What You See is What You Reach: Towards Spatial Navigation with High-Level Human Instructions

Lingfeng Zhang[1,2,3,*], **Haoxiang Fu**[4,*], **Xiaoshuai Hao**[3,†,‡], **Shuyi Zhang**[5], **Qiang Zhang**[6], **Rui Liu**[7], **Long Chen**[3], **Wenbo Ding**[1,†]

[1] Tsinghua Shenzhen International Graduate School, Tsinghua University
[2] Peng Cheng Laboratory
[3] Xiaomi EV
[4] National University of Singapore
[5] Institute of Automation, CAS
[6] HKUSTGZ
[7] Inner Mongolia University
lfzhang715@gmail.com, haoxiaoshuai@xiaomi.com, ding.wenbo@sz.tsinghua.edu.cn

Abstract

Embodied navigation is a fundamental capability that enables embodied agents to effectively interact with the physical world in various complex environments. However, a significant gap remains between current embodied navigation tasks and real-world requirements, as existing methods often struggle to integrate high-level human instructions with spatial understanding. To address this gap, we propose a new task of embodied navigation called *spatial navigation*, which encompasses two key components: *spatial object navigation* (***SpON***) for object-specific guidance and *spatial area navigation* (***SpAN***) for navigating to designated areas. Specifically, ***SpON*** guides agents to specific objects by leveraging spatial relationships and contextual understanding, while ***SpAN*** focuses on navigating to defined areas within complex environments. Together, these components significantly enhance agents' navigation capabilities, enabling more effective interactions in real-world scenarios. To support this task, we have generated a ***spatial navigation dataset*** consisting of 10K trajectories within the simulator. This dataset includes high-level human instructions, detailed observations, and corresponding navigation actions, providing a comprehensive resource to enhance agent training and performance. Building on the *spatial navigation dataset*, we introduce ***SpNav***, a hierarchical navigation framework. Specifically, ***SpNav*** employs vision-language model (VLM) to interpret high-level human instructions and accurately identify goal objects or areas within the observation range, achieving precise point-to-point navigation using a map and enhancing the agent's ability to operate effectively in complex environments by bridging the gap between perception and action. Extensive experiments show that ***SpNav*** achieves state-of-the-art (SOTA) performance in spatial navigation tasks across both simulated and real-world environments, validating the effectiveness of our method.

Introduction

Embodied navigation is crucial for autonomous agents operating in physical environments, serving as the foundation for advanced robotic tasks such as mobile manipulation, exploration, and human-robot interaction (Hao et al. 2025; Zhang et al. 2025a; Tang et al. 2025b; Team et al. 2025; Li et al. 2024a). This capability allows agents to navigate and interact effectively within real-world spaces, enabling them to perform complex tasks across a diverse range of scenarios. (Wu et al. 2025, 2024b; Zheng et al. 2024; Zhang et al. 2024d; Tang et al. 2022; Xu et al. 2024; Liu et al. 2025c; Liu, Guo, and Cangelosi 2025; Ma et al. 2024; Liu et al. 2025b)

Existing embodied navigation research primarily focuses on two main paradigms: vision-language navigation (VLN) (Zhang et al. 2025c; Gao et al. 2025; Cheng et al. 2025; Zhang et al. 2024a) and object-goal navigation (ObjectNav) (Zhang et al. 2025d, 2024c; Wu et al. 2024a; Long et al. 2025; Gong et al. 2025). VLN approaches require agents to execute detailed sequential commands, such as "turn left, go through the doorway, and keep going," demanding meticulous spatial understanding but often relying on overly specific instructions that do not reflect natural human communication patterns. In contrast, ObjectNav centers on recognizing predetermined object categories (e.g., "find any chair") and making conclusions based on representative instances, independent of contextual location or specific user needs. However, real human instructions frequently embody abstract intent and necessitate complex reasoning and environmental perception. For instance, instructions like "Wait for me around the door" or "Help me get the fruit from the left side of the tea area" require not only basic goal navigation but also a nuanced understanding of spatial relationships among objects and regions. These discrepancies between current navigation frameworks and practical applications highlight a critical need for advancements in the field, ***strongly motivating the design of a navigation task that can effectively comprehend complex human instructions and reason about spatial relationships***, thereby enhancing the

[*] These authors contributed equally.
[†] Corresponding Authors
[‡] Project Leader

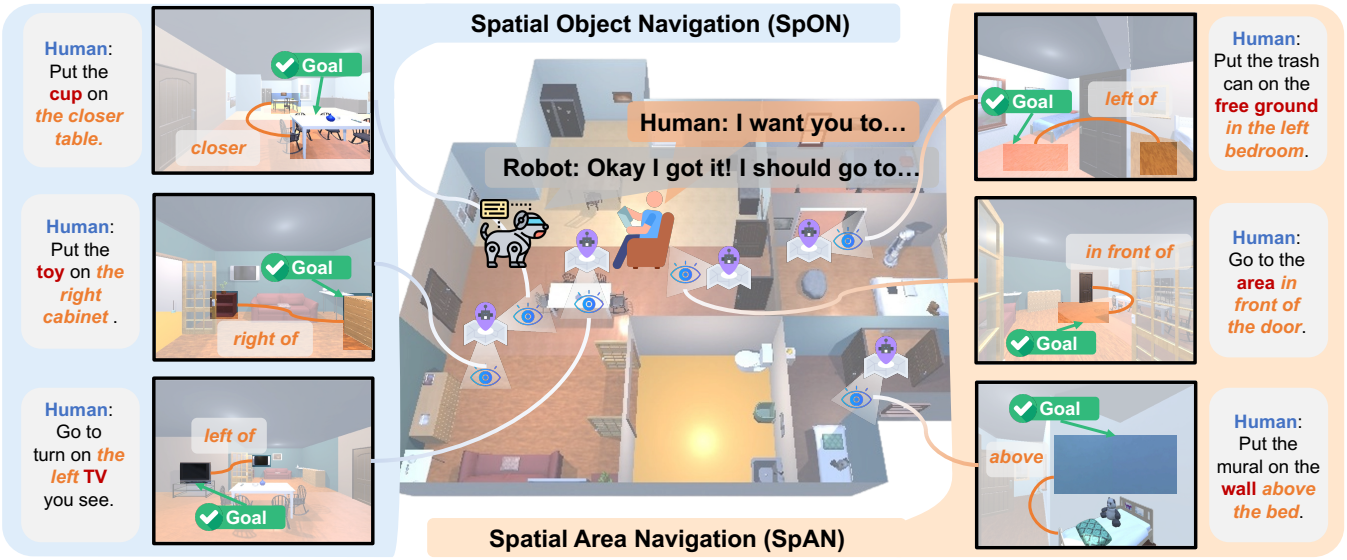

Figure 1: **Overview of Spatial Navigation.** Our spatial navigation requires the agent to understand high-level human instructions and navigate to a specific object or area. This task encompasses two key components: Spatial Object Navigation (***SpON***), where agents are tasked with locating specific objects based on spatial constraints (*e.g.*, "Put the cup on the closer table"), and Spatial Area Navigation (***SpAN***), which involves navigating to designated areas defined by spatial relationships (*e.g.*, "Go to the area in front of the door").

capabilities of embodied agents in real-world scenarios.

