ESCAPE: Episodic Spatial Memory and Adaptive Execution Policy for Long-Horizon Mobile Manipulation

Jingjing Qian[1,2] Zeyuan He[1] Chen Shi[1] Lei Xiao[3] Li Jiang[1,2,†]

[1]The Chinese University of Hong Kong, Shenzhen
[2]Shenzhen Loop Area Institute [3]LiAuto Inc.

Abstract. Coordinating navigation and manipulation with robust performance is essential for embodied AI in complex indoor environments. However, as tasks extend over long horizons, existing methods often struggle due to catastrophic forgetting, spatial inconsistency, and rigid execution. To address these issues, we propose **ESCAPE** (**E**pisodic **S**patial memory **C**oupled with an **A**daptive **P**olicy for **E**xecution), operating through a tightly coupled perception-grounding-execution workflow. For robust perception, ESCAPE features a Spatio-Temporal Fusion Mapping module to autoregressively construct a depth-estimation-free, persistent 3D spatial memory, and a Memory-Driven Target Grounding module for precise interaction mask generation. To achieve flexible action, our Adaptive Execution Policy dynamically orchestrates proactive global navigation and reactive local manipulation to seize opportunistic targets. ESCAPE achieves state-of-the-art performance on the ALFRED benchmark, reaching 65.09% and 60.79% success rates in test seen and unseen environments with step-by-step instructions. By reducing redundant exploration, our ESCAPE substantially improves path-length-weighted metrics and remains robust (61.24%/56.04%) even without detailed guidance for long-horizon tasks.

Keywords: Embodied Mobile Manipulation · Scene Spatial Memory · Adaptive Action Planning

1 Introduction

The vision of autonomous robots capable of assisting humans in everyday household tasks has long been a driving force in robotics research [3,18,25,28]. Recent advances in Embodied AI [7, 16, 31, 36, 38, 43] have demonstrated remarkable progress in enabling artificial agents to perceive, navigate, and interact with their surroundings. Within this context, embodied Long-Horizon Mobile Manipulation [34,46] emerges as a crucial next step, requiring agents to coordinate both long-term navigation and sophisticated object manipulation while maintaining robust performance in dynamic, complex indoor environments.

† Corresponding Author.

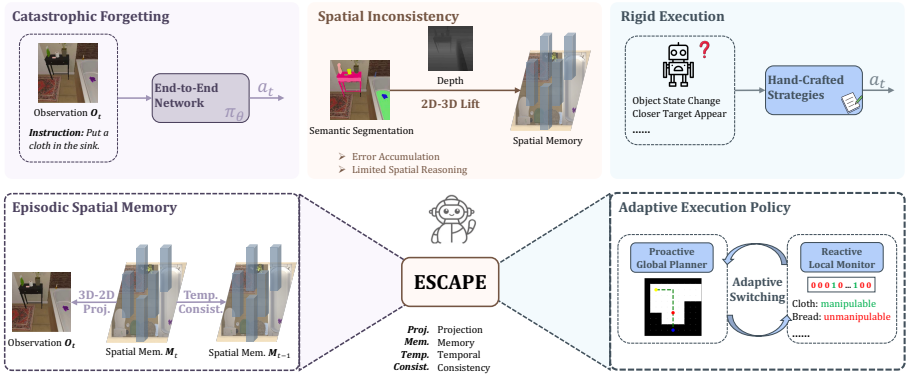

Fig. 1: Existing methods struggle with long-horizon tasks due to catastrophic forgetting, spatial inconsistency, and rigid execution. ESCAPE addresses these issues through a persistent episodic spatial memory that captures spatio-temporal relationships, coupled with an adaptive execution policy that seamlessly coordinates proactive global navigation and reactive local manipulation.

Embodied Long-Horizon Mobile Manipulation poses key challenges for autonomous agents, particularly as tasks shift from isolated actions to complex, *long-horizon* goals. These tasks demand robust spatial reasoning to understand complex and dynamic 3D environments from limited egocentric views, as well as sophisticated planning to coordinate long-term navigation with precise object manipulation while maintaining goal-oriented behavior. Real-time adaptation to environmental feedback further adds to the complexity. ALFRED [34] benchmark serves as a standard testbed for evaluating these challenges, requiring agents to follow natural language instructions to complete long-horizon household tasks involving extensive exploration and opportunistic interaction. In this work, following the ALFRED [34] and HomeRobot [46] style setting, manipulation refers to high-level object interactions with predicted object masks, rather than low-level continuous control or dexterous manipulation.

Despite recent progress, executing long-horizon mobile manipulation tasks remains exceptionally challenging, exposing three critical limitations in existing methods, as illustrated in Fig. 1. First, **catastrophic forgetting** occurs in end-to-end approaches (e.g., Seq2Seq [34], MOCA [35], and E.T. [28]) that directly map current observations to actions without maintaining a persistent spatial memory. This forces agents to redundantly re-explore previously visited areas when searching for target objects, drastically reducing long-horizon efficiency.

Second, **spatial inconsistency** plagues modular methods like FILM [25], CAPEAM [15], DISCO [44], and ThinkBot [23]. These methods process semantics and affordances independently in the 2D image before back-projecting them into 3D space using imperfect depth estimation. In long-horizon tasks, these perception errors gradually accumulate due to inaccurate depth-semantic lift-

ing. Furthermore, restricting feature interactions entirely to the 2D image plane fundamentally limits the agent's 3D spatial reasoning capabilities.

Third, **rigid execution** limits the generalization of methods (e.g., FILM [25], Prompter [11], and CAPEAM [15]) that rely heavily on hand-crafted strategies for specific scenarios, such as object state changes and closer target appearances. By blindly following fixed long-term trajectories, these methods fail to react promptly and cannot preemptively interrupt navigation when encountering a target earlier than expected, leading to suboptimal execution paths.

To address these key issues, we propose **ESCAPE**, a memory-centric framework designed for long-horizon mobile manipulation, as shown in Fig. 1. ESCAPE operates through a tightly coupled perception → grounding → execution workflow. In the **perception** phase, our *Spatio-Temporal Fusion Mapping* module leverages spatio-temporal cross-attention operating directly in 3D space to fuse current observation with historical memory, autoregressively constructing a persistent episodic spatial memory that captures temporal consistency and spatial-semantic relationships. Notably, by using precise 3D-to-2D projection to extract visual features, this module eliminates depth estimation dependence, enhancing the agent's 3D spatial reasoning capabilities. For the subsequent **grounding** phase, a *Memory-Driven Target Grounding* module uses the 3D geometric features stored in the memory to query current 2D visual observations, extracting precise masks for physical interaction. Finally, in the **execution** phase, our *Adaptive Execution Policy* resolves the conflict between long-term planning and short-term reacting. It employs a proactive global planner for memory-based navigation, while running a reactive local monitor concurrently to seize immediate manipulation targets, seamlessly shifting the execution horizon as needed. For fair evaluation on ALFRED [34], we adopt [25]'s language parsing module while focusing on advancing navigation and manipulation abilities.

