STEP: Spatial Thinking and Egocentric Pointing for Embodied Instruction Following

Hanxuan Li[1,2], Bin Fu[1,2], Zeyuan Lin[1,2], Ruiping Wang[1,2(✉)], and Xilin Chen[1,2]

[1] Key Laboratory of AI Safety of CAS, Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
{hanxuan.li,bin.fu,zeyuan.lin}@vipl.ict.ac.cn, {wangruiping,xlchen}@ict.ac.cn

Abstract. Embodied Instruction Following requires agents to execute natural-language tasks via holistic perception and sound planning. Current LLM-driven agents suffer from two primary bottlenecks: Spatial Myopia, where limited egocentric views hinder global topological understanding, and Granularity Imbalance, where planning is polarized between inefficient atomic actions that lack interpretability and ambiguous subgoals that fail to align with low-level control. To address these issues, we propose STEP, a framework that empowers agents with Spatial Thinking and Egocentric Pointing. For perception, STEP integrates Hybrid Map-Egocentric Perception, fusing multi-view streams into Bird's-Eye-View maps to maintain both global context and local visual cues. For planning, STEP introduces Point-Level Planning, which derives spatial waypoints through explicit reasoning chains, achieving highly interpretable, intermediate-granularity execution. We also design STEP-CoT, a scalable data engine, to generate a 400k Chain-of-Thought dataset that aligns diverse modalities—views, maps, actions, and instructions—through reasoning chains. Results on ALFRED and AI2-THOR-Nav, along with real-world deployments, demonstrate STEP's superior performance and generalization. Ultimately, STEP achieves a unified paradigm: **Look at the View, Think with the Map, and Point to the Goal.**

Keywords: Embodied Agent · Instruction Following · VLM-based Planning

1 Introduction

Embodied Instruction Following focuses on enabling agents to complete diverse embodied tasks by precisely following natural language instructions. This objective necessitates a holistic perception of the environment and sound planning grounded in observations. Given the remarkable zero-shot reasoning of Large-Language Models (LLMs), recent work [3, 5, 11, 24] has shifted toward LLM-driven embodied decision-making. As illustrated in Fig. 1, categorized by the perceptual scope and planning granularity, LLMs primarily serve in two roles:

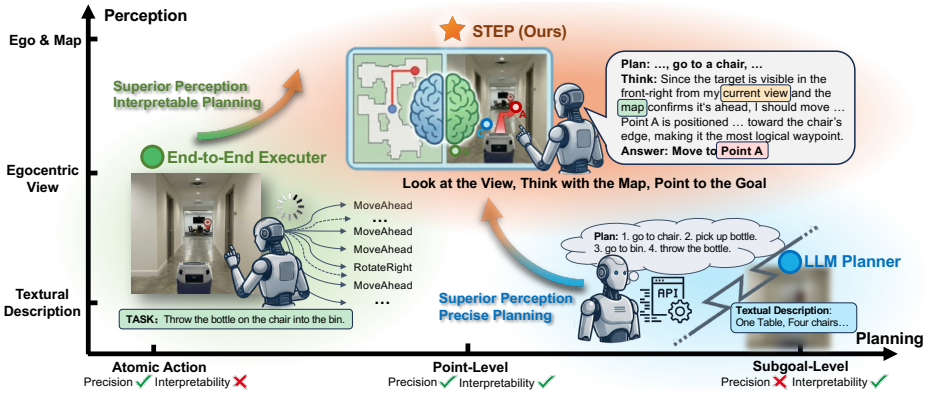

Fig. 1: Positioning of STEP across perception and planning dimensions. The x-axis represents planning granularity: moving from atomic actions to subgoal-level, interpretability increases while precision decreases. The y-axis denotes perception scope, which scales from textual descriptions to a hybrid map-egocentric view. STEP achieves a superior perception and appropriate planning level.

a) High-level Planners [7, 22, 24], where LLMs decompose instructions via scene abstractions, while execution relies on predefined APIs; and b) End-to-end Executors [33, 39], which map egocentric visual streams directly to atomic actions.

Despite their success in specific scenarios, these paradigms suffer from either shared or specific limitations. Along the dimensions of perception and planning, as shown in Fig. 1, we summarize these limitations as two primary bottlenecks: **1) Spatial Myopia in Perception:** Comprehensive perception is fundamental to robust decision-making, yet current agents remain information-constrained [34]. While end-to-end executors typically rely exclusively on egocentric views, high-level planners frequently lack direct visual perception. Such fragmented perception hinders spatial memory, preventing unified topological understanding. **2) Granularity Imbalance in Planning:** Planning granularity remains polarized between two extremes. On one hand, action-level outputs (e.g., move-ahead) suffer from low planning efficiency and poor interpretability due to their excessively fine scale. On the other hand, subgoal-level instructions (e.g., go to the kitchen) are too ambiguous, leaving a critical gap [3] between high-level reasoning and low-level control. Diverging from existing paradigms, Tolman's theory of "Cognitive Maps" [26] shows that human efficiency stems from a global mental layout rather than isolated local cues. Furthermore, human planning operates through semantic-spatial anchoring: we simulate "where to go" by anchoring semantic targets to the physical environment.

Inspired by these insights, we introduce **STEP**, an interpretable framework that empowers **S**patial **T**hinking for comprehensive perception and **E**gocentric **P**ointing for sound planning. STEP addresses the aforementioned two bottlenecks through two core mechanisms:

- **Hybrid Perception**: With a Map Encoder, STEP jointly reasons over egocentric views and BEV maps, broadening the perception scope.
- **Point-Level Planning**: STEP predicts waypoints on egocentric images via Chain-of-Thought (CoT) reasoning, balancing physical precision with semantic interpretability.

To provide robust data support, we developed the scalable STEP-CoT engine to automatically align instructions, images, maps, and actions via Chain-of-Thought (CoT). This pipeline yields a 400k+ dataset, which, integrated with our proposed modules, empowers the agent to: Look at the View, Think with the Map, and Point to the Goal.

To enable hybrid perception, STEP simplifies the scene via Environment Filtering, generating MLLM-interpretable Bird's-Eye-View (BEV) maps. Both the BEV map and the egocentric view serve as joint perceptual inputs. To process these modalities, STEP uses a tri-encoder architecture: a Map Encoder for topological maps, a Visual Encoder for egocentric views, and a Text Encoder for instructions. By projecting these features into a unified space, they are fused into a joint representation for planning. This integration grants the agent global spatial awareness beyond its immediate sight.

To balance interpretability with planning precision, we adopt a Point-Level Planning paradigm. While MLLMs effectively ground salient objects, floor-based localization remains challenging due to the visual homogeneity of traversable surfaces. To address this, STEP incorporates Egocentric Traversability Grounding, leveraging Grounded-SAM [19] to detect reachable areas and a Set-of-Mark [29] mechanism for visual prompting. By labeling each traversable region, we facilitate spatial grounding, enabling the model to generate a precise waypoint alongside an explicit reasoning trace via these visual markers. This intermediate granularity unites rigorous precision with semantic interpretability, transcending the limitations of both atomic actions and vague subgoals.

