manipulation task. As shown in Fig. 3, Claude-3.7-Sonnet [1] achieved the highest average performance (63.6%), excelling in discriminative reasoning (92%) and contextual planning (88%). GPT-4o [25] showed strong temporal reasoning, leading in future prediction (75% vs. 68%) and past description (62% vs. 48%). Smaller models, Qwen2.5-VL-7B [33] and LLaVA-NeXT-7B [20], averaged only 28.6% and 26.6%, with poor temporal reasoning (8-14%). In complex tasks, Claude-3.7-Sonnet (32%) and GPT-4o (25%) outperformed smaller models (10% and 6%), underscoring the importance of model scale.

**Perception** We assessed seven VLMs on five perception tasks. As shown in Tab. 1, Claude-3.7-Sonnet [1] excelled in navigation trajectory summarization (72.3±2.1) and trajectory analysis. GPT-4o led in navigation map understanding (85.7±1.8), object affordance recognition (AP: 37.2), and spatial pointing accuracy (22.17). Among open-source models, Qwen2-VL-72B [33] significantly outperformed smaller variants, achieving a score of 45.2±3.8 compared to 35.7±4.1 for the 7B version, demonstrating a clear scaling advantage. Closed-source models outperformed open-source ones by 20-40% on most tasks, highlighting model scale importance.

**Execution** We evaluated models on execution tasks, including navigation (R2R-CE [15]) and manipulation (CALVIN [23]). As shown in Tab. 2, While GPT-4o [25] and Claude-3.7-Sonnet [1] excel in planning and perception, their execution capabilities are limited, with success rates of 6.5% and 7.2% in navigation tasks. Fine-tuned VLM methods show advantages: Navid [37] achieves the highest success rate (40.8%) and SPL (38.7%), while MapNav [38] excels in navigation error reduction (4.89 NE) and obstacle success rate (53.6% OSR). For manipulation tasks, As shown in Tab. 3, RoboVLMs [16] achieved the best performance in continuous task completion (3.28 in Avg. Len), highlighting the need to bridge high-level reasoning with low-level control.

## 4.3 Ablation Study

**Chain-of-Thought Enhancement** To enhance planning and perception in VLMs, we introduce a Chain of Thoughts (CoT) approach to structure reasoning into steps. For complex planning tasks, CoT identifies required objects, locates them, plans high-level tasks into subtasks, and generates continuous <Navigation> and <Manipulation> instructions through step-by-step reasoning, considering the current position after each action. For navigation trajectory summarization, CoT analyzes semantic objects, identifies key frames, and splits trajectories into smaller sub-trajectories for progressive reasoning. As shown in Tab. 4, we achieved significant improvements: in GPT-4o enhanced with CoT, efficiency in complex planning tasks increased from 25.4% to 32.0% (26% improvement), and navigation trajectory summary improved from 68.5±2.4 to 74.5±2.1 (9% improvement). This approach effectively enables VLMs to focus on key information and current states, enhancing planning capabilities.

**Table 4: Ablation Experiments on Chain-of-Thought.**

| Models | Complex Planning Tasks | Navigation Trajectory Summarization |
|---|---|---|
| | Similarity Score ↑ | Similarity Score ↑ |
| Qwen2.5-VL-7B [33] | 10.2 | 35.7±4.1 |
| GPT-4o [25] | 25.4 | 68.5±2.4 |
| **GPT-4o-CoT (Ours)** | **32.0 (26%↑)** | **74.5±2.1 (9%↑)** |

**Table 5: Ablation Experiments on Hybrid Training.**

| Models | Navigation R2R-CE | Manipulation CALVIN |
|---|---|---|
| | SR ↑ | Avg. Len ↑ |
| RoboVLMs-Navi.[16] | 28.2 | - |
| RoboVLMs-Mani.[16] | - | 3.28 |
| **Uni-Execution (Ours)** | **31.9 (13%↑)** | **3.76 (15%↑)** |

