

HitMem: Hierarchical Temporal 3D Memory with Multi-Modal Context-Aware Retrieval for Dynamic Environments

Ruijie Tang^{1,2}, Chenye Zou³, Guoquan Wu^{1,2(✉)},
Jun Wei^{1,2}, Wei Chen^{1,2}, and Jiaxin Zhu^{1,2}

¹ Institute of Software, Chinese Academy of Sciences (ISCAS), Beijing, China

² University of Chinese Academy of Sciences, Beijing, China

³ Alibaba Group, Shanghai, China

Abstract. Executing long-term tasks in dynamic environments requires embodied agents to maintain robust and adaptive 3D scene representations. However, most existing 3D memory frameworks rely on static world assumptions. When objects are displaced by human activities or unobserved events, agents encounter memory-observation conflicts and often require costly geometric recomputations or inefficient global re-exploration. To address this, we propose HitMem, a hierarchical temporal 3D memory framework with a multi-modal context-aware retrieval mechanism. Through continuous perception, HitMem unifies semantic and spatial information into a lightweight topological graph that captures support relationships, while a temporal decay mechanism dynamically regulates memory activeness to mitigate the impact of stale representations. In addition, the multi-modal context-aware retrieval mechanism defaults to filtering candidates using integrated semantic, spatial, and temporal memory features, and activates a specialized two-stage retrieval process when object displacement is detected. This process combines spatial constraints inferred from external agent trajectories with semantic common sense grounded in class affinities, efficiently identifying high-probability candidate regions. Extensive evaluations on our constructed Dyna-THOR benchmark demonstrate that HitMem significantly improves object relocation accuracy, reduces exploration costs, and enhances task execution performance in dynamic environments.

Keywords: Dynamic 3D Memory · Context-Aware Retrieval · Embodied AI

1 Introduction

Performing complex and long-term tasks within real-world environments is a fundamental objective of modern embodied agents. Such agents must operate continuously, respond to user instructions, autonomously identify targets, and execute corresponding actions [29, 34, 37]. A cornerstone of achieving these objectives is the integration of memory, which enables the agent to store and effec-

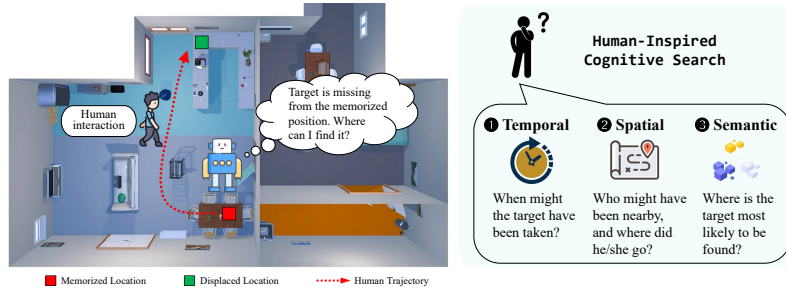


Fig. 1: Design intention of HitMem. Existing methods are constrained by static environment assumptions or resort to blind exploration when targets are displaced. In contrast, we construct a hierarchical temporal memory and leverage multi-modal environmental information for retrieval, effectively enhancing adaptability to dynamic environments and improving exploration efficiency.

tively retrieve environmental information to support task execution. Many recent methods have successfully transformed sensory inputs into 3D memories via dense reconstruction or scene graph modeling, equipping embodied agents with the requisite spatial awareness for navigation and manipulation [7,9,16,21,22,35].

However, existing methods are fundamentally limited by their reliance on a static world assumption. Real-world environments are intrinsically dynamic, where objects are frequently displaced by human activities or unobserved events. This discrepancy causes memory conflicts when an agent navigates to a location and finds the target missing. Current memory frameworks offer limited support for dynamic updates. Specifically, the re-inference of relational structures or the recomputation of geometric clusters is both computationally inefficient and prone to errors [8,31,35]. Moreover, the absence of temporal dynamics modeling leaves agents highly vulnerable to obsolete information [3,6,19,28]. These limitations restrict the effective resolution of memory discrepancies, which fundamentally impairs the agent’s adaptability in dynamic environments.

Furthermore, existing methods relying on exhaustive global search or single-modality information are fundamentally inadequate for object relocation in dynamic environments [17,23,32,36]. Inspired by human cognitive search behaviors, as shown in Fig. 1, individuals locate misplaced objects by leveraging multi-modal contextual cues rather than resorting to random or exhaustive environmental scanning. This process involves synthesizing physical spatial dynamics, such as the influence of recent nearby human interactions on object placement, together with semantic relevance. For example, it recognizes that a missing bowl from a dining table is far more likely to be relocated to a kitchen sink or refrigerator rather than to a semantically unrelated bathroom. Therefore, integrating multi-modal contextual knowledge that encompasses both semantic common-sense and spatio-temporal dynamics into the memory retrieval loop is essential for achieving efficient dynamic object relocation.