In summary, our main contributions are three-fold:

- We propose a novel Episodic Spatial Memory framework tailored for long-horizon tasks, featuring a Spatio-Temporal Fusion Mapping module for depth-estimation-free, consistent 3D spatial memory construction, and a Memory-Driven Target Grounding module for precise interaction mask generation.
- We introduce an Adaptive Execution Policy that dynamically orchestrates proactive global navigation and reactive local manipulation, improving the operational efficiency and adaptability of the embodied agent.
- ESCAPE achieves superior performance on the highly competitive ALFRED benchmark. Notably, by reducing redundant exploration, it attains substantial improvements in path-length-weighted metrics (PLWSR and PLWGC), demonstrating highly efficient long-horizon execution.

2 Related Work

Embodied Mobile Manipulation. Enabling mobile robots to navigate and manipulate objects in complex environments remains a major embodied AI challenge. Various simulators and benchmarks have emerged to advance this

field [1, 8, 17, 37, 39]. While early works focused on navigation in static scenes, recent benchmarks require agents to manipulate objects in dynamic environments with evolving semantics. Some support physical robot testing [12, 46], while most use simulated environments for reproducible evaluation [6, 45]. ALFRED [34] stands out by integrating navigation and manipulation, challenging agents to follow natural language instructions for *long-horizon* household tasks. Its dynamic environments, which require extensive exploration and opportunistic interaction, provide an ideal testbed for our proposed ESCAPE framework, which addresses these challenges by coupling a persistent episodic spatial memory with an adaptive execution policy for flexible navigation and manipulation.

Scene Spatial Memory. Spatial memory is crucial for long-horizon mobile manipulation. Some approaches rely on explicit memory structures like 2D semantic maps or topological graphs [48, 50]. While intuitive, these discrete representations struggle to preserve fine-grained 3D geometry and suffer from error accumulation during prolonged navigation. Conversely, implicit representations like Neural Radiance Fields (NeRF) [21, 24, 51] and 3D Gaussian Splatting (3DGS) [5, 13, 32, 47] enable high-fidelity 3D reconstruction and rendering. However, their intensive optimization and implicit formulation complicate real-time semantic querying for physical interaction. To bridge this gap, Bird's-Eye-View (BEV) representations [4, 10, 20, 22, 27, 30, 41] offer an effective top-down spatial memory paradigm. Utilizing spatio-temporal transformers [19] for historical feature aggregation, BEV enables rich spatial-semantic fusion. Building on this, ESCAPE constructs a persistent episodic spatial memory, elegantly balancing 3D spatial reasoning with efficient 2D target grounding.

Adaptive Action Planning. In long-horizon mobile manipulation, agents must robustly translate complex perceptual states into executable actions [42, 49]. FILM [25] utilizes neural networks to generate actions from semantic maps, though such map-based methods often lack precision for object grounding. To address this, DISCO [44] introduces fine-action networks that improve alignment and reduce ambiguity. However, despite these perceptual and grounding advancements, both methods rely on rigid sequential planning. This inflexibility prevents agents from preemptively interrupting navigation to seize opportunistic manipulation targets. To address this issue, our Adaptive Execution Policy couples a proactive global planner with a reactive local monitor, dynamically shifting between long-horizon navigation and immediate manipulation.

3 Method

In this work, we present our proposed ESCAPE. The overall architecture is shown in Fig. 2. To address the catastrophic forgetting and rigid execution typically encountered in long-horizon embodied tasks, ESCAPE is built upon two core pillars: a unified **Episodic Spatial Memory** system for robust perception, and an **Adaptive Execution Policy** for flexible action.

Specifically, this memory-centric framework forms a tightly coupled sequential perception → grounding → execution workflow. First, the *Spatio-Temporal*

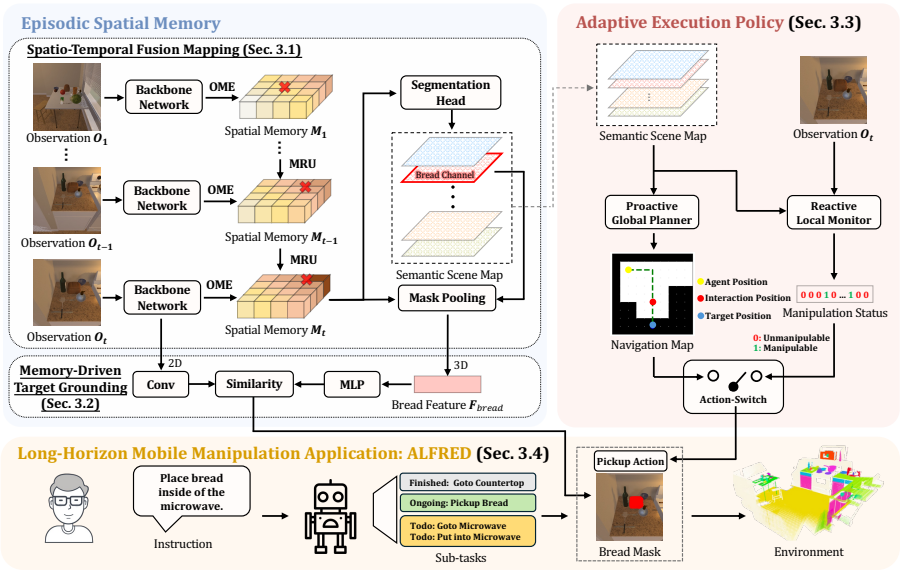

Fig. 2: Our ESCAPE framework features two core pillars: a unified Episodic Spatial Memory that processes observations via Spatio-Temporal Fusion Mapping module and Memory-Driven Target Grounding module to generate semantic scene maps and precise interaction masks, and an Adaptive Execution Policy that dynamically shifts between Proactive Global Planner and Reactive Local Monitor. The application shows placing bread inside a microwave, where the agent (marked by ✗) navigates and manipulates objects based on instructions.

Fusion Mapping module (Sec. 3.1) autoregressively constructs and updates a persistent 3D spatial memory from egocentric observations. Based on this globally consistent memory, the *Memory-Driven Target Grounding* module (Sec. 3.2) uses 3D object features derived from the memory to query 2D visual observations, extracting precise masks for physical interactions. Finally, the *Adaptive Execution Policy* (Sec. 3.3) leverages the spatial memory for proactive long-horizon navigation trajectories, while a reactive local monitor simultaneously seizes immediate manipulation opportunities. We conclude this section by detailing ESCAPE's application on the ALFRED [34] benchmark (Sec. 3.4).

