NavA^3: Understanding Any Instruction, Navigating Anywhere, Finding Anything

Anonymous ACL submission

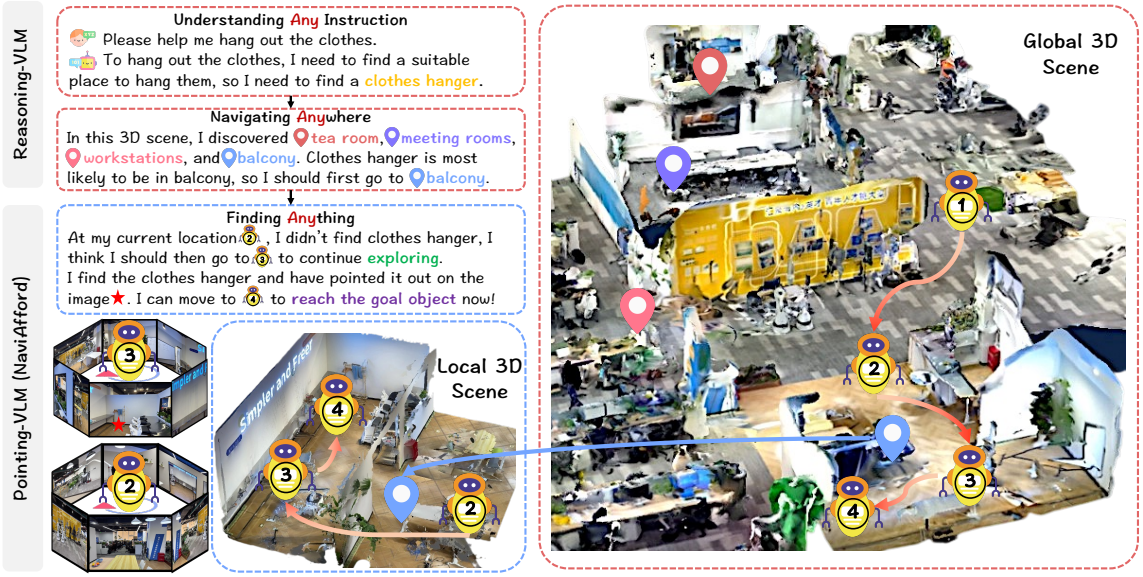


Figure 1: **Execution Process of NavA^3.** The global policy employs ***Reasoning-VLM*** to interpret high-level instructions (*e.g.*, "hang out the clothes" $\rightarrow$ clothes hanger) and identify the target location (balcony) using 3D scene understanding. The local policy uses ***Pointing-VLM*** to navigate waypoints and perform precise object localization with our ***NaviAfford*** model, which leverages spatial affordance understanding to accurately locate the target object (clothes hanger).

Abstract

Embodied navigation is a fundamental capability of embodied intelligence, enabling robots to move and interact within physical environments. However, existing navigation tasks primarily focus on predefined object navigation or instruction following, which significantly differs from human needs in real-world scenarios involving complex, open-ended scenes. To bridge this gap, we introduce a challenging *long-horizon navigation task* that requires understanding high-level human instructions and performing spatial-aware object navigation in real-world environments. Existing embodied navigation methods struggle with such tasks due to their limitations in comprehending high-level human instructions and localizing objects with an open vocabulary. In this paper, we propose NavA^3, a hierarchical framework divided into two stages: global and local policies. In global policy, we leverage the reasoning capabilities of ***Reasoning-VLM*** to parse high-level human instructions and integrate them with global 3D scene views. This allows us to reason and navigate to regions most likely to contain the goal object. In local policy, we collect a dataset of 1.0M samples of spatial-aware object affordances to train the ***NaviAfford*** model (***Pointing-VLM***), which provides robust open-vocabulary object localization and spatial awareness for precise goal identification and navigation in complex environments. Extensive experiments demonstrate that NavA^3 achieves SOTA results in navigation performance and can successfully complete *long-horizon navigation tasks* across different robot embodiments in real-world settings, paving the way for universal embodied navigation. The dataset and code will be made available.

1 Introduction

Embodied navigation (Zheng et al., 2024a; Morad et al., 2021) is a foundational capability of embodied intelligence, essential for robots to perform complex tasks in physical environments. This capability enables autonomous agents to navigate and interact within real-world spaces, forming the basis for more sophisticated embodied behaviors such as manipulation, exploration, and human-robot col-

laboration (Zhang et al., 2025c; Wu et al., 2025b). Despite significant advancements in this field, existing research primarily focuses on relatively low-level tasks, such as instruction following and basic object navigation, which do not fully capture the nuances of human needs in dynamic environments. Current embodied navigation methods can be broadly categorized into two main approaches: vision and language navigation (VLN) (Hong et al., 2021; Zhang et al., 2025a) and object navigation (ObjectNav) (Qi et al., 2025; Gao et al., 2025; Gong et al., 2025). VLN tasks require agents to follow detailed, step-by-step instructions, such as "turn left, go out the door, and then go straight." While these tasks necessitate precise spatial understanding, they often rely on overly specific directives that are seldom provided by humans in natural settings. Conversely, ObjectNav tasks aim to locate predefined object categories (*e.g.*, "find any chair in the scene") and succeed upon encountering any instance of the target object, regardless of the spatial context or specific requirements.

However, real-world human instructions frequently involve high-level intentions that demand complex reasoning and spatial perception. For instance, requests like "I want a cup of coffee" or "I want to eat the fruit on the left side of the tea room" necessitate not only an understanding of the underlying targets but also reasoning about the spatial relationships among objects. This highlights a fundamental gap between current navigation tasks and real-world needs, which significantly hampers the development of embodied agents capable of advanced human-computer interaction.

To address the challenges of long-horizon navigation, we propose ***Nav***A^3, a novel hierarchical framework that decomposes this complex problem into two stages: *global policy and local policy*. As shown in Fig. 1, the global policy leverages the powerful reasoning capabilities of the vision language model (VLM) (Ji et al., 2025; O'Neill et al., 2024; Tan et al., 2025; Zhai et al., 2024), termed ***Reasoning-VLM***, to parse high-level human instructions. Reasoning-VLM identifies key objects to locate based on these instructions and determines the most probable space for the goal object using an annotated global 3D scene. For example, when given the instruction "I want a cup of coffee," the global policy infers that the coffee machine is likely located in the pantry, guiding the agent to this high-probability area. Upon completion of the global policy, the local policy takes over, focusing on exploration and precise object localization within the identified target area. The VLM, referred to as ***Pointing-VLM***, selects waypoints from the local 3D scene for exploration. At each waypoint, we perform panoramic perception and utilize our specially trained ***NaviAfford*** model (an implementation of Pointing-VLM) for accurate target object identification. If the target object is detected, we transform its location from the agent's perspective to the robot's coordinate system, enabling navigation to the final goal. The ***NaviAfford*** model is trained on a spatial object affordance dataset comprising 1.0 million sample pairs, facilitating spatial-aware object and affordance localization. This allows the model to understand complex spatial relationships, such as "cup by the window" or "empty space on the left side of the table." Extensive experimental evaluations demonstrate that ***Nav***A^3 achieves state-of-the-art performance in long-horizon navigation tasks across large-scale real-world environments. Additionally, our system exhibits excellent cross-embodiment capabilities, making it adaptable to various robot instances and highlighting its potential for practical applications.

Our contributions are summarized as follows:

- We introduce a challenging and realistic long-horizon navigation task that requires agents to comprehend high-level human instructions and locate open vocabulary objects with complex spatial relationships in intricate indoor environments.

- We propose ***Nav***A^3, a novel hierarchical framework leveraging both global and local policies. This framework enables the understanding of diverse high-level instructions, navigation across various areas, and the ability to find any object.

- We have collected a dataset of 1.0 million samples of spatial-aware object affordances to train the ***NaviAfford*** model, enabling it to effectively understand complex spatial relationships and perform accurate object pointing.

- Extensive experiments demonstrate that our approach achieves SOTA navigation performance compared to existing methods, paving the way for development of general embodied navigation systems in real-world scenarios.

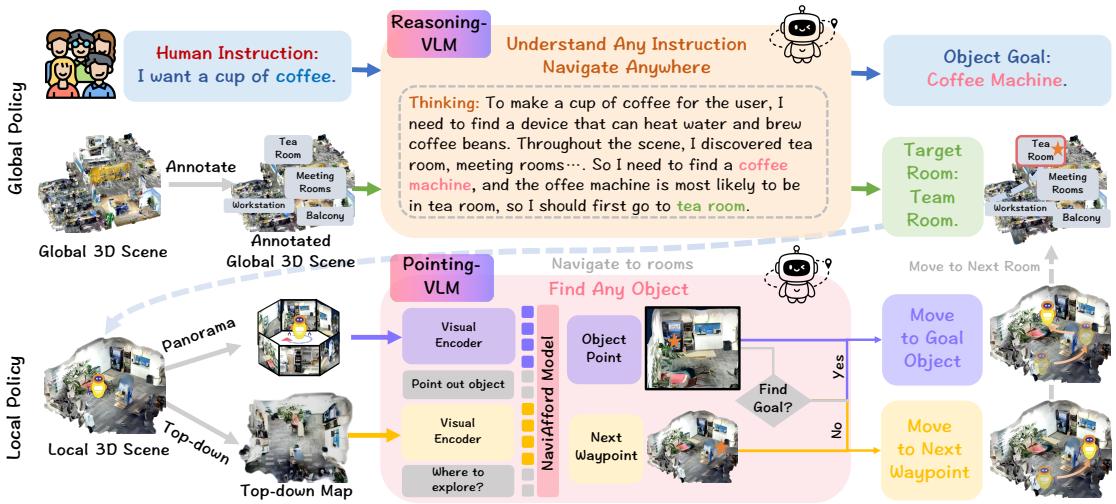


Figure 2: **Overview of the *NavA*^{3} Framework.** Our hierarchical approach has two stages: the global policy uses Reasoning-VLM to interpret high-level instructions and identify target areas in the 3D scene. The local policy then uses Pointing-VLM to search for the goal object at each waypoint. If not found, it predicts the next waypoint; if detected, it marks the object on the egocentric image and navigates to the destination.