To address these challenges, we introduce HitMem, a hierarchical temporal 3D memory framework with multi-modal context-aware retrieval for dynamic environments. HitMem unifies semantic, spatial, and temporal information within a multi-level representation. At the high level, it constructs a lightweight semantic topological graph that encodes only categorical carrier-carried relationships. This bypasses the computationally expensive instance-level correspondence inference and geometric clustering mandated by conventional methods. At the low level, HitMem preserves the detailed 3D attributes of objects, ensuring precise spatial matching and supporting downstream 3D manipulation tasks. To capture the inherently dynamic nature of the environment, we introduce a temporal decay mechanism. It dynamically modulates the activeness scores of memory nodes over time, proactively penalizing outdated observations and preventing the agent from relying on obsolete information during task execution.

Crucially, HitMem incorporates a multi-modal context-aware retrieval mechanism. By sequentially leveraging the semantic and spatial information of the hierarchical memory graph with temporal activeness scores, the system can efficiently filter out invalid targets. When a target is determined to be displaced, we activate a two-stage retrieval strategy that tightly integrates multi-modal environmental context. First, we extract spatial constraints from the historical trajectories of external agents, significantly pruning the search space. Second, we conduct a class affinity analysis that evaluates the probabilistic correlation between the target and candidate carriers. By combining physical trajectory cues with semantic common sense, HitMem avoids blind global exploration and achieves highly targeted, context-driven spatial reasoning.

Considering that existing benchmarks fail to adequately capture the dynamic challenges of real-world environments, we construct a novel benchmark, Dyna-Thor, built upon the AI2-THOR [14] simulator. Dyna-Thor introduces external agents that actively manipulate and relocate task-relevant objects during execution, simulating the unexpected interactions caused by humans or other agents in real-world settings. Extensive experiments on Dyna-THOR demonstrate that HitMem significantly improves long-term task execution, showing strong adaptability to dynamic disturbances while substantially reducing exploration costs compared to existing baselines.

In summary, our main contributions are as follows:

- We propose HitMem, a hierarchical temporal 3D memory that structurally unifies semantic, spatial, and fine-grained geometric information. A temporal decay mechanism dynamically regulates memory activeness, enabling robust adaptation to environmental changes and filtering of obsolete observations.
- We design a multi-modal context-aware retrieval mechanism. By integrating physical spatial cues (external agent trajectories) with semantic common sense (class affinity analysis) in a two-stage strategy, the proposed method efficiently and accurately relocates displaced objects.
- We introduce Dyna-Thor, a novel benchmark that incorporates interactions between external agents and objects to simulate unpredictable environmental

dynamics. Extensive experiments show that HitMem improves relocation accuracy, exploration efficiency, and adaptability in dynamic environments.

2 Related Work

3D Scene Representations. 3D scene representation is crucial for the task execution of embodied agents. Early dense 3D reconstruction approaches [9, 42] capture detailed geometric structures but lack high-level semantics, limiting their effectiveness in tasks involving semantic queries. While integrating Vision-Language Models (VLMs) [15, 25] embeds 2D semantic features into 3D structures for language-guided object localization and manipulation [7, 17, 22, 39], these methods remain computationally expensive and incapable of real-time updates in dynamic environments. 3D Scene Graph-based methods further simplify scene representation [16, 19, 32, 35], offering stronger semantic interpretability through nodes and edges. Accurate scene characterization requires precise edge correspondences, typically achieved through geometric clustering [8, 21] or inference via Large Language Models (LLMs) [11, 31, 35]. However, these methodologies are structurally unstable in dynamic settings, where object displacements necessitate error-prone and computationally prohibitive recomputations of clusters and edges. Moreover, lacking temporal dynamics modeling leaves agents reliant on obsolete information, degrading task accuracy and efficiency. In this work, we streamline the 3D scene representation by focusing on categorical carrier-carried relationships. This topological choice bypasses complex edge inference to enable continuous memory updates while preserving fine-grained 3D attributes. Coupled with a temporal decay mechanism, our approach substantially improves agent adaptability and execution efficiency in dynamic environments.

