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

Anonymous ACL submission

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.

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) (Cangelosi et al., 2015; Liu et al., 2025; Wu et al., 2025). 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 (Liu et al., 2024; Song et al., 2023; Wang et al., 2024a), perception (Duan et al., 2022; Zhao et al., 2024; Ling and Qianqian, 2025), and execution (Zhang et al., 2024a; Driess et al., 2023). ***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*** 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) (Bai et al., 2023; Beyer et al., 2024; Ji et al., 2025; Peng et al., 2023) show significant promise for embodied intelligence applications (Zhang et al., 2024b, 2025a). 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

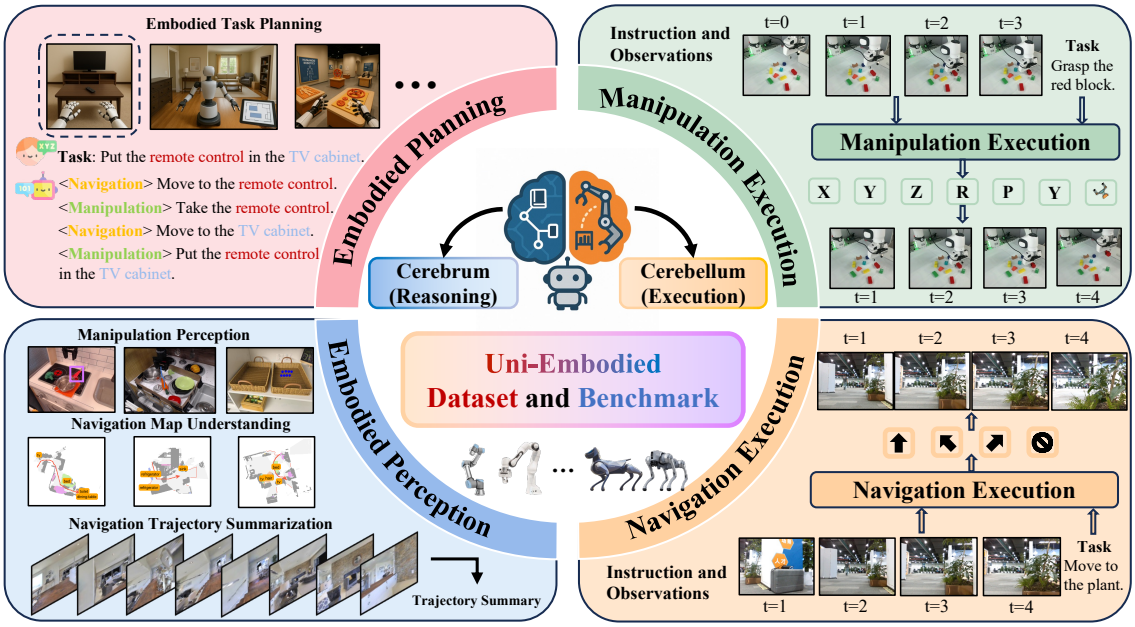

Figure 1: **Overview of the Uni-Embodied Benchmark.** It encompasses three key dimensions: planning, perception, and execution.

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 planning and execution capabilities.

The contributions of this paper are as follows:

- We propose the ***Uni-Embodied Benchmark***, the first comprehensive benchmark for evaluating three core embodied capabilities, which are 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*** for complex planning task and ***hybrid training*** for embodied navigation and manipulation.

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 (Zhang et al., 2025b) introduces a three-stage process—observation, thinking, and action—for spatial reasoning and task planning. EgoCOT (Mu et al., 2023) is a chained reasoning dataset designed for complex task execution. In perception, the ShareRobot dataset (Ji et al., 2025) supports affordance detection and semantic understanding, while Where2Place (Yuan et al., 2024) uses large-scale scene annotations to assist agents in identifying object placements. For execution, Open X-Embodiment (O'Neill et al., 2024) consolidates millions of robot trajectories across 22 platforms for cross-platform learning, and RH20T (Fang et al., 2024) 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 (Mu et al., 2023) connects natural language instructions to executable plans using thought chain reasoning, while Embodied-Reasoner (Zhang et al., 2025b) improves planning for interactive tasks through deep reasoning. For perception, RoboBrain (Ji et al., 2025) uses large-scale datasets and three-stage training to enhance embodied perception, and RoboPoint (Yuan et al., 2024) improves spatial key point prediction with synthetic instruction tuning data. In execution, NaVid (Zhang et al., 2024a) employs video frames for end-to-end navigation predictions, while MapNav (Zhang et al., 2025a) utilizes structured annotated semantic maps to optimize memory usage. RT-2 (Brohan et al., 2023) excels in manipulation by predicting action sequences through autoregressive poses, and OpenVLA (Kim et al., 2024) enhances task generalization via large-scale pre-training on the Open-X-Embodiment (O'Neill et al., 2024) 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), perception (125k), and execution (20.5k). 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 (OpenAI, 2024) 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

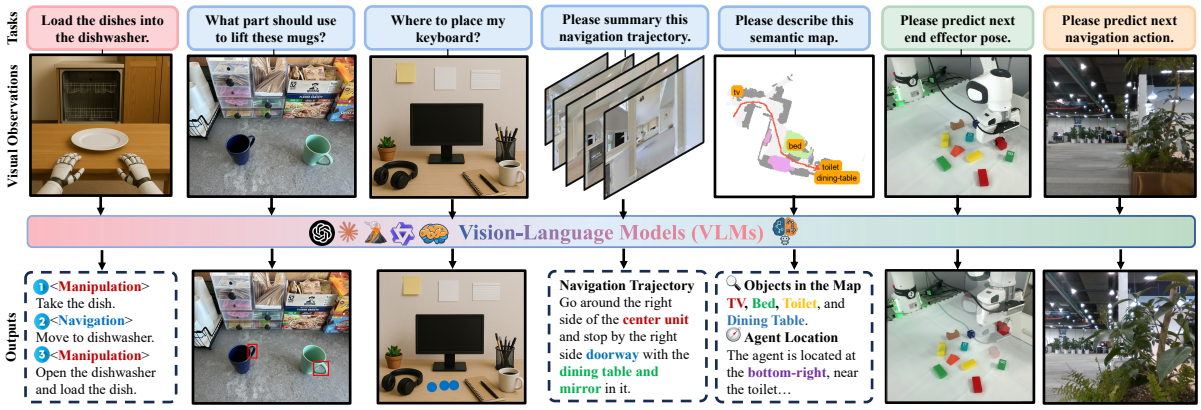

Figure 2: Tasks in Uni-Embodied Benchmark including planning, perception and execution.