2 Related Work

Embodied Navigation Embodied navigation research encompasses two paradigms: visual-language navigation (VLN) and object-target navigation (ObjectNav)(Zhang et al., 2024c; Gong et al., 2025). In VLN, systems like NavGPT(Zhou et al., 2024) use GPT-4o (Hurst et al., 2024) for action generation, while DiscussNav (Long et al., 2024) reduces human involvement. InstructNav (Long et al., 2025) breaks navigation into subtasks, and Nav-CoT (Lin et al., 2025) employs chain-of-thought reasoning. MapNav (Zhang et al., 2025a) improves memory with spatial representations, and NaVid (Zhang et al., 2024b) maintains temporal context. For ObjectNav, PirlNav (Ramrakhya et al., 2023) and XGX (Wasserman et al., 2024) mimic human demonstrations, while L3MVN (Yu et al., 2023) and Uni-NaVid (Zhang et al., 2024a) create semantic maps or utilize VLMs. However, these methods primarily focus on detailed instructions, struggling with high-level commands and spatial-aware localization of open-vocabulary objects, limiting their long-horizon navigation effectiveness.

Spatial Reasoning with VLMs Spatial reasoning is crucial for robots interacting with the physical world (Beyer et al., 2024; Liu et al., 2024; Doveh et al., 2024). Researchers have developed methods to enhance the spatial understanding of VLMs by extracting spatial information from images. For example, SpatialVLM (Chen et al., 2024) converts images into object-centered point clouds, and SpatialRGPT (Cheng et al., 2024) improves reasoning using spatial scene graphs. RoboPoint (Yuan et al., 2025) offers a synthetic dataset for accurate predictions, while SpatialBot (Cai et al., 2024) utilizes RGB-D data for spatial understanding. Recent advancements like Spatial-CoT (Liu et al., 2025) and VILASR (Wu et al., 2025a) optimize reasoning processes. However, these methods still struggle with open-vocabulary object pointing and long-horizon navigation, which are vital for practical applications.

3 Methodology

As shown in Fig. 2, our *NavA*^{3} framework employs a hierarchical global-to-local approach that integrates semantic reasoning with spatial localization for long-view navigation tasks. The global policy utilizes *Reasoning-VLM* to interpret high-level instructions (*e.g.*, "I want a cup of coffee"), inferring the target object (coffee machine) and likely room (*e.g.*, tea room, kitchen). Upon reaching this room, the local policy employs our ***NaviAfford*** model (*Pointing-VLM*) to analyze panoramic RGB observations and the local map at each waypoint. It checks for the target object; if found, it points to its location, and if not, it predicts the next waypoint or consults Reasoning-VLM to continue exploring until the target is located.

3.1 Preliminaries

Problem Definition We define the long-horizon embodied navigation task as follows: given a high-

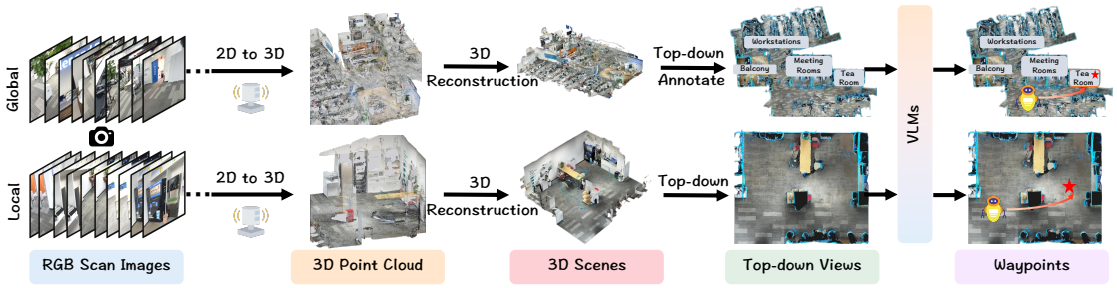


Figure 3: **Construction Process of 3D Scenes.** We reconstruct 3D scenes from RGB scan images using 2D-to-3D reconstruction techniques. These scenes are then transformed into annotated top-down views, which are subsequently processed by Vision-Language Models (VLMs) for navigation planning. This approach enhances the accuracy and efficiency of navigation tasks.

level instruction I (*e.g.*, I want coffee" or Help me hang the clothes on the balcony"), the agent must navigate a large indoor environment E to locate and reach a target object O that meets the semantic and spatial requirements of I. The agent begins at an arbitrary position p_0 and has access to RGB-D observations o_t and a global 3D scene representation S. Unlike traditional ObjectNav tasks that terminate upon finding any object category, our task requires multi-step reasoning to identify specific target objects from instructions (*e.g.*, inferring coffee machine" from I want coffee"), determining likely spatial locations (*e.g.*, kitchen), and navigating to the precise instance that meets the contextual requirements. Success is defined as the agent reaching within 1 meter of the target object O while maintaining line of sight, demonstrating accurate semantic understanding and spatial navigation in complex environments.

3D Scenes Construction To enable effective navigation in real-world environments, we construct a hierarchical 3D scene representation using a straightforward reconstruction pipeline, as illustrated in Fig. 3. Our process begins with a sequence of RGB images captured from multiple viewpoints, which are processed through a 2D-to-3D reconstruction pipeline. Using a mobile device equipped with a LiDAR sensor, we generate a dense point cloud represented by:

$$P = \{p_i | p_i \in R^3\}_{i=1}^{N}, \tag{1}$$

where each point p_i represents a 3D coordinate in the scene. The reconstruction employs a feature point matching algorithm to establish correspondences between consecutive frames, followed by mesh reconstruction to generate coherent 3D geometry. A 3D scanner application is used to streamline this process and ensure high-quality results.

The reconstructed 3D scene is converted into a top-down view for global and local policies. For global policy, we use MapNav's (Zhang et al., 2025a) annotation method to provide room- and region-level semantic annotations, such as "tea room", "conference room", "balcony", et al. This allows VLM to effectively understand spatial semantics and reason about object locations. The annotated global scene is represented as:

$$S_{\text{global}} = \{R_j, A_j\}_{j=1}^{M}, \tag{2}$$

where R_j represents the geometric region and A_j the corresponding semantic annotation. For the local strategy, we use the top-down map M_{local} directly, without annotations.

3.2 Global Policy

The global policy utilizes the advanced reasoning capabilities of the vision language model (Reasoning-VLM) to bridge the semantic gap between high-level human instructions and navigation goals. As shown in Fig. 2, given human instructions I and an annotated global 3D scene S_{global}, we treat the global reasoning task as a multimodal problem where Reasoning-VLM performs both semantic object reasoning and spatial location prediction.

To support systematic reasoning, we designed a structured prompt template to effectively guide Reasoning-VLM:

> "You need to complete the human instruction: I. Now given this top-down scene view S_{global} and several optional regions, please think about what object you should find to complete the instruction and where you should look for this object. Please show your thinking process and give your answer at the end."

The Reasoning-VLM processes textual instructions and the visual representation of the annotated global scene to enable hierarchical reasoning. It first infers the target object O^* needed to fulfill the instruction through semantic decomposition: $O^* = f_{\text{semantic}}(I)$. The model then analyzes spatial semantic relations to identify the target region R^*, where the object is most likely located, defined by $R^* = \arg\max_{R_j \in S_{\text{global}}} P(O^*|R_j, A_j)$, with $P(O^*|R_j, A_j)$ representing the conditional probability of finding O^* with annotation A_j in region R_j.

After identifying the target region R^*, we randomly sample a waypoint $w \in R^*$ within its local boundary and use Pointing-VLM to guide the agent. This strategy promotes robust exploration while effectively narrowing the search space to relevant subregions where the target object is likely located, enhancing the efficiency of the search process.

3.3 Local Policy

NaviAfford Model To achieve precise spatial object localization, we developed NaviAfford (Pointing-VLM), as shown in Fig. 4. During training, we curated a dataset of approximately 50K images and 1.0M QA pairs from the LVIS (Gupta et al., 2019) and Where2Place datasets (Yuan et al., 2025). We converted instance segmentation masks into an object detection format with bounding box coordinates (x_1, y_1, x_2, y_2) and sampled 5-8 representative points within each box to enhance spatial granularity and improve localization accuracy, supporting Reasoning-VLM's capabilities.

Our dataset construction systematically generates two types of affordance annotations for comprehensive spatial understanding. For ***object affordance***, we compute directional relations (up, down, left, right, et al.) to identify target objects in specific contexts. For example, given the query "find the TV in front of the sofa," we determine the goal object and its spatial relationship to reference objects. For ***spatial affordance***, we identify free spaces that satisfy these constraints, enabling the model to understand available areas for navigation and placement. This dual-affordance approach creates training samples that capture complex spatial relations necessary for real-world navigation.

The NaviAfford model architecture follows a vision-language framework, processing the input question Q and RGB image V through separate tokenizer and visual encoder paths. The architecture is expressed as:

$$Model(Q, V) = f_{\text{LLM}}(f_{\text{text}}(Q), f_{\text{proj}}(f_{\text{vision}}(V))), \tag{3}$$

where f_{text} processes the text query, f_{vision} encodes the visual input, and f_{proj} maps visual features to the LLM embedding space. The function f_{LLM} generates text point coordinates. The training objective uses supervised fine-tuning (SFT) with the loss function:

$$\mathcal{L} = -\sum_{i=1}^{N} \log P(t_i|t_{<i}, Q, V), \tag{4}$$

where t_i represents the i-th token containing point coordinates in the target text sequence. In the local policy of ***NaviAfford***, we input the self-centered RGB view and the target object query based on spatial relationships, deploying the model in real-world environments with zero samples. The model outputs accurate point coordinates, with specific usage detailed in the local policy.