To address this critical gap, as shown in Fig. 1, we introduce the spatial navigation task, which requires agents to understand high-level human instructions and navigate to specific objects or areas. This task is divided into two complementary components: *Spatial Object Navigation* (***SpON***) and *Spatial Area Navigation* (***SpAN***). The ***SpON*** component enables agents to locate specific objects through an understanding of spatial relationships and contextual reasoning, while ***SpAN*** focuses on navigating to designated environmental areas within complex environments. Together, these components significantly enhance the navigation capabilities of agents in real-world scenarios. To support this novel task, we developed a comprehensive ***spatial navigation dataset*** comprising 10,000 trajectories generated in the AI2THOR simulator, with 5,000 trajectories dedicated to ***SpON*** and 5,000 to ***SpAN***. Each trajectory contains a sequence of navigation actions spanning 10-30 timesteps and includes high-level human instructions, egocentric observations, and corresponding navigation action sequences, thereby laying a solid foundation for agent training and evaluation. Based on this dataset, we propose ***SpNav***, a hierarchical navigation framework aiming to achieve the principle of "***what you see is what you get.***" Specifically, our hierarchical approach decomposes the spatial navigation task into structured stages: first, a general Vision-Language Model (VLM) is employed to reason about human instructions and extract goal objects or regions (*e.g.*, "clean the right window → right window"); subsequently, our dedicated ***NaviPoint*** is utilized for accurate visual goal pointing. After identifying the target point from the egocentric image, we employ ***Map-to-Action***, based on a constructed map, to perform coordinate transformation and precise point-to-point navigation, enabling the agent to reach its final destination. Extensive experiments demonstrate that ***SpNav*** exhibits superior performance in spatial navigation tasks, outperforming all baseline methods. The main contributions of our work can be summarized as follows:

- We introduce a challenging spatial navigation task with two components: *Spatial Object Navigation* (***SpON***) and *Spatial Area Navigation* (***SpAN***). The agent must interpret high-level human instructions to navigate to objects or areas while considering spatial relationships.
- For (***SpON***) and (***SpAN***), we collected and generated 5,000 episodes for each, totaling 10,000 episodes. Each episode includes high-level human instructions, goal locations, and navigation action sequences.
- We propose a novel hierarchical framework, ***SpNav***, which leverages VLMs for instruction reasoning and spatial-aware goal pointing. Additionally, we introduce ***Map-to-Action*** to facilitate precise point-to-point navigation to the final destination.
- Extensive experiments show that ***SpNav*** achieves state-of-the-art performance in spatial navigation, surpassing all baseline methods and paving the way for future advancements in embodied navigation systems.

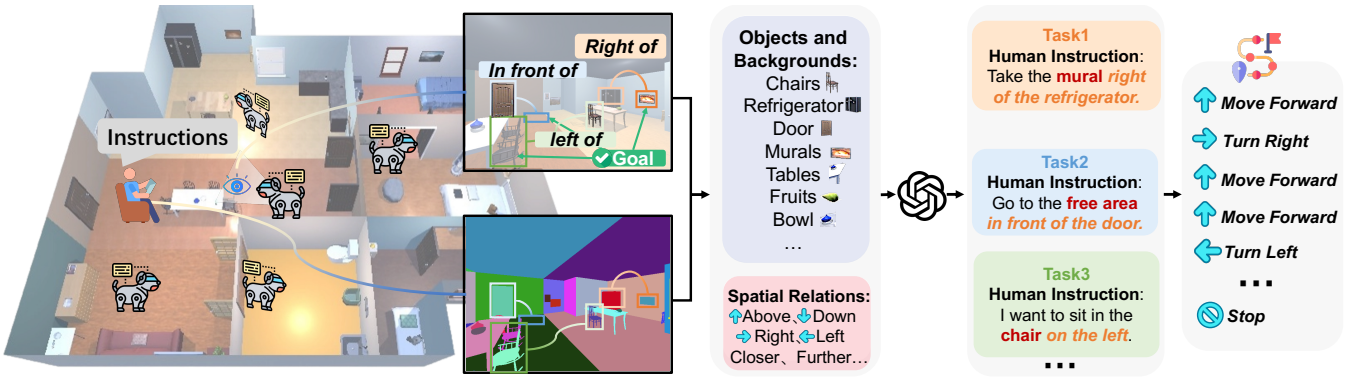

Figure 2: **Dataset Generation Progress for Spatial Navigation Tasks.** We collect egocentric observations to extract spatial relationship triplets (goal object, spatial relation, reference object). These triplets are processed by vision-language models (VLMs) to generate diverse high-level human instructions, which are paired with expert trajectory sequences, resulting in a comprehensive spatial navigation dataset for robotic training.

Related Work

Embodied Navigation Research in embodied navigation primarily focuses on vision-language navigation (VLN) and object-goal navigation (ObjectNav) (Zhang et al. 2025e,b; Liu et al. 2025a). Key VLN efforts include NavGPT (Zhou, Hong, and Wu 2024), which leverages GPT-4o for action generation; DiscussNav (Long et al. 2024), which minimizes human supervision; and Nav-CoT (Lin et al. 2025), enhancing environmental understanding through thought chaining. MapNav (Zhang et al. 2025c) optimizes memory with map-based spatial representations, while NaVid (Zhang et al. 2024b) maintains temporal context. In ObjectNav, PirlNav (Ramrakhya et al. 2023) and XGX (Wasserman et al. 2024) mimic human demonstrations, and semantic mapping approaches like InstructNav (Long et al. 2025) enable zero-shot navigation. However, current methods primarily emphasize step-by-step instruction following and predefined object categories, limiting their ability to interpret high-level instructions and complex spatial relationships, which constrains their applicability to our proposed navigation tasks.