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 (Krantz et al., 2020) 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 10k samples from diverse indoor navigation episodes to ensure comprehensive coverage of room types, layouts, and object distributions. Using Claude (Anthropic, 2025) and GPT-4o (OpenAI, 2024), 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 (Ramanathan et al., 2023), 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 (OpenAI, 2024) 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 (Yuan et al., 2024), featuring two datasets: RoboRefIt (Lu et al., 2023) and Where2Place (Yuan et al., 2024). RoboRefIt (Lu et al., 2023) 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 (Ji et al., 2025), 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 (Krantz et al., 2020). 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 π 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 (Mees et al., 2022), 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 be expressed as:

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

where ϕ 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.

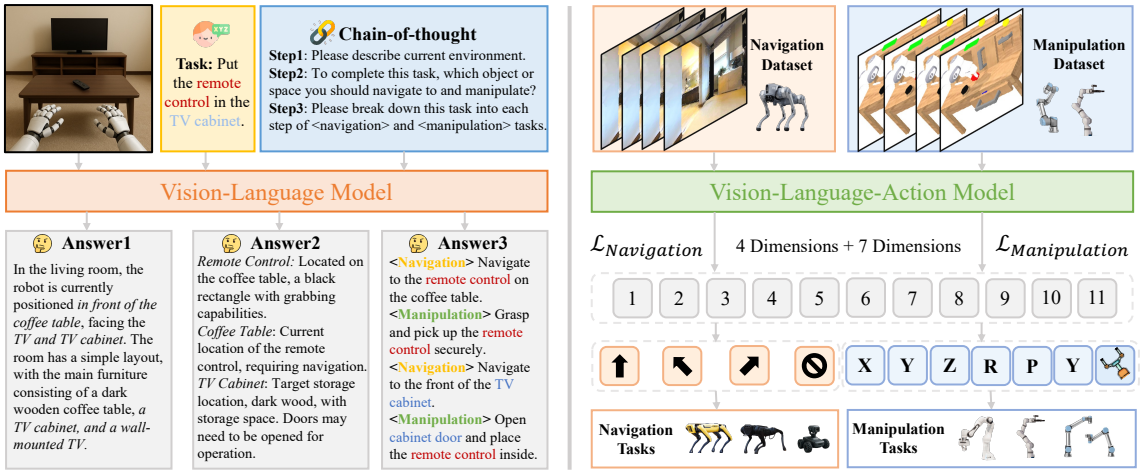

Figure 3: **Our effective strategies for enhancing different embodied capabilities, including chain-of-thought (left) and hybrid training (right).** Chain-of-thought enhance the planning capability of VLM while hybrid training enhance the execution capability of VLA.

Planning, Navigation Trajectory Summarization, and Map Understanding For these tasks, we use GPT-4o (OpenAI, 2024) to compute similarity scores between model predictions and ground truth. Following RoboBrain (Ji et al., 2025), 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 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.

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 (Mees et al., 2022) 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.

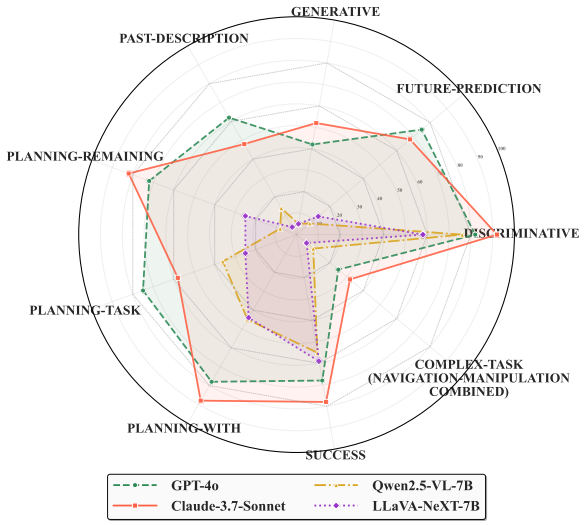

Figure 4: Performance Comparison of VLMs on Planning Tasks.

4 Effective Strategies for Enhancing Embodied Capabilities

Chain-of-Thought for Complex Planning Complex embodied tasks that combine navigation and manipulation require advanced reasoning to break down high-level instructions into executable action sequences. To enhance planning performance of VLMs, we introduce a structured Chain of Thought (CoT) prompting strategy that guides the model through incremental reasoning. Our approach divides the planning process into three stages: (1) En-

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 (OpenAI, 2024)	68.5±2.4	**85.7±1.8**	**37.2**	**22.17**	0.158	0.145	0.112
	Claude-3.7-Sonnet (Anthropic, 2025)	**72.3±2.1**	82.1±2.2	35.8	20.85	**0.142**	**0.128**	**0.095**
	Qwen-VL-Max (Wang et al., 2024b)	54.1±3.5	72.8±3.1	28.9	17.29	0.172	0.156	0.124
Open-source	Qwen2-VL-72B (Wang et al., 2024b)	**45.2±3.8**	**63.4±3.6**	**24.1**	15.18	**0.189**	**0.174**	**0.139**
	Qwen2.5-VL-7B (Wang et al., 2024b)	35.7±4.1	55.2±3.8	22.1	**16.22**	0.188	0.178	0.142
	Qwen2-VL-7B (Wang et al., 2024b)	33.4±4.3	48.7±4.0	19.6	15.86	0.203	0.185	0.155
	LLaVA-NeXT-7B (Liu et al., 2023)	30.2±4.6	52.6±4.2	18.2	14.73	0.215	0.194	0.168

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

Models	Params	Fine-tuned	Navigation (R2R-CE)			
			NE↓	OSR↑	SR↑	SPL↑
GPT-4o (OpenAI, 2024)	-	✗	12.5	10.2	6.5	3.9
Claude-3.7-Sonnet (Anthropic, 2025)	-	✗	11.6	9.8	7.2	4.6
Qwen2-VL (Wang et al., 2024b)	7B	✗	15.8	6.0	3.2	1.3
Navid (Zhang et al., 2024a)	7B	✓	5.12	52.3	**40.8**	**38.7**
MapNav (Zhang et al., 2025a)	7B	✓	**4.89**	**53.6**	39.4	37.1

vironment Perception, where the model analyzes the scene, identifies relevant objects, and understands spatial relationships; (2) Object and Space Recognition, where it determines which objects to manipulate and which spaces to traverse; and (3) Subtask Generation, where it synthesizes these analyses to create a structured sequence of <navigation> and <manipulation> actions. This approach updates the agent's position after each action, ensuring coherent multi-step planning.