Navigation Process In the local policy, our system employs a fine-grained object localization and navigation strategy based on systematic waypoint exploration. As shown in Fig. 2, the agent captures panoramic RGB views at each waypoint via rotational scanning. The NaviAfford model processes these views to detect and accurately localize the target object. Upon detection, the model outputs multiple point coordinates, and we select the center point by averaging for robust localization.

To convert pixel coordinates to robot coordinates, we use the camera intrinsic function:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} \frac{(u-c_x)\cdot d}{f_x} \\ \frac{(v-c_y)\cdot d}{f_y} \\ d \end{bmatrix}, \tag{5}$$

where f_x and f_y are the focal lengths, c_x and c_y are the principal points, and d is the depth at pixel (u, v). This ensures effective navigation to the target object.

Next, we transform camera coordinates to robot coordinates using rotation and translation:

$$\begin{bmatrix} x_{world} \\ y_{world} \end{bmatrix} = \begin{bmatrix} \cos\theta_r & -\sin\theta_r \\ \sin\theta_r & \cos\theta_r \end{bmatrix} \begin{bmatrix} x_{robot} \\ y_{robot} \end{bmatrix} + \begin{bmatrix} x_r \\ y_r \end{bmatrix}, \tag{6}$$

where (x_r, y_r, θ_r) is the robot's world pose, and (x_{robot}, y_{robot}) are derived from camera coordinates: $x_{robot} = Z_{cam}, y_{robot} = -X_{cam}$.

If the goal object is not detected, the system follows a two-stage decision process. First,

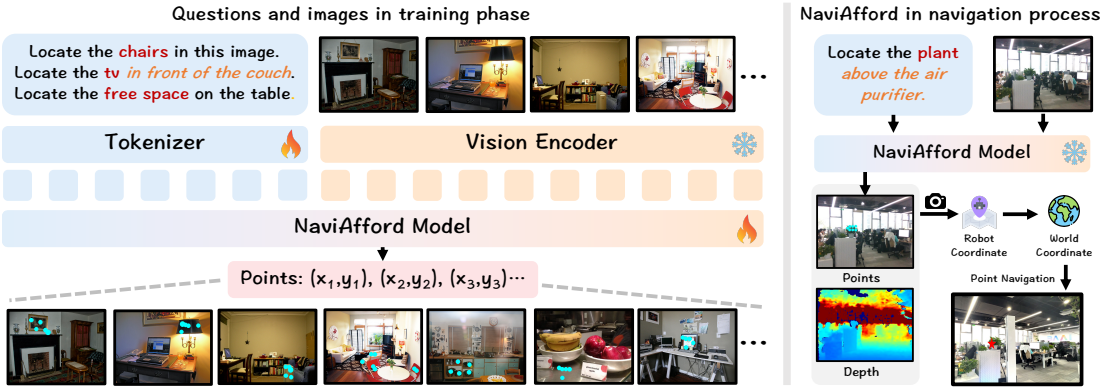


Figure 4: ***NaviAfford* Model Training and Deployment Process.** The ***NaviAfford*** model learns object and spatial affordances from various indoor scenes to output precise point coordinates. During navigation, it performs real-time object localization and generates target points, which the local policy converts into robot coordinates for effective navigation to goal objects.

Reasoning-VLM analyzes the local 3D scene and historical exploration data to decide whether to continue exploring the current area or transition to a new one. If it opts to continue, the NaviAfford model identifies the next best exploration point. Otherwise, it selects the most promising room or space to explore based on previous searches, facilitating efficient transitions.

4 Experiments

4.1 Experimental Details

Evaluation Benchmark To evaluate long-horizon navigation, we established a benchmark with five scenes: Meeting Room A, Meeting Room B, Tea Room, Workstation, and Balcony, totaling 50 tasks (10 per scene). Each task underwent 10 rollouts to minimize randomness, with human experts defining high-level instructions and unique goal objects. Tasks were tested five times under different starting conditions for reliability. The agent interacted with the environment using egocentric RGB-D perception and waypoint selection. For Pointing-VLM evaluation, we used 1,000 images not included in the training set.

Evaluation Metrics We use two metrics in embodied navigation: Navigation Error (NE) and Success Rate (SR). NE measures the distance (in meters) from the agent's final position to the target, with lower values indicating better performance. SR reflects the percentage of successful navigation events within 1 meter of the target, computed over 50 trials and reported as average SR (Avg. SR). For Pointing-VLM evaluation, we measure accuracy (Acc) as the ratio of correctly predicted points within the ground truth mask to total predicted points.

Implementation Details For Reasoning-VLM, we utilize GPT-4o to interpret high-level instructions and make spatial decisions. The Pointing-VLM uses the NaviAfford model, trained on the 1.0M Spatial Perception Object Affordances dataset, initialized with pre-trained Qwen2.5-VL-7B weights and fine-tuned as described in (Zheng et al., 2024b). Experiments are conducted on four H100 GPUs with AdamW as the optimizer, a learning rate of 10^{-5} for one epoch, a batch size of 4 per GPU, and gradient accumulation set to 2 steps for an effective batch size of 32. To validate cross-embodiment capability, we deploy the system on both the RealMan wheeled robot and the Unitree Go2 quadruped robot, each equipped with Intel RealSense D435i cameras for RGB-D perception.

Baseline Models Existing navigation methods often face challenges with long-horizon tasks that involve high-level instructions. To ensure fair comparisons, we refine their task formulations for clarity: "Complete the following: I want to drink coffee. Find the target object and stop near it." We provide baseline models with top-down global 3D scene information and evaluate three model types: (1) closed-source general VLMs, including GPT-4o (Hurst et al., 2024), Claude-3.5-Sonnet (Anthropic, 2024), and Qwen-VL-Max (Bai et al., 2025); (2) open-source general-purpose VLMs like Janus-Pro-7B (Chen et al., 2025), Qwen2.5-VL-7B (Bai et al., 2025), and LLaVA-Next-7B (Li et al., 2024); and (3) navigation-specific methods, such as NaVid (Zhang et al., 2024b), NaVILA (Cheng et al., 2025), and MapNav (Zhang et al., 2025a), which require adaptation for long-horizon navigation tasks.

Table 1: **Navigation Performance Comparison with SOTA methods.** * denotes that we modify the method to allow it to complete our task. Our *NavA*^{3} outperforms all the SOTA methods on the navigation performance.

Category	Methods	Meeting Room A		Meeting Room B		Tea Room		Workstation		Balcony		Avg. SR↑
		NE↓	SR↑	NE↓	SR↑	NE↓	SR↑	NE↓	SR↑	NE↓	SR↑	
Closed-source	GPT-4o (Hurst et al., 2024)	12.45	2.0%	13.78	0.0%	14.12	2.0%	11.89	4.0%	10.45	2.0%	2.0%
	Claude-3.5-Sonnet (Anthropic, 2024)	11.18	6.0%	12.56	0.0%	13.94	2.0%	10.67	4.0%	11.31	2.0%	2.8%
	Qwen-VL-Max (Bai et al., 2025)	13.67	0.0%	15.01	0.0%	16.45	0.0%	14.12	2.0%	12.89	0.0%	0.4%
Open-source	Janus-Pro-7B (Chen et al., 2025)	16.42	0.0%	17.89	0.0%	18.23	0.0%	15.98	0.0%	16.54	0.0%	0.0%
	Qwen2.5-VL-7B (Bai et al., 2025)	18.98	0.0%	19.34	0.0%	20.78	0.0%	17.45	0.0%	18.12	0.0%	0.0%
	LLaVA-Next-7B (Li et al., 2024)	17.98	0.0%	18.34	0.0%	19.78	0.0%	16.45	0.0%	17.12	0.0%	0.0%
Navigation-specific	NaVid* (Zhang et al., 2024b)	8.14	18.0%	9.31	12.0%	10.52	10.0%	7.89	16.0%	8.76	18.0%	14.8%
	NaVILA* (Cheng et al., 2025)	7.93	20.0%	8.67	16.0%	9.84	12.0%	7.45	18.0%	8.22	16.0%	16.4%
	MapNav* (Zhang et al., 2025a)	7.21	26.0%	7.94	24.0%	9.12	26.0%	6.78	28.0%	7.45	22.0%	25.2%
	NavA^3 (Ours)	**1.23**	**72.0%**	**1.45**	**64.0%**	**1.89**	**60.0%**	**1.56**	**76.0%**	**1.34**	**60.0%**	**66.4%**

Table 2: **Ablation Study on Annotation Components.**

Annotation Variants	Tea Room		Workstation		Avg. SR↑
	NE↓	SR↑	NE↓	SR↑	
NavA^3 w/o Map	4.13	32.0%	5.88	40.0%	36.0%
NavA^3 w/o Annotation	3.32	36.0%	2.78	44.0%	40.0%
NavA^3 w/o Room-level Annotation	3.21	36.0%	3.01	40.0%	38.0%
NavA^3 w/ Full Annotation (Ours)	**1.89**	**60.0%**	**1.56**	**76.0%**	**68.0%**