Memory for Embodied Tasks. Embodied agent memory takes various forms. Short-term memory approaches typically employ VLMs for scene analysis and action prediction based on instantaneous observations [2, 38, 43–45]. The absence of a stable and persistent knowledge representation makes it difficult to ensure reliable long-term task execution. Some methods store historical images as long-term memory for embodied reasoning and exploration [4, 20, 33], but they often suffer from low efficiency and poor adaptability to dynamic environments. A common approach is to leverage 3D scene representations by directly converting extracted semantic and spatial information into agent memory [3, 6, 10, 40]. When needed, the memory can be retrieved to support LLM-based task understanding and planning. Such methods significantly enhance environmental comprehension but are strictly limited to static applications. To address dynamics, recent studies incorporate updates [17, 27, 32] or temporal awareness [23, 36]. However, for object relocation, these methods still resort to global matching or scoring. By relying on exhaustive search and failing to fully exploit environmental context, these approaches suffer from severe inefficiencies in memory retrieval. In contrast, we propose a multi-modal context-aware retrieval mechanism leveraging

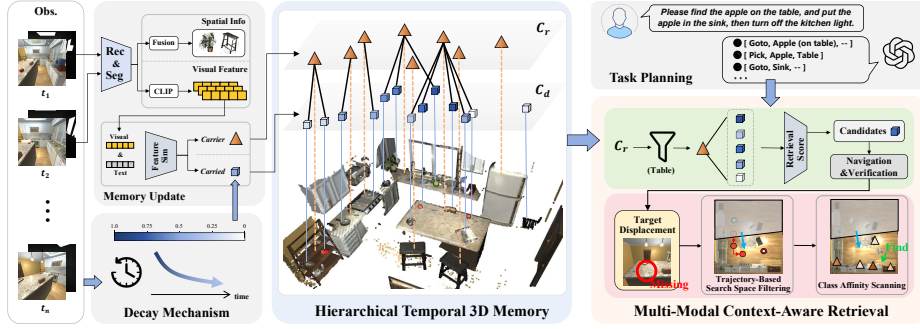


Fig. 2: Overview of HitMem. HitMem is a hierarchical 3D memory framework that integrates spatial-semantic information with a decay mechanism to model temporal dynamics. It employs multi-modal context-aware retrieval for efficient target localization. When objects are displaced, it leverages external agent trajectories with class affinity to rapidly identify candidate regions, enhancing robustness in dynamic environments.

our hierarchical memory structure with temporal dynamics for rapid target localization. To relocate displaced objects, we design a two-stage retrieval strategy that integrates external agent trajectories and semantic class-affinity scanning, significantly reducing the search space and improving retrieval efficiency.

3 Method

Our objective is to enable embodied agents to execute instructions via continuous RGB-D observations while robustly handling unexpected object displacements in dynamic environments. Fig. 2 illustrates an overview of the proposed HitMem framework. We construct a hierarchical 3D representation that fuses spatial and semantic information, integrated with a decay mechanism to model temporal dynamics. To efficiently locate targets and mitigate memory inconsistencies, HitMem employs a multi-modal context-aware retrieval mechanism. When a target is displaced, it activates a two-stage strategy that synthesizes spatial cues (external agent trajectories) and semantic common sense (class affinities) to rapidly pinpoint candidate regions, ensuring adaptability and robustness.

3.1 Hierarchical Temporal 3D Memory

Topological and Semantic Representations. By continuously perceiving the environment, our approach constructs and dynamically updates an object-centered hierarchical memory graph, denoted as $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$.

Node Instantiation ($\mathcal{V}_t$): Given a continuous stream of observations $\mathcal{I} = \{I_1, I_2, \dots\}$ from the ego-agent, where each I_n comprises an RGB-D image and the corresponding camera pose, we leverage an open-vocabulary recognition and

detection method (e.g., RAM-Grounded-SAM2 [13, 18, 41]) to obtain instance-level bounding boxes and segmentation masks. These extractions form the vertex set $\mathcal{V}_t$, with each node $v_i \in \mathcal{V}_t$ encapsulating both semantic and spatial attributes. Specifically, we derive visual embeddings F_i by processing cropped images and masks through CLIP [25]. Concurrently, we associate each node with precise 3D geometric attributes, including the reconstructed point cloud pcd_i , global coordinates pos_i , and physical dimensions.

Carrier-Carried Edges ($\mathcal{E}_t$): To constrain the search space for downstream tasks, we organize nodes into a hierarchical structure by classifying them into carriers $\mathcal{C}_r$ (e.g., tables, shelves) and carried objects $\mathcal{C}_d$ (e.g., cups, apples). A node v_i is assigned to $\mathcal{C}_r$ if the cosine similarity between its visual feature F_i and the textual embedding of “furniture or appliances for holding objects” exceeds a semantic threshold θ_{sem} . For remaining non-carrier nodes v_j , we evaluate their spatial relationship with all carriers $v_i \in \mathcal{C}_r$ via a comprehensive support score:

$$S_{support}(v_i, v_j) = \omega_1 R_p + \omega_2 R_c + \omega_3 R_v + \omega_4 S_h \quad (1)$$

where $\sum_{k=1}^4 \omega_k = 1$. This aggregate score encapsulates four geometric metrics derived from the instance-level point clouds: spatial proximity (R_p), horizontal coverage (R_c), vertical support (R_v), and height consistency (S_h). A directed edge $E(v_i, v_j) \in \mathcal{E}_t$ is established, and v_j is formally assigned to the carried set $\mathcal{C}_d$, only if the maximum support score exceeds a predefined threshold θ_{sup} :

$$v_j \rightarrow \mathcal{C}_d \cup \{E(\hat{v}_i, v_j) \rightarrow \mathcal{E}_t\}, \text{ if } \max_{v_i \in \mathcal{C}_r} S_{support}(v_i, v_j) > \theta_{sup} \quad (2)$$