VLMs for Spatial Reasoning Effective spatial reasoning enhances robotic systems' performance in navigating and manipulating physical environments. (Liu, Emerson, and Collier 2023; Wang et al. 2024; Zha et al. 2025; Zhou et al. 2024; Tang et al. 2025a; Xiao et al. 2025) Recent research has aimed to improve the spatial understanding capabilities of vision-language models (VLMs) through various approaches. For instance, SpatialVLM (Chen et al. 2024) utilizes metric depth estimation to convert visual input into object-centric point cloud representations for spatial visual question answering (VQA). SpatialRGPT (Cheng et al. 2024) advances region-based spatial reasoning with a proposal mechanism and spatial scene graph construction. RoboPoint (Yuan et al. 2025) offers a synthetic dataset for unrestricted spatial reference and accurate action point prediction. SpatialBot (Cai et al. 2024) and RoboRefer (Zhou et al. 2025) enhance spatial understanding by integrating RGB-D modalities. However, existing methods often lack spatial-aware instance localization as well as seamless integration with navigation frameworks, which limits their effectiveness in real-world robotic navigation scenarios.

Methodology

Overview

As shown in Fig. 2, we leverage the ground truth semantic graph from the AI2THOR simulator to establish spatial relationships between objects and regions, identifying relationships like ***"left," "right," "above," and "below."*** Using a VLM, we generate high-level human instructions such as "pick up the mural on the right side of the refrigerator." Our ***SpNav*** framework, illustrated in Fig. 3, operates in two stages: during training, we create question-answer pairs from spatial relationship data to enhance NaviPoint's pointing ability; during inference, given a high-level instruction, ***NaviPoint*** processes RGB-D observations to locate the goal. We then use ***Map-to-Action*** to convert image coordinates to map coordinates, enabling effective path planning and successful navigation to complete the spatial navigation task.

Preliminaries

Problem Definition In our spatial navigation task, an agent starts at a random position and orientation in an unknown indoor environment. Given a high-level human instruction, the agent must navigate to either a specific object with spatial constraints (*SpON*) or a designated area (*SpAN*). At each timestep t, the system processes multimodal inputs: RGB-D observations o_t, natural language instruction l_t, and the robot's pose state r_t. The agent generates navigation actions as follows: *Spatial Navigation Agent:* $(o_t, l_t, r_t) \rightarrow a_{t+1}$, where $o_t = (rgb_t, depth_t)$, l_t specifies object or area constraints, and r_t denotes the current pose. The predicted action $a_{t+1} \in \{MoveAhead, RotateLeft, RotateRight, \text{Done}\}$. Success is achieved when the agent executes the *Done* action and reaches a final position p_{final} such that $\|p_{final} - g\|_2 \leq 1.0$ meters, facilitating consistent evaluation across both object-centric and area-centric navigation scenarios.

Map Construction Our map construction module, based on (Zhang et al. 2024c) and (Zhang et al. 2025d), creates a

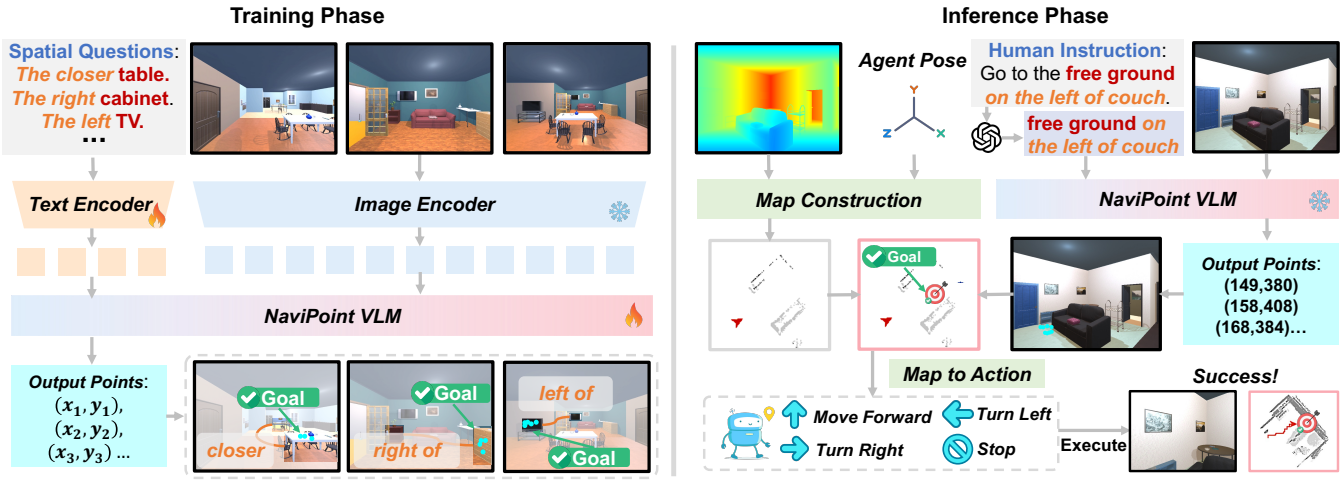

Figure 3: **Overview of our *SpNav* framework.** During training, we develop our ***NaviPoint*** model using spatial question-answer pairs to enhance goal-pointing capabilities. In the inference phase, we use a vision-language model (VLM) to interpret human instructions, determine egocentric goal coordinates with ***NaviPoint***, construct maps from RGB-D observations and pose data, and execute point-to-point navigation to reach the destination.

comprehensive environmental representation by integrating multimodal sensor observations from the AI2THOR simulator. We initialize a multi-channel map $\mathcal{M} \in \mathbb{R}^{C\times H\times W}$, where C is the number of geometric channels and $H \times W$ defines the spatial resolution. The map consists of layers $\mathcal{M}_{obs}$ (obstacle occupancy), $\mathcal{M}_{exp}$ (explored areas), and $\mathcal{M}_{free}$ (traversable areas), with each cell corresponding to a 5cm $\times$ 5cm real-world area.

At each timestep t, we receive RGB images I_t^{rgb}, depth maps D_t, and agent poses $p_t = (x_t, y_t, \theta_t)$. We convert depth data into a 3D point cloud and project it onto a 2D map. The map is updated using a probabilistic fusion approach, where obstacle occupancy and explored areas are adjusted based on sensor data, allowing us to accurately represent traversable space. This construction is crucial for effective spatial navigation tasks.