The STEP framework establishes an end-to-end paradigm for reasoning across instructions, egocentric observations, and BEV maps. However, substantial domain gaps—such as mapping egocentric objects to BEV locations—often hinder effective joint reasoning. To bridge these gaps, we developed STEP-CoT, a scalable data engine that distills expert trajectories into structured Chain-of-Thought (CoT) sequences. By design, CoT explicitly interlinks instructions, observations, maps, and actions, providing a vital supervisory signal for cross-modal alignment. We ensure high fidelity with a dual-stage verification mechanism, combining logical consistency and VLM scoring. Furthermore, STEP-CoT extracts primitive skills data, such as topological reasoning and pose estimation, to boost spatial awareness. This scalable pipeline yields 400k+ multimodal samples for reliable embodied decision-making.

STEP demonstrates exceptional performance on ALFRED and exhibits strong cross-task generalization in AI2-THOR Navigation. We further validate its practical utility through extensive real-world deployments. Our results demonstrate that STEP enables joint reasoning across egocentric views and BEV maps, yielding interpretable and precise point-level planning.

2 Related Works

2.1 Embodied Task Planning with LLMs

Embodied Task Planning requires agents to translate natural-language directives into actionable decisions [10]. Conventional RL-based or supervised methods [18, 25] often struggle to generalize to unseen environments [8, 15]. Benefiting from the zero-shot reasoning of LLMs, recent approaches have increasingly adopted LLMs as the core engine for embodied planning. Existing paradigms generally diverge into two granularities. One line of research employs LLMs as high-level planners [7,13,22,24,38], where the model decomposes complex instructions into abstract subgoals or PDDL descriptions [14], leaving the execution to pre-defined low-level controllers. For instance, SayCan [7] leverages LLMs to propose task steps while grounding them with a vision-based affordance function. Despite their logical clarity, these methods often face an execution gap [3] because the subgoals are too vague for precise control. Conversely, end-to-end executors [33, 39] attempt to map observations directly to atomic actions. RT-2 [39] treats robotic actions as text tokens in a unified vision-language-action transformer, while Navid [33] generates continuous control signals directly from video streams. While precise, such fine granularity lacks semantic interpretability, resulting in "black-box" decision-making. Unlike these polarized paradigms, by predicting point-level targets through explicit reasoning, STEP establishes an intermediate granularity that connects high-level intent with precise control.

2.2 Embodied Spatial Perception

Comprehensive perception is a cornerstone of Embodied AI. However, many current agents [18, 39] suffer from Spatial Myopia by relying mostly on limited egocentric views, which hinders the formation of long-term memory and global understanding. To mitigate this, structured representations such as topological maps [20, 28, 32] and semantic maps [6, 31, 35] have been introduced. Specifically, methods like MC-GPT [32] and VoroNav [28] utilize topological maps to capture viewpoints and spatial relationships, while InstructNav [13] and VLFM [30] utilize value maps to select waypoints. Yet these representations are often insufficient for Multimodal LLMs (MLLMs) to interpret. To address this, recent advancements [34,37] have explored Bird's-Eye-View (BEV) projections with semantic annotations for better MLLM understanding. However, effectively aligning egocentric perspectives with global layouts remains difficult, as MLLMs frequently struggle with spatial geometry and orientation. STEP addresses this by augmenting the MLLM with a dedicated Map Encoder to process BEV features and introducing a specialized data engine to align egocentric views and maps. By reinforcing cross-modal correspondence, our approach enables the model to jointly consider both the map and egocentric views for reasoning.

3 Method

The STEP framework is designed to overcome Spatial Myopia and Granularity Imbalance in Embodied Instruction Following (EIF). We achieve this by unifying global spatial context and local egocentric observations into a point-level planning paradigm, as shown in Fig. 2. Our methodology consists of three main components: **Hybrid Map-Egocentric Perception** (Sec. 3.1), which constructs an annotated Bird's-Eye-View (BEV) map and grounds traversable regions; **Cognitive Point-Level Planning** (Sec. 3.2), a multi-modal architecture that reasons over maps and views to predict Chain-of-Thought traces and waypoints; and a **Scalable CoT Data Engine** (Sec. 3.3), which addresses the scarcity of spatial reasoning traces through automated trajectory annotation and primitive task synthesis.

3.1 Hybrid Map-Egocentric Perception

To mitigate spatial myopia, STEP enhances perception by enabling the agent to process global map information and local egocentric streams concurrently. However, two fundamental challenges persist: (1) interpreting the semantics of maps, and (2) grounding spatial targets from the map onto their corresponding pixel-level coordinates in the egocentric view. Such cross-modal alignment remains a formidable barrier even for state-of-the-art MLLMs. STEP introduces **Environment Filtering** and **Egocentric Traversability Grounding**, transforming raw, heterogeneous observations into structured, MLLM-digestible formats.

Egocentric Traversability Grounding. For precise spatial grounding, STEP aligns instructions with the physical environment. While MLLMs effectively ground salient objects, localizing targets on the ground is challenging due to the homogeneous appearance of floor regions. To mitigate this, STEP incorporates an Egocentric Traversability Grounding process that applies visual prompting to navigable areas, thereby reducing the complexity of spatial reasoning. We employ a grounded segmentation pipeline on the egocentric observation I_t. Given text prompts $\mathcal{T}_{ground} = \{\text{ground, carpet}\}$, we use an open-set detector $\Psi(\cdot)$ and a Segment Anything Model SAM$(\cdot)$ [9] to generate a binary semantic mask $\mathcal{M}_{sem} \in \{0,1\}^{H\times W}$, where H and W denote the image height and width:

$$\mathcal{M}_{sem} = \text{SAM}(I_t, \Psi(I_t, \mathcal{T}_{ground})). \tag{1}$$

To discretize the action space, we project a 3D ground grid $\mathcal{P}_w = \{(n \cdot s, m \cdot s, 0, 1)^T \mid n \in \mathbb{Z}^+, m \in \mathbb{Z}\}$ onto the image plane, where s denotes the step size, while n and m represent the longitudinal depth and lateral offset, respectively. Each point $\mathbf{P}_i \in \mathcal{P}_w$ is mapped to pixel coordinates $\mathbf{u}_i$ via the camera intrinsics $\mathbf{K}$ and extrinsics $\mathbf{T}_c$, and the final candidate waypoints $\mathcal{V}_{pts}$ are obtained by filtering these projections through the semantic mask $\mathcal{M}_{sem}$:

$$\lambda_i \mathbf{u}_i = \mathbf{K}\mathbf{T}_c\mathbf{P}_i, \quad \mathcal{V}_{pts} = \{\mathbf{u}_i \mid \mathcal{M}_{sem}(\mathbf{u}_i) = 1, \forall \mathbf{P}_i \in \mathcal{P}_w\}, \tag{2}$$