If v_j fails to satisfy the support condition for any existing carrier, it is temporarily categorized as an unassigned object ($v_j \in \mathcal{C}_u$). These nodes are retained in memory and dynamically re-evaluated for support relationships whenever new carrier nodes are incrementally discovered during subsequent exploration. By abstracting the environment into this dynamic $\mathcal{C}_r \rightarrow \mathcal{C}_d$ topological structure, updates can be executed efficiently via localized node insertion or deletion, circumventing the computational burden of global graph reconstruction.

Temporal Decay Mechanism. A fundamental challenge in dynamic environments is memory obsolescence, where over-reliance on stale information inevitably leads to memory-observation conflicts and task failures. To explicitly quantify temporal validity, we introduce a memory decay mechanism that assigns a dynamic activeness score $\mathcal{A}(v_i, t) \in (0, 1]$ to each node $v_i \in \mathcal{V}_t$. This score decays exponentially with the time elapsed since the node’s last observation:

$$\mathcal{A}(v_i, t) = \exp(-\lambda(v_i) \cdot (t - T_{last})) \quad (3)$$

Crucially, instead of using a uniform static constant, we define the decay rate $\lambda(v_i)$ as an intrinsic property of the object’s structural role, modulated by the environment’s spatial scale. Since carrier objects ($\mathcal{C}_r$) are relatively stationary, their decay rate is set to zero to ensure stable structural memory. Conversely,

carried objects ($\mathcal{C}_d$) are assigned a baseline rate λ_{base} , which is scaled by a scene factor $\eta(S_{scene})$ that grows with the mapped area or room count. Specifically, $\lambda(v_i) = \frac{\lambda_{base}}{\eta(S_{scene})}$. This inverse scaling ensures slower decay in larger environments, preventing premature forgetting during long-term tasks. As a result, a high activeness score indicates reliable memory, while a lower score reduces retrieval priority, encouraging the agent to discount outdated observations.

In real-world scenarios, environments are often perturbed by external agents, such as humans or other autonomous entities, whose interactions alter spatial configurations and introduce unpredictable dynamics. Since these agents provide no explicit notification of displacement, the ego-agent must infer changes solely through observation. Our decay mechanism efficiently supports the tracking of these dynamics. Entities identified as external agents (labeled “human/robot”) by the perception module are independently recorded as a node set $\mathcal{V}^{agent}$. By ordering these nodes chronologically according to their activeness scores, we reconstruct the external agent’s motion to generate a coherent trajectory:

$$\mathcal{T} = (v_{k_1}^{agent}, v_{k_2}^{agent}, \dots, v_{k_m}^{agent}), \quad \text{s.t.} \quad \mathcal{A}(v_{k_j}^{agent}) < \mathcal{A}(v_{k_{j+1}}^{agent}) \quad (4)$$

Visual instance tracking can be integrated to assign identities, allowing the system to partition external agent nodes into subsets and construct independent trajectories for multiple external agents. These trajectories serve as critical physical contexts, guiding multi-modal memory retrieval and target relocation.

Continuous Memory Fusion. To ensure a consistent and deduplicated scene representation, incoming observations must be tightly fused with existing nodes. When integrating new observations, conventional Iterative Closest Point (ICP) matching is prone to segmentation inaccuracies and localization noise, which often absorb small objects into larger ones, leading to the loss of memory entries. To address this, we introduce a Bidirectional ICP matching process requiring both forward and backward alignment scores to pass thresholds ($S_{ICP}^{fwd} \geq \sigma_{fwd} \wedge S_{ICP}^{bwd} \leq \sigma_{bwd}$). Following a successful match, the new point cloud is fused into the stored node v_i , its temporal state is refreshed ($\mathcal{A}(v_i, t) = 1$), and its visual feature is updated via a weighted average based on the observation count.