Hybrid Training for Embodied Execution Existing vision-language-action (VLA) models often specialize in either navigation or manipulation tasks. Recognizing the shared capabilities in vision-language alignment, spatial reasoning, and sequential decision-making, we propose a hybrid training strategy that integrates both tasks within a single VLA model. The main challenge is reconciling different action spaces: navigation uses discrete commands (e.g., go forward, turn), while manipulation relies on a continuous 7-DOF end-effector pose $[x, y, z, r, p, y, g]$. We address this by designing a unified action head architecture with a shared backbone and task-specific output branches, allowing the model to learn universal representations while retaining task-specific controls. Our training paradigm optimizes navigation datasets (e.g., R2R-CE) and manipulation datasets (e.g., CALVIN) using a combined loss function:

$$\mathcal{L} = \mathcal{L}_{Navigation}(4D) + \mathcal{L}_{Manipulation}(7D).$$

This approach enables knowledge transfer between tasks, as spatial understanding from navigation improves manipulation planning, while fine control from manipulation enhances navigation accuracy.

5 Experiments

5.1 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 (Mees et al., 2022) 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.

5.2 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.4, Claude-3.7-Sonnet(Anthropic, 2025) achieved the highest average performance (63.6%), excelling in discriminative reasoning (92%) and contextual planning (88%). GPT-4o (OpenAI, 2024) demonstrated strong temporal reasoning, leading in future prediction (75% vs. 68%) and past description (62% vs. 48%). Smaller models, Qwen2.5-VL-7B (Wang et al., 2024b) and LLaVA-NeXT-7B (Liu et al., 2023), 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%) significantly outperformed smaller models (10% and 6%), highlighting the importance of model scale.

Perception We assessed 7 VLMs on five perception tasks. As shown in Tab. 1, Claude-3.7-Sonnet (Anthropic, 2025) excelled in navigation trajectory summarization (72.3±2.1) and trajectory analysis. GPT-4o led in navigation map un-

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

Models	Params	Manipulation (CALVIN)					
		1↑	2↑	3↑	4↑	5↑	Avg. Len.↑
RT-1 (Brohan et al., 2022)	35M	51.8	20.7	8.9	3.2	1.1	0.85
RoboFlamingo (Li et al., 2023)	3B	79.8	58.4	43.2	30.7	21.8	2.34
GR-1 (Wu et al., 2023)	195M	82.7	67.9	56.2	46.4	37.6	2.91
OpenVLA (Kim et al., 2024)	7B	**88.9**	74.1	58.6	49.3	41.2	3.12
RoboVLMs (Li et al., 2024)	1.7B	86.4	**77.5**	**70.8**	**64.9**	**57.8**	**3.28**

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

Models	Complex Planning Tasks
	Similarity Score ↑
Qwen2.5-VL-7B (Wang et al., 2024b)	10.2
GPT-4o (OpenAI, 2024)	25.4
GPT-4o w/ CoT (Ours)	**32.0 (26%↑)**

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

Models	Navigation R2R-CE	Manipulation CALVIN
	SR ↑	Avg. Len ↑
RoboVLMs-Navi.(Li et al., 2024)	28.2	-
RoboVLMs-Mani.(Li et al., 2024)	-	3.28
Uni-Execution (Ours)	**31.9 (13%↑)**	**3.76 (15%↑)**

derstanding (85.7±1.8), object affordance recognition (AP: 37.2), and spatial pointing accuracy (22.17). Among open-source models, Qwen2-VL-72B (Wang et al., 2024b) significantly outperformed smaller variants, achieving 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 (Krantz et al., 2020)) and manipulation (CALVIN (Mees et al., 2022)). As shown in Tab. 2, While GPT-4o (OpenAI, 2024) and Claude-3.7-Sonnet (Anthropic, 2025) 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 (Zhang et al., 2024a) achieves the highest success rate (40.8%) and SPL (38.7%), while MapNav (Zhang et al., 2025a) excels in navigation error reduction (4.89 NE) and obstacle success rate (53.6% OSR). For manipulation tasks, As shown in Tab. 3, RoboVLMs (Li et al., 2024) 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.

5.3 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, identifies semantic objects, and generates continuous <Navigation> and <Manipulation> instructions through step-by-step reasoning, considering the current position after each action. 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). This approach effectively enables VLMs to focus on key information and current states, enhancing planning capabilities.

Hybrid Training We investigate the effectiveness of hybrid training for navigation and manipulation within a unified VLM backbone. Using RoboVLMs (Li et al., 2024) 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 model achieves a 31.9% success rate in R2R-CE (Krantz et al., 2020) (a 13% improvement) and an average task length of 3.76 in CALVIN (Mees et al., 2022) (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.

6 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.

Limitations

While Uni-Embodied provides a comprehensive benchmark for evaluating embodied intelligence, it also has several limitations. First, execution evaluation is limited by hardware dependencies and computational resources. The benchmark relies on simulated environments (Habitat and CALVIN) for large-scale execution testing. While necessary for scalability, this introduces a gap between simulation and reality and may not perfectly reflect real-world performance. Second, our evaluation framework primarily assesses the planning and perception capabilities of VLMs through VQA-like tasks, but cannot directly measure end-to-end execution capabilities of vision-language-action (VLA) models in a unified manner. This separation means that models that excel at high-level reasoning may not be able to translate their capabilities to low-level control, and our benchmark cannot fully capture this integration challenge in a single evaluation pipeline. Future work should expand hardware diversity and develop more unified evaluation protocols to seamlessly evaluate entire process from reasoning to execution.

References

Anthropic. 2025. Claude 3.7 sonnet and claude code.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, and 1 others. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, and 1 others. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*.

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

Angelo Cangelosi, Josh Bongard, Martin H Fischer, and Stefano Nolfi. 2015. Embodied intelligence. *Springer handbook of computational intelligence*, pages 697–714.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, and 1 others. 2023. Palm-e: An embodied multimodal language model.

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):230–244.

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*, pages 653–660.

Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, and 1 others. 2025. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. *arXiv preprint arXiv:2502.21257*.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, and 1 others. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.

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*, pages 104–120.

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*.

Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, and 1 others. 2023. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378*.

Luo Ling and Bai Qianqian. 2025. Endowing embodied agents with spatial reasoning capabilities for vision-and-language navigation. *arXiv preprint arXiv:2504.08806*.

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

Huaping Liu, Di Guo, and Angelo Cangelosi. 2025. Embodied intelligence: A synergy of morphology, action, perception and learning. *ACM Computing Surveys*.