Dataset Construction

As shown in Fig. 2, our spatial navigation dataset is built through multimodal data collection, spatial relationship extraction, high-level instruction generation, expert trajectory verification. For each scene $s \in \mathcal{S}$, we randomly sample a proxy location $p \in \mathcal{P}_s$ and capture multimodal observations:

$$OBS = \{I_{rgb}, I_{depth}, I_{seg}, M_{semantic}\} = \phi(s, p, \theta), \quad (1)$$

where I_{rgb} is the RGB image, I_{depth} is the depth map, I_{seg} indicates instance segmentation, and $M_{semantic}$ contains semantic annotations. We collect 50,000 observation pairs.

From these pairs, we extract spatial relationships between objects. For two objects with centers (x_1, y_1) and (x_2, y_2), the relationship is determined by:

$$r_{ij} = \arg\max_{r\in R} \mathbb{I}[\psi_r((x_i, y_i), (x_j, y_j))], \quad (2)$$

where $R = \{left\_of, right\_of, above, below, near, next\_to\}$. Each spatial relationship results in a triple (o_i, r_{ij}, o_j). We then use GPT-4 to generate natural language instructions:

$$I_{instruction} = \mathcal{G}(o_i, r_{ij}, o_j, C_{context}), \quad (3)$$

ensuring semantic coherence while preserving spatial information, yielding commands like "navigate to the free area left of the sofa." This process generates 10,000 question-answer pairs, split evenly between 5,000 *SpON* tasks (object navigation) and 5,000 *SpAN* tasks (area navigation).

To create realistic navigation sequences, we employ the A* algorithm to plan optimal trajectories from the agent's initial position to the goal:

$$\tau^* = A^*(p_{start}, p_{goal}, \mathcal{M}_{collision}), \quad (4)$$

where $\tau^* = \{(s_0, a_0), (s_1, a_1), \ldots, (s_T, a_T)\}$ represents the state-action sequence, T ranges from 10 to 30 time steps, p_{start} is the agent's starting position, p_{goal} is the goal object's 3D position, and $\mathcal{M}_{collision}$ is a collision-free navigation mesh. Validity is confirmed through successful path completion. Each navigation trajectory comprises observations, expert actions, spatial annotations, and instructions for effective training and evaluation of spatial understanding.

SpNav

Training Phase In the training phase, we developed the ***NaviPoint*** model for accurate object pointing capabilities using a dataset of 200,000 spatial relation question-answer pairs (100,000 for object goals and 100,000 for region goals) collected from 200 indoor scenes in the AI2THOR simulator. Each pair includes a spatial question, such as "point out the free area on the table to the left of the chair," and a corresponding ground truth point derived from precise annotations. The model takes input as a tuple (I, Q)—where $I \in \mathbb{R}^{H\times W\times 3}$ is the RGB observation and $Q = \{q_1, q_2, \ldots, q_L\}$ is the tokenized spatial question—and generates text-based point coordinates

through the mapping function $f_\theta : (I, Q) \rightarrow Y$, where $Y =$ "$[(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)]$. The architecture combines a visual encoder and a language model to understand spatial queries and produce coordinate responses, fine-tuned using a cross-entropy loss function, enabling precise spatial reasoning and effective navigation:

$$\mathcal{L}_{SFT} = -\sum_{i=1}^{|Y|} \log P(y_i | y_{<i}, I, Q; \theta), \tag{5}$$

where y_i represents the ith token in the goal sequence Y. This enables the model to learn precise spatial reasoning and coordinate pointing capabilities from the rich spatial relationship annotations in the dataset, laying the foundation for effective spatial navigation.

Inference Phase In the inference phase, our SpNav framework performs spatial navigation tasks by integrating high-level instruction reasoning, precise goal localization, map construction, and dynamic path planning to achieve robust navigation performance in complex indoor environments. The inference process begins with a VLM for reasoning, which reasons on the input high-level human instruction $I_{instruction}$ to extract the goal description, (*e.g.*, "go to the open space on the left side of the sofa" $\rightarrow$ "the open space on the left side of the sofa"). Subsequently, we input the current RGB observations and the extracted goal description into the trained NaviPoint to obtain the goal coordinates observed from the egocentric observation. At the same time, we use the RGBD observations and pose information from the AI2THOR simulator to build and continuously update the map $\mathcal{M}^t$ described in ***Section Preliminaries***.

During ***Map-to-Action***, the key coordinate transformation uses the transformation matrix $P_{global} = T_{ego \rightarrow global} \cdot P_{ego}$ to transform the goal point observed from egocentric observations into the global map coordinates:

$$T_{ego \rightarrow global} = \begin{bmatrix} \cos\theta_t & -\sin\theta_t & x_t \\ \sin\theta_t & \cos\theta_t & y_t \\ 0 & 0 & 1 \end{bmatrix}. \tag{6}$$

Then we map the global coordinates to the spatial grid using $P_{map}[u, v] = \lfloor P_{global} \cdot \text{resolution}^{-1} \rfloor$, where resolution denotes the map's spatial resolution. Using the goal map coordinates, we apply a fast marching method (FMM) path planner to compute the optimal trajectory over the traversable areas of the spatial map, generating waypoints $W = \{w_1, w_2, \ldots, w_k\}$ that define the navigation path. These waypoints are then converted into discrete actions via a policy function $a_t = \pi(w_{current}, w_{next}, \theta_{agent})$, with action selection following the formula:

$$a_t = \begin{cases} \textit{MOVE\_FORWARD} & \textit{if } |\theta_{relative}| < 15^\circ \\ \textit{TURN\_LEFT} & \textit{if } \theta_{relative} < -15^\circ \\ \textit{TURN\_RIGHT} & \textit{if } \theta_{relative} > 15^\circ \\ \textit{STOP} & \textit{if } \text{reached goal} \end{cases}, \tag{7}$$

where $\theta_{relative}$ denotes the angle between the agent's direction and the next waypoint. During navigation, we continuously update $\mathcal{M}^t$ with new RGB-D observations and poses, conduct real-time obstacle detection by marking obstacles in $\mathcal{M}_{obs}$, and replan the trajectory as needed to adapt to dynamic changes in the environment. This approach ensures robust navigation performance until the agent reaches the goal location within the specified proximity threshold.