3.2 Multi-Modal Context-aware Retrieval

Efficient memory retrieval is fundamental for executing high-level task plans in dynamic environments. To address the challenge of target object displacement, we introduce a closed-loop, multi-modal context-aware retrieval method, as shown in Fig. 3. The system initially performs a hierarchical semantic query within the memory graph $\mathcal{G}_t$, integrating temporal information to filter candidate targets before conducting active visual verification. Upon detecting a memory-observation conflict signaling object displacement, the system proactively prunes the obsolete node. It then activates a two-stage relocation strategy, synthesizing external agent trajectory and class affinity to perform active reasoning.

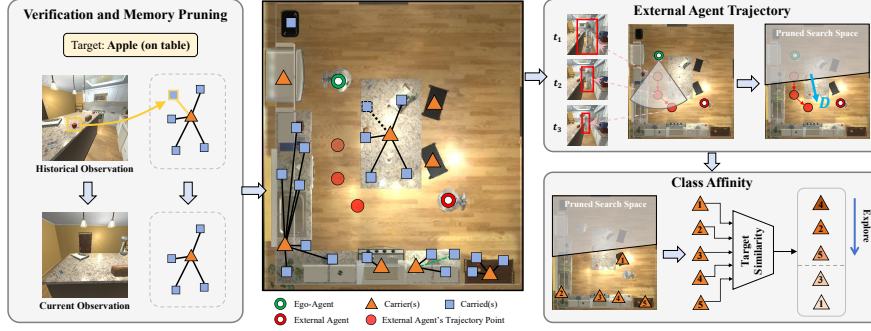


Fig. 3: Retrieval strategy of HitMem. Upon navigating to the retrieved target object location, the ego-agent actively verifies its visual observation. If the target has been displaced, the corresponding obsolete memory node is removed, and a two-stage retrieval process is then initiated. First, the local trajectories of external agents are analyzed to infer potential displacement directions and narrow the search space. Second, class affinity scores between the target and candidate carriers are calculated to prioritize subsequent exploration.

Hierarchical Semantic Querying and Active Verification. Leveraging the natural language understanding capabilities of LLMs, user instructions are parsed into structured semantic queries (e.g., $\langle \text{Pick, Apple, Table} \rangle$). To minimize computational overhead, HitMem capitalizes on the hierarchical structure of the memory graph $\mathcal{G}_t$. The system initially filters relevant carrier nodes within $\mathcal{C}_r$. Subsequently, it restricts the fine-grained search to the carried objects $v_i \in \mathcal{C}_d$ that are topologically linked to the filtered carriers. In scenarios where the instruction lacks a specific carrier context, the hierarchical topology still accelerates the process by enabling a top-down structured traversal.

We define a comprehensive retrieval score for each candidate carried object: $S_{\text{retrieval}}(v_i) = \alpha_1 \cdot \text{sim}(F_i, F_{\text{query}}) + \alpha_2 \cdot \mathcal{A}(v_i, t)$. Here, F_{query} denotes the textual embedding of the queried target object, and $\text{sim}(\cdot, \cdot)$ computes the cosine similarity between the visual and textual embeddings. Both the similarity metric and the activeness score are normalized to $[0, 1]$ using min-max scaling. In our implementation, α_1 and α_2 are set to 1. Candidates satisfying the requirement $S_{\text{retrieval}}(v_i) > \theta_{\text{target}}$ are ranked in descending order. Up to three top-ranked nodes are then selected as primary exploration targets, effectively prioritizing objects that possess both semantic relevance and temporal recency.

Upon navigating to the expected location of the top-ranked candidate, the ego-agent verifies whether the current visual observation matches the anticipated target memory. If the target is successfully detected, the retrieval concludes and the downstream manipulation task proceeds. Conversely, if the target is visually absent, HitMem identifies a memory-observation conflict indicating displacement and executes active memory pruning to keep the scene representation consistent. The obsolete node v_i and its associated carrier-carried edge are deleted from the graph ($\mathcal{V}_t \leftarrow \mathcal{V}_t \setminus \{v_i\}$). This proactive mechanism promptly purges verified stale

information from memory, ensuring sustained alignment between the ego-agent’s internal representation and the evolving physical environment.

Two-Stage Retrieval for Relocated Objects. When the target object is displaced, the ego-agent initiates a two-stage retrieval strategy. This strategy fully leverages multi-modal environmental context to efficiently re-localize the displaced target, transitioning from blind exploration to guided spatial reasoning.