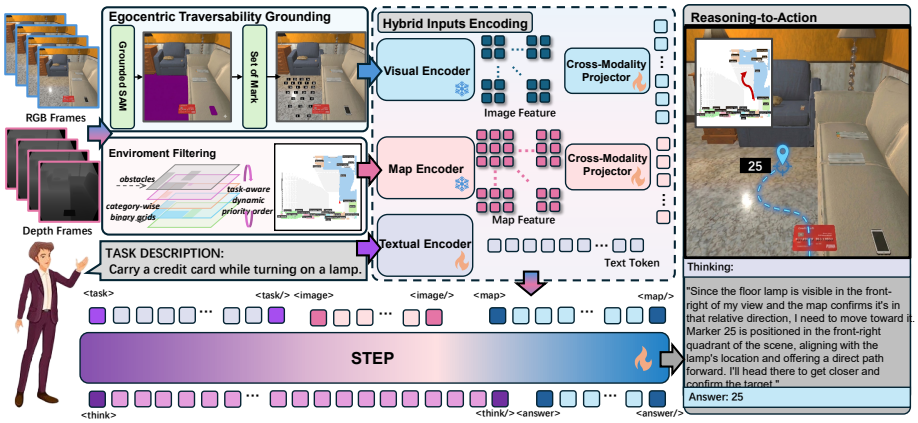

Fig. 2: Framework Overview. STEP integrates Egocentric Traversability Grounding and Environment Filtering to refine inputs. Map, ego-view, and instruction are encoded and projected into a unified latent space, enabling Reasoning-to-Action Prediction.

where λ_i is the depth scale factor. To treat these regions as discrete interactive entities for the MLLM, we overlay each $\mathbf{u}_i \in \mathcal{V}_{pts}$ with a unique numerical ID using a Set-of-Mark (SoM) strategy. By assigning these discrete identifiers to the ground, we enable the MLLM to leverage its inherent object-grounding capabilities to select a specific waypoint (e.g., "Marker 5"). This reduces reasoning complexity and ensures physical compliance.

Environment Filtering. To mitigate spatial myopia, STEP employs an Environment Filtering pipeline that distills raw observations into an annotated semantic map. At each timestep t, RGB-D inputs $\{I_t, D_t\}$ are first processed to construct a persistent semantic map. We perform pixel-wise segmentation on I_t and back-project each semantic pixel into 3D space using the depth map D_t and camera parameters $\{\mathbf{K}, \mathbf{T}_c\}$. The resulting point cloud is voxelized and orthographically projected along the Z-axis (height dimension) to generate a 2D multi-channel semantic map $M_t \in \{0,1\}^{(C+2)\times H\times W}$. Here, C represents object categories, with two additional channels for obstacles and exploration status.

To facilitate MLLM reasoning, we transform the high-dimensional map into an Annotated Semantic Map (ASM) $\mathcal{G}_t$ through a streamlined abstraction pipeline. First, a dilation kernel K_{dila} expands obstacle boundaries by δ to ensure safety. To resolve "many-to-one" occlusions (e.g., small objects on surfaces), a goal-aware aggregate function $A(\cdot)$ filters the map based on instruction $\mathcal{I}$, prioritizing task-relevant and fine-grained entities. Finally, a rendering function $\mathcal{R}(\cdot)$ generates the ASM via force-directed label placement:

$$\mathcal{G}_t = \mathcal{R}\left(A(K_{dila}(M_t), \mathcal{I})\right). \tag{3}$$

By simplifying the complex 3D environment into a linguistically indexed 2D overview, this representation significantly reduces the MLLM's cognitive load.

3.2 Cognitive Point-Level Planning

To mitigate granularity imbalances in existing embodied agents, we introduce a CoT-Finetuning paradigm that enables MLLMs to perform explicit CoT planning across hybrid ego-map inputs to predict egocentric waypoints. STEP integrates three primary modalities: the task instruction $\mathcal{T}$, the Semantic Contextual Map $\mathcal{G}_t$, and the egocentric observation I_t enriched with navigational markers. We present this paradigm through two key perspectives: **Hybrid Input Encoding**, which details the cross-modal projection of these multi-modal signals into a unified latent space, and **Reasoning-to-Action Prediction**, which defines a structured output format for deriving deterministic waypoints through explicit reasoning chains.

Hybrid Input Encoding. Built upon the Qwen2.5-VL framework, STEP formalizes the transformation of task instructions $\mathcal{T}$, Semantic Contextual Maps $\mathcal{G}_t$, and egocentric observations I_t into a synchronized latent representation. For the instruction $\mathcal{T}$ containing L tokens, a text encoder f_θ produces hidden states $\mathbf{X}_{inst} \in \mathbb{R}^{L\times d}$, where d is the model's hidden dimension. Simultaneously, the egocentric view $I_t \in \mathbb{R}^{3\times H\times W}$ and the semantic map $\mathcal{G}_t$ are processed by specialized encoders f_ϕ and f_ψ to yield visual and spatial features:

$$\mathbf{X}_{vis} = f_\phi(I_t) \in \mathbb{R}^{N\times d_v}, \quad \mathbf{X}_{map} = f_\psi(\mathcal{G}_t) \in \mathbb{R}^{K\times d_m}, \tag{4}$$

where d_v and d_m represent the raw visual and map feature dimensions, respectively, and K denotes the number of map tokens. To bridge the domain gap, we employ modality-specific projection layers g_v and g_m to align these features into the d-dimensional latent space:

$$\mathbf{Z}_{vis} = g_v(\mathbf{X}_{vis}) \in \mathbb{R}^{N\times d}, \quad \mathbf{Z}_{map} = g_m(\mathbf{X}_{map}) \in \mathbb{R}^{K\times d}. \tag{5}$$

The final input sequence $\mathbf{H}_t$ is constructed by interleaving these aligned features with learned modality-specific boundary tokens s_{inst}, s_{map}, and s_{view}. The unified representation is defined as the concatenation of these components:

$$\mathbf{H}_t = [\, s_{inst}\,;\, \mathbf{X}_{inst}\,;\, s_{map}\,;\, \mathbf{Z}_{map}\,;\, s_{view}\,;\, \mathbf{Z}_{vis}\,], \tag{6}$$

where the semicolon ; denotes the concatenation operation along the sequence length dimension. The resulting composite sequence $\mathbf{H}_t$ is then fed into the multimodal LLM backbone. This formulation enables the model to perform joint cross-modal reasoning.