Table 3: **Ablation Study on Different Reasoning-VLMs.**

Reasoning-VLMs	Tea Room		Workstation		Avg. SR↑
	NE↓	SR↑	NE↓	SR↑	
Open-source					
NavA^3 w/ Qwen2.5-VL-72B (Bai et al., 2025)	2.67	52.0%	2.23	66.0%	59.0%
NavA^3 w/ Qwen2.5-VL-7B (Bai et al., 2025)	3.12	38.0%	2.89	42.0%	40.0%
NavA^3 w/ Janus-Pro-7B (Chen et al., 2025)	3.78	30.0%	3.56	34.0%	32.0%
NavA^3 w/ LLaVA-NeXT-7B (Li et al., 2024)	3.45	34.0%	3.23	38.0%	36.0%
Closed-source					
NavA^3 w/ Claude-3.5-Sonnet (Anthropic, 2024)	2.01	58.0%	1.68	72.0%	65.0%
NavA^3 w/ Qwen-VL-Max (Bai et al., 2025)	2.34	54.0%	2.12	68.0%	61.0%
NavA^3 w/ GPT-4o (Hurst et al., 2024) (Ours)	**1.89**	**60.0%**	**1.56**	**76.0%**	**68.0%**

4.2 Comparisons with SOTA Methods

As shown in Tab.1, *NavA*^{3} significantly outperforms existing state-of-the-art methods across all evaluation scenarios, achieving a 41.2 percentage point increase in success rate (SR) with an average of 66.4%, compared to the best baseline, MapNav(Zhang et al., 2025a), at 25.2%. Specifically, *NavA*^{3} improves SR by 46.0% in Conference Room A (72.0% *vs.* 26.0%), 40.0% in Conference Room B (64.0% *vs.* 24.0%), 34.0% in Tea Room (60.0% *vs.* 26.0%), 48.0% in Workstation (76.0% *vs.* 28.0%), and 38.0% in Balcony (60.0% *vs.* 22.0%). It also significantly reduces navigation error (NE) in all scenarios: by 5.98m (1.23m *vs.* 7.21m) in Conference Room A, 6.49m (1.45m *vs.* 7.94m) in Conference Room B, 7.23m (1.89m *vs.* 9.12m) in Tea Room, 5.22m (1.56m *vs.* 6.78m) in Workstation, and 6.11m (1.34m *vs.* 7.45m) in Balcony. While general-purpose VLMs (both closed-source and open-source) often achieve near-zero success rates in this challenging long-horizon navigation task, our hierarchical approach effectively bridges the gap between high-level command understanding and accurate spatial navigation.

4.3 Ablation Study

Effect of Annotation To evaluate our annotation strategy, we conducted ablation studies in the tea room and workstation. Results in Tab. 2 demonstrate the importance of semantic annotation in long-horizon navigation. Compared to the *NavA*^{3} w/o map, *NavA*^{3} w/ full annotation (ours) shows a 28.0% improvement in tea room (60.0% *vs.* 32.0%) and a 36.0% improvement in the workstation (76.0% *vs.* 40.0%), resulting in an average SR improvement of 32.0% (68.0% *vs.* 36.0%). When compared to *NavA*^{3} w/o annotation, we see a 24.0% improvement in the tea room (60.0% *vs.* 36.0%) and a 32.0% improvement in the workstation (76.0% *vs.* 44.0%), leading to an average enhancement of 28.0%. Against *NavA*^{3} w/o room-level annotation, our strategy yields a 24.0% improvement in the tea room (60.0% *vs.* 36.0%) and a 36.0% improvement in the workstation (76.0% *vs.* 40.0%), with an average enhancement of 30.0%. These findings confirm that detailed semantic annotations improve Reasoning-VLMs' understanding of spatial relationships.

Effect of Reasoning-VLMs To evaluate the impact of different Reasoning-VLM models on navigation performance, we conducted ablation studies in the tea room and workstation scenarios. Results in Tab. 3 reveal significant differences among VLM architectures. Our GPT-4o (Hurst et al., 2024)-based Reasoning-VLM achieves the highest success rate (SR) of 68.0%. Closed-source models like Claude-3.5-Sonnet (Anthropic, 2024) and Qwen-VL-Max (Bai et al., 2025) show decreases of 3.0% (65.0%) and 7.0% (61.0%), respectively. Open-

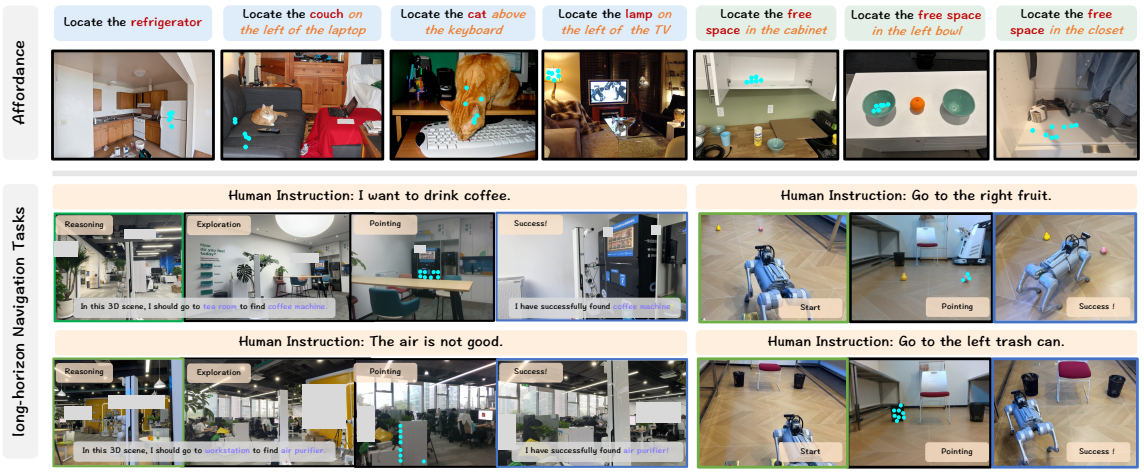


Figure 5: **Qualitative analysis on *NaviAfford* and *NavA*^{3}.** Affordance visualization includes performance of ***NaviAfford*** model on object and spatial affordance. Long-horizon navigation tasks visualization includes performance of *NavA*^{3} in real-world environments and its cross-embodiment deployment capabilities.

Table 4: **Ablation Study on Different Pointing-VLMs.**

Pointing-VLMs	Params	Obj. Aff.↑	Spa. Aff.↑	Avg. Acc↑	Nav. SR↑
Closed-source					
NavA^3 w/ GPT-4o (Hurst et al., 2024)	-	21.3%	25.1%	23.2%	32.0%
NavA^3 w/ Claude-3.5-Sonnet (Anthropic, 2024)	-	20.1%	22.8%	21.5%	28.0%
NavA^3 w/ Qwen-VL-Max (Bai et al., 2025)	-	18.8%	21.5%	20.2%	20.0%
Open-source					
NavA^3 w/ Qwen2.5-VL (Bai et al., 2025)	72B	15.2%	18.7%	17.0%	16.0%
NavA^3 w/ Qwen2.5-VL (Bai et al., 2025)	7B	6.2%	15.3%	10.8%	12.0%
NavA^3 w/ Janus-Pro (Chen et al., 2025)	7B	2.8%	3.1%	2.95%	6.0%
NavA^3 w/ LLaVA-NeXT (Li et al., 2024)	7B	3.1%	0.8%	2.0%	4.0%
Specific					
NavA^3 w/ RoboPoint (Yuan et al., 2025)	13B	55.9%	44.5%	50.2%	57.5%
NavA^3 w/ NaviAfford (Ours)	**7B**	**70.8%**	**55.6%**	**63.2%**	**68.0%**

source models, such as Qwen2.5-VL-72B (Bai et al., 2025), drop by 9.0% (59.0%). Smaller 7B models, including Qwen2.5-VL-7B (Bai et al., 2025), Janus-Pro-7B (Chen et al., 2025), and LLaVA-NeXT-7B (Li et al., 2024), exhibit declines of 28.0%, 32.0%, and 36.0%, respectively. These findings underscore the importance of reasoning capabilities in complex spatial tasks.

Effect of Pointing-VLMs To evaluate the effectiveness of different Pointing-VLMs for object localization, we compare NaviAfford with baseline methods. Results in Tab.4 demonstrate NaviAfford's superior performance on the affordance understanding benchmark, achieving a 13.0% improvement in average affordance accuracy over the previous state-of-the-art RoboPoint(Yuan et al., 2025) (70.8% *vs.* 55.9%). This strong affordance understanding translates to enhanced navigation performance, with NaviAfford showing a 10.5% increase in success rate (SR) over RoboPoint (68.0% *vs.* 57.5%), a 36.0% improvement over GPT-4o (Hurst et al., 2024) (68.0% *vs.* 32.0%), and a 52.0% boost over the best open-source model, Qwen2.5-VL-72B (Bai et al., 2025) (68.0% *vs.* 16.0%). These results indicate that our spatial affordance training effectively bridges accurate object localization and practical navigation execution.

4.4 Qualitative Analysis

We qualitatively evaluate *NavA*^{3}'s capabilities in affordance understanding, navigation, and cross-embodiment deployment, as shown in Fig. 5. Affordance visualizations demonstrate spatial awareness, accurately identifying references like "the sofa to the left of the laptop" and localizing objects in cluttered environments. Long-horizon navigation visualizations illustrate a systematic approach, tracing reasoning from instruction parsing (*e.g.*, "I want coffee") to goal achievement in multi-room settings. Cross-embodiment experiments highlight versatility, with consistent performance on quadruped robots in tasks like "walk to the fruit on the right." These findings confirm our approach's adaptability across various robotic platforms.

5 Conclusion

This paper introduces *NavA*^{3}, a hierarchical framework that connects embodied navigation with human needs, enabling robots to interpret high-level instructions and navigate complex environments. It features a two-stage approach: a global policy with Reasoning-VLM for instruction parsing and a local strategy using the ***NaviAfford*** model for object localization. Experiments show *NavA*^{3} outperforms state-of-the-art methods in complex spatial relations and open-vocabulary pointing. Its deployment on wheeled and quadruped robots demonstrates versatility. Future work will enhance adaptability in dynamic environments and integrate additional sensory inputs.