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*.

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*, pages 976–983.

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):7327–7334.

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*, 36:25081–25094.

OpenAI. 2024. Hello gpt-4o.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, and 1 others. 2024. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *IEEE International Conference on Robotics and Automation*, pages 6892–6903.

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*.

Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, and 1 others. 2023. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7141–7151.

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*, pages 2998–3009.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

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*.

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*.

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*.

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

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

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

Wenqi Zhang, Mengna Wang, Gangao Liu, Xu Huixin, Yiwei Jiang, Yongliang Shen, Guiyang Hou, Zhe Zheng, Hang Zhang, Xin Li, and 1 others. 2025b. Embodied-reasoner: Synergizing visual search, reasoning, and action for embodied interactive tasks. *arXiv preprint arXiv:2503.21696*.

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*.

Appendix

This supplementary material provides additional details on the proposed method and experimental results that could not be included in the main manuscript due to page limitations. Specifically, this appendix is organized as follows:

- Sec. A offers a detailed overview of the dataset construction process.
- Sec. B outlines the detailed hybrid training strategy and unified action space design.
- Sec. C describes our real-world experiment setup process and data collection protocols.
- Sec. D presents additional qualitative results in VQA-tasks, embodied execution in both simulator and real-world environments.

A Dataset Construction Details

Complex Planning Task The construction of complex planning tasks involves a multi-stage process designed to generate diverse, realistic, and challenging scenarios requiring coordinated navigation and manipulation. We use Claude-3.7-Sonnet (Anthropic, 2025) as the primary generator, creating high-level task instructions and corresponding realistic subtask sequences. The generation process is structured using carefully crafted prompts that specify multiple contextual dimensions, including viewpoint (first-person or third-person), environment type (simulator-based, real indoor or outdoor), robot embodiment (single-arm wheeled, dual-arm wheeled, or humanoid), and task complexity (number of objects, required action sequence, and spatial reasoning requirements). Each generated example consists of a high-level task instruction in natural language (e.g., "Install the lightbulb on the door leading to the ceiling"), a detailed scene description including room layout and object locations, and a ground-truth subtask sequence alternating between "<navigate>" and "<manipulate>" actions. This process generates 1,200 candidate scenarios, systematically sampling to ensure diversity across all contextual dimensions.

To create realistic visual representations corresponding to these text descriptions, we leveraged GPT-4o's (OpenAI, 2024) image generation capabilities. Using Claude-generated scene descriptions as input prompts, we generated 1024×1024 pixel images that accurately depicted all mentioned objects, clearly displayed the specified robot morphology, and maintained spatial accuracy with the text descriptions.

We implemented a two-stage validation process, consisting of automatic verification using an object detection model to check the presence of all objects, followed by manual inspection to ensure that spatial relationships matched the descriptions. Images that failed in either stage were regenerated up to three times. All generated scene-task-answer triples were rigorously validated by three human experts with backgrounds in robotics and embodied AI, who assessed task feasibility, logical consistency, visual-text alignment, appropriateness of difficulty, and spatial reasoning requirements. Each triple was independently reviewed by two experts, with disagreements resolved through discussion with a third expert, resulting in an inter-annotator agreement of Cohen's = 0.82. Of the initial 1,200 candidate proposals, 634 (52.8%) were accepted without modification, 366 (30.5%) were modified and accepted, and 200 (16.7%) were rejected, resulting in a dataset of 1,000 high-quality complex planning scenarios. Common reasons for modification included incorrect object positioning in the generated imagery (45%), missing intermediate navigation steps (28%), unrealistic operational requirements (18%), and ambiguous spatial references (9%). The detailed prompt for task and image generation are as follows:

GPT-4o Image Generation Prompt

Generate a photorealistic image based on the following scene description:
Scene Description: [Insert Claude-generated scene description]
Image Requirements:

- Resolution: 1024×1024 pixels
- Style: [Photorealistic for real environments / Simulator-style for simulated scenes]
- Perspective: [First-person view from robot / Third-person view showing robot in scene]
- Include all objects mentioned: [List of objects from scene description]
- Robot embodiment: [Single-arm wheeled/Dual-arm wheeled/Humanoid] - clearly visible
- Maintain spatial accuracy: Object positions and relationships must match the textual description
- Ensure scene is well-lit and all objects are clearly identifiable

Critical: The generated image must accurately depict the spatial layout described in the text, with all objects in their specified locations.

Claude Generation Prompt for Complex Planning Tasks

Context: [Perspective: First-person/Third-person], [Environment: Simulator/Indoor/Outdoor], [Robot Type: Single-arm wheeled/Dual-arm wheeled/Humanoid]
Task: Generate a complex embodied task that requires both navigation and manipulation.
Requirements:

- Generate a high-level task instruction in natural language
- Provide a detailed scene description including:
  - Room layout and dimensions
  - All object locations with spatial relationships
  - Robot's initial position and orientation
  - Environmental features (furniture, obstacles, landmarks)
- Generate ground-truth subtask sequence using <Navigation> and <Manipulation> tags
- Ensure the task involves 3-8 action steps and 2-5 objects
- Verify logical feasibility and physical plausibility for the specified robot

Output Format:

- **High-level Task:** [Natural language instruction]
- **Scene Description:** [Detailed textual description]
- **Subtasks:**
  - <Navigation> [Target location and approach]
  - <Manipulation> [Specific action and object]
  - ...

Navigation Map Understanding The navigation map understanding task evaluates the spatial reasoning capabilities of the VLM using annotated semantic maps constructed from RGB images, depth maps, and odometry data. We collected 10,000 samples from various indoor navigation scenarios in the Habitat simulator, ensuring comprehensive coverage of different room types (bedroom, kitchen, living room, bathroom), layout configurations (open-plan, corridor-style, multi-room), and object distribution (sparse to dense). For each navigation scenario, we selected the optimal timestep where the semantic map provides the most informative spatial context—typically the midpoint of a trajectory, when the agent has explored enough area to construct a meaningful map while still having navigable space ahead. The semantic maps are generated using the MapNav method, which constructs a top-down, two-dimensional representation using color-coded semantic categories to represent different object types and room regions. We generated comprehensive textual annotations for each map, covering multiple aspects of spatial intelligence, using Claude (Anthropic, 2025) and GPT-4o (OpenAI, 2024). The annotation process uses text boxes directly on the map image to highlight and describe key spatial elements, creating a richly annotated visual text dataset. Each annotation includes a list of visible objects and their semantic categories, the spatial relationships between objects and rooms (e.g., "the sofa is adjacent to the TV stand"), room properties (such as size and connectivity), the agent's current position and orientation on the map, and an analysis of historical exploration paths (where available). All annotations were rigorously reviewed by human experts to ensure accuracy, completeness, and consistency across the dataset. The detailed prompt for annotation generation is as follows:

Annotation Generation Tips for Navigation Map Understanding

Task: Generate comprehensive annotations for a given semantic navigation map, highlighting key spatial information using text boxes. **Input:** Semantic map image showing color-coded object, room, and agent locations **Annotation requirements:**

- **Object list:** List all visible objects in the map and their semantic categories
  - Format: "Objects in the map: [object 1], [object 2],..."
  - Includes furniture, appliances, and other noteworthy items
- **Spatial relations:** Describes the spatial relationships between objects and rooms
  - Uses directional terms: left, right, above, below, adjacent, opposite
  - Example: "The bed is in the upper left corner of the room, adjacent to the nightstand"
- **Room attributes:** Identify room types and their characteristics
  - Room types: bedroom, kitchen, bathroom, living room, etc.
  - Connectivity: Which rooms are connected to each other.
- **Agent location:** Describes the agent's current location and orientation.
  - Using map coordinates: "The agent is located in the lower right corner, near the restroom."
  - Direction: "Facing the dining table."
- **History path (if visible):** Analyzes exploration trajectories.
  - Describes visited areas.
  - Identifies unexplored areas.

Output format: Generates natural language annotations with clear section headings. Annotations should be comprehensive yet concise, focusing on information relevant to navigation planning.

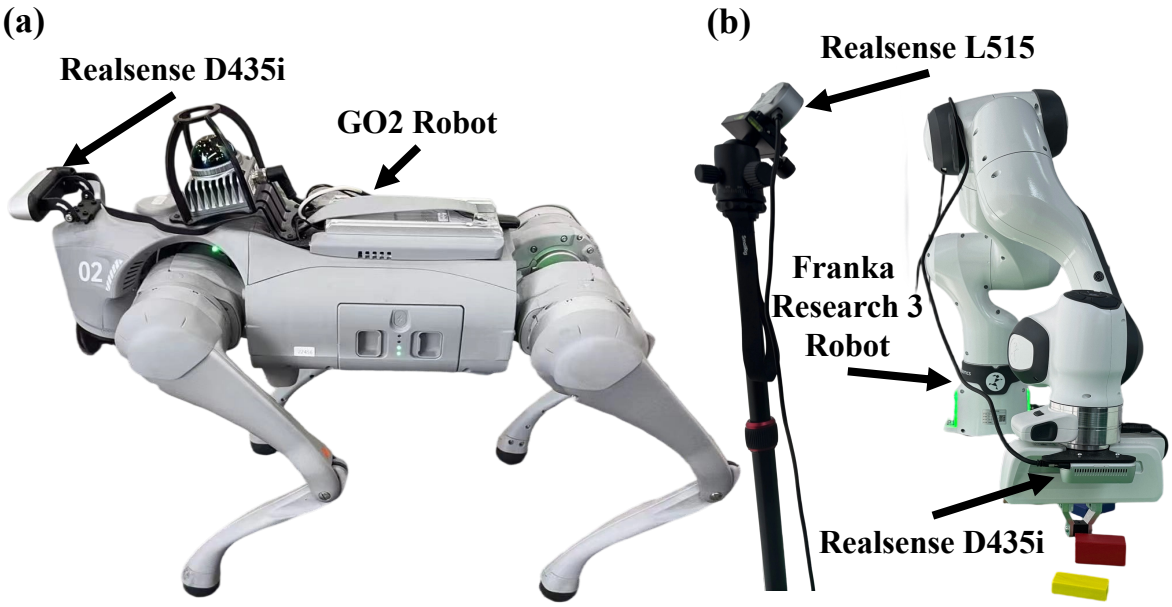

Figure 5: **Real-World Robot Platforms**. (a) GO2 robot platform for navigation tasks. (b) Franka 3 robot platform for manipulation tasks.

B Hybrid Training Details

Our hybrid training strategy integrates navigation and manipulation datasets for unified embodied intelligence learning. We construct a comprehensive hybrid dataset by combining samples from the R2R-CE navigation benchmark (Krantz et al., 2020) and the CALVIN manipulation benchmark (Mees et al., 2022). We implement a balanced sampling strategy to ensure that no single task dominates the training process. The hybrid dataset contains approximately 500,000 navigation trajectory samples from R2R-CE and 500,000 manipulation demonstration samples from the CALVIN A/B/C cohorts, for a total of approximately 1 million training samples. Each sample in the hybrid dataset contains task-specific observations and corresponding action labels. Navigation samples consist of egocentric RGB images (224×224 pixels) with discrete 4D motions, while manipulation samples consist of dual camera views (200×200 static and 84×84 wrist pixels) with continuous 7-DOF motion sequences. To enable seamless training across domains, we design a unified 11-dimensional action space:

$$a_{unified} = [a_{nav}, a_{manip}] \in R^{11},$$

where $a_{nav} \in R^4$ represents navigation actions (forward, turn left, turn right, stop), encoded as a one-hot vector; $a_{manip} \in R^7$ represents manipulation actions, encoded as continuous values (position delta $[\Delta x, \Delta y, \Delta z]$, orientation $[\alpha, \beta, \gamma]$, and gripper state g). During training, task-specific conditional masks are applied to ensure the correct actions are generated for each domain. For navigation tasks, actions are masked to $a_{exec} = [a_{nav}, \mathbf{0}_7]$; for manipulation tasks, actions are masked to $a_{exec} = [\mathbf{0}_4, a_{manip}]$. This masking mechanism enables the model to learn shared visual-language representations while maintaining task-specific action generation capabilities. Our balanced sampling protocol alternates between navigation and manipulation samples during batch construction, maintaining a 1:1 sampling ratio, with a batch size of 32 per GPU (16 navigation samples + 16 manipulation samples). The task-adaptive loss function weights the contribution of each domain:

$$\mathcal{L}_{unified} = \lambda_{nav}\mathcal{L}_{nav} + \lambda_{manip}\mathcal{L}_{manip},$$

where $\mathcal{L}_{nav} = \mathcal{L}_{focal}(a^{1:4}_{pred}, a^{1:4}_{gt})$ Focal Loss is used to address the class imbalance problem in navigation, $\mathcal{L}_{manip} = \mathcal{L}_{huber}(a^{5:10}_{pred}, a^{5:10}_{gt}) + \beta\mathcal{L}_{BCE}(a^{11}_{pred}, a^{11}_{gt})$ combines the pose estimation loss with the gripper state prediction loss. Through ablation experiments with different $\lambda_{nav}/\lambda_{manip}$ ratios, we found that a ratio of 1.0 achieves the best performance in both tasks, significantly outperforming independent training methods. This hybrid training strategy enables effective knowledge transfer between navigation and manipulation domains, leveraging their complementary cognitive demands to enhance overall embodied intelligence capabilities.