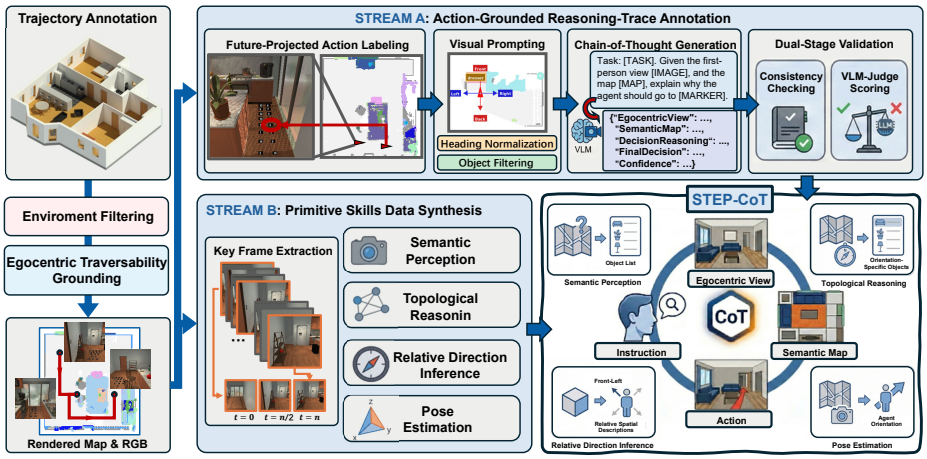

Fig. 3: Overview of STEP-CoT. STREAM A performs Action-Grounded Reasoning-Trace Annotation; concurrently, STREAM B facilitates Primitive Skills Data Synthesis.

Reasoning-to-Action Prediction. To address the granularity imbalance and operational inefficiency inherent in conventional agents, STEP adopts a Reasoning-to-Action paradigm. Traditional frameworks that predict atomic actions (e.g., *move_forward*) suffer from low information density; such a fragmented action space often requires excessive decision cycles to reach a semantic goal and lacks an explicit correspondence to a verifiable thinking process. In contrast, STEP mandates the generation of a Chain-of-Thought (CoT) reasoning trace C_t prior to selecting a point-level waypoint. Given the unified multimodal input $\mathbf{H}_t$, the joint generation of the reasoning chain $C_t = \{c_1, \ldots, c_{|C_t|}\}$ and the final decision y_t is formalized as:

$$P(C_t, y_t \mid \mathbf{H}_t) = P(y_t \mid C_t, \mathbf{H}_t) \prod_{i=1}^{|C_t|} P(c_i \mid c_{<i}, \mathbf{H}_t). \tag{7}$$

The action space $\mathcal{Y}$ encompasses discrete markers $v_i \in \mathcal{V}_{pts}$ and commands $\mathcal{A}_{base} = \{\text{forward, left, right}\}$. Since these navigable markers reside within the agent's immediate egocentric field of view, reaching them constitutes a highly reliable local task. The final executable action sequence $\mathbf{a}_{seq}$ is derived as:

$$\mathcal{Y} = \mathcal{V}_{pts} \cup \mathcal{A}_{base}, \quad \mathbf{a}_{seq} = \begin{cases} \Phi_{\text{FMM}}(\mathbf{p}_i, M_t) & \text{if } y_t = v_i \\ \{y_t\} & \text{if } y_t \in \mathcal{A}_{base} \end{cases}, \tag{8}$$

where Φ_{FMM} utilizes the Fast Marching Method over the local occupancy matrix M_t to compute optimal trajectories for point-level targets $\mathbf{p}_i$.

3.3 Scalable CoT Data Engine

Point-level planning faces significant cross-modal disparities between instructions $\mathcal{T}$, maps $\mathcal{G}_t$, observations I_t, and actions y_t. To bridge these gaps, we utilize a CoT framework to interlink these modalities into a cohesive reasoning sequence, as in Fig. 3. Given the scarcity of CoT annotations, we introduce STEP-CoT, an automated engine that distills high-fidelity reasoning traces from expert trajectories via **Action-Grounded Reasoning-Trace Annotation** (STREAM A). To further bolster map comprehension and ego-map alignment, the engine concurrently performs **Primitive Skills Data Synthesis** (STREAM B). STEP-CoT generates 400k+ multi-modal samples, providing essential supervision to strengthen the model's reasoning and planning capabilities.

Action-Grounded Reasoning-Trace Annotation. Formally, given a task instruction $\mathcal{T}$ and expert trajectory $\tau = \{(s_t, a_t, \mathbf{T}_t)\}_{t=0}^{T}$, where s_t is the state and $\mathbf{T}_t$ is the global pose, we dynamically reconstruct Semantic Contextual Maps $\{\mathcal{G}_t\}_{t=1}^{T}$. To define the target waypoint v^*, we employ a look-ahead window Δ and select the most distant future pose index δ^*:

$$\delta^* = \max\{\delta \leq \Delta \mid \text{Proj}(s_{t+\delta}, \mathbf{T}_t) \in \mathcal{V}_{pts}\}. \tag{9}$$

The target waypoint is then assigned as $v^* = \text{Proj}(s_{t+\delta^*}, \mathbf{T}_t)$. To facilitate CoT Generation, we address the challenge of concurrent egocentric-global map interpretation. MLLMs often confuse egocentric and world reference frames, particularly regarding directional descriptors. To mitigate this, we introduce two strategies: Heading Normalization, which aligns the agent's coordinate system with the image plane to eliminate directional ambiguity, and Object Filtering, which isolates task-critical entities to suppress semantic hallucinations. Then, a powerful teacher MLLM f_{teacher} (Qwen3-VL-235B-A22B-Instruct [1]) synthesizes reasoning chains C_t based on instruction $\mathcal{T}$, prompted observation I_t, semantic map $\mathcal{G}_t$, and oracle label v^*, yielding the data quintuple $\mathcal{D}_t = (\mathcal{T}, I_t, \mathcal{G}_t, C_t, v^*)$. To ensure fidelity, we implement a Dual-Stage Validation pipeline. A sample is retained in the final dataset $\mathcal{D}_{\text{STEP-CoT}}$ only if it passes both the Consistency Checking module Φ_{cons} and the VLM-Judge verification Φ_{eval}:

$$\mathcal{D}_{\text{STEP-CoT}} = \{\mathcal{D}_t \mid \Phi_{cons}(\mathcal{D}_t) = 1 \wedge \Phi_{eval}(\mathcal{D}_t) = 1\}, \tag{10}$$

where Φ_{cons} verifies that the parsed reasoning matches the oracle label v^*, and Φ_{eval} utilizes f_{teacher} to perform a final binary audit on the data quality.

Primitive Skills Data Synthesis. To bolster spatial awareness, we decouple the essential capabilities required by STEP into primitive skills. This decoupling addresses two core challenges: Internal Spatial Reasoning on abstract maps and Cross-modal Alignment between global and egocentric perspectives. To bridge these gaps, we automatically extract keyframes and map states to synthesize task-target pairs $(\mathcal{Q}, \mathcal{A})$ as part of our automated data pipeline.

Table 1: Overview of the STEP-CoT Dataset. $\mathcal{T}$, I, and $\mathcal{G}$ denote the text, egocentric observation, and map, respectively.