Limitations

While our ***Nav***A^3 framework demonstrates impressive performance on long-view navigation tasks, several limitations must be acknowledged. First, our system relies on precise depth information from RGB-D sensors for accurate object localization and coordinate transformation. Depth estimation may be less reliable in environments with reflective surfaces, transparent objects, or poor lighting conditions, potentially impacting navigation accuracy. Second, the current framework requires a two-stage hierarchical design with independent global and local policies, rather than an end-to-end action prediction model. While interpretable and effective, this modular approach may introduce latency between high-level reasoning and low-level control and could be simplified in future work through unified action generation. Future work could explore more robust depth estimation techniques, end-to-end learning methods, and adaptive mechanisms for handling dynamic environments.

References

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *https://docs.anthropic.com/zh-CN/release-notes/claude-apps*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

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*.

Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. 2024. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*.

Wenzhe Cai, Siyuan Huang, Guangran Cheng, Yuxing Long, Peng Gao, Changyin Sun, and Hao Dong. 2023. Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill. *arXiv preprint arXiv:2309.10309*.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Bıyık, Hongxu Yin, Sifei Liu, and Xiaolong Wang. 2025. Navila: Legged robot vision-language-action model for navigation. In *Robotics: Science and Systems*.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093.

Sivan Doveh, Shaked Perek, M Jehanzeb Mirza, Wei Lin, Amit Alfassy, Assaf Arbelle, Shimon Ullman, and Leonid Karlinsky. 2024. Towards multimodal in-context learning for vision and language models. In *European Conference on Computer Vision*, pages 250–267. Springer.

Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. 2023. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23171–23181.

Chen Gao, Liankai Jin, Xingyu Peng, Jiazhao Zhang, Yue Deng, Annan Li, He Wang, and Si Liu. 2025. Octonav: Towards generalist embodied navigation. *arXiv preprint arXiv:2506.09839*.

Zeying Gong, Rong Li, Tianshuai Hu, Ronghe Qiu, Lingdong Kong, Lingfeng Zhang, Yiyi Ding, Leying Zhang, and Junwei Liang. 2025. Stairway to success: Zero-shot floor-aware object-goal navigation via llm-driven coarse-to-fine exploration. *arXiv preprint arXiv:2505.23019*.

Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364.

Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. 2021. Vln bert: A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 1643–1653.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1

others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1724–1734.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Bingqian Lin, Yunshuang Nie, Ziming Wei, Jiaqi Chen, Shikui Ma, Jianhua Han, Hang Xu, Xiaojun Chang, and Xiaodan Liang. 2025. Navcot: Boosting llm-based vision-and-language navigation via learning disentangled reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sichao Liu, Jianjing Zhang, Robert X Gao, Xi Vincent Wang, and Lihui Wang. 2024. Vision-language model-driven scene understanding and robotic object manipulation. In *IEEE 20th International Conference on Automation Science and Engineering*, pages 21–26. IEEE.

Yuecheng Liu, Dafeng Chi, Shiguang Wu, Zhanguang Zhang, Yaochen Hu, Lingfeng Zhang, Yingxue Zhang, Shuang Wu, Tongtong Cao, Guowei Huang, and 1 others. 2025. Spatialcot: Advancing spatial reasoning through coordinate alignment and chain-of-thought for embodied task planning. *arXiv preprint arXiv:2501.10074*.

Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. 2025. Instructnav: Zero-shot system for generic instruction navigation in unexplored environment. In *Conference on Robot Learning*, pages 2049–2060.

Yuxing Long, Xiaoqi Li, Wenzhe Cai, and Hao Dong. 2024. Discuss before moving: Visual language navigation via multi-expert discussions. In *IEEE International Conference on Robotics and Automation*, pages 17380–17387. IEEE.

Steven D Morad, Roberto Mecca, Rudra PK Poudel, Stephan Liwicki, and Roberto Cipolla. 2021. Embodied visual navigation with automatic curriculum learning in real environments. *IEEE Robotics and Automation Letters*, 6(2):683–690.

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. IEEE.

Dean A Pomerleau. 1988. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1.

Zhangyang Qi, Zhixiong Zhang, Yizhou Yu, Jiaqi Wang, and Hengshuang Zhao. 2025. Vln-r1: Vision-language navigation via reinforcement fine-tuning. *arXiv preprint arXiv:2506.17221*.

Ram Ramrakhya, Dhruv Batra, Erik Wijmans, and Abhishek Das. 2023. Pirlnav: Pretraining with imitation and rl finetuning for objectnav. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17896–17906.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

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*.

Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. 2019. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6629–6638.

Justin Wasserman, Girish Chowdhary, Abhinav Gupta, and Unnat Jain. 2024. Exploitation-guided exploration for semantic embodied navigation. In *IEEE International Conference on Robotics and Automation*, pages 2901–2908. IEEE.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. 2025a. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *arXiv preprint arXiv:2506.09965*.

Yujie Wu, Huaihai Lyu, Yingbo Tang, Lingfeng Zhang, Zhihui Zhang, Wei Zhou, and Siqi Hao. 2025b. Evaluating gpt-4o's embodied intelligence: A comprehensive empirical study. *Authorea Preprints*.

Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. 2024a. Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In *IEEE International Conference on Robotics and Automation*, pages 42–48. IEEE.

Naoki Yokoyama, Ram Ramrakhya, Abhishek Das, Dhruv Batra, and Sehoon Ha. 2024b. Hm3d-ovon: A dataset and benchmark for open-vocabulary object goal navigation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5543–5550. IEEE.

Bangguo Yu, Hamidreza Kasaei, and Ming Cao. 2023. L3mvn: Leveraging large language models for visual target navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3554–3560. IEEE.

Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. 2025. Robopoint: A vision-language model for spatial affordance prediction in robotics. In *Conference on Robot Learning*, pages 4005–4020.

Simon Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Peter Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, and 1 others. 2024. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. *Advances in neural information processing systems*, 37:110935–110971.

Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. 2024a. Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. *arXiv preprint arXiv:2412.06224*.

Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. 2024b. Navid: Video-based vlm plans the next step for vision-and-language navigation. In *Robotics: Science and Systems*.

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. In *The 63rd Annual Meeting of the Association for Computational Linguistics*.

Lingfeng Zhang, Hao Wang, Erjia Xiao, Xinyao Zhang, Qiang Zhang, Zixuan Jiang, and Renjing Xu. 2025b. Multi-floor zero-shot object navigation policy. In *IEEE International Conference on Robotics and Automation*. IEEE.

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

Shuyi Zhang, Xiaoshuai Hao, Yingbo Tang, Lingfeng Zhang, Pengwei Wang, Zhongyuan Wang, Hongxuan Ma, and Shanghang Zhang. 2025c. Video-cot: A comprehensive dataset for spatiotemporal understanding of videos based on chain-of-thought. *arXiv preprint arXiv:2506.08817*.

Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. 2024a. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13624–13634.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024b. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

Gengze Zhou, Yicong Hong, and Qi Wu. 2024. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7641–7649.

Kaiwen Zhou, Kaizhi Zheng, Connor Pryor, Yilin Shen, Hongxia Jin, Lise Getoor, and Xin Eric Wang. 2023. Esc: Exploration with soft commonsense constraints for zero-shot object navigation. In *International Conference on Machine Learning*, pages 42829–42842. PMLR.

Ziyu Zhu, Xilin Wang, Yixuan Li, Zhuofan Zhang, Xiaojian Ma, Yixin Chen, Baoxiong Jia, Wei Liang, Qian Yu, Zhidong Deng, and 1 others. 2025. Move to understand a 3d scene: Bridging visual grounding and exploration for efficient and versatile embodied navigation. *arXiv preprint arXiv:2507.04047*.

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 provides a comprehensive overview of NaviAfford dataset construction, including data sources, object and spatial affordance generation, and quality control procedures.
- Sec. B outlines the detailed training procedures for the NaviAfford model, including base model selection, hyperparameters, and computational resources.
- Sec. C presents additional experimental results on standard navigation benchmarks (HM3D ObjectNav and HM3D-OVON) that further validate NaviAfford's generalization.
- Sec. D describes our real-world robot deployment setup, including hardware configuration, sensor specifications, and system integration across different robot embodiments.
- Sec. E provides extensive qualitative visualizations of ***Nav***A^3 deployed across diverse real-world environments, demonstrating robustness across different scenarios, commands, and robot platforms.

A *NaviAfford* Dataset Generation

This section fully describes the construction of the NaviAfford dataset, which contains approximately one million question-answer pairs and is designed to train models for spatially aware object localization. The dataset is systematically divided into two complementary categories: *Object Affordance* and *Spatial Affordance*, each targeting different aspects of spatial understanding required for embodied navigation.

A.1 Data Sources and Preprocessing

Our dataset construction leverages two primary sources: the LVIS dataset (Gupta et al., 2019), which provides high-quality instance segmentation annotations covering a variety of indoor and outdoor scenes (contributing approximately 40,000 images); and the Where2Place (Yuan et al., 2025) dataset, which focuses on position reasoning and spatial relationships (contributing approximately 10,000 images). During preprocessing, we convert instance segmentation masks into bounding box coordinates (x_1, y_1, x_2, y_2) in normalized image coordinates. To enhance spatial granularity and improve the model's pointing accuracy, we sample 5-8 representative points within each bounding box, rather than using only box corners or centers. The distribution of these sampled points captures the spatial extent of objects, enabling the model to learn more precise localization patterns, which is crucial for real-world navigation tasks.