Stage 1: Trajectory-Based Search Space Filtering. The initial step leverages the historical trajectory $\mathcal{T}$ of external agents to infer the displacement direction of the target. We first identify the interaction point $\mathcal{P}_{closest} \in \mathcal{T}$, defined as the trajectory node spatially closest to the object’s last known position. A local trajectory segment, denoted as τ_{local} , is then extracted, consisting of a subsequence of k consecutive points following $\mathcal{P}_{closest}$. To capture the stable motion trend rather than local fluctuations, we estimate the displacement vector $\mathbf{D}$ as a weighted average of the subsequent trajectory segments:

$$\mathbf{D} = \sum_{m=1}^k [\mathcal{A}(v_{c+m}^{agent}) - \mathcal{A}(v_c^{agent})] \cdot [\text{pos}(v_{c+m}^{agent}) - \text{pos}(v_{c+m-1}^{agent})] \quad (5)$$

where c is the index of $\mathcal{P}_{closest}$ in $\mathcal{T}$, and the weight is computed based on the activeness score difference. Guided by $\mathbf{D}$, carrier objects within a sector of angle θ_R are selected to form the candidate exploration set $\mathcal{C}_{cand}$, formally defined as:

$$\mathcal{C}_{cand} = \left\{ v_i \in \mathcal{C}_r \mid \frac{\mathbf{D} \cdot (\text{pos}_i - \text{pos}_{last})}{\|\mathbf{D}\|_2 \cdot \|\text{pos}_i - \text{pos}_{last}\|_2} \geq \cos\left(\frac{\theta_R}{2}\right) \right\} \quad (6)$$

This process bounds the directional exploration, significantly narrowing the physical search space for the subsequent semantic scanning.

Stage 2: Semantic Class Affinity Scanning. The second stage prioritizes the candidate carriers in $\mathcal{C}_{cand}$ based on class affinities. To quantify the class affinity between the target category Cat_t and each candidate carrier $v_j \in \mathcal{C}_{cand}$, we populate Cat_t into a descriptive text template (e.g., “A receptacle where a $[Cat_t]$ is most often placed”) to extract a contextualized textual embedding $E_{text}(Cat_t)$. The raw semantic affinity $A_{affinity}(Cat_t, v_j)$ is then evaluated by computing the cosine similarity between the textual and visual embeddings of the candidate carrier. Alternatively, this score can be replaced by a commonsense prior inferred by an LLM. To form a valid probability distribution for sequential exploration, these values are normalized across the entire candidate set $\mathcal{C}_{cand}$ using a standard softmax function:

$$P(v_j \mid Cat_t) = \frac{\exp(A_{affinity}(Cat_t, v_j))}{\sum_{v_k \in \mathcal{C}_{cand}} \exp(A_{affinity}(Cat_t, v_k))} \quad (7)$$

By directing the ego-agent to explore candidate carriers prioritized by $P(v_j \mid Cat_t)$, this mechanism significantly reduces redundant exploration.

Table 1: Comparison of existing embodied AI benchmarks

	Vis. Obs.	Movable Objects	Long-term Tasks	Dynamism
R2R [1]	Ego	✗	✗	✗
VirtualHome [24]	3 rd Person	✗	✓	✗
ALFRED [26]	Ego	✓	✗	✗
Dyna-Thor (Ours)	Ego	✓	✓	✓

4 Experiments

To comprehensively evaluate our HitMem framework, we first introduce Dyna-Thor, a novel benchmark constructed to address the critical lack of interaction-driven environmental dynamics. We conduct comparative experiments to assess the performance of various methods in dynamic and long-term task scenarios. Furthermore, we design ablation studies and extension experiments to validate the effectiveness and scalability of our approach across diverse settings. Qualitative visualizations are provided to intuitively interpret the operational process.

4.1 Dyna-Thor Design

Existing embodied AI datasets and benchmarks primarily focus on navigation, reasoning, and manipulation tasks within static environments [1, 12, 24, 26, 30]. As shown in Table 1, existing benchmarks are typically constructed in non-interactive, static environments or overlook the impact of object manipulations performed by other agents (e.g., humans). Consequently, they fail to effectively simulate the challenges encountered by robots operating in real-world settings, where unanticipated human interactions frequently alter object locations during long-term task execution. This limitation results in insufficient evaluation of a method’s robustness and adaptability under realistic dynamic conditions.

To bridge this critical gap, we introduce Dyna-THOR, a novel benchmark built upon the AI2-THOR simulator [14]. Its core evaluation mechanism extends beyond merely requiring the *ego-agent* to accomplish predefined tasks. Specifically, we incorporate *external agents* that interact with environmental objects to simulate realistic human behaviors. These external agents deliberately manipulate task-relevant objects by altering their physical states or relocating them prior to the ego-agent’s execution. Crucially, these interventions occur without providing any explicit notification to the ego-agent. By explicitly modeling the root causes of real-world environmental changes, this design intentionally decouples the ground-truth state from the ego-agent’s internal memory, thereby naturally inducing memory-observation conflicts through unpredictable dynamics during task execution, enabling a highly rigorous and precise evaluation of various methods regarding their robustness and adaptability in dynamic settings.