Task Group	Description	Size	Modality
STREAM A: CoT Reasoning			
Action-Grounded Trace	Explicit CoT reasoning sequence C_t for point-level planning	229,544	$\mathcal{T}, I, \mathcal{G}$
STREAM B: Primitive Skills			
Pose Estimation	Estimate agent heading $\theta_t \in [0, 360^\circ)$ relative to map $\mathcal{G}$	45,623	$I, \mathcal{G}$
Relative Direction	Predict relative bearing of target object o_j (e.g., "back-left")	18,172	$\mathcal{G}, \mathcal{T}$
Semantic Perception	Catalog environmental entities $\mathcal{O}_t = \{o_1, \ldots, o_n\}$ within $\mathcal{G}$	9,069	$\mathcal{G}$
Topological Reasoning	Infer spatial layouts and landmarks at orientation θ	72,552	$\mathcal{G}, \mathcal{T}$
OTHERS			
Task Breakdown	Decompose instruction $\mathcal{T}$ into subgoal sequences	45,321	$\mathcal{T}$
Total		**420,281**	

We incorporate Semantic Perception, Topological Reasoning, and Relative Direction Inference for Internal Spatial Reasoning, alongside Pose Estimation for Cross-modal Alignment. These tasks are automatically synthesized from trajectory keyframes within our data pipeline. Detailed task specifications and data scales are summarized in Tab. 1. Additionally, we incorporate Task Breakdown data to strengthen the model's decomposition ability. For complex instructions, STEP first decomposes instructions into subgoals, and then performs point-level planning based on a unified model. By integrating these primitive tasks, we strengthen the model's capabilities in both Internal Spatial Reasoning and Cross-modal Alignment, establishing a robust foundation for STEP.

4 Experiments

4.1 Experiment Setup

Benchmark. We conduct a comprehensive evaluation on the ALFRED [21] benchmark, a rigorous testbed for Embodied Instruction Following. This benchmark encompasses seven distinct types of long-horizon tasks that demand a seamless integration of navigation and interaction. We report results on the standard ALFRED validation splits, including Val Seen and Val Unseen. The experimental protocol strictly complies with the standard settings, incorporating low-level action execution, maximum step constraints, and predefined failure thresholds.

Implementation Details. STEP is developed based on the Qwen2.5-VL-7B-Instruct [2] foundation model, where the original visual encoder is utilized as the Map Encoder to leverage its robust multimodal perception. To preserve generalization and prevent overfitting, the Map and Visual Encoders remain frozen during training. Conversely, we perform full-parameter fine-tuning on the Large Language Model (LLM), the Map Projector, and the Visual Projector. We train on a random 80% split of the STEP-CoT dataset (345,288 samples), derived from the ALFRED [21] training set. For fair comparison, all STEP ablation variants are trained on the same data split with identical optimization hyperparameters

Table 2: Main results on ALFRED. SR and GC are short for success rate and goal-conditioned success rate. Δ SR and Δ GC represent the performance gap between the seen and the unseen environments.

Method	Training Mode	Val Seen		Val Unseen		Δ SR ↓	Δ GC ↓
		SR ↑	GC ↑	SR ↑	GC ↑		
Specialists, only for ALFRED tasks							
E.T. [18]	from scratch	46.59	52.92	7.32	20.87	39.27	32.05
HiTUT [36]	from scratch	25.24	34.85	12.44	23.71	12.80	11.14
M-TRACK [23]	from scratch	26.70	33.21	17.29	28.98	9.41	4.23
FILM [15]	from scratch	24.63	37.20	20.10	32.45	4.53	4.75
LEBP [12]	from scratch	27.63	35.76	22.36	29.58	5.27	6.18
Generalists, based on LLMs							
SayCan [7]	few-shot	12.30	24.52	9.88	22.54	2.42	1.98
LLM-P (GPT) [24]	few-shot	16.45	30.11	15.36	29.88	1.09	0.23
R2C-GPT-4 [3]	zero-shot	20.00	28.46	24.00	28.24	4.00	0.22
R2C-Llama-7B [3]	fine-tune	20.83	29.60	18.99	29.69	1.84	0.09
R2C-Mistral-7B [3]	fine-tune	22.31	32.40	22.35	31.97	0.04	0.43
Qwen2.5-VL-7B-Instruct [2]	zero-shot	9.96	18.59	7.87	18.26	2.09	0.33
Qwen3-VL-32B-Instruct [1]	zero-shot	15.20	26.69	11.76	24.54	3.44	2.15
GPT-4o [16]	zero-shot	16.00	27.13	16.00	25.00	0.00	2.13
GPT-5.4 [17]	zero-shot	18.00	28.68	16.00	25.38	2.00	3.30
STEP-7B (ours)	fine-tune	**30.28**	**39.69**	**26.77**	**39.36**	3.51	0.33

and hardware configurations. All models are trained for 2 epochs with a global batch size of 128. The fine-tuning experiments are conducted on a cluster of 8 NVIDIA H800 GPUs, while all evaluations are performed on 4 NVIDIA A40 GPUs. The code is available on our project page[3].

4.2 Main Results

Tab. 2 presents the evaluation results on ALFRED. We compare STEP-7B against two primary categories: traditional robotic learning methods (Specialists) and LLM-based approaches (Generalists). STEP-7B achieves state-of-the-art performance among all Generalist methods, significantly outperforming existing LLM-based solutions in both seen and unseen environments.

Compared to Generalist methods like SayCan [7], LLM-P [24], and R2C [3], which primarily utilize LLMs for high-level planning or text-based reasoning, STEP integrates Hybrid Map-Egocentric Perception with Cognitive Point-Level Planning to enable a more fine-grained understanding of the physical environment. Compared to the baseline model, Qwen2.5-VL-7B-Instruct [2], which only achieves 9.96%/7.87% SR on the seen/unseen splits, STEP-7B demonstrates a substantial performance leap. This sharp improvement directly validates the effectiveness of our specialized fine-tuning process. Furthermore, even when compared to significantly larger models such as Qwen3-VL-32B-Instruct (15.20%/11.76% SR) [1] and proprietary models like GPT-4o (16.00%/16.00% SR) [16] and GPT-5.4 (18.00%/16.00% SR) [17], STEP-7B still maintains a

[3] `https://github.com/VIPL-VSU/STEP`

Table 3: Ablation study on ALFRED. **RGB**: Egocentric modality; **Marker**: visual prompting markers; **Map**: Map modality; **CoT**: Chain-of-Thought Annotation; **Prim.**: Primitive skills. The last two rows denote performance using Oracle information.