3.1 Spatio-Temporal Fusion Mapping

For long-horizon mobile manipulation tasks, an effective memory representation must capture spatial relationships (such as relative object positions) and semantic information, while maintaining persistent temporal consistency across dynamic environments to avoid redundant exploration. Inspired by recent BEV representations [10, 19] that employ spatio-temporal transformers, we propose the Spatio-Temporal Fusion Mapping module. Unlike original BEV methods us-

ing multi-view images, our module autoregressively aggregates temporal images from a single egocentric camera to construct a persistent 3D spatial memory.

To achieve accurate and efficient memory construction, we introduce an Observation-to-Memory Encoding (OME) mechanism implemented via spatial cross attention [19] to update the memory feature from the current observation. Additionally, to integrate historical memory and acquire spatial reasoning ability in 3D space, we propose a Memory Retrieval and Update (MRU) mechanism utilizing temporal cross attention. Finally, we employ a 3D map semantic segmentation head to convert memory features into a semantic map.

Observation-to-Memory Encoding (OME). As shown in Fig. 3 (a) and (b), from a bird's-eye perspective, we partition the entire room into an $H \times W$ grid, where each grid cell is initialized with a memory query $\mathbf{Q}_p \in \mathbb{R}^{1\times C}$ located at $p = (x, y)$ with C dimension. These queries are then lifted into pillar-like queries, where each 2D grid $p = (x, y)$ is transformed into $\mathbf{N}_{\text{ref}}$ 3D reference points (x, y, z_i) sampled along the height dimension. For each reference point, we project it onto the current observation through the projection matrix of the agent's camera, which can be written as:

$$z_i' \cdot \begin{bmatrix} x_i' & y_i' & 1 \end{bmatrix}^T = \mathbf{P} \cdot \begin{bmatrix} x & y & z_i & 1 \end{bmatrix}^T . \tag{1}$$

Here, (x_i', y_i') is the 2D image plane coordinate projected from 3D point (x, y, z_i), $\mathbf{P} \in \mathbb{R}^{3\times 4}$ is the known projection matrix of the agent's camera, and z_i' represents the depth factor of the projected point in the camera coordinate system. This projection yields a 2D reference point in the image plane, around which we sample and aggregate image features through weighted summation to obtain the output of OME, formulated as follows:

$$\text{OME}(\mathbf{Q}_p, \mathbf{F}_t) = \sum_{i=1}^{\mathbf{N}_{\text{ref}}} \text{DeformAttn}(\mathbf{Q}_p, (x_i', y_i'), \mathbf{F}_t), \tag{2}$$

where i indexes the reference points and $\mathbf{F}_t$ is the multi-level semantic features extracted from the current frame $\mathbf{O}_t$ using a ResNet50 backbone [9]. After applying this computation process to all queries in the spatial memory, we finish updating the global memory feature using the current observation.

Here, two points require clarification. First, vanilla multi-head attention [40] in our OME is computationally expensive when applied globally across the entire spatial grid. Following [19], we adopt deformable attention (DeformAttn in Eq. 2) [52], where each memory query $\mathbf{Q}_p$ only interacts with its regions of interest in the observation. Second, previous methods like FILM [25], DISCO [44] and ThinkBot [23] back-project 2D image features to 3D space using depth estimation, causing error accumulation in long-horizon mobile manipulation tasks. Our OME instead projects 3D points to the image plane to extract 2D features for 3D spatial memory updates, eliminating depth estimation dependence and improving memory fidelity.

Memory Retrieval and Update (MRU). As shown in Fig. 3 (b) and (c), given the memory query $\mathbf{Q}$ at current timestep t and the cached spatial memory

$\mathbf{M}_{t-1}$ from timestep $t-1$, each grid query $\mathbf{Q}_p$ at position $p=(x,y)$ retrieves relevant features from $\mathbf{M}_{t-1}$ to prevent catastrophic forgetting while interacting with neighboring features in $\mathbf{Q}$ to enable spatial reasoning. For computational efficiency, we adopt deformable attention to implement MRU, similar to OME, which can be formulated as:

$$\mathrm{MRU}(\mathbf{Q}_p, \{\mathbf{Q}, \mathbf{M}_{t-1}\}) = \sum_{\mathbf{V}\in\{\mathbf{Q},\mathbf{M}_{t-1}\}} \mathrm{DeformAttn}(\mathbf{Q}_p, p, \mathbf{V}). \tag{3}$$

Our MRU leverages temporal cross attention to adaptively sample and aggregate memory features from the previous timestep, maintaining long-term consistency. Meanwhile, each grid query interacts with surrounding memory features in 3D space, which better captures geometric relationships and enhances spatial reasoning such as relative positions and distances.

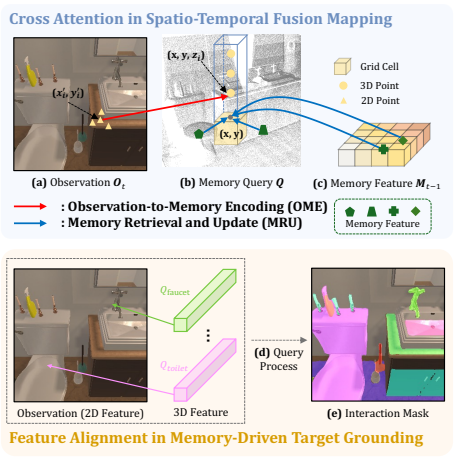

Fig. 3: Demonstration of cross attention in the Spatio-Temporal Fusion Mapping and 2D-3D feature alignment in the Memory-Driven Target Grounding.

3D Map Semantic Segmentation Head. As shown in Fig. 2, our Spatio-Temporal Fusion Mapping module processes the current observation via a backbone to generate image features, then uses OME to extract features for memory updates while MRU fuses temporal and 3D spatial information, producing the updated episodic memory $\mathbf{M}_t \in \mathbb{R}^{H\times W\times C}$, where $H\times W$ denotes grid cells and C represents feature dimension. Based on this memory feature, we introduce a 3D map semantic segmentation head to predict object category distribution on the grid map $\mathbf{D}_t$, trained with ground-truth labels $\mathbf{L}_t$ through:

$$\mathcal{L}_{\mathrm{map}} = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} \mathrm{BCE}(\mathbf{D}_t^{h,w}, \mathbf{L}_t^{h,w}), \tag{4}$$

where $\mathcal{L}_{\mathrm{map}}$ is the map segmentation loss, $\mathbf{L}_t$ is a multi-hot vector for each grid cell indicating the object categories present in that cell and BCE denotes binary cross-entropy loss measuring grid-wise discrepancy between predicted probabilities and ground-truth labels per category.