Dyna-THOR comprises 12 interactive scenes across four distinct room categories, each featuring five long-term tasks. In these tasks, the ego-agent must interpret and sequentially execute multiple natural language instructions to

achieve specified objectives, while simultaneously adapting to environmental dynamics induced by external agents. More details about the Dyna-Thor dataset are provided in the supplementary materials.

4.2 Experimental Setup

Task and Environment Configurations. We conduct experiments in the AI2-THOR simulator. The ego-agent processes egocentric RGB-D observations at a resolution of 400×400 pixels. For task execution, we enforce a strict maximum limit of 500 steps. If the ego-agent exceeds this threshold, the episode is immediately terminated, and the final environmental state is evaluated against the expected goal. The action space of the ego-agent includes navigation actions (MoveAhead-0.25m, RotateLeft-90°, RotateRight-90°), manipulation actions (e.g., PickupObject, PutObject, OpenObject, CloseObject, ...), and a termination action (Done).

Baselines. We compare our method with the following baselines: 1) DELTA [19]: This method leverages a label-based scene graph for task understanding and planning with LLMs. 2) ConceptGraphs [6]: It constructs a static open-vocabulary 3D scene graph for object-level semantic mapping, enabling feature-based matching and querying. 3) DovSG [32]: It maintains a dynamic 3D scene graph that supports localized updates of spatial and semantic relationships. 4) DynaMem [17]: It proposes an online updatable spatio-temporal semantic point cloud memory, which dynamically reconstructs spatial coordinates in response to environmental changes and supports open-vocabulary object localization.

To ensure a fair comparison, all methods requiring pre-constructed memory are provided with an identical initial exploration trajectory. Every method employs the identical recognition and segmentation models for visual processing and utilizes the GPT-4o for all LLM-based reasoning and planning tasks. All methods are restricted to navigating via memory-retrieved spatial coordinates rather than relying on simulator-defined ground-truth object identifiers.

Evaluation Metrics. We quantitatively assess task performance using three primary metrics. *Success Rate (SR)*: This metric measures the overall task completion. A trial is deemed successful only if all necessary actions are executed and the final environmental state strictly matches the ground-truth goal state. *Success weighted by Path Length (SPL)*: This metric evaluates the physical spatial exploration efficiency by calculating the ratio of the theoretically shortest path to the actual executed path length. *Ground-truth Completion Rate (GCR)*: We additionally employ this metric to measure the consistency of task execution across various natural language instructions within each long-term task.

4.3 Comparison Result

Comparison with Baselines. Table 2 presents a detailed comparison on Dyna-THOR. The results show that HitMem significantly outperforms all baselines in both success rate and execution efficiency, demonstrating its strong capability for robust and efficient long-term task execution in dynamic environments.

Table 2: Detailed results on the Dyna-Thor.

Method	Kitchen			Living-room			Bedroom			Bathroom			Overall		
	SR↑	SPL↑	GCR↑	SR↑	SPL↑	GCR↑	SR↑	SPL↑	GCR↑	SR↑	SPL↑	GCR↑	SR↑	SPL↑	GCR↑
DELTA [19]	0.00	0.00	15.56	6.67	5.63	22.22	0.00	0.00	31.89	6.67	5.24	23.89	3.33	2.72	23.39
ConceptGraph [6]	6.67	2.32	23.89	13.33	8.19	28.89	13.33	6.95	36.89	13.33	10.71	17.78	11.67	7.04	26.86
DovSG [32]	13.33	8.03	32.22	13.33	10.94	38.89	26.67	18.89	50.22	20.00	14.50	37.22	18.33	13.09	39.64
DynaMem [17]	33.33	14.78	43.33	40.00	17.29	52.22	40.00	17.20	58.00	26.67	18.16	37.22	35.00	16.86	47.69
HitMem(Ours)	40.00	24.97	56.67	53.33	36.80	65.56	66.67	46.32	73.00	46.67	28.86	61.11	51.67	34.24	64.08

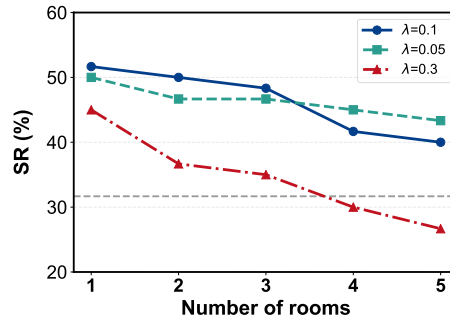
Static memory approaches (DELTA and ConceptGraphs) lack effective memory update mechanisms and rely primarily on outdated scene information despite environmental changes. Given that each long-term task in Dyna-THOR involves object displacement introduced by external agents, these methods struggle to adapt, leading to low success rates. Their few successful cases occur only when object displacement is minor and the moved object remains visible from its original memorized position. ConceptGraph achieves higher performance than DELTA by utilizing feature-based matching. This suggests that feature-based methods offer greater robustness against misclassifications arising from recognition models compared to label-based methods.