Method	Components							Val Seen		Val Unseen	
	RGB	Marker	Map	CoT	Prim.	GT Seg.	GT Goal	SR	GC	SR	GC
Qwen2.5-VL-7B (Base)								9.96	18.59	7.87	18.26
+ RGB Data	✓							12.75	23.44	13.92	22.38
+ Marker	✓	✓						19.68	30.66	18.43	30.61
+ Map Modality	✓	✓	✓					21.12	31.25	19.22	29.85
+ Chain-of-Thought	✓	✓	✓	✓				28.40	40.35	23.72	33.33
+ Primitive. (**STEP-7B**)	✓	✓	✓	✓	✓			**30.28**	**39.69**	**26.77**	**39.36**
+ Oracle Seg.	✓	✓	✓	✓	✓	✓		38.55	46.76	41.57	50.45
+ Oracle Goal	✓	✓	✓	✓	✓	✓	✓	**45.20**	**52.50**	**55.69**	**61.52**

clear advantage on both splits. This underscores that the high-quality embodied alignment provided by the STEP-CoT data engine outweighs the benefits of pure parameter scaling.

Furthermore, STEP-7B demonstrates competitive performance against Specialist models like FILM [15] and LEBP [12]. While these specialist models are explicitly trained from scratch for the ALFRED environment and often excel in seen scenes (e.g., E.T. reaching 46.59% SR), they frequently suffer from severe performance degradation in unseen environments due to overfitting on specific visual distributions. In contrast, STEP-7B exhibits superior robustness, with a much smaller performance gap (Δ SR of 3.51%). This stability stems from its Markovian formulation, which treats each observation as an independent decision state, effectively preventing scene memorization and forcing the model to rely on generalized spatial reasoning. Qualitative results are shown in Fig. 4.

4.3 Ablation Study

We conduct a comprehensive ablation study on the ALFRED benchmark to evaluate the effectiveness of our proposed modules and data engine (Tab. 3).

Impact of Modalities and Basic Training. Starting from Qwen2.5-VL-7B-Instruct, RGB action data improves SR from 9.96%/7.87% to 12.75%/13.92% on the Seen/Unseen splits. Markers further raise SR to 19.68%/18.43%, confirming their importance for action grounding. The Map modality then increases SR to 21.12%/19.22%, showing consistent gains across both splits. However, the modest gain suggests that simple image concatenation is insufficient for the model to develop robust map reasoning or achieve precise ego-map alignment.

The Power of STEP-CoT and Primitive Tasks. With CoT traces, Seen/Unseen SR improves from 21.12%/19.22% to 28.40%/23.72%. Seen/Unseen GC also improves from 31.25%/29.85% to 40.35%/33.33%. The CoT paradigm enables the model to learn complex joint reasoning and align egocentric observations with the global map effectively. Finally, by incorporating primitive capability data, which focuses on fundamental spatial reasoning and map-ego alignment,

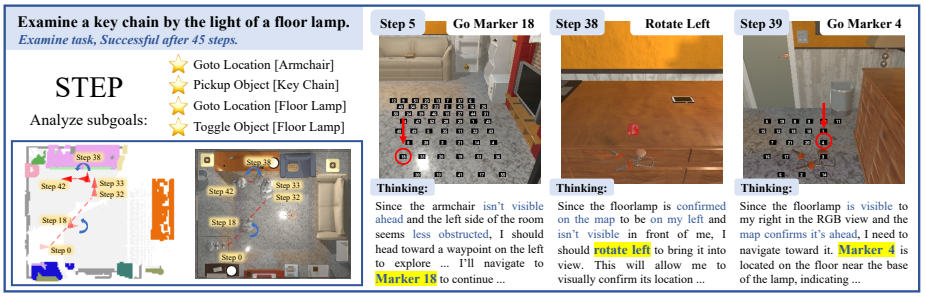

Fig. 4: Qualitative Case Study of STEP-7B on ALFRED. We showcase the decomposed subgoals and the agent's spatial reasoning during key execution steps.

the full non-oracle version of STEP reaches Seen/Unseen SR of 30.28%/26.77% and GC of 39.69%/39.36%.

Oracle Analysis. To isolate the model's decision-making capabilities from external factors, we conduct two oracle experiments. When Oracle Seg. is used to eliminate the impact of mapping errors and perception noise, Seen/Unseen SR reaches 38.55%/41.57%, while Seen/Unseen GC reaches 46.76%/50.45%. Furthermore, we employ the Oracle Goal setting, which provides ground-truth subgoal decomposition and addresses discrepancies between natural language instructions and the benchmark's internal evaluation logic. This configuration yields the highest Seen/Unseen SR of 45.20%/55.69% and GC of 52.50%/61.52%, underscoring STEP's strong potential as an embodied agent when provided with idealized perception and task decomposition.

Efficiency Analysis. We evaluate the computational efficiency of STEP compared to LLM-based baselines in Tab. 4. Due to the hybrid-modality input, STEP incurs a higher input token overhead per call. However, its reasoning output is more concise than R2C [3] owing to a more focused CoT process. Crucially, our point-level planning strategy allows the agent to navigate long-range trajectories with fewer decision-making cycles. This significantly reduces the inference frequency, leading to a lower total computational cost per task.

Failure Modes. We categorize the failure cases of STEP on ALFRED into four types: (i) Segmentation Failure, where inaccurate masks degrade the fidelity of the semantic map; (ii) Instruction Misalignment, where subgoals do not match instructions or environments; (iii) Object Search Failure, particularly for small or occluded items, often resulting in incomplete task execution; and (iv) Interaction Failure, caused by visibility or proximity constraints in the simulator. Fig. 5 shows the failure distribution; see the supplementary material for further details.

4.4 Results on Navigation

To further evaluate the model's generalization capability across distinct embodied tasks, we extend STEP-7B to the AI2-THOR ObjectNav benchmark [27]

Table 4: Comparison of computational efficiency. Inference speed and frequency denote latency per call and average calls per episode, respectively.

Model	LLM-P [24]	R2C [3]	STEP
Input Prompt Length (tokens)	∼1k	∼1k	∼1.5k
Output Seq. Length (tokens)	∼20	∼1.5k	∼0.5k
Inference Speed (seconds)	5	3.79	0.978
Inference Freq. (times per ep)	<10	∼50	∼23

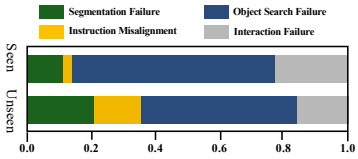

Fig. 5: The percentage of different failure types on ALFRED.

Table 5: Performance of STEP on the AI2-THOR object navigation task.

Method	SR (%) ↑ ‖ SPL (%) ↑									
	Kitchen		Living room		Bedroom		Bathroom		Average	
Random [4]	10.40	4.80	11.47	3.40	13.07	8.20	21.60	11.13	14.13	6.88
SAVN [27]	43.60	17.80	21.60	7.71	29.20	8.65	69.60	28.49	40.86	16.15
S2P [4]	45.50	31.05	28.50	19.00	50.40	25.29	60.25	36.70	46.16	28.01
R2C [3]	45.00	27.10	46.00	26.27	33.00	17.31	**85.00**	47.38	52.25	29.52
STEP-7B (ours)	**50.00**	**34.38**	**56.00**	**34.87**	**55.00**	**33.41**	82.00	**52.43**	**60.75**	**38.77**

without any additional fine-tuning. As shown in Tab. 5, STEP-7B achieves superior performance across the majority of metrics, reaching an average SR of 60.75% and an SPL of 38.77%. Specifically, STEP-7B maintains consistent performance in cluttered living rooms and bedrooms, while baselines like S2P [4] and R2C [3] degrade in these complex layouts. Notably, the consistent SPL improvements across all categories reflect superior navigation efficiency. This success stems from precise spatial-semantic grounding across egocentric cues and BEV maps, as well as efficient point-level planning.