3.2 Memory-Driven Target Grounding

While the Spatio-Temporal Fusion Mapping module effectively constructs a global 3D spatial and semantic memory, it lacks the fine-grained 2D localization

required for physical manipulation. In long-horizon tasks, accurate identification and localization of target objects from the agent's egocentric view are crucial. To address this, we introduce the Memory-Driven Target Grounding module, which extracts interaction masks by leveraging the rich, object-specific features already stored within our episodic memory. Inspired by recent works [29, 33] on 2D-3D feature alignment, we propose generating target object masks by querying the current observation with abstract geometric features from the spatial memory.

As shown in Fig. 2, for the **Pickup Bread** sub-task, we perform mask pooling over the memory feature $\mathbf{M}_t$ using the bread channel from the semantic scene map. We aggregate features from bread-containing grid cells via averaging to compute $\mathbf{F}_{bread}$, then apply a multi-layer perceptron (MLP) to generate the bread query $\mathbf{Q}_{bread}$. The general formula for computing object queries is:

$$\begin{aligned} \mathbf{F}_{object} &= \frac{1}{|\mathcal{G}_{object}|} \sum_{(i,j)\in\mathcal{G}_{object}} \mathbf{M}_t^{(i,j)}, \\ \mathbf{Q}_{object} &= \mathrm{MLP}(\mathbf{F}_{object}), \end{aligned} \tag{5}$$

where $\mathcal{G}_{object}$ denotes grid cells where the object channel equals 1. Since object queries are derived from the episodic memory, they encode global geometric properties and are termed 3D object features.

Simultaneously, convolutional layers are used to map backbone image features to new 2D object features. The similarity computation between 3D object queries and 2D object features produces object masks, precisely localizing objects for interaction. The toilet scene in Fig. 3 (d) and (e) further shows this 2D-3D feature alignment mechanism. To train the Memory-Driven Target Grounding module, we introduce a 2D image semantic segmentation loss $\mathcal{L}_{\text{img}}$ as follows:

$$\mathcal{L}_{\text{img}} = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} \mathrm{BCE}(\mathbf{P}_t^{h,w}, \mathbf{G}_t^{h,w}), \tag{6}$$

where $\mathbf{P}_t^{h,w}$ represents the predicted image mask at pixel (h, w) at time-step t, and $\mathbf{G}_t^{h,w}$ denotes the corresponding ground-truth segmentation label.

Joint Training Objective and Implementation. To enable the episodic memory to capture spatial relationships while empowering accurate target grounding, we propose joint supervision combining 3D map segmentation and 2D image segmentation losses. The total loss $\mathcal{L}$ is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{map}} + \lambda \mathcal{L}_{\text{img}}, \tag{7}$$

where λ is a hyperparameter that balances the two segmentation losses, and we set $\lambda = 1.0$ in all experiments.

Following DISCO [44], we collect training instructions, images, semantic map and image labels from ALFRED [34] expert trajectories. For temporal continuity, semantic map labels at timestep t incorporate object distributions from step 1 to current timestep. We use a ResNet50 [9] backbone, OME and MRU layers

with deformable attention, a map segmentation head and an image segmentation head. On ALFRED, each scene is modeled as a $25m \times 25m$ room with $25cm \times 25cm$ grids, resulting in memory queries of size $H \times W \times C$ where $H = W = 100$ and $C = 256$. These memory queries serve as learnable parameters that are jointly optimized during training. The model is trained using AdamW with BCE loss for 25 epochs, batch size 64 and learning rate 2e-4. Training completes in 24 hours on 8 RTX 4090 GPUs.

3.3 Adaptive Execution Policy

In long-horizon mobile manipulation tasks, there is a fundamental conflict between executing a prolonged navigation plan and reacting to suddenly appearing environmental opportunities. As shown in Fig. 4 (a), during a **Pickup Potato** sub-task, after detecting a potato and proactively planning a long route, the agent might discover another potato in a previous blind spot after turning right. Rigidly following the initial plan reduces execution efficiency. To address this, we propose the Adaptive Execution Policy, a dynamic execution mechanism combining a proactive global planner and a reactive local monitor.

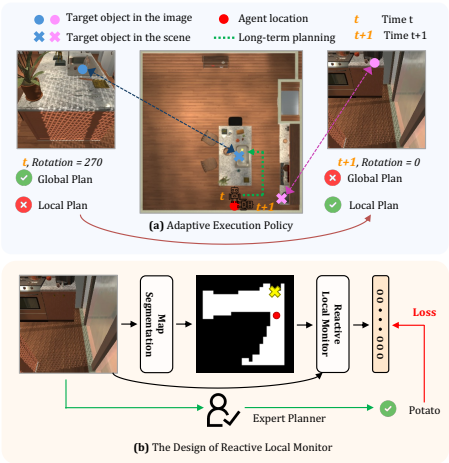

Fig. 4: Demonstration of the Adaptive Execution Policy and our design of reactive local monitor.

Random Exploration. Before potato detection, the agent explores based on a navigable map. In the semantic map generated by the spatial memory, a grid is considered **navigable** if only the floor category has a value of 1 while all other object categories are 0, thereby generating the navigable map. The agent randomly selects destinations from navigable cells and plans paths using the BFS (Breadth-First Search) algorithm.

Proactive Global Planner. After potato detection, the agent utilizes its persistent spatial memory for proactive goal navigation. It locates the potato at the highest probability grid, expands with navigable grids within a 1-meter radius, and uses BFS to generate long-horizon navigation waypoints (grid coordinates) considering potato region, agent position, and navigable area. Thus, the proactive global planner is a rule-based module requiring no training.

Reactive Local Monitor. Due to the agent's limited field of view, the global planner may overlook information in occluded areas, causing unnecessary travel. To address this, we introduce a reactive local monitor that continuously screens the short-horizon environment for immediate manipulation opportunities while executing the long-horizon navigation plan. As shown in Fig. 4 (a), when discovering a potato within range during navigation, the local monitor preempts the

global planner, bypassing the remainder of the long-horizon path to execute an immediate pickup action.

As shown in Fig. 4 (b), our reactive local monitor takes current observations and semantic scene maps as inputs, outputting a binary vector indicating whether each object category is manipulable (1=manipulable, 0=unmanipulable). An object is manipulable only if semantically manipulable and within operational range, requiring precise understanding of semantic information in observation and spatial relationships in the semantic scene map. To train this monitor, we design an expert planner with full scene knowledge for data collection. Following ALFRED's [34] data generation process, our expert planner annotates manipulable target categories per frame, yielding 332,432 training images.