C Real-world Experimental Details

To validate VLA's performance in real-world environments, we conducted experiments on navigation and manipulation tasks using various robotic platforms. As shown in Fig. 5, for navigation, we deployed a Unitree GO2 quadruped robot equipped with a Realsense D435i camera, which provides egocentric RGB observation data at a resolution of 640×480. The robot was evaluated in five different environments (office, conference room, corridor, tea room, and appliance area). Each scenario consisted of five command execution tasks, each tested across ten deployments, for a total of 50 navigation scenarios. We employed a zero-shot transfer protocol, directly deploying a simulation-trained model without fine-tuning, to evaluate its generalization capabilities. For manipulation, we used a 7-DOF Franka Research 3 robotic arm equipped with two cameras: a static Realsense L515 (third-person view, 1024×768) and a wrist-worn Realsense D435i (end-effector view, 640×480). Using kinesthetic instruction, we collected 200 real-world trajectories across five task categories: object grasping, relocalization, drawer manipulation, drawer placement, and container placement. These trajectories (160 for training and 40 for validation) were used to fine-tune all models to ensure fair comparison. Each manipulation task category was evaluated using 10 rolling tests covering various object positions, lighting conditions, and backgrounds, for a total of 50 manipulation scenarios. Success criteria for navigation included reaching the target within 1 meter with the correct orientation and collision-free execution; while manipulation required successful grasping, accurate placement (within 5 cm), appropriate interaction, and correct gripper state. Both systems ran at 10 Hz on the NVIDIA Jetson AGX Orin embedded platform, with an inference time of 80-100 milliseconds per timestep. Safety measures included emergency stops and speed limits (0.5 m/s for navigation and 0.3 m/s for manipulation).

D Additional Visualization Results

As shown in Fig. 6- 13, this section presents comprehensive visualizations of many tasks in our Uni-Embodied benchmark, demonstrating the model's capabilities in planning, perception, and execution. For complex planning tasks, we visualize various scenes with different robot implementations and viewpoints, showing how the model decomposes high-level instructions into structured navigation and manipulation sequences. The navigation trajectory summary example presents 8-10 keyframes and their corresponding natural language descriptions, demonstrating spatial understanding and temporal reasoning. The semantic map understanding visualization presents a colored top-down map with text annotations, covering the object inventory, spatial relationships, and agent position. Object affordance recognition results present RGB images with predicted bounding boxes, highlighting graspable or manipulable areas for various household objects and interaction types. The manipulation trajectory analysis visualization overlays the predicted 2D coordinate sequence onto a tabletop scene, with the predicted path highlighted in red. Simulated execution results include navigation trajectories in the Habitat environment from an egocentric perspective and manipulation sequences in CALVIN from a third-person camera perspective, demonstrating the ability to successfully complete tasks under different instructions. Real-world deployment visualizations show a Unitree GO2 quadruped robot executing navigation commands in office and corridor environments, and a Franka Research 3 robotic arm performing manipulation tasks under various background and lighting conditions, validating simulation-to-reality transfer and robust performance in unstructured real-world environments. These comprehensive visualizations provide qualitative evidence demonstrating the effectiveness of our unified framework across all dimensions of embodied intelligence.

Figure 6: More Complex Planning Task Visualization.