**Hybrid Training** We investigate the effectiveness of hybrid training for navigation and manipulation within a unified VLM backbone. Using RoboVLMs [16] as the base architecture, we compare single-task models (RoboVLMs-Navi. and RoboVLMs-Mani.) with our unified execution model (Uni-Execution), trained jointly on both tasks. As shown in Tab. 5, the hybrid training approach demonstrates mutual enhancement: our Uni-Execution model achieves a 31.9% success rate in R2R-CE [15] (a 13% improvement) and an average task length of 3.76 in CALVIN [23] (a 15% improvement over single-task training). These results suggest that navigation and manipulation tasks are complementary, and shared vision-language alignment and planning capabilities enhance both modalities when trained simultaneously in a consistent architecture.

# 5 Conclusion

This paper introduces ***Uni-Embodied***, the first unified benchmark for evaluating vision-language models (VLMs) across three key dimensions of embodied intelligence: planning, perception, and execution. We extensively assess state-of-the-art open-source and closed-source VLMs on nine tasks, including task planning, navigation trajectory summarization, semantic graph understanding, object affordance recognition, spatial pointing, manipulation trajectory analysis, and execution. Our experiments reveal the strengths and limitations of current methods, highlighting effective enhancement strategies such as CoT enhancement and hybrid training. ***Uni-Embodied*** paves the way for developing general embodied intelligence models that integrate planning, perception, and execution seamlessly.

# References


[1] Anthropic. 2025. Claude 3.7 Sonnet and Claude Code. https://www.anthropic.com/news/claude-3-7-sonnet

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609* (2023).

[3] Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. 2020. Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171* (2020).

[4] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726* (2024).

[5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818* (2023).

[6] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. 2022. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817* (2022).

[7] Angelo Cangelosi, Josh Bongard, Martin H Fischer, and Stefano Nolfi. 2015. Embodied intelligence. *Springer handbook of computational intelligence* (2015), 697–714.

[8] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. 2023. Palm-e: An embodied multimodal language model. (2023).

[9] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. 2022. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence* 6, 2 (2022), 230–244.

[10] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. 2024. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. In *IEEE International Conference on Robotics and Automation*. 653–660.

[11] Jonathan Francis, Nariaki Kitamura, Felix Labelle, Xiaopeng Lu, Ingrid Navarro, and Jean Oh. 2022. Core challenges in embodied vision-language planning. *Journal of Artificial Intelligence Research* 74 (2022), 459–515.

[12] Xiaoshuai Hao, Yunfeng Diao, Mengchuan Wei, Yifan Yang, Peng Hao, Rong Yin, Hui Zhang, Weiming Li, Shu Zhao, and Yu Liu. 2025. MapFusion: A Novel BEV Feature Fusion Network for Multi-modal Map Construction. *arXiv preprint arXiv:2502.04377* (2025).

[13] Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, et al. 2025. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. *arXiv preprint arXiv:2502.21257* (2025).

[14] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246* (2024).

[15] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. 2020. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*. 104–120.

[16] Xinghang Li, Peiyan Li, Minghuan Liu, Dong Wang, Jirong Liu, Bingyi Kang, Xiao Ma, Tao Kong, Hanbo Zhang, and Huaping Liu. 2024. Towards generalist robot policies: What matters in building vision-language-action models. *arXiv preprint arXiv:2412.14058* (2024).

[17] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. 2023. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378* (2023).

[18] Luo Ling and Bai Qianqian. 2025. Endowing Embodied Agents with Spatial Reasoning Capabilities for Vision-and-Language Navigation. *arXiv preprint arXiv:2504.08806* (2025).

[19] Huaping Liu, Di Guo, and Angelo Cangelosi. 2025. Embodied Intelligence: A Synergy of Morphology, Action, Perception and Learning. *Comput. Surveys* (2025).

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.

[21] Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. 2024. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *arXiv preprint arXiv:2407.06886* (2024).