A.2 Object Affordance Generation

Object affordance annotation focuses on teaching the model to identify and localize specific objects based on their spatial relationships to reference objects or environmental context, and includes approximately 600,000 question-answer pairs. We use the object's bounding box center and spatial configuration to compute directional relationships between objects, including cardinal directions (up, down, left, right), diagonal directions (upper left, upper right, lower left, lower right), depth relationships (in front of, behind), and proximity relationships (beside, beside, near, adjacent). For each object pair, we compute the relative position and generate a natural language description, such as "find the TV in front of the sofa," along with the corresponding point coordinates.

In addition to pairwise relationships, we also generate annotations for objects in complex spatial environments involving multiple reference objects, such as "find the lamp above the table to the left of the TV" or "find the plant between the window and chair." These multi-constraint queries require the model to reason about multiple spatial relationships simultaneously, thereby enhancing its ability to understand complex scenes. We also incorporate object properties such as color, size, state, and position to create more specific localization tasks, such as queries like "find the red cushion on the sofa," "find the open laptop on the table," and "find the tall plant in the corner." This comprehensive approach ensures the model can handle a wide range of real-world queries with varying levels of specificity and complexity.

A.3 Spatial Affordance Generation

Spatial affordance annotation teaches the model to identify free space and suitable placement areas that satisfy spatial constraints, comprising approximately 400,000 question-answer pairs. This

capability is crucial for navigation tasks that require spatial reasoning about navigable or interactable areas. We leverage depth information and object occupancy analysis to identify and annotate free space, encompassing both surface free space (empty areas on tables, desks, shelves, and countertops) and ground free space (navigable areas suitable for robot movement).

To generate ground truth annotations, we employ a systematic pipeline: first, we construct a binary occupancy map based on instance segmentation masks to label occupied and unoccupied pixels. Then, we use connected component analysis to extract contiguous free regions and apply a minimum area threshold to filter out noise. Within each valid free space region, we sample 5-8 representative points to ensure their spatial distribution within the available area. For spatially constrained queries, such as "find the free space to the right of the keyboard" or "find the empty space to the left of the bowl," we perform geometric verification to verify that the sampled points satisfy orientation constraints relative to the reference object.

A.4 Question Template Design and Quality Control

To ensure linguistic and natural language diversity, we designed multiple question templates for each annotation type. Object affordance templates include variations such as locate [object] [spatial relationship] [reference]," find [attribute] [object] [location description]," and point to [object] at [spatial constraint]." Spatial affordance templates include locate free space [location description]," find empty space [spatial constraint]," and where can I place an object [location constraint]?" This diversity ensures the model can learn to understand a variety of representations of spatial concepts.

Quality control is achieved through multiple validation procedures. All sampled points are verified to lie within the corresponding object mask or free space area, and their orientation relationships are verified through geometric calculations to ensure correctness. We manually review a subset of 5,000 samples to identify and correct annotation errors and analyze linguistic and spatial diversity metrics to ensure comprehensive coverage of object categories, spatial relationships, and scene types.

A.5 Dataset Statistics

The final NaviAfford dataset contains approximately 1.0 million question-answer pairs, of which 600,000 samples (60%) are for object affordances and 400,000 samples (40%) are for spatial affordances. The dataset covers over 1,200 unique object categories from the LVIS taxonomy and includes over 25 orientations and contextual spatial relationships. Each annotation contains an average of 6.2 points, and the dataset is drawn from approximately 50,000 different indoor scenes. This comprehensive dataset enables the NaviAfford model to develop strong capabilities in accurate object localization and spatial reasoning, laying the foundation for effective long-viewpoint embodied navigation.

B *NaviAfford* Training Details

This section provides comprehensive details about the training procedure for the NaviAfford model, including the base model selection, training strategy, hyperparameters, and computational resources.

B.1 Base Model and Training Strategy

We use Qwen2.5-VL-7B (Bai et al., 2025) as the base model, which has 7 billion parameters and provides strong visual-language understanding capabilities. The model architecture consists of 28 hidden layers, each with a hidden size of 3584, and utilizes 28 attention heads and 4 key-value heads for efficient attention computation. The visual encoder uses a Transformer with a depth of 32, a hidden layer size of 1280, and a patch size of 14. Images are processed via spatial pooling with a pool size of 2. Our training strategy prioritizes maximizing spatial pointing performance over maintaining general visual-language capabilities. Unlike typical fine-tuning methods, which mix domain-specific data with general question-answering datasets to preserve broad functionality, we train exclusively on the NaviAfford spatial affordance dataset. This focused training approach forces the model to focus on precise object localization and spatial reasoning, intentionally sacrificing general conversational capabilities in exchange for superior pointing accuracy. This design choice is consistent with our specific use case, where the model serves as a dedicated Pointing-VLM component within the hierarchical *NavA*^{3} framework rather than a general-purpose helper.

B.2 Training Configuration

This model was trained for a full epoch on a dataset of 1 million examples using full parameter fine-tuning techniques on 8 NVIDIA A800 GPUs (80GB of memory each). Total training time

Table 5: Results of HM3D object navigation benchmark.

Method	HM3D	
	SR↑	SPL↑
Randomly	0	0
CoW (Gadre et al., 2023)	32.0	18.1
ESC (Zhou et al., 2023)	38.5	22.0
PixNav (Cai et al., 2023)	37.9	20.5
L3MVN (Yu et al., 2023)	50.4	23.1
NaVid (Zhang et al., 2024b)	32.5	21.6
MFNP (Zhang et al., 2025b)	54.3	20.5
VLFM (Yokoyama et al., 2024a)	52.5	30.4
VLFM w/ NaviAfford (Ours)	**58.5**	**32.4**

Table 6: Evaluation of HM3D-OVON benchmark.

Method	Val-Unseen	
	SR↑	SPL↑
BC (Pomerleau, 1988)	5.4	1.9
DAgger (Ross et al., 2011)	5.4	1.9
RL (Schulman et al., 2017)	10.2	4.7
DAgRL (Chen et al., 2019)	18.3	7.9
BCRL (Wang et al., 2019)	18.6	7.5
DAgRL+OD (Yokoyama et al., 2024b)	37.1	19.8
Uni-NaVid (Zhang et al., 2024a)	39.5	19.8
MTU3D (Zhu et al., 2025)	40.8	12.1
VLFM (Yokoyama et al., 2024a)	35.2	19.6
VLFM w/ NaviAfford (Ours)	**42.3**	**20.4**

was approximately 24 hours, consuming 192 GPU hours (8 GPUs × 24 hours) of compute resources. We used the AdamW optimizer with a learning rate of 1e-5, which achieves stable convergence without complex learning rate scheduling. Each GPU processed a batch size of 4 examples, and gradient accumulation was set to 2 steps, resulting in an effective global batch size of 64 (8 GPUs × 4 batch sizes × 2 accumulation steps). Bfloat16 mixed precision was used during training to reduce memory usage and speed up computation while maintaining numerical stability. We used the standard supervised fine-tuning (SFT) objective function with a cross-entropy loss, as defined in Equation 4 in the main paper. The model handles variable-length sequences, with a maximum positional embedding of 128,000 tokens, although typical training examples are significantly shorter. No dropout was applied ($attention_{dropout}$ = 0.0) as the dataset is large enough to provide sufficient regularization to prevent overfitting.

C Additional Experiments on *NaviAfford*

To further validate the generalization capability and robustness of our NaviAfford model beyond the long-horizon navigation task, we conduct extensive experiments on standard embodied navigation benchmarks. Specifically, we evaluate NaviAfford's performance on Object Navigation (ObjectNav) and Open-Vocabulary Object Navigation (OVON) tasks, demonstrating its effectiveness as a general-purpose spatial localization module.

Analysis on HM3D Object Navigation Performance As shown in Tab. 5, results on the HM3D object navigation benchmark demonstrate that NaviAfford significantly advances the state-of-the-art in object localization in somatosensory navigation space. Our NaviAfford-enhanced VLFM achieves a 58.5% success rate and 32.4% SPL, surpassing all previous methods, including L3MVN (50.4% SR, 23.1% SPL), MFNP (54.3% SR, 20.5% SPL), and the original VLFM baseline (52.5% SR, 30.4% SPL). The SPL improvement of 11.9 percentage points over MFNP (32.4% vs. 20.5%) is particularly significant, demonstrating that NaviAfford's point-based localization approach not only achieves a higher success rate but also leads to more efficient navigation trajectories. This superior performance stems from our spatial affordance training strategy: by learning from 1 million samples covering diverse spatial relations, NaviAfford provides more accurate object localization, reduces unnecessary exploration, and facilitates direct path planning towards target objects.