Experiments

Experimental Details

Evaluation Benchmark To comprehensively evaluate the performance of various methods on spatial navigation tasks, we created a challenging benchmark dataset in the AI2THOR simulator (Kolve et al. 2017), featuring previously unseen indoor scenes. This dataset includes 1,315 navigation trajectories, evenly split between 713 SpON tasks and 602 SpAN tasks, ensuring complete separation between the training and evaluation environments. Each trajectory contains high-level human instructions, such as "navigate to the empty space on the left side of the sofa," along with the corresponding ground truth goal location. Our benchmark assesses the agent's ability to perform real-time interactions, generate appropriate navigation actions, and successfully reach the specified goal.

Evaluation Metrics We assess the performance of our methods using four standard metrics commonly employed in embodied navigation research: navigation error (NE), path length (PL), success-weighted path length (SPL), and success rate (SR). Navigation error measures the Euclidean distance between the agent's final position and the goal, with lower values indicating better performance. Path length represents the total distance traveled during navigation. SPL combines success rate and efficiency, defined as $SPL = \frac{1}{N}\sum_{i=1}^{N} S_i \times \left(\frac{L_i}{\max(P_i, L_i)}\right)$, where S_i indicates success, L_i is the shortest path length, P_i is the actual path length, and N is the total number of scenarios. Success rate reflects the percentage of scenarios in which the agent successfully reaches within 1.0 meter of the goal. We evaluate these metrics separately for SpON (spatial object navigation) and SpAN (spatial area navigation), reporting the average performance over all 1,315 trajectories to provide a comprehensive assessment of spatial navigation capabilities across different task types.

Implementation Details For goal extraction from high-level human instructions, we utilize GPT-4o (Hurst et al. 2024), a generic reasoning VLM. The NaviPoint model, responsible for precise goal pointing, is initialized with pre-trained Qwen2.5-VL-7B (Bai et al. 2025) weights and undergoes supervised fine-tuning using a standard instruction-following protocol. Training is conducted on four A100 GPUs with the AdamW (Kingma and Ba 2014) optimizer and a learning rate of 10^{-5} per cycle. The batch size per GPU is set to 4, with 2 steps of gradient accumulation, resulting in an effective batch size of 32. During inference, the spatial map module maintains a 480×480 grid at a 5 cm resolution, while the FMM path planner (Engheta et al. 1992) operates on a binary occupancy map to generate collision-free trajectories. Coordinate transformations between egocentric image coordinates and global map coordinates leverage the agent's real-time pose information from the AI2THOR simulator (Kolve et al. 2017).

Category	Methods	Spatial Navigation											
		SpON				SpAN				Average			
		NE↓	PL↓	SPL↑	SR↑	NE↓	PL↓	SPL↑	SR↑	NE↓	PL↓	SPL↑	SR↑
Closed-source	GPT-4o (Hurst et al. 2024)	4.78	12.34	12.3	4.8%	5.23	13.67	9.1	3.2%	5.01	13.01	10.7	4.0%
	Claude-3.5-Sonnet (Anthropic 2024)	3.89	11.45	15.8	6.1%	4.34	12.78	12.4	4.7%	4.12	12.12	14.1	5.4%
	Qwen-VL-Max (Bai et al. 2025)	5.45	13.89	8.7	3.4%	6.12	15.23	6.9	2.8%	5.79	14.56	7.8	3.1%
Open-source	Janus-Pro-7B (Chen et al. 2025)	6.67	18.45	0.0	0.0%	7.23	21.34	0.0	0.0%	6.95	19.90	0.0	0.0%
	Qwen2.5-VL-7B (Bai et al. 2025)	6.34	22.78	0.0	0.0%	6.89	25.67	0.0	0.0%	6.62	24.23	0.0	0.0%
	LLaVA-Next-7B (Li et al. 2024b)	5.89	20.12	0.0	0.0%	6.45	23.45	0.0	0.0%	6.17	21.79	0.0	0.0%
Navigation-specific	NaVid* (Zhang et al. 2024b)	3.23	8.67	28.9	14.2%	3.82	9.89	25.7	11.8%	3.53	9.28	27.3	13.0%
	NaVILA* (Cheng et al. 2025)	2.78	7.94	32.4	18.6%	3.25	8.73	29.1	15.9%	3.02	8.34	30.8	17.3%
	MapNav* (Zhang et al. 2025c)	2.21	6.89	33.7	22.4%	2.87	7.56	30.5	19.7%	2.54	7.23	32.1	21.1%
	SpNav (Ours)	**1.02**	**3.14**	**35.4**	**42.3%**	**1.18**	**3.67**	**34.1**	**39.8%**	**1.10**	**3.41**	**34.8**	**41.1%**

Table 1: **Comparison with SOTA methods on spatial navigation task.** * denotes that we adapt their task formulation to complete our spatial navigation task.

Baseline Methods Since existing navigation methods are not tailored for spatial relation instructions, we adapt their task formulation by modifying the prompt format to include explicit spatial guidance. For fair comparison, we adjust the instruction prompt to read: "Follow the following instructions to navigate to the goal: Wait for me in the empty space to the left of the sofa. Stop near the goal point." All baseline models are constrained to the same discrete action space and provided with equivalent observation information. We evaluate three types of baseline models: (1) closed-source general VLMs, including GPT-4o (Hurst et al. 2024), etc; (2) open-source general VLMs, such as Qwen2.5-VL-7B (Bai et al. 2025), etc; and (3) navigation-specific methods, including NaVid (Zhang et al. 2024b), etc.

Comparisons with SOTA Methods

We compare our SpNav framework against SOTA baseline methods across three categories on the spatial navigation benchmark, demonstrating significant performance gains across all evaluation metrics. As shown in Tab. 1, our method achieves notable improvements over existing approaches. Compared to the best-performing closed-source model, Claude-3.5-Sonnet (Anthropic 2024), SpNav realizes a 661% increase in success rate (SR), achieving 41.1% versus 5.4%. Additionally, it reduces navigation error (NE) by 73% (1.10 *vs.* 4.12) and improves success-weighted path length (SPL) by 147% (34.8 *vs.* 14.1). The open-source general-purpose VLM performs poorly, registering a 0% success rate, underscoring the difficulty of understanding spatial relationships without specialized training. Among navigation-specific methods, our framework outperforms the strongest baseline, MapNav (Zhang et al. 2025c), by 95% in SR (41.1% *vs.* 21.1%), reduces NE by 57% (1.10 *vs.* 2.54), and improves SPL by 8% (34.8 *vs.* 32.1). Our framework consistently excels in both SpON and SpAN tasks, demonstrating its effectiveness in handling both object-centric (SpON) and area-based (SpAN) spatial navigation scenarios. These results validate the efficacy of our hierarchical approach, which combines general reasoning with specialized spatial pointing for complex navigation tasks.