[22] Yuhao Lu, Yixuan Fan, Beixing Deng, Fangfu Liu, Yali Li, and Shengjin Wang. 2023. Vl-grasp: a 6-dof interactive grasp policy for language-oriented objects in cluttered indoor scenes. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*. 976–983.

[23] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. 2022. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters* 7, 3 (2022), 7327–7334.

[24] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. 2023. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *Advances in Neural Information Processing Systems* 36 (2023), 25081–25094.

[25] OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/

[26] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. 2024. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *IEEE International Conference on Robotics and Automation*. 6892–6903.

[27] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824* (2023).

[28] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. 2023. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7141–7151.

[29] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2023. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2998–3009.

[30] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. 2025. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752* (2025).

[31] Yingbo Tang, Shuaike Zhang, Xiaoshuai Hao, Pengwei Wang, Jianlong Wu, Zhongyuan Wang, and Shanghang Zhang. 2025. Affordgrasp: In-context affordance reasoning for open-vocabulary task-oriented grasping in clutter. *arXiv preprint arXiv:2503.00778* (2025).

[32] Liuyi Wang, Zongtao He, Ronghao Dang, Mengjiao Shen, Chengju Liu, and Qijun Chen. 2024. Vision-and-language navigation via causal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13139–13150.

[33] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024).

[34] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. 2023. Unleashing large-scale video generative pre-training for visual robot manipulation. *arXiv preprint arXiv:2312.13139* (2023).

[35] Yujie Wu, Huaihai Lyu, Yingbo Tang, Lingfeng Zhang, Zhihui Zhang, Wei Zhou, and Siqi Hao. 2025. Evaluating GPT-4o's Embodied Intelligence: A Comprehensive Empirical Study. *TechRxiv preprint techrxiv.174495686.69962588/v1* (2025).

[36] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. 2024. Robopoint: A vision-language model for spatial affordance prediction for robotics. *arXiv preprint arXiv:2406.10721* (2024).

[37] Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. 2024. Navid: Video-based vlm plans the next step for vision-and-language navigation. *arXiv preprint arXiv:2402.15852* (2024).

[38] Lingfeng Zhang, Xiaoshuai Hao, Qinwen Xu, Qiang Zhang, Xinyao Zhang, Pengwei Wang, Jing Zhang, Zhongyuan Wang, Shanghang Zhang, and Renjing Xu. 2025. Mapnav: A novel memory representation via annotated semantic maps for vlm-based vision-and-language navigation. *arXiv preprint arXiv:2502.13451* (2025).

[39] Lingfeng Zhang, Hao Wang, Erjia Xiao, Xinyao Zhang, Qiang Zhang, Zixuan Jiang, and Renjing Xu. 2024. Multi-Floor Zero-Shot Object Navigation Policy. *arXiv preprint arXiv:2409.10906* (2024).

[40] Lingfeng Zhang, Qiang Zhang, Hao Wang, Erjia Xiao, Zixuan Jiang, Honglei Chen, and Renjing Xu. 2024. Trihelper: Zero-shot object navigation with dynamic assistance. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*. 10035–10042.

[41] Wenqi Zhang, Mengna Wang, Gangao Liu, Xu Huixin, Yiwei Jiang, Yongliang Shen, Guiyang Hou, Zhe Zheng, Hang Zhang, Xin Li, et al. 2025. Embodied-reasoner: Synergizing visual search, reasoning, and action for embodied interactive tasks. *arXiv preprint arXiv:2503.21696* (2025).

[42] Xinxin Zhao, Wenzhe Cai, Likun Tang, and Teng Wang. 2024. ImagineNav: Prompting Vision-Language Models as Embodied Navigator through Scene Imagination. *arXiv preprint arXiv:2410.09874* (2024).

[43] Xinyu Zheng, Yangfan He, Yuhao Luo, Lingfeng Zhang, Jianhui Wang, Tianyu Shi, and Yun Bai. 2025. Railway Side Slope Hazard Detection System Based on Generative Models. *IEEE Sensors Journal* (2025).