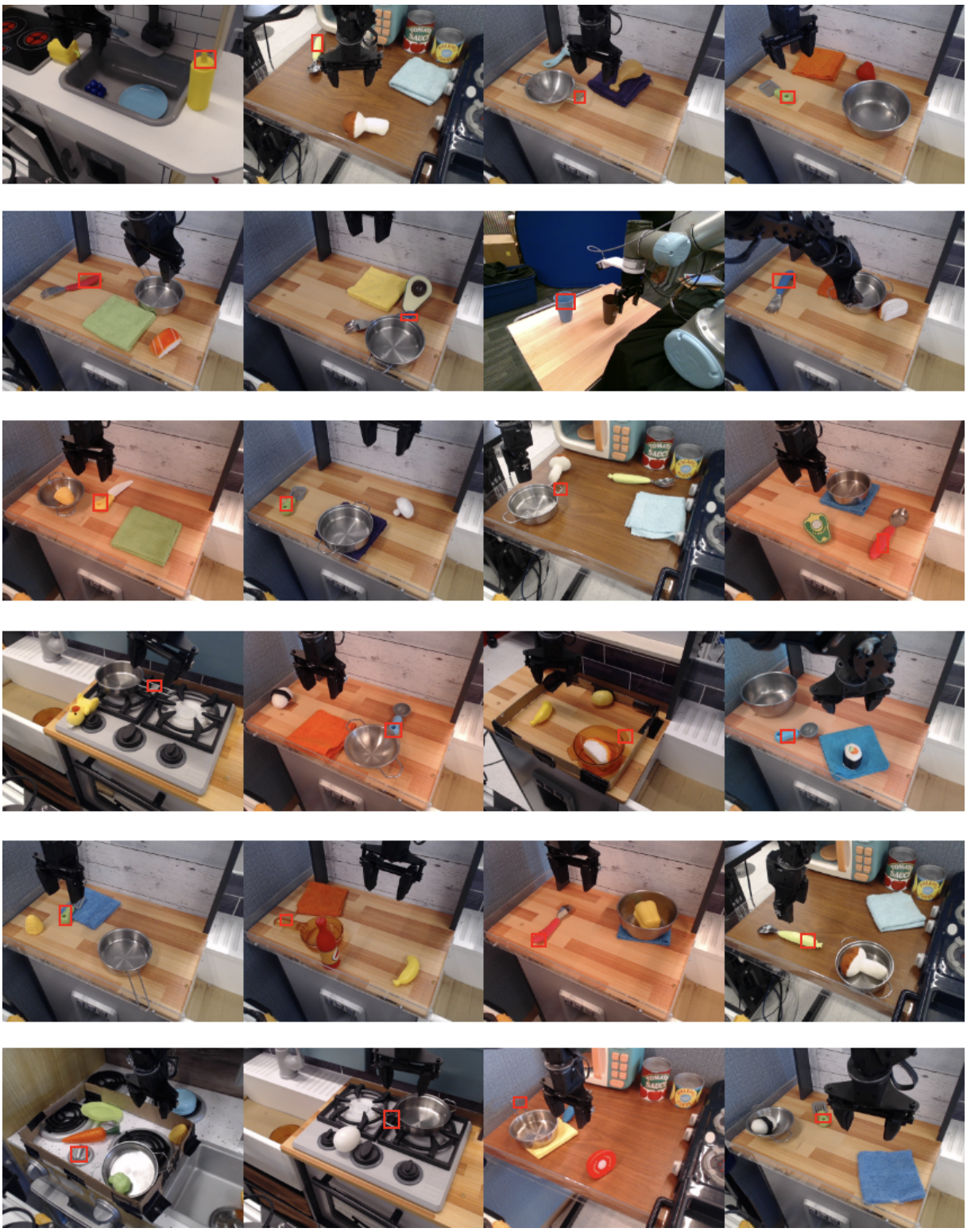

Figure 7: More Object Affordance Recognition Task Visualization.

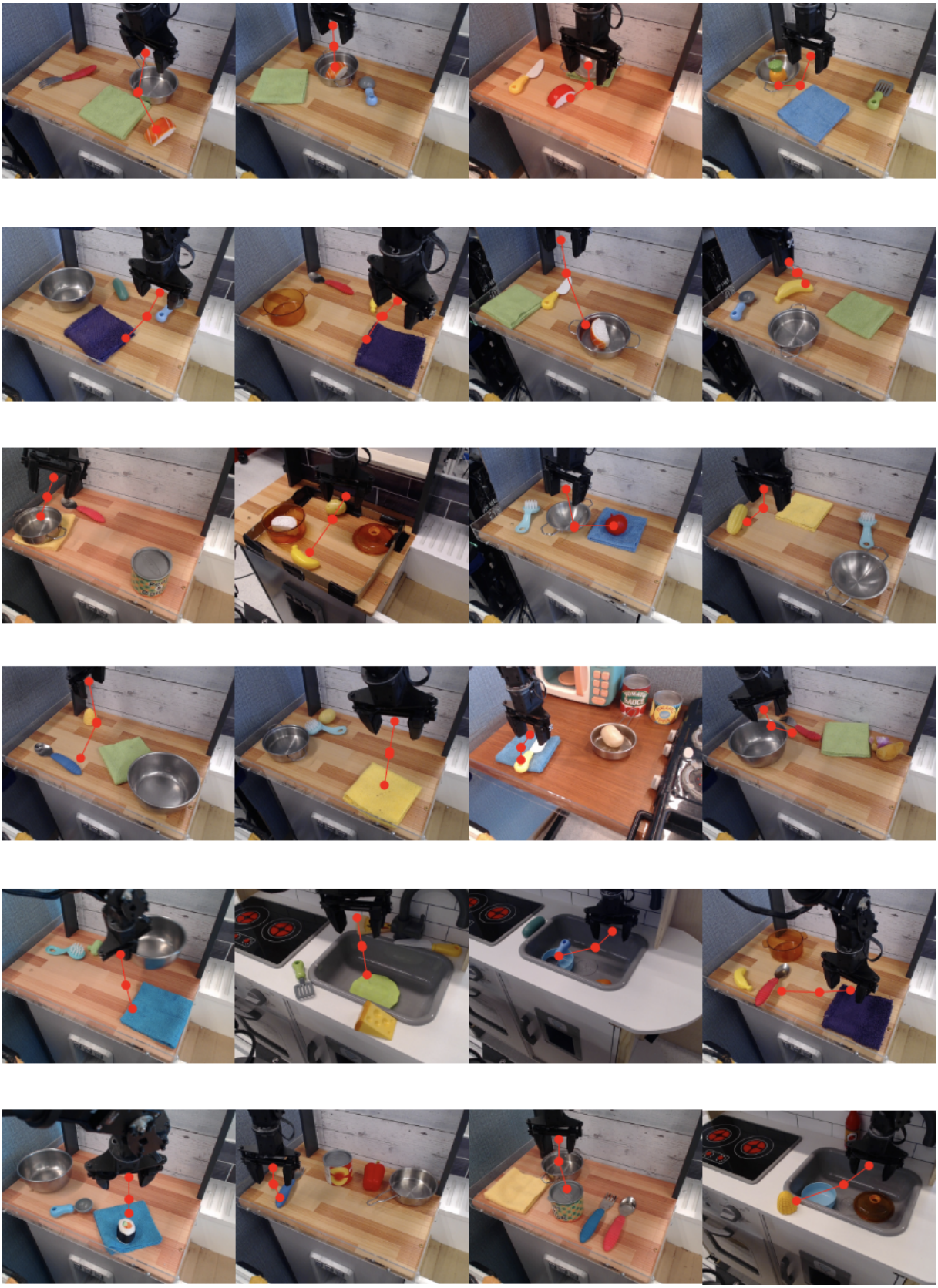

Figure 8: More Manipulation Trajectary Analysis Task Visualization.

Figure 9: More Navigation Semantic Map Understanding Visualization.

Figure 10: More Navigation Trajectory Summarization Visualization.