The reactive local monitor employs ResNet50 [9] as the backbone for visual feature and map feature extraction, followed by a linear classifier predicting manipulable object categories. The model is trained using BCE loss based on manipulation labels from the expert planner. Training utilizes the AdamW optimizer with learning rate of 5×10^{-5} and batch size of 128 for 40 epochs to achieve optimal performance. Training completes in 2 hours on 8 RTX 4090 GPUs.

Our Adaptive Execution Policy systematically generates actions by leveraging both modules. For multi-instance targets (potatoes, watches, apples), the reactive local monitor opportunistically discovers new goals to improve efficiency. For unique large targets (fridges, sidetables, sofas), the episodic spatial memory provides precise localization, making the proactive global routes highly accurate. When the reactive local monitor fails due to occlusion or inappropriate pose, our hierarchical fallback strategy continues executing existing long-horizon plans or generates new navigation plans to reachable targets.

3.4 Long-Horizon Mobile Manipulation Application: ALFRED

We evaluate our approach on ALFRED [34], a widely-adopted benchmark for long-horizon embodied instruction following. More details about the benchmark can be found in Appendix A. Each ALFRED task contains high-level goal descriptions and step-by-step instructions. Appendix A.5 shows an ALFRED task **Put a cooked potato slice on the counter** where our agent only needs to place any cooked potato slice on any counter to complete the task. This allows our approach to focus on semantic segmentation to identify object categories rather than distinguishing individual instances.

Fig. 2 illustrates the ESCAPE pipeline on ALFRED. Based on received instruction, the agent converts it into executable sub-tasks. For fair comparison with baselines [15,25,44], we adopt [25]'s instruction processing module. For each sub-task, the agent relies on its Episodic Spatial Memory to perceive the environment and generate accurate scene maps and interaction masks. Simultaneously, it employs the Adaptive Execution Policy to seamlessly alternate between proactive long-horizon planning and reactive short-horizon interaction for improved task efficiency. Details are in Appendix B.

Table 1: Performance comparison on ALFRED benchmark [34]. ✓/ ✗ denotes whether step-by-step instructions are used. Bold values indicate the best results for each metric.

		Test Seen				Test Unseen			
	Ins.	SR	GC	PLWSR	PLWGC	SR	GC	PLWSR	PLWGC
Seq2Seq [34]	✓	4.00	9.40	2.00	6.30	0.40	7.00	0.10	4.30
MOCA [35]	✓	26.81	33.20	19.52	26.33	7.65	15.73	4.21	11.24
E.T. [28]	✓	38.42	45.44	27.78	34.93	8.57	18.56	4.21	11.46
ABP [14]	✓	44.55	51.13	3.88	4.92	15.43	24.76	1.08	2.22
FILM [25]	✓	27.67	38.51	11.23	15.06	26.49	36.37	10.55	14.30
LGS-RPA [26]	✓	40.05	48.66	21.28	28.97	35.41	45.24	15.68	22.76
Prompter [11]	✓	51.17	60.22	25.12	30.21	45.32	56.57	20.79	25.80
CAPEAM [15]	✓	51.79	60.50	21.60	25.88	46.11	57.33	19.45	24.06
DISCO [44]	✓	59.50	66.10	40.60	47.40	56.50	66.80	36.50	44.50
ThinkBot [23]	✓	62.69	71.64	32.02	37.01	57.82	67.75	26.93	30.73
ESCAPE (Ours)	✓	**65.09**	**72.88**	**52.42**	**58.29**	**60.79**	**68.60**	**46.82**	**52.57**
HLSM [2]	✗	25.11	35.79	6.69	11.53	16.29	27.24	4.34	8.45
FILM [25]	✗	25.77	36.15	10.39	14.17	24.46	34.75	9.67	13.13
LGS-RPA [26]	✗	33.01	41.71	16.65	24.49	27.80	38.55	12.92	20.01
Prompter [11]	✗	47.95	56.98	23.29	28.42	41.53	53.69	18.84	24.20
CAPEAM [15]	✗	47.36	54.38	19.03	23.78	43.69	54.66	17.64	22.76
DISCO [44]	✗	58.00	64.90	39.60	46.50	54.70	**65.50**	35.50	43.60
ESCAPE (Ours)	✗	**61.24**	**68.74**	**46.89**	**53.22**	**56.04**	64.79	**41.82**	**48.04**

4 Experiments

4.1 Dataset and Metrics

We conduct experiments on ALFRED [34], a benchmark dataset for vision-language navigation and long-horizon interaction tasks. The dataset comprises 25,726 episodes: 21,023 for training, 1,641 for validation, and 3,062 for testing. Both validation and test sets are divided into seen and unseen environments, with validation containing 820/821 episodes and test containing 1,533/1,529 episodes for seen/unseen scenes. We evaluate our method against competitive baselines using the test split, while using the validation split for analytical studies.

We employ four metrics: Success Rate (**SR**) measures task completion proportion; Goal Condition (**GC**) measures the percentage of met goal conditions for partial task success; Path Length Weighted Success Rate (**PLWSR**) and Goal Condition (**PLWGC**) incorporate efficiency by comparing paths to expert demonstrations. SR and GC assess task effectiveness, while PLW variants assess efficiency. Higher values indicate better performance.

4.2 Comparison with the State of the Art

We evaluate against competitive ALFRED baselines: Seq2Seq [34], MOCA [35], E.T. [28], HLSM [2], ABP [14], FILM [25], LGS-RPA [26], Prompter [11], CAPEAM [15], DISCO [44] and ThinkBot [23]. All methods use RGB inputs and language instructions during inference. For fair comparison, we present results

Table 2: Ablation study results on ALFRED benchmark [34]. w/ GT masks denotes using ground-truth semantic segmentation masks from simulator, w/ SQ denotes using static query for interaction mask generation, and w/ Grid 80/120 denotes scene grid resolutions of 80×80 and 120×120. w/o OME, w/o MRU, and w/o AEP represent the removal of Observation-to-Memory Encoding, Memory Retrieval and Update, and Adaptive Execution Policy, respectively. Standard deviations over 3 runs are provided in Appendix E.