Analysis on Open-Vocabulary Object Navigation (OVON) Performance As shown in Tab. 6, the open-vocabulary object navigation results on HM3D-OVON further demonstrate NaviAfford's generalization capabilities in challenging scenarios involving novel object categories. Our method achieves 42.3% SR and 20.4% SPL, surpassing previous state-of-the-art methods including Uni-NaVid (39.5% SR, 19.8% SPL) and MTU3D (40.8% SR, 12.1% SPL). The 7.1 percentage point SR improvement over the VLFM baseline (42.3% vs. 35.2%) demonstrates stronger zero-shot transfer to unseen objects compared to the standard ObjectNav setting, indicating that our training on 1,200+ object categories with 25+ spatial relation types enables generalizable spatial reasoning patterns. The comparison with MTU3D is particularly revealing: while achieving similar success rates (42.3% vs. 40.8%), NaviAfford's 8.3 percentage point SPL advantage (20.4% vs. 12.1%) shows

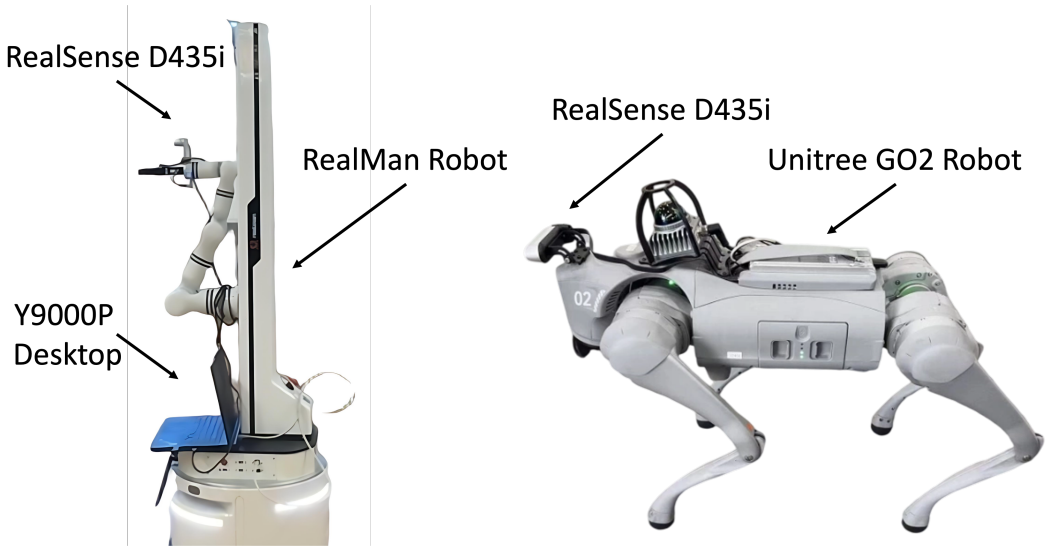


Figure 6: Real-world Deployment Embodiments.

that our model can quickly and accurately identify novel objects without extensive exploration, confirming the practical value of integrating vision-language understanding with specialized spatial affordance training for real-world embodied navigation applications.

D Real-World Deployment Details

This section provides comprehensive implementation details for deploying *NavA*^{3} on physical robot platforms in real-world environments, including hardware configuration, software architecture, and system integration.

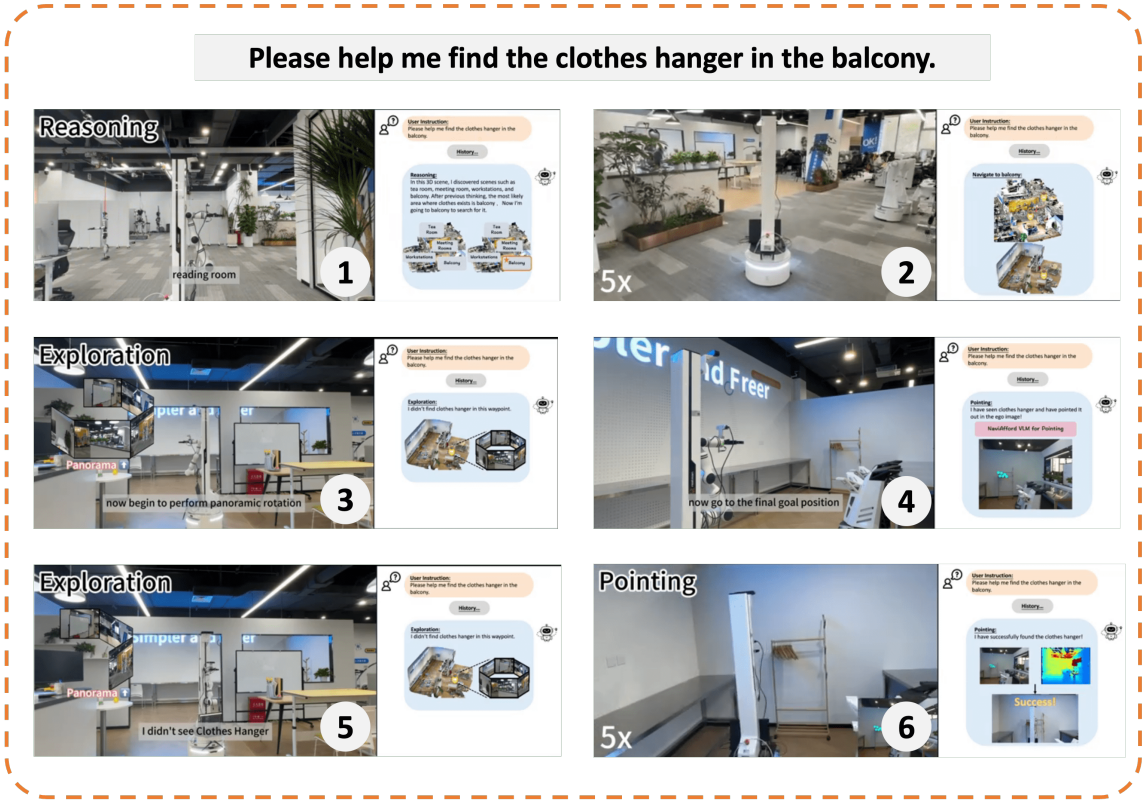
Robot Platform and Sensor Configuration We deploy *NavA*^{3} on two distinct robot platforms to demonstrate cross-embodiment capabilities: a RealMan wheeled mobile manipulator and a Unitree Go2 quadruped robot. The RealMan robot provides a stable wheeled platform for long-horizon navigation tasks in indoor environments, equipped with a single robotic arm for potential manipulation tasks, though our current navigation experiments focus primarily on mobility. The Unitree Go2 quadruped robot offers dynamic legged locomotion capabilities, enabling navigation in more complex terrains and demonstrating the versatility of our framework across different embodiments. Both platforms are equipped with Intel RealSense D435i cameras as the primary perception sensor. On the RealMan robot, the camera is mounted at a fixed height of 1.0 meter above the ground, positioned facing forward with a slight downward tilt to capture both the immediate surroundings and the floor area for obstacle detection. On the Unitree Go2, the camera is fixed to the robot's head module, maintaining a similar forward-facing orientation with adaptive height based on the robot's posture. The RealSense D435i provides synchronized RGB and depth streams at 640×480 resolution and 30 FPS, with a depth range of 0.3 to 3.0 meters and a wide field of view (87° × 58°) suitable for comprehensive scene coverage. For panoramic perception at waypoints, both robots perform rotational scanning by rotating in place while capturing RGB-D observations at 45-degree intervals, generating 8 viewpoints that collectively provide 360-degree coverage of the local environment for accurate object localization.

Computing Infrastructure and System Integration The computing architecture is distributed, with dedicated hardware for different *NavA*^{3} components. A Lenovo Y9000P laptop powered by an NVIDIA RTX 4090 mobile GPU serves as the master control station, responsible for robot control, sensor data acquisition, cloud-based API reasoning for the global policy (GPT-4o (Hurst et al., 2024) for Reasoning-VLM), and overall system coordination. The global policy API typically responds within 2-5 seconds, which is acceptable given that spatial reasoning rarely occurs during region transitions. For local policy, a separate desktop workstation powered by an NVIDIA RTX 4090 GPU is deployed, dedicated to running the 7B NaviAfford model at bfloat16 precision, achieving a frame rate of approximately 2-3 FPS for panoramic analysis. Communication between the control laptop and the inference workstation occurs over the local network using a custom RPC protocol. When the

robot reaches a waypoint, RGB images from all eight panoramic views are sent to the workstation along with a goal query. NaviAfford returns the point coordinates, which are converted to robot coordinates using camera intrinsics and the robot's current pose. This distributed architecture remains consistent across both robot platforms, requiring only the underlying motion control interface to be adapted for each implementation, demonstrating the versatility of our navigation framework.

E More Qualitative Visualizations

This section presents a number of qualitative visualizations of ***Nav***A^3 deployed in real-world environments, demonstrating its robustness across diverse scenarios, commands, and robot implementations. As shown in Fig. 7-Fig. 11, these visualizations cover multiple indoor environments, including a conference room, a tea room, a workstation, and a balcony. Each environment presents unique spatial layouts and object arrangements, posing challenges for navigation systems. The figure traces the complete navigation trajectory from command understanding to global reasoning, local exploration, and final goal achievement, highlighting key decision points for the Reasoning-VLM to identify the target region and the NaviAfford model to perform precise object pointing. Notably, the visualizations cover experiments with the RealMan wheeled manipulator and the Unitree Go2 quadruped robot, demonstrating consistent performance across different modalities despite their distinct motion mechanisms and kinematic constraints. Successfully completing a variety of tasks across diverse environment complexity, command types, and robot platforms demonstrates the versatility and practical applicability of our hierarchical navigation framework for long-horizon navigation task.