VLMs for Reasoning	Spatial Navigation			
	NE↓	PL↓	SPL↑	SR↑
SpNav (w/o Reasoning)	4.12	11.3	8.3	12.4%
Open-source				
SpNav (w/ Qwen2.5-VL-72B) (Bai et al. 2025)	2.12	6.78	24.3	28.9%
SpNav (w/ Qwen2.5-VL-7B) (Bai et al. 2025)	2.89	8.45	19.7	23.4%
SpNav (w/ LLaVA-NeXT-7B) (Li et al. 2024b)	3.34	9.23	16.8	19.6%
SpNav (w/ Janus-Pro-7B) (Chen et al. 2025)	3.78	10.67	14.2	16.8%
Closed-source				
SpNav (w/ Claude-3.5-Sonnet) (Anthropic 2024)	1.34	4.12	31.2	37.5%
SpNav (w/ Qwen-VL-Max) (Bai et al. 2025)	1.48	4.67	29.6	35.8%
SpNav (w/ GPT-4o (Hurst et al. 2024)) (Ours)	**1.10**	**3.41**	**34.8**	**41.1%**

Table 2: Ablation study on different VLMs for reasoning.

VLMs for Goal Pointing	Params	Spatial Navigation			
		NE↓	PL↓	SPL↑	SR↑
Open-source					
SpNav (w/ Qwen2.5-VL) (Bai et al. 2025)	72B	4.23	12.67	12.1	16.7%
SpNav (w/ Qwen2-VL) (Bai et al. 2025)	7B	5.67	15.89	8.9	12.4%
SpNav (w/ Janus-Pro) (Chen et al. 2025)	7B	6.78	18.45	5.2	8.1%
SpNav (w/ LLaVA-NeXT) (Li et al. 2024b)	7B	7.23	20.34	3.8	6.5%
Closed-source					
SpNav (w/ GPT-4o) (Hurst et al. 2024)	-	2.89	8.45	18.7	25.4%
SpNav (w/ Claude-3.5-Sonnet) (Anthropic 2024)	-	3.12	9.23	16.2	22.8%
SpNav (w/ Qwen-VL-Max) (Bai et al. 2025)	-	3.45	10.12	14.5	20.3%
Specific					
SpNav (w/ RoboPoint) (Yuan et al. 2025)	13B	1.45	4.78	28.9	35.2%
SpNav (w/ NaviPoint) (Ours)	**7B**	**1.10**	**3.41**	**34.8**	**41.1%**

Table 3: Ablation study on different VLMs for goal pointing.

Ablation Study

Effect of VLMs for Reasoning We conducted ablation studies to assess the impact of various VLMs on reasoning high-level human commands and extracting goal objects within our SpNav framework. The choice of reasoning VLM significantly influences overall navigation performance. Among open-source models, larger architectures generally yield better results, SpNav w/ Qwen2.5-VL-72B (Bai et al. 2025) achieving the highest success rate (SR) at 28.9%. SpNav w/ GPT-4o-based (Ours) (Hurst et al. 2024) reasoning enhances SR by 42% (41.1% *vs.* 28.9%) and reduces navigation error (NE) by 48% (1.10 *vs.* 2.12)

Methods	Spatial Navigation			
	NE↓	PL↓	SPL↑	SR↑
Random	7.23	10.67	8.4	1.0%
Direct-and-Avoid	4.15	7.34	15.2	15.0%
PointNav	3.87	5.89	18.7	28.0%
Map-to-Action (Ours)	**1.10**	**3.41**	**34.8**	**41.1%**

Table 4: Ablation on point-to-point navigation methods.

compared to the best open-source model, underscoring the importance of high-level reasoning capabilities. The performance gap between open-source and closed-source models is substantial, with GPT-4o's (Hurst et al. 2024) SR being 76% higher than the best 7B open-source model Qwen2.5-VL-7B (Bai et al. 2025) , highlighting the critical role of advanced language understanding and spatial reasoning in processing complex spatial relationship instructions.

Effect of VLMs for Pointing We evaluate various VLMs on egocentric goal pointing to validate our NaviPoint. As shown in Tab 3, NaviPoint significantly outperforms general VLMs and existing pointing methods. Among closed-source models, GPT-4o (Hurst et al. 2024) achieves the best performance with 25.4% SR. However, our method surpasses it with 62% improvement in SR (41.1% *vs.* 25.4%) and 62% reduction in NE (1.10 *vs.* 2.89), highlighting the importance of task-specific training. Compared to the best open-source model Qwen2.5-VL-72B (Bai et al. 2025) (16.7% SR), our approach shows 146% improvement in SR and 74% reduction in NE. Most importantly, our 7B NaviPoint outperforms the state-of-the-art 13B RoboPoint (Yuan et al. 2025), achieving 17% improvement in SR (41.1% *vs.* 35.2%), 24% reduction in NE, and 20% improvement in SPL.

Effect of Map-to-Action

To verify Map-to-Action's importance, we conducted an ablation study comparing spatial map-based navigation with other strategies. As shown in Tab. 4, our method significantly outperforms all baselines. Compared to random navigation, we achieve 4010% improvement in success rate (SR) (41.1% vs. 1.0%) and 85% reduction in navigation error (NE) (1.10 vs. 7.23). Against direct avoidance with basic obstacle avoidance, our method improves SR by 174% (41.1% vs. 15.0%) and reduces NE by 73% (1.10 vs. 4.15). Most notably, compared to learning-based PointNav, our Map-to-Action enhances SR by 47% (41.1% vs. 28.0%), reduces NE by 72% (1.10 vs. 3.87), and improves SPL by 86% (34.8 vs. 18.7). These results demonstrate that our incremental map construction—integrating obstacle tracking, exploration labeling, and traversability analysis—provides crucial environmental context for more efficient navigation than direct point-to-point approaches.