	Valid Seen				**Valid Unseen**			
	SR	GC	PLWSR	PLWGC	SR	GC	PLWSR	PLWGC
ESCAPE	62.32	66.57	44.24	49.39	61.27	69.01	38.06	42.63
w/ GT masks	62.56	67.52	44.66	49.93	63.82	72.31	39.69	44.30
w/ SQ	60.98	65.15	42.38	47.04	58.83	66.18	36.18	40.37
w/ Grid 80	52.68	60.08	40.62	46.22	52.25	60.94	30.40	34.47
w/ Grid 120	53.54	59.60	38.98	44.65	52.86	62.40	33.94	39.18
w/o OME	53.54	62.30	38.48	48.03	52.74	65.47	35.70	40.51
w/o MRU	42.80	53.53	19.56	30.51	27.16	40.42	9.51	18.43
w/o AEP	57.59	62.74	38.44	43.63	57.13	63.39	34.57	38.07

based on whether step-by-step instructions are used. By default, ESCAPE uses high-level goals, but can incorporate step-by-step instructions when available.

We present comprehensive comparisons with state-of-the-art methods on ALFRED benchmark in Table 1. Our ESCAPE framework demonstrates superior performance across most evaluation metrics in both instruction settings. With step-by-step instructions, ESCAPE achieves 65.09% and 60.79% success rates on seen and unseen test splits respectively, surpassing the previous best method ThinkBot [23] by 2.40% and 2.97%. Without instructions, ESCAPE maintains strong performance with 61.24% and 56.04% success rates, representing improvements of 3.24% and 1.34% over DISCO [44].

Notably, ESCAPE exhibits substantial improvements in path-length-weighted metrics, achieving 52.42% PLWSR and 58.29% PLWGC on seen environments, and 46.82% PLWSR and 52.57% PLWGC on unseen environments. These significant gains validate the effectiveness of our Adaptive Execution Policy in seizing opportunistic interactions to enhance long-horizon efficiency. The consistent superior performance across almost all metrics and instruction settings demonstrates ESCAPE's robust generalization capability, indicating strong potential for real-world applications where detailed instructions may not be available.

4.3 Ablation Study

We conduct ablation studies on ALFRED validation splits to evaluate each component in ESCAPE (Table 2). Results are averaged over 3 runs with different random seeds. Detailed experimental configurations are in Appendix C.

Spatio-Temporal Fusion Mapping. We first examine Observation-to-Memory Encoding (OME) and Memory Retrieval and Update (MRU). Removing OME causes moderate drops (8.78% SR on seen, 8.53% on unseen), confirming its role

Table 3: Additional quantitative metrics. mIOU scores for map and image semantic segmentation are presented, while efficiency factors (EF) are compared between ESCAPE and w/o Adaptive Execution Policy (AEP).

mIOU	Seen	Unseen	**EF**	Seen SR	Seen GC	Unseen SR	Unseen GC
Map Seg.	0.758	0.654	ESCAPE	0.710	0.742	0.621	0.618
Image Seg.	0.869	0.779	w/o AEP	0.668	0.695	0.605	0.601

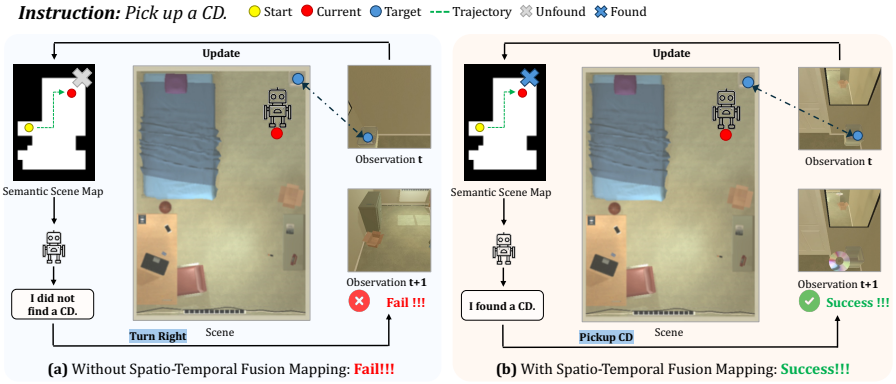

Fig. 5: Qualitative analysis of our Spatio-Temporal Fusion Mapping module. **Left:** Failure due to missing spatial memory updates. **Right:** Success with the persistent 3D spatial memory.

in fusing visual observations into 3D memory. Without MRU, severe drops occur (19.52% SR on seen, 34.11% on unseen), proving temporal memory retrieval and update are critical for scene understanding given the agent's limited camera view. As shown in Table 3, our map semantic segmentation achieves 0.758 mIOU on seen and 0.654 on unseen environments, indicating accurate spatial localization and robustness across novel scenarios. We also test grid resolutions of 80×80, 100×100, and 120×120. 100×100 performs best, while 80×80 and 120×120 cause significant drops (9.64%/8.78% SR on seen, 9.02%/8.41% on unseen). This confirms 100×100 best matches ALFRED's 25m×25m rooms with 25cm steps, where each grid cell equals one agent step. Both OME and MRU are essential, with MRU being more critical to avoid catastrophic forgetting.

Memory-Driven Target Grounding. Ablation results show the effectiveness of our Memory-Driven Target Grounding module in generating precise interaction masks. First, image semantic segmentation achieves 0.869 mIOU on seen and 0.779 on unseen environments (Table 3), showing accurate object localization. Second, comparing with GT masks shows minimal gaps (0.24% SR on seen, 2.55% on unseen), confirming our masks are not the bottleneck for task success. Moreover, using scene-specific 3D object features beats static queries (1.34% SR on seen, 2.44% on unseen), indicating that dynamically derived memory queries capture object specificity better. Furthermore, failure analysis in Appendix D

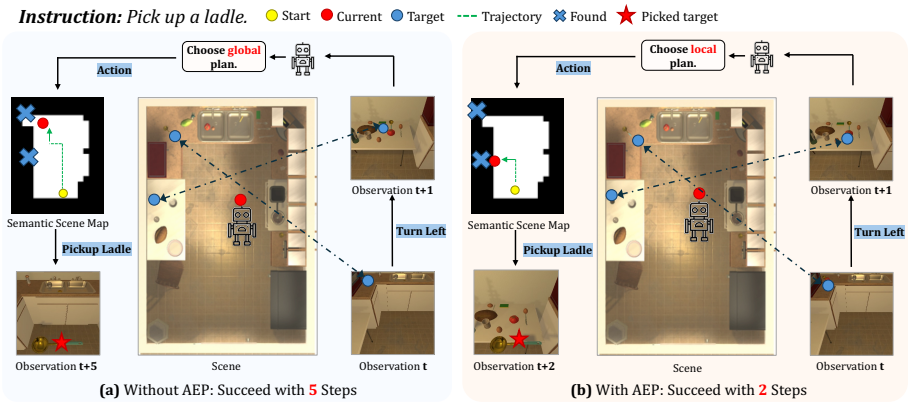

Fig. 6: Qualitative comparison of the Adaptive Execution Policy (AEP) effectiveness. **Left:** Pickup task completed in 5 steps without AEP. **Right:** The same task accomplished in 2 steps with our proposed AEP.

also shows our method reduces interaction failures by 12.6% on seen and 24.3% on unseen tasks versus DISCO [44].