4.5 Application in Real-World Scenarios

To evaluate the sim-to-real transferability of the STEP framework, we deploy STEP-7B on a LoCoBot platform within a custom indoor environment. Our real-world evaluation comprises three distinct scenarios: two object-goal navigation tasks and a room-exit task. Without any fine-tuning on real-world data, STEP-7B shows promising zero-shot transfer to raw egocentric visual streams and semantic maps in the real world. As illustrated in Fig. 6, the agent accurately aligns semantic instructions with physical entities and generates efficient, point-level navigational waypoints. These results underscore STEP's robust generalization and its significant utility for real-world embodied applications. Further details are included in the supplementary material.

5 Conclusion

In this paper, we introduce STEP, a novel framework for Embodied Instruction Following that empowers agents with Spatial Thinking and Egocentric Pointing.

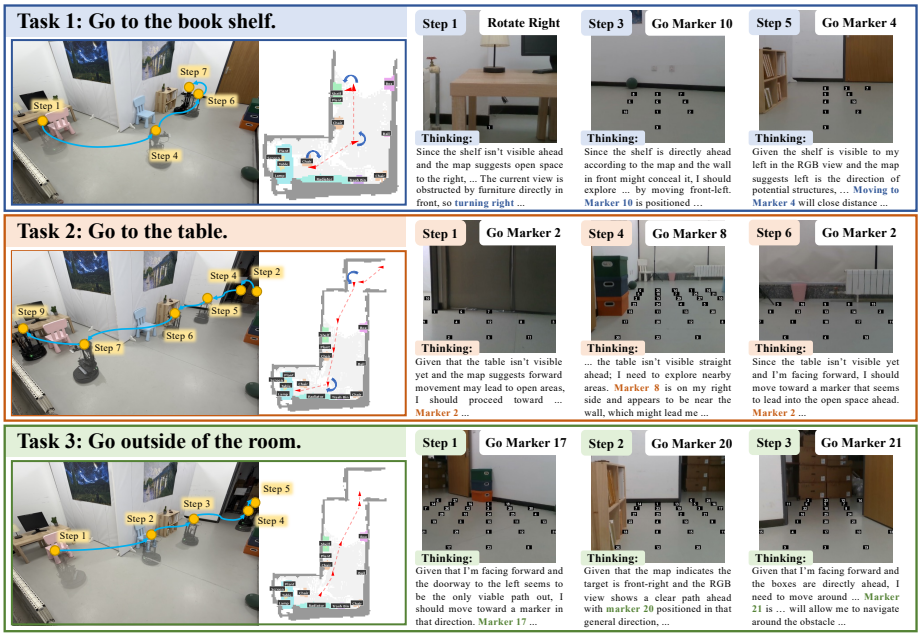

Fig. 6: Real-World Deployment on LoCoBot. We visualize STEP-7B's spatial reasoning process during key execution steps in a custom indoor environment.

Supported by our STEP-CoT data engine, the framework enables agents to perform joint reasoning across egocentric views and global maps. By integrating a Hybrid Map-Egocentric Perception module and Cognitive Point-Level Planning, STEP directly outputs point-level plans on egocentric images alongside explicit reasoning traces. This approach successfully realizes "Look at the View, Think with the Map, and Point to the Goal". Despite its robust performance, several limitations provide avenues for future research. In extremely cluttered environments, where a large number of objects are densely stacked, the quality of the current annotation maps can be adversely affected. Additionally, the high cost of data generation remains a limitation, as maintaining high-quality reasoning traces currently necessitates the use of powerful frontier models such as Qwen3-VL-235B or the GPT-4 series. In future work, we intend to explore more suitable map representations to enhance perception in complex layouts. Furthermore, we aim to investigate training paradigms based on reinforcement learning to alleviate the current dependency on expensive, high-quality CoT data while further boosting the agent's decision-making autonomy.

Acknowledgements This work is partially supported by Natural Science Foundation of China under contracts Nos. 62495082, 62461160331, and Beijing Municipal Natural Science Foundation Nos. L257009, L242025.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-VL Technical Report. arXiv preprint arXiv:2511.21631 (2025)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-VL Technical Report. arXiv preprint arXiv:2502.13923 (2025)
3. Bai, Z., Li, H., Fu, B., Xiong, C., Wang, R., Chen, X.: R2C: Mapping room to chessboard to unlock LLM as low-level action planner. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19456–19466 (2025)
4. Buoso, D., Robinson, L., Averta, G., Torr, P., Franzmeyer, T., Martini, D.D.: Select2Plan: Training-free ICL-based planning through VQA and memory retrieval. IEEE Robotics and Automation Letters **10**(11), 11267–11274 (2025)
5. Chen, J., Lin, B., Liu, X., Ma, L., Liang, X., Wong, K.Y.K.: Affordances-oriented planning using foundation models for continuous vision-language navigation. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 23568–23576 (2025)
6. Hong, Y., Zhou, Y., Zhang, R., Dernoncourt, F., Bui, T., Gould, S., Tan, H.: Learning navigational visual representations with semantic map supervision. In: IEEE/CVF International Conference on Computer Vision. pp. 3055–3067 (2023)
7. Ichter, B., Brohan, A., Chebotar, Y., Finn, C., Hausman, K., Herzog, A., Ho, D., Ibarz, J., Irpan, A., Jang, E., Julian, R., Kalashnikov, D., Levine, S., Lu, Y., Parada, C., Rao, K., Sermanet, P., Toshev, A.T., Vanhoucke, V., Xia, F., Xiao, T., Xu, P., Yan, M., Brown, N., Ahn, M., Cortes, O., Sievers, N., Tan, C., Xu, S., Reyes, D., Rettinghouse, J., Quiambao, J., Pastor, P., Luu, L., Lee, K.H., Kuang, Y., Jesmonth, S., Joshi, N.J., Jeffrey, K., Ruano, R.J., Hsu, J., Gopalakrishnan, K., David, B., Zeng, A., Fu, C.K.: Do As I Can, Not As I Say: Grounding language in robotic affordances. In: Conference on Robot Learning. vol. 205, pp. 287–318 (2023)
8. Kim, B., Kim, J., Kim, Y., Min, C., Choi, J.: Context-aware planning and environment-aware memory for instruction following embodied agents. In: IEEE/CVF International Conference on Computer Vision. pp. 10936–10946 (2023)
9. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
10. LaValle, S.M.: Planning Algorithms. Cambridge University Press (2006)
11. Lin, B., Nie, Y., Wei, Z., Chen, J., Ma, S., Han, J., Xu, H., Chang, X., Liang, X.: NavCoT: Boosting LLM-based vision-and-language navigation via learning disentangled reasoning. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(7), 5945–5957 (2025)
12. Liu, H., Liu, Y., He, H., Yang, H.: LEBP – language expectation & binding policy: A two-stream framework for embodied vision-and-language interaction task learning agents. arXiv preprint arXiv:2203.04637 (2022)