Qualitative Analysis

Fig. 4 demonstrates the effectiveness of our SpNav framework and NaviPoint in various spatial navigation scenarios. In the simulator, our system accurately interprets complex spatial commands and identifies goal locations, as shown in the progression from observation to goal recognition with

Figure 4: Qualitative analysis on our ***SpNav*** framework.

precise pointing coordinates. Importantly, our framework exhibits strong zero-shot transfer capabilities, functioning in real-world settings without additional training. In the real-world scenario depicted, our system processed commands like "I want to sit on the blue chair on the left" and "Wait for me in the open space to the left of the plant." These results confirm that our spatial relationship understanding and goal pointing can effectively transition from simulation to practical deployment in diverse indoor environments.

Conclusion

This paper presents spatial navigation, a novel embodied task requiring intelligent agents to interpret high-level human commands and navigate based on spatial relationships, bridging the gap between current research and practical applications. We introduce SpNav, a layered framework that integrates a visual-language model for command reasoning, a specialized NaviPoint for precise target pointing, and Map-to-Action for efficient navigation. Extensive experiments show that SpNav achieves state-of-the-art performance, significantly surpassing existing methods, and demonstrates effective zero-shot transfer from simulated to real environments, confirming its practical applicability. In future work, we will release datasets, benchmarks, and source code to support research in spatially aware embodied navigation and enhance the development of intelligent agents for natural human-computer interaction.

References

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

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Cai, W.; Ponomarenko, I.; Yuan, J.; Li, X.; Yang, W.; Dong, H.; and Zhao, B. 2024. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*.

Chen, B.; Xu, Z.; Kirmani, S.; Ichter, B.; Sadigh, D.; Guibas, L.; and Xia, F. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14455–14465.

Chen, X.; Wu, Z.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; and Ruan, C. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Cheng, A.-C.; Ji, Y.; Yang, Z.; Gongye, Z.; Zou, X.; Kautz, J.; Bıyık, E.; Yin, H.; Liu, S.; and Wang, X. 2025. Navila: Legged robot vision-language-action model for navigation. In *Robotics: Science and Systems*.

Cheng, A.-C.; Yin, H.; Fu, Y.; Guo, Q.; Yang, R.; Kautz, J.; Wang, X.; and Liu, S. 2024. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37: 135062–135093.

Engheta, N.; Murphy, W. D.; Rokhlin, V.; and Vassiliou, M. S. 1992. The fast multipole method (FMM) for electromagnetic scattering problems. *IEEE Transactions on Antennas and Propagation*, 40(6): 634–641.

Gao, C.; Jin, L.; Peng, X.; Zhang, J.; Deng, Y.; Li, A.; Wang, H.; and Liu, S. 2025. OctoNav: Towards Generalist Embodied Navigation. *arXiv preprint arXiv:2506.09839*.

Gong, Z.; Li, R.; Hu, T.; Qiu, R.; Kong, L.; Zhang, L.; Ding, Y.; Zhang, L.; and Liang, J. 2025. Stairway to Success: Zero-Shot Floor-Aware Object-Goal Navigation via LLM-Driven Coarse-to-Fine Exploration. *arXiv preprint arXiv:2505.23019*.

Hao, P.; Zhang, C.; Li, D.; Cao, X.; Hao, X.; Cui, S.; and Wang, S. 2025. Tla: Tactile-language-action model for contact-rich manipulation. *arXiv preprint arXiv:2503.08548*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kolve, E.; Mottaghi, R.; Han, W.; VanderBilt, E.; Weihs, L.; Herrasti, A.; Deitke, M.; Ehsani, K.; Gordon, D.; Zhu, Y.; et al. 2017. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*.

Li, D.; Jin, Y.; Sun, Y.; Yu, H.; Shi, J.; Hao, X.; Hao, P.; Liu, H.; Sun, F.; Zhang, J.; et al. 2024a. What foundation models can bring for robot learning in manipulation: A survey. *arXiv preprint arXiv:2404.18201*.

Li, F.; Zhang, R.; Zhang, H.; Zhang, Y.; Li, B.; Li, W.; Ma, Z.; and Li, C. 2024b. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Lin, B.; Nie, Y.; Wei, Z.; Chen, J.; Ma, S.; Han, J.; Xu, H.; Chang, X.; and Liang, X. 2025. Navcot: Boosting llm-based vision-and-language navigation via learning disentangled reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, F.; Emerson, G.; and Collier, N. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11: 635–651.

Liu, H.; Guo, D.; and Cangelosi, A. 2025. Embodied intelligence: A synergy of morphology, action, perception and learning. *ACM Computing Surveys*, 57(7): 1–36.

Liu, P.; Zhang, Q.; Peng, D.; Zhang, L.; Qin, Y.; Zhou, H.; Ma, J.; Xu, R.; and Ji, Y. 2025a. Toponav: Topological graphs as a key enabler for advanced object navigation. *arXiv preprint arXiv:2509.01364*.

Liu, Y.; Cao, X.; Chen, T.; Jiang, Y.; You, J.; Wu, M.; Wang, X.; Feng, M.; Jin, Y.; and Chen, J. 2025b. From screens to scenes: A survey of embodied AI in healthcare. *Information Fusion*, 119: 103033.

Liu, Y.; Chen, W.; Bai, Y.; Liang, X.; Li, G.; Gao, W.; and Lin, L. 2025c. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *IEEE/ASME Transactions on Mechatronics*.

Long, Y.; Cai, W.; Wang, H.; Zhan, G.; and Dong, H. 2025. InstructNav: Zero-shot System for Generic Instruction Navigation in Unexplored Environment. In *Conference on Robot Learning*, 2049–2060.

Long, Y.; Li, X.; Cai, W.; and Dong, H. 2024. Discuss before moving: Visual language navigation via multi-expert discussions. In *IEEE International Conference on Robotics and Automation*, 17380–17387. IEEE.

Ma, Y.; Song, Z.; Zhuang, Y.; Hao, J.; and King, I. 2024. A survey on vision-language-action models for embodied ai. *arXiv preprint arXiv:2405.14093*.

Ramrakhya, R.; Batra, D.; Wijmans, E.; and Das, A. 2023. Pirlnav: Pretraining with imitation and rl finetuning for objectnav. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17896–17906.

Tang, Y.; Zhang, L.; Zhang, S.; Zhao, Y.; and Hao, X. 2025a. RoboAfford: A Dataset and Benchmark for Enhancing Object and Spatial Affordance Learning in Robot Manipulation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 12706–12713.

Tang, Y.; Zhang, S.; Hao, X.; Wang, P.; Wu, J.; Wang, Z.; and Zhang, S. 2025b. Affordgrasp: In-context affordance reasoning for open-vocabulary task-oriented grasping in clutter. *arXiv preprint arXiv:2503.00778*.