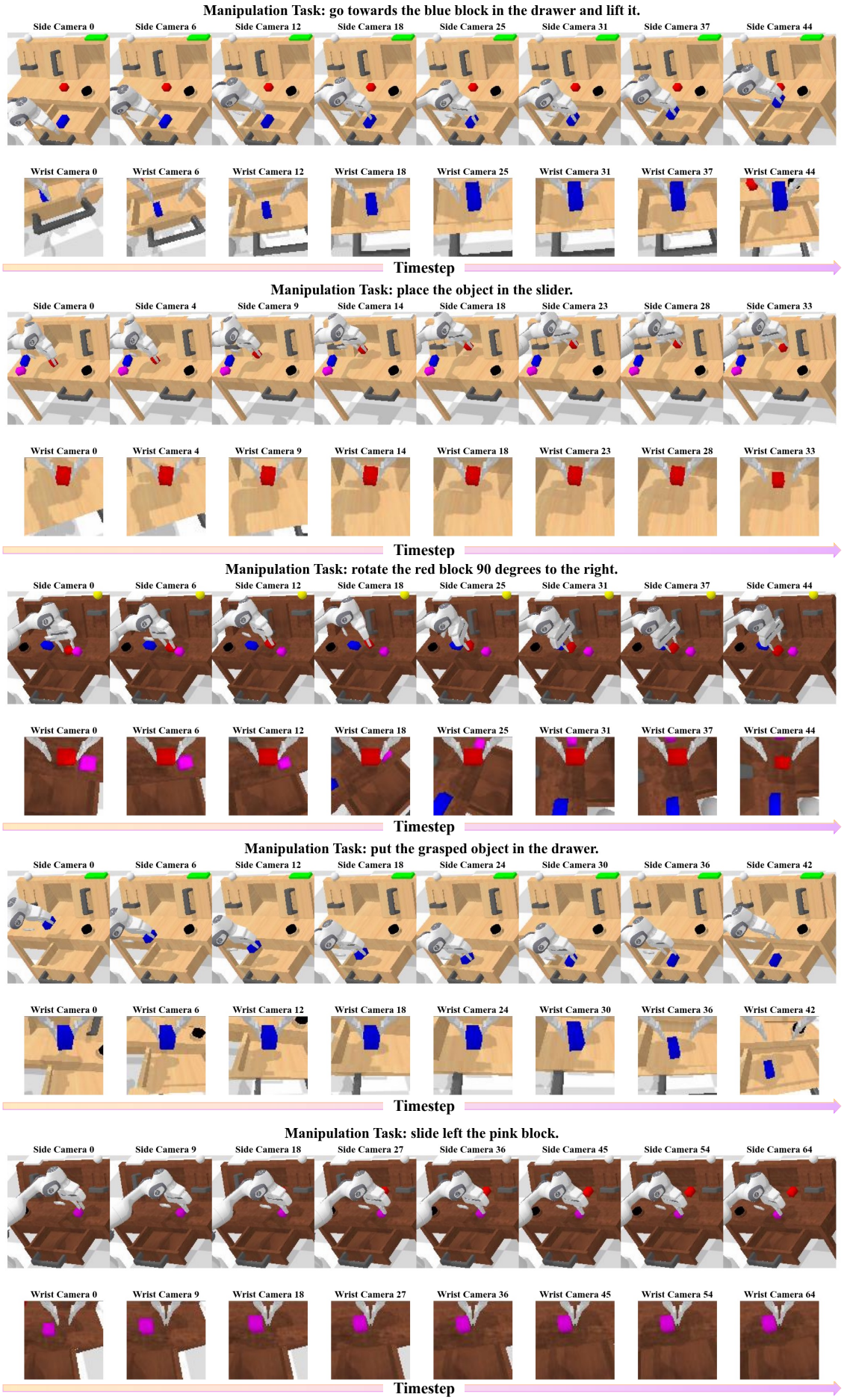

Figure 11: More Simulated Embodied Manipulation Visualization.

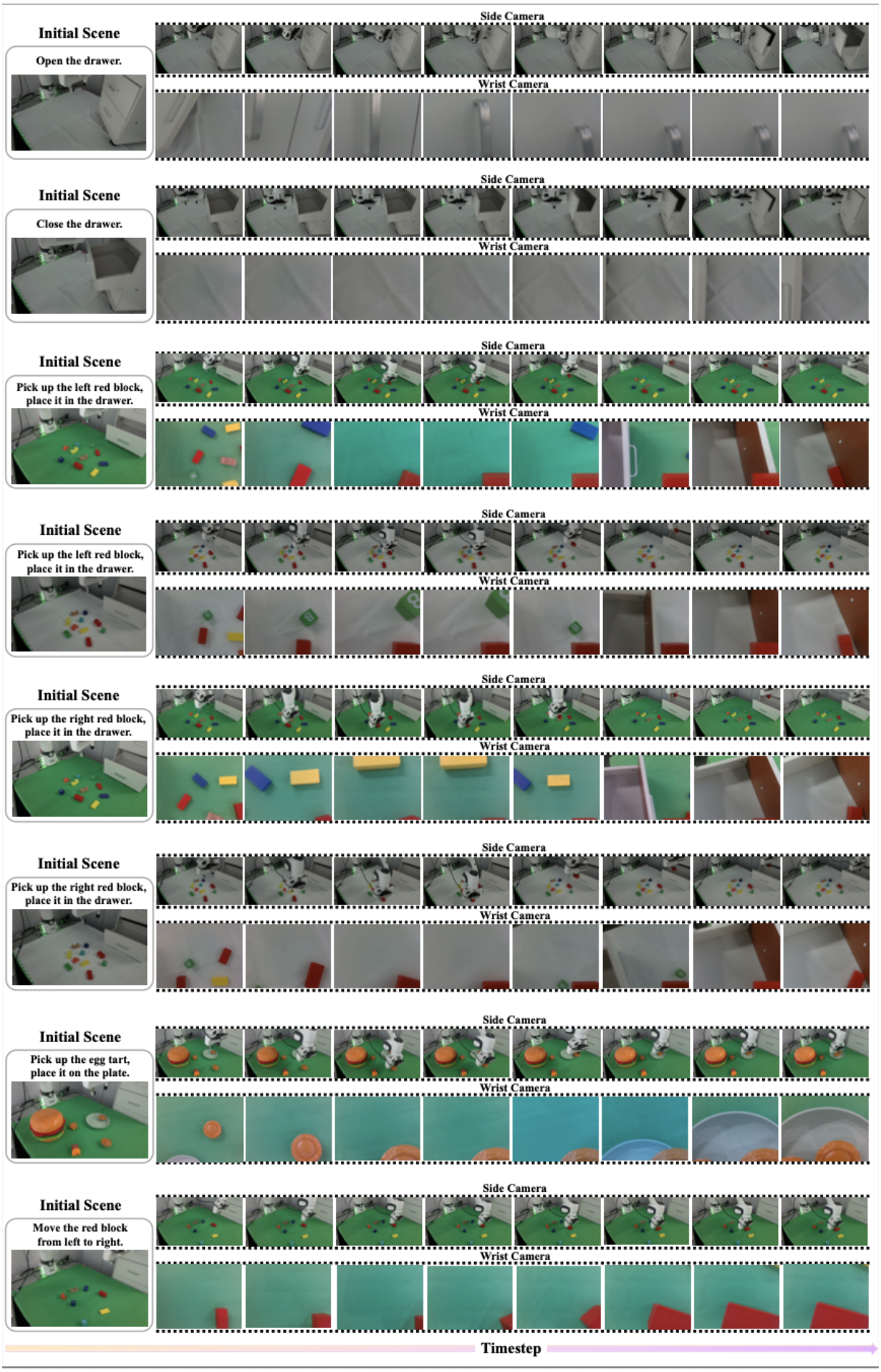

Figure 12: More Real-world Embodied Manipulation Demos.

Figure 13: More Real-world Embodied Navigation Demos.