Adaptive Execution Policy. Our Adaptive Execution Policy (AEP) significantly improves both task success and efficiency. Removing it causes drops: SR falls 4.73% on seen and 4.14% on unseen settings. To measure efficiency, we define the **Efficiency Factor (EF)**: EF = Expert Length/Agent Length = PLWSR/SR. Higher EF indicates more efficient task completion with shorter paths compared to expert trajectories. Table 3 shows consistent EF gains across all metrics when employing AEP. These results demonstrate that dynamically shifting between global proactive planning and local opportunistic interaction yields highly efficient long-horizon execution.

4.4 Qualitative Results

The visualization result in Fig. 5 illustrates the efficacy of our Spatio-Temporal Fusion Mapping module. In scenario (a), prior approaches fail to detect the CD within the bin, demonstrating their limitations with contained objects. In contrast, scenario (b) demonstrates how our memory-centric agent successfully localizes and retrieves the CD from the bin, highlighting the module's capability in handling complex spatial relationships and containment scenarios.

Fig. 6 shows the effectiveness of our Adaptive Execution Policy. Without AEP, agents strictly follow fixed long-term plans, blindly bypassing closer optimization opportunities (scenario a). In contrast, our reactive local monitor continuously screens for immediate manipulation targets and preemptively interrupts long-term routes when viable interactions are detected, greatly improving ESCAPE's efficiency (**5 steps versus 2 steps**). This behavioral improvement perfectly corroborates the substantial gains in PLW metrics shown in Table 1.

5 Conclusion

We present ESCAPE, a memory-centric framework for long-horizon mobile manipulation that addresses the fundamental challenges of long-horizon efficiency and dynamic execution. By coupling a persistent Episodic Spatial Memory with an Adaptive Execution Policy, ESCAPE mitigates catastrophic forgetting, spatial inconsistency, and rigid execution constraints. Extensive experiments demonstrate that ESCAPE achieves state-of-the-art performance in both seen and unseen environments on ALFRED benchmark, with significant gains in path-weighted efficiency metrics. Our ablation studies and thorough visualizations validate the contribution of each module.

Limitation and Future Directions. ESCAPE is currently trained with simulator derived supervision, including semantic maps, image masks, and manipulability labels. In future work, this dependency could be reduced by using masks generated by foundation models, offline pseudo labels, or self-training from real robot data. Further evaluation on real-world robots would also be valuable for studying sim-to-real transfer and improving deployment robustness.

Acknowledgement. This work is supported by the Shenzhen Science and Technology Program (Grant No. ZDCY20250901113000001), the Guangdong Provincial Key Laboratory of Mathematical Foundations for Artificial Intelligence (Grant No. 2023B1212010001), and the Shenzhen Loop Area Institute under Grant No. FPF10120250006.

References

1. Batra, D., Gokaslan, A., Kembhavi, A., Maksymets, O., Mottaghi, R., Savva, M., Toshev, A., Wijmans, E.: Objectnav revisited: On evaluation of embodied agents navigating to objects. arXiv:2006.13171 (2020) 4
2. Blukis, V., Paxton, C., Fox, D., Garg, A., Artzi, Y.: A persistent spatial semantic representation for high-level natural language instruction execution. In: Conference on Robot Learning. pp. 706–717. PMLR (2022) 11
3. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al.: Rt-1: Robotics transformer for real-world control at scale. arXiv:2212.06817 (2022) 1
4. Can, Y.B., Liniger, A., Paudel, D.P., Van Gool, L.: Structured bird's-eye-view traffic scene understanding from onboard images. In: ICCV. pp. 15661–15670 (2021) 4
5. Chen, G., Wang, W.: A survey on 3d gaussian splatting. arXiv:2401.03890 (2024) 4
6. Cheng, Z., Li, R., Hu, J., Tu, Y., Dai, S., Hu, S., Shi, Y., Shi, L., Sun, M.: Embodiedeval: Evaluate multimodal llms as embodied agents. In: CVPR. pp. 11420–11432 (2026) 4
7. Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., Song, S.: Diffusion policy: Visuomotor policy learning via action diffusion. The International Journal of Robotics Research **44**(10-11), 1684–1704 (2025) 1
8. Ehsani, K., Gupta, T., Hendrix, R., Salvador, J., Weihs, L., Zeng, K.H., Singh, K.P., Kim, Y., Han, W., Herrasti, A., et al.: Spoc: Imitating shortest paths in

simulation enables effective navigation and manipulation in the real world. In: CVPR. pp. 16238–16250 (2024) 4
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 6, 8, 10
10. Huang, J., Huang, G., Zhu, Z., Ye, Y., Du, D.: Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv:2112.11790 (2021) 4, 5
11. Inoue, Y., Ohashi, H.: Prompter: Utilizing large language model prompting for a data efficient embodied instruction following. arXiv:2211.03267 (2022) 3, 11
12. Jaafar, A., Raman, S.S., Wei, Y., Juliani, S.E., Wernerfelt, A., Idrees, I., Liu, J.X., Tellex, S.: LaNMP: A multifaceted mobile manipulation benchmark for robots. In: RSS 2024 Workshop: Data Generation for Robotics (2024) 4
13. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023) 4
14. Kim, B., Bhambri, S., Singh, K.P., Mottaghi, R., Choi, J.: Agent with the big picture: Perceiving surroundings for interactive instruction following. In: CVPR Embodied AI Workshop (2021) 11
15. Kim, B., Kim, J., Kim, Y., Min, C., Choi, J.: Context-aware planning and environment-aware memory for instruction following embodied agents. In: ICCV. pp. 10936–10946 (2023) 2, 3, 10, 11
16. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E.P., Sanketi, P.R., Vuong, Q., et al.: Openvla: An open-source vision-language-action model. In: Conference on Robot Learning. pp. 2679–2713. PMLR (2025) 1
17. Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Deitke, M., Ehsani, K., Gordon, D., Zhu, Y., et al.: Ai2-thor: An interactive 3d environment for visual ai. arXiv:1712.05474 (2017) 4
18. Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J., Quillen, D.: Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. The International journal of robotics research **37**(4-5), 421–436 (2018) 1
19. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., Dai, J.: Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(3), 2020–2036 (2024) 4, 5, 6
20. Liang, T., Xie, H., Yu, K., Xia, Z., Lin, Z., Wang, Y., Tang, T., Wang, B., Tang, Z.: Bevfusion: A simple and robust lidar-camera fusion framework. NeurIPS **35**, 10421–10434 (2022) 4
21. Liu, F., Zhang, C., Zheng, Y., Duan, Y.: Semantic ray: Learning a generalizable semantic field with cross-reprojection attention. In: CVPR. pp. 17386–17396 (2023) 4
22. Liu, Y., Wang, T., Zhang, X., Sun, J.: Petr: Position embedding transformation for multi-view 3d object detection. In: ECCV. pp. 531–548. Springer (2022) 4
23. Lu, G., Wang, Z., Liu, C., Lu, J., Tang, Y.: Thinkbot: Embodied instruction following with thought chain reasoning. In: ICLR (2025) 2, 6, 11, 12
24. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021) 4
25. Min, S.Y., Chaplot, D.S., Ravikumar, P., Bisk, Y., Salakhutdinov, R.: Film: Following instructions in language with modular methods. In: ICLR (2022) 1, 2, 3, 4, 6, 10, 11