Tang, Y.; Zhao, C.; Wang, J.; Zhang, C.; Sun, Q.; Zheng, W. X.; Du, W.; Qian, F.; and Kurths, J. 2022. Perception and navigation in autonomous systems in the era of learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12): 9604–9624.

Team, B. R.; et al. 2025. Robobrain 2.0 technical report. *arXiv preprint arXiv:2507.02029*.

Wang, Y.; Chen, W.; Han, X.; Lin, X.; Zhao, H.; Liu, Y.; Zhai, B.; Yuan, J.; You, Q.; and Yang, H. 2024. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*.

Wasserman, J.; Chowdhary, G.; Gupta, A.; and Jain, U. 2024. Exploitation-guided exploration for semantic embodied navigation. In *IEEE International Conference on Robotics and Automation*, 2901–2908. IEEE.

Wu, P.; Mu, Y.; Wu, B.; Hou, Y.; Ma, J.; Zhang, S.; and Liu, C. 2024a. VoroNav: voronoi-based zero-shot object navigation with large language model. In *Proceedings of the 41st International Conference on Machine Learning*, 53737–53775.

Wu, Y.; Lyu, H.; Tang, Y.; Zhang, L.; Zhang, Z.; Zhou, W.; and Hao, S. 2025. Evaluating GPT-4o's Embodied Intelligence: A Comprehensive Empirical Study. *Authorea Preprints*.

Wu, Y.; Zhang, P.; Gu, M.; Zheng, J.; and Bai, X. 2024b. Embodied navigation with multi-modal information: A survey from tasks to methodology. *Information Fusion*, 112: 102532.

Xiao, E.; Zhang, L.; Tang, Y.; Cheng, H.; Xu, R.; Ding, W.; Zhou, L.; Chen, L.; Ye, H.; and Hao, X. 2025. Team Xiaomi EV-AD VLA: Learning to Navigate Socially Through Proactive Risk Perception-Technical Report for IROS 2025 RoboSense Challenge Social Navigation Track. *arXiv preprint arXiv:2510.07871*.

Xu, Z.; Wu, K.; Wen, J.; Li, J.; Liu, N.; Che, Z.; and Tang, J. 2024. A survey on robotics with foundation models: toward embodied ai. *arXiv preprint arXiv:2402.02385*.

Yuan, W.; Duan, J.; Blukis, V.; Pumacay, W.; Krishna, R.; Murali, A.; Mousavian, A.; and Fox, D. 2025. RoboPoint: A Vision-Language Model for Spatial Affordance Prediction in Robotics. In *Conference on Robot Learning*, 4005–4020.

Zha, J.; Fan, Y.; Yang, X.; Gao, C.; and Chen, X. 2025. How to enable llm with 3d capacity? a survey of spatial reasoning in llm. *arXiv preprint arXiv:2504.05786*.

Zhang, C.; Hao, P.; Cao, X.; Hao, X.; Cui, S.; and Wang, S. 2025a. Vtla: Vision-tactile-language-action model with preference learning for insertion manipulation. *arXiv preprint arXiv:2505.09577*.

Zhang, J.; Wang, K.; Wang, S.; Li, M.; Liu, H.; Wei, S.; Wang, Z.; Zhang, Z.; and Wang, H. 2024a. Uni-NaVid: A Video-based Vision-Language-Action Model for Unifying Embodied Navigation Tasks. In *Robotics: Science and Systems*.

Zhang, J.; Wang, K.; Xu, R.; Zhou, G.; Hong, Y.; Fang, X.; Wu, Q.; Zhang, Z.; and Wang, H. 2024b. Navid: Video-based vlm plans the next step for vision-and-language navigation. In *Robotics: Science and Systems*.

Zhang, L.; Hao, X.; Tang, Y.; Fu, H.; Zheng, X.; Wang, P.; Wang, Z.; Ding, W.; and Zhang, S. 2025b. $NavA^3$: Understanding Any Instruction, Navigating Anywhere, Finding Anything. *arXiv preprint arXiv:2508.04598*.

Zhang, L.; Hao, X.; Xu, Q.; Zhang, Q.; Zhang, X.; Wang, P.; Zhang, J.; Wang, Z.; Zhang, S.; and Xu, R. 2025c. 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*.

Zhang, L.; Wang, H.; Xiao, E.; Zhang, X.; Zhang, Q.; Jiang, Z.; and Xu, R. 2025d. Multi-floor zero-shot object navigation policy. In *IEEE International Conference on Robotics and Automation*. IEEE.

Zhang, L.; Xiao, E.; Zhang, Y.; Fu, H.; Hu, R.; Ma, Y.; Ding, W.; Chen, L.; Ye, H.; and Hao, X. 2025e. Team Xiaomi EV-AD VLA: Caption-Guided Retrieval System for Cross-Modal Drone Navigation-Technical Report for IROS 2025 RoboSense Challenge Track 4. *arXiv preprint arXiv:2510.02728*.

Zhang, L.; Zhang, Q.; Wang, H.; Xiao, E.; Jiang, Z.; Chen, H.; and Xu, R. 2024c. Trihelper: Zero-shot object navigation with dynamic assistance. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 10035–10042. IEEE.

Zhang, Y.; Ma, Z.; Li, J.; Qiao, Y.; Wang, Z.; Chai, J.; Wu, Q.; Bansal, M.; and Kordjamshidi, P. 2024d. Vision-and-Language Navigation Today and Tomorrow: A Survey in the Era of Foundation Models. *Transactions on Machine Learning Research*.

Zheng, D.; Huang, S.; Zhao, L.; Zhong, Y.; and Wang, L. 2024. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13624–13634.

Zhou, E.; An, J.; Chi, C.; Han, Y.; Rong, S.; Zhang, C.; Wang, P.; Wang, Z.; Huang, T.; Sheng, L.; et al. 2025. RoboRefer: Towards Spatial Referring with Reasoning in Vision-Language Models for Robotics. *arXiv preprint arXiv:2506.04308*.

Zhou, G.; Hong, Y.; and Wu, Q. 2024. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 7, 7641–7649.

Zhou, X.; Liu, M.; Yurtsever, E.; Zagar, B. L.; Zimmer, W.; Cao, H.; and Knoll, A. C. 2024. Vision language models in autonomous driving: A survey and outlook. *IEEE Transactions on Intelligent Vehicles*.