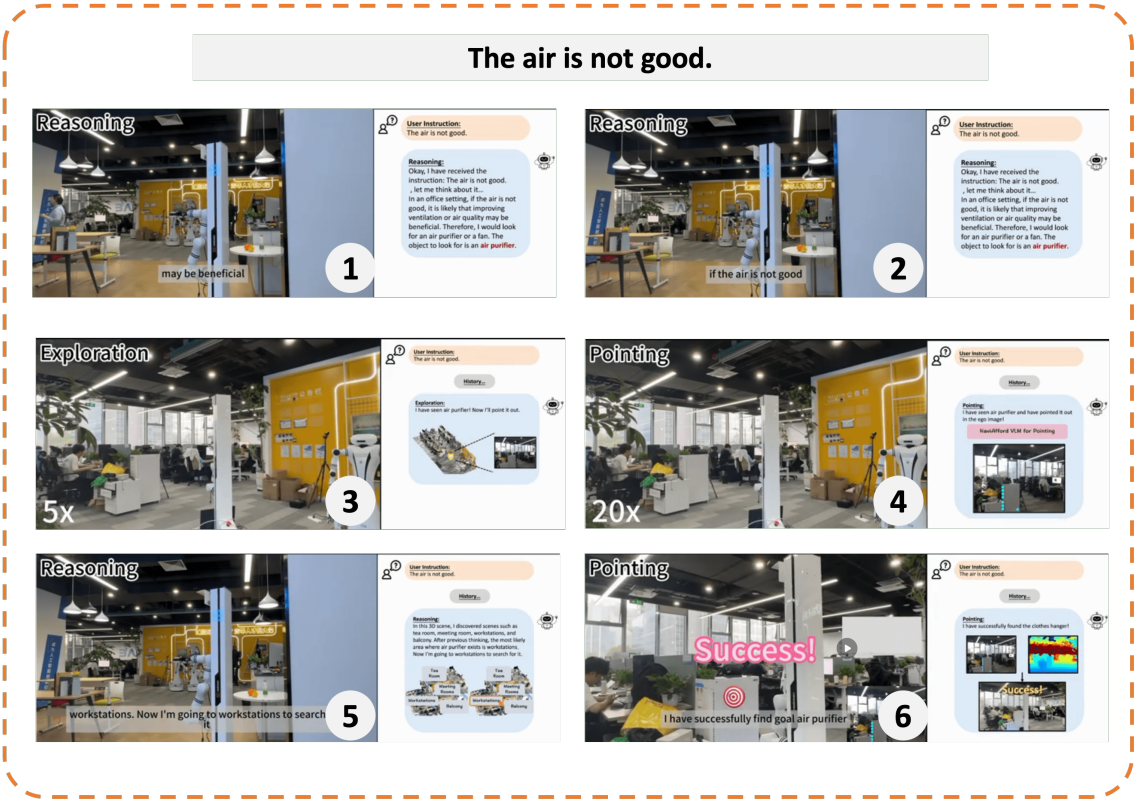


Figure 7: More $NavA^3$ Real-world Deployment Demos.

Figure 8: More *NavA*^{3} Real-world Deployment Demos.

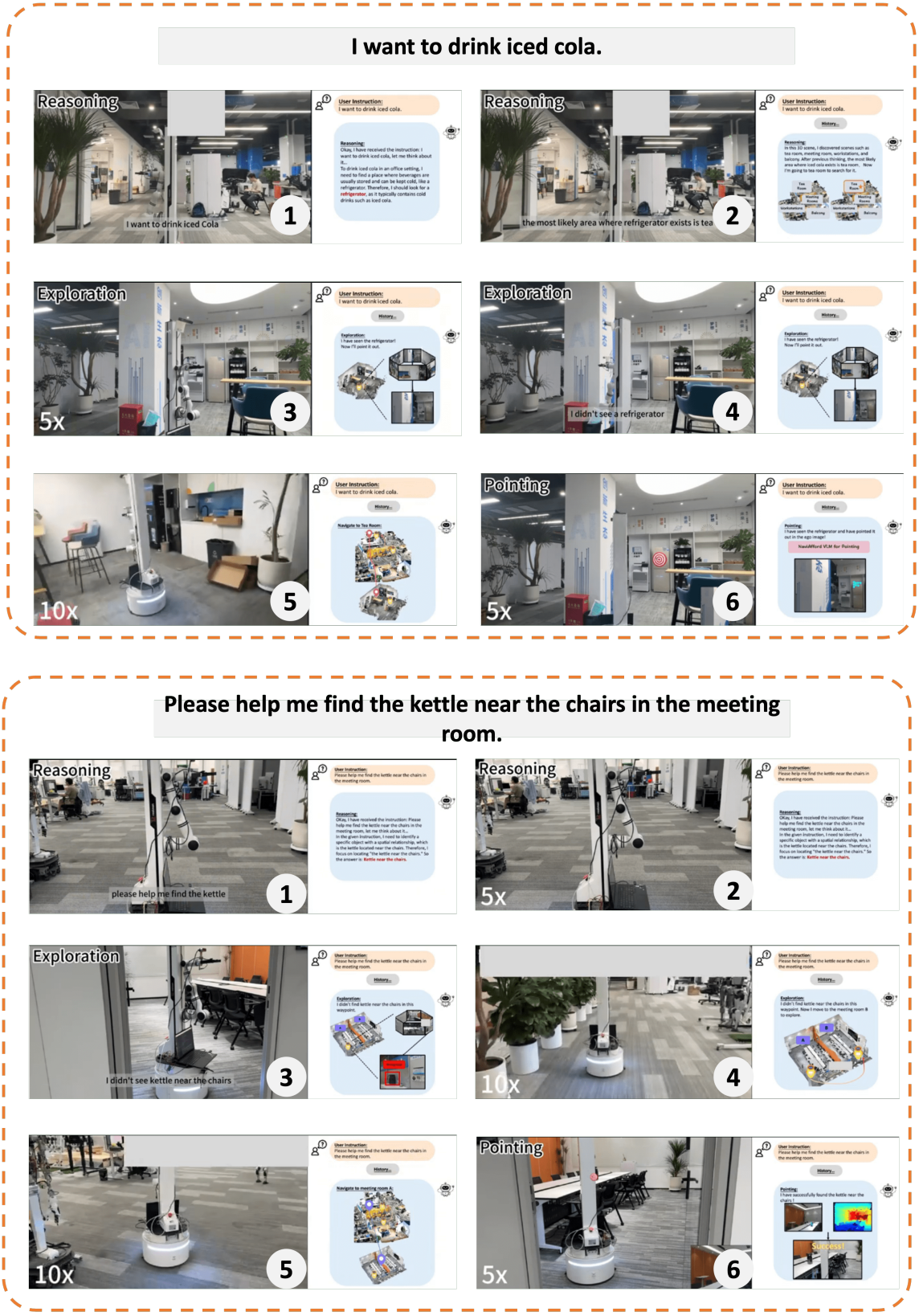


Figure 9: More *NavA*^{3} Real-world Deployment Demos.

Figure 10: More *NavA*^{3} Real-world Deployment Demos.

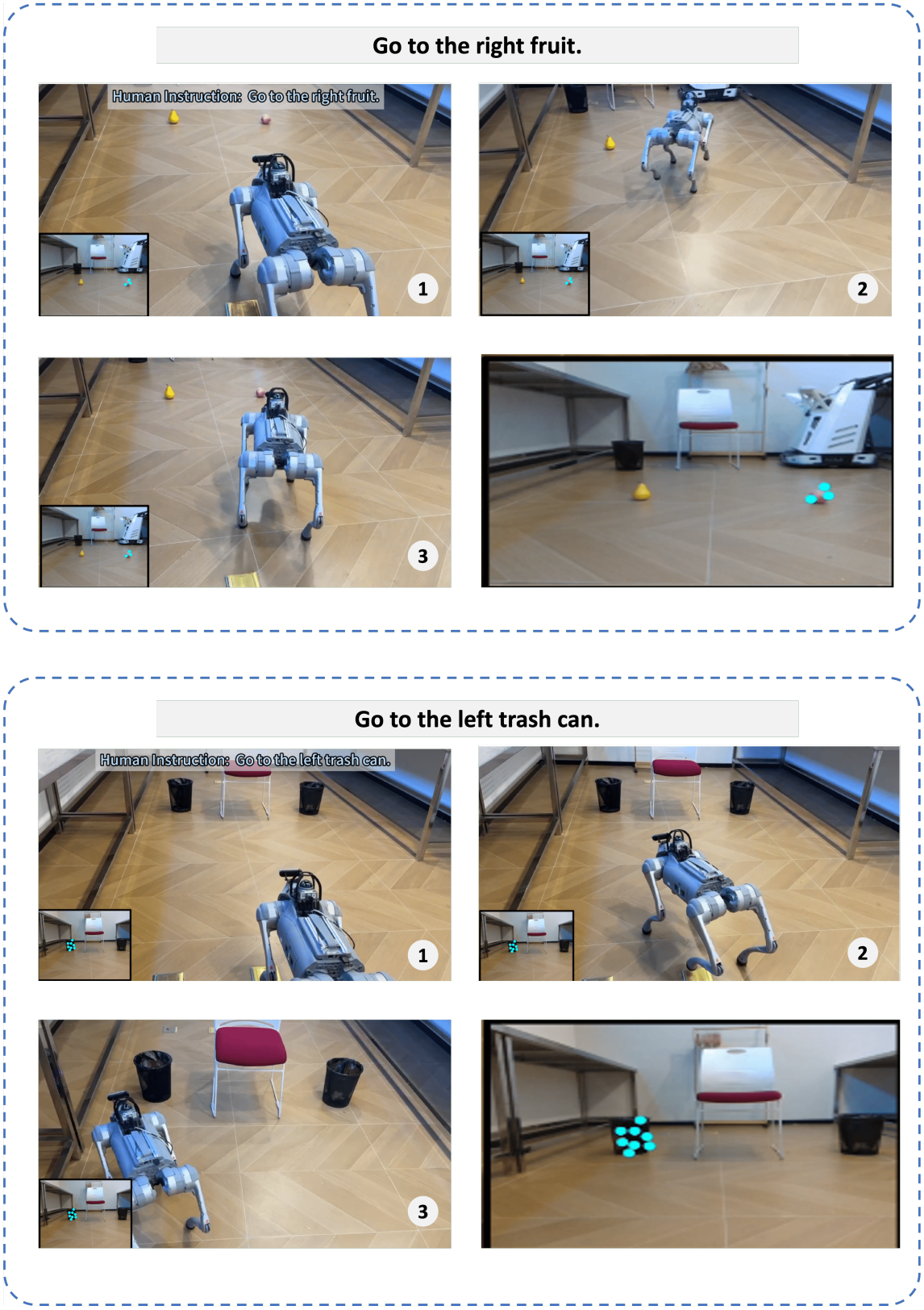


Figure 11: More *NavA*^{3} Real-world Deployment Demos.