26. Murray, M., Cakmak, M.: Following natural language instructions for household tasks with landmark guided search and reinforced pose adjustment. IEEE Robotics and Automation Letters **7**(3), 6870–6877 (2022) 11
27. Pan, B., Sun, J., Leung, H.Y.T., Andonian, A., Zhou, B.: Cross-view semantic segmentation for sensing surroundings. IEEE Robotics and Automation Letters **5**(3), 4867–4873 (2020) 4
28. Pashevich, A., Schmid, C., Sun, C.: Episodic transformer for vision-and-language navigation. In: ICCV. pp. 15942–15952 (2021) 1, 2, 11
29. Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T., et al.: Openscene: 3d scene understanding with open vocabularies. In: CVPR. pp. 815–824 (2023) 8
30. Philion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In: ECCV. pp. 194–210. Springer (2020) 4
31. Prabhudesai, M., Tung, H.Y.F., Javed, S.A., Sieb, M., Harley, A.W., Fragkiadaki, K.: Embodied language grounding with 3d visual feature representations. In: CVPR. pp. 2220–2229 (2020) 1
32. Qian, J., Han, B., Shi, C., Xiao, L., Yang, L., Shi, S., Jiang, L.: Geopredict: Leveraging predictive kinematics and 3d gaussian geometry for precise vla manipulation. In: CVPR. pp. 13529–13539 (2026) 4
33. Schult, J., Engelmann, F., Hermans, A., Litany, O., Tang, S., Leibe, B.: Mask3d: Mask transformer for 3d semantic instance segmentation. In: ICRA. pp. 8216–8223. IEEE (2023) 8
34. 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: CVPR. pp. 10740–10749 (2020) 1, 2, 3, 4, 5, 8, 10, 11, 12
35. Singh, K.P., Bhambri, S., Kim, B., Mottaghi, R., Choi, J.: Factorizing perception and policy for interactive instruction following. In: ICCV. pp. 1888–1897 (2021) 2, 11
36. 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: ICCV. pp. 2998–3009 (2023) 1
37. Srivastava, S., Li, C., Lingelbach, M., Martín-Martín, R., Xia, F., Vainio, K.E., Lian, Z., Gokmen, C., Buch, S., Liu, K., et al.: Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In: Conference on robot learning. pp. 477–490. PMLR (2022) 4
38. Stepputtis, S., Campbell, J., Phielipp, M., Lee, S., Baral, C., Ben Amor, H.: Language-conditioned imitation learning for robot manipulation tasks. NeurIPS **33**, 13139–13150 (2020) 1
39. Szot, A., Clegg, A., Undersander, E., Wijmans, E., Zhao, Y., Turner, J., Maestre, N., Mukadam, M., Chaplot, D., Maksymets, O., Gokaslan, A., Vondrus, V., Dharur, S., Meier, F., Galuba, W., Chang, A., Kira, Z., Koltun, V., Malik, J., Savva, M., Batra, D.: Habitat 2.0: Training home assistants to rearrange their habitat. In: NeurIPS (2021) 4
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017) 6
41. Wang, Y., Guizilini, V.C., Zhang, T., Wang, Y., Zhao, H., Solomon, J.: Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In: Conference on robot learning. pp. 180–191. PMLR (2022) 4
42. Wu, Z., Wang, Z., Xu, X., Yin, H., Liang, Y., Ma, A., Lu, J., Yan, H.: Embodied instruction following in unknown environments. In: IROS. pp. 21825–21832. IEEE (2025) 4

43. Xiao, L., Li, J., Gao, J., Ye, F., Jin, Y., Qian, J., Zhang, J., Wu, Y., Yu, X.: Ava-vla: Improving vision-language-action models with active visual attention. In: CVPR. pp. 13453–13463 (2026) 1
44. Xu, X., Luo, S., Yang, Y., Li, Y.L., Lu, C.: Disco: Embodied navigation and interaction via differentiable scene semantics and dual-level control. In: ECCV. pp. 108–125. Springer (2024) 2, 4, 6, 8, 10, 11, 12, 14
45. Yang, R., Chen, H., Zhang, J., Zhao, M., Qian, C., Wang, K., Wang, Q., Koripella, T.V., Movahedi, M., Li, M., et al.: Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. arXiv:2502.09560 (2025) 4
46. Yenamandra, S., Ramachandran, A., Yadav, K., Wang, A., Khanna, M., Gervet, T., Yang, T.Y., Jain, V., Clegg, A.W., Turner, J., Kira, Z., Savva, M., Chang, A., Chaplot, D.S., Batra, D., Mottaghi, R., Bisk, Y., Paxton, C.: Homerobot: Open vocab mobile manipulation. In: Conference on Robot Learning (2023) 1, 2, 4
47. Yu, Z., Chen, A., Huang, B., Sattler, T., Geiger, A.: Mip-splatting: Alias-free 3d gaussian splatting. In: CVPR. pp. 19447–19456 (2024) 4
48. Zhang, S., Song, X., Yu, X., Bai, Y., Guo, X., Li, W., Jiang, S.: Hoz++: Versatile hierarchical object-to-zone graph for object navigation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025) 4
49. Zhao, Q., Fu, H., Sun, C., Konidaris, G.: Epo: Hierarchical llm agents with environment preference optimization. EMNLP (2024) 4
50. Zhu, F., Liang, X., Zhu, Y., Yu, Q., Chang, X., Liang, X.: Soon: Scenario oriented object navigation with graph-based exploration. In: CVPR. pp. 12689–12699 (2021) 4
51. Zhu, J., Huo, Y., Ye, Q., Luan, F., Li, J., Xi, D., Wang, L., Tang, R., Hua, W., Bao, H., et al.: I2-sdf: Intrinsic indoor scene reconstruction and editing via raytracing in neural sdfs. In: CVPR. pp. 12489–12498 (2023) 4
52. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: ICLR (2021) 6