13. Long, Y., Cai, W., Wang, H., Zhan, G., Dong, H.: InstructNav: Zero-shot system for generic instruction navigation in unexplored environment. In: Conference on Robot Learning. vol. 270, pp. 2049–2060 (2025)
14. McDermott, D., Ghallab, M., Howe, A., Knoblock, C., Ram, A., Veloso, M., Weld, D., Wilkins, D.: PDDL – The planning domain definition language. Tech. Rep. CVC TR-98-003, Yale Center for Computational Vision and Control (1998)
15. Min, S.Y., Chaplot, D.S., Ravikumar, P., Bisk, Y., Salakhutdinov, R.: FILM: Following instructions in language with modular methods. In: International Conference on Learning Representations (2022)
16. OpenAI: GPT-4o System Card. arXiv preprint arXiv:2410.21276 (2024)
17. OpenAI: Introducing GPT-5.4 (2026), `https://openai.com/index/introducing-gpt-5-4/`, accessed 1 July 2026
18. Pashevich, A., Schmid, C., Sun, C.: Episodic transformer for vision-and-language navigation. In: IEEE/CVF International Conference on Computer Vision. pp. 15942–15952 (2021)
19. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., Zhang, L.: Grounded SAM: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024)
20. Shah, D., Osiński, B., Ichter, B., Levine, S.: LM-Nav: Robotic navigation with large pre-trained models of language, vision, and action. In: Conference on Robot Learning. vol. 205, pp. 492–504 (2023)
21. Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., Fox, D.: ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10740–10749 (2020)
22. Silver, T., Hariprasad, V., Shuttleworth, R.S., Kumar, N., Lozano-Pérez, T., Kaelbling, L.P.: PDDL planning with pretrained large language models. In: NeurIPS Foundation Models for Decision Making Workshop (2022)
23. Song, C.H., Kil, J., Pan, T.Y., Sadler, B.M., Chao, W.L., Su, Y.: One step at a time: Long-horizon vision-and-language navigation with milestones. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15482–15491 (2022)
24. Song, C.H., Wu, J., Washington, C., Sadler, B.M., Chao, W.L., Su, Y.: LLM-Planner: Few-shot grounded planning for embodied agents with large language models. In: IEEE/CVF International Conference on Computer Vision. pp. 2998–3009 (2023)
25. Suglia, A., Gao, Q., Thomason, J., Thattai, G., Sukhatme, G.: Embodied BERT: A transformer model for embodied, language-guided visual task completion. arXiv preprint arXiv:2108.04927 (2021)
26. Tolman, E.C.: Cognitive maps in rats and men. Psychological Review **55**(4), 189–208 (1948)
27. Wortsman, M., Ehsani, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Learning to learn how to learn: Self-adaptive visual navigation using meta-learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6750–6759 (2019)
28. Wu, P., Mu, Y., Wu, B., Hou, Y., Ma, J., Zhang, S., Liu, C.: VoroNav: Voronoi-based zero-shot object navigation with large language model. In: International Conference on Machine Learning. vol. 235, pp. 53757–53775 (2024)
29. Yang, J., Zhang, H., Li, F., Zou, X., Li, C., Gao, J.: Set-of-Mark prompting unleashes extraordinary visual grounding in GPT-4V. arXiv preprint arXiv:2310.11441 (2023)

30. Yokoyama, N., Ha, S., Batra, D., Wang, J., Bucher, B.: VLFM: Vision-language frontier maps for zero-shot semantic navigation. In: IEEE International Conference on Robotics and Automation. pp. 42–48 (2024)
31. Yu, B., Kasaei, H., Cao, M.: L3MVN: Leveraging large language models for visual target navigation. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 3554–3560 (2023)
32. Zhan, Z., Yu, L., Yu, S., Tan, G.: MC-GPT: Empowering vision-and-language navigation with memory map and reasoning chains. arXiv preprint arXiv:2405.10620 (2024)
33. Zhang, J., Wang, K., Xu, R., Zhou, G., Hong, Y., Fang, X., Wu, Q., Zhang, Z., Wang, H.: NaVid: Video-based VLM plans the next step for vision-and-language navigation. In: Robotics: Science and Systems (2024)
34. Zhang, L., Hao, X., Xu, Q., Zhang, Q., Zhang, X., Wang, P., Zhang, J., Wang, Z., Zhang, S., Xu, R.: MapNav: A novel memory representation via annotated semantic maps for VLM-based vision-and-language navigation. In: Annual Meeting of the Association for Computational Linguistics. pp. 13032–13056 (2025)
35. Zhang, L., Zhang, Q., Wang, H., Xiao, E., Jiang, Z., Chen, H., Xu, R.: TriHelper: Zero-shot object navigation with dynamic assistance. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 10035–10042 (2024)
36. Zhang, Y., Chai, J.: Hierarchical task learning from language instructions with unified transformers and self-monitoring. In: Findings of the Association for Computational Linguistics. pp. 4202–4213 (2021)
37. Zhong, L., Gao, C., Ding, Z., Liao, Y., Ma, H., Zhang, S., Zhou, X., Liu, S.: TopV-Nav: Unlocking the top-view spatial reasoning potential of MLLM for zero-shot object navigation. arXiv preprint arXiv:2411.16425 (2024)
38. Zhou, G., Hong, Y., Wu, Q.: NavGPT: Explicit reasoning in vision-and-language navigation with large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 7641–7649 (2024)
39. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., Vuong, Q., Vanhoucke, V., Tran, H., Soricut, R., Singh, A., Singh, J., Sermanet, P., Sanketi, P.R., Salazar, G., Ryoo, M.S., Reymann, K., Rao, K., Pertsch, K., Mordatch, I., Michalewski, H., Lu, Y., Levine, S., Lee, L., Lee, T.W.E., Leal, I., Kuang, Y., Kalashnikov, D., Julian, R., Joshi, N.J., Irpan, A., Ichter, B., Hsu, J., Herzog, A., Hausman, K., Gopalakrishnan, K., Fu, C., Florence, P., Finn, C., Dubey, K.A., Driess, D., Ding, T., Choromanski, K.M., Chen, X., Chebotar, Y., Carbajal, J., Brown, N., Brohan, A., Arenas, M.G., Han, K.: RT-2: Vision-language-action models transfer web knowledge to robotic control. In: Conference on Robot Learning. vol. 229, pp. 2165–2183 (2023)