Although DovSG supports dynamic updates of the 3D scene graph, its update strategy is limited to local observations. Specifically, memory is refreshed only before manipulation actions, without continuous updating during navigation and execution. As a result, it can handle only object displacements that occur near their original locations. While this design offers some improvement over fully static methods, the performance gain remains limited. DynaMem supports continuous dynamic memory updates and can trigger re-exploration when a target is lost. However, it is constrained by the inefficiency of its global search strategy. This exhaustive exploration often causes the agent to exceed the maximum step limit during the relocation phase. In contrast, HitMem leverages multi-modal context-aware retrieval to drastically narrow the search space and locate targets rapidly. Consequently, our method exhibits superior performance in both overall task completion and execution efficiency.

Ablation Study. Table 3 quantifies the individual contributions of HitMem’s core components. Removing the hierarchical structure (M_H) noticeably degrades success rate and navigation efficiency. The topology provides essential high-level semantics for task goal understanding. Without this semantic stratification, the agent cannot effectively filter the search space using relevant carrier nodes, which increases exploration path length and impairs re-localization of displaced objects. Disabling the temporal decay mechanism (M_D) results in a severe drop in the SPL metric. Without activeness scoring, the agent treats stale and fresh observations equally, leading to repeated navigation to outdated coordinates. This results in wasted steps on invalid targets, such as “ghost entries” of objects displaced by external agents. The most substantial degradation occurs when

Table 3: Ablation studies on different components.

Method	SR $\uparrow$	SPL $\uparrow$	GCR $\uparrow$
w/o Hierarchical Structure (M_H)	46.67	26.71	61.43
w/o Decay Mechanism (M_D)	43.33	20.45	52.97
w/o Retrieval Strategy & M_D	31.67	12.66	42.97
HitMem (Ours)	51.67	34.24	64.08

**Fig. 4:** Performance across different scene scales and decay rate settings.

both the context-aware retrieval strategy and M_D are removed. Our full strategy leverages external agent trajectories and semantic affinity to predict probable locations, effectively simulating human reasoning. Conversely, lacking this guidance forces the agent into blind global searches. Such ineffective wandering exhausts the step limit, critically compromising overall task execution.

4.4 Scalability Analysis and Visualizations

Scalability Analysis. We further utilize the multi-room environments provided by RoboTHOR [5] to evaluate the robustness of HitMem across varying spatial scales and to analyze the impact of the temporal decay rate (λ) on overall performance. Fig. 4 illustrates the success rate trends under different room quantities and decay configurations.

In small-scale environments, performance variations across different decay rates remain relatively minor, as frequent observations enable timely memory updates. However, the choice of λ becomes critical in larger spatial configurations. Values of λ that are optimal in single-room scenarios may yield performance degradation when applied to multi-room settings, where reduced revisit frequencies for the same object cause historical memories to continuously decay, thereby lowering their retrieval priority. Adopting a smaller λ can alleviate this issue by enhancing memory persistence. Conversely, an excessively large λ leads the agent to over-prioritize recent observations, resulting in narrowly-focused searches that fail to track substantially displaced targets and consequently cause

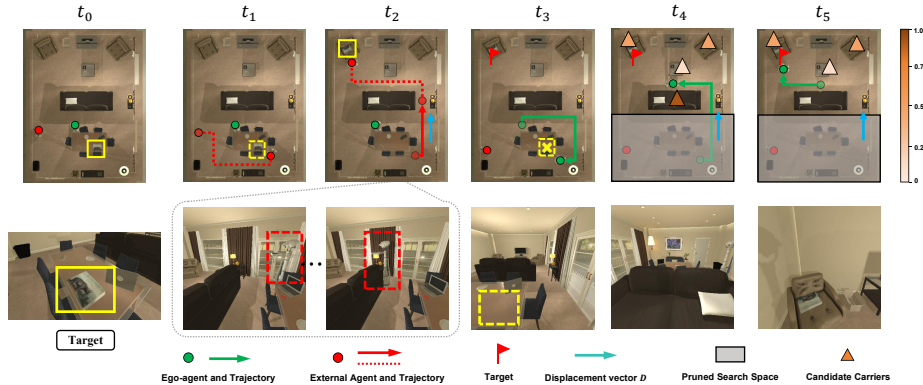


Fig. 5: Visualization of a representative dynamic object relocation process.

a sharp decline in success rates. Therefore, dynamically adjusting the decay rate according to the spatial scale can effectively balance the persistence of memory with the adaptability required for dynamic environments.

Under identical decay settings, expanding the exploration space from a single-room to multi-room environments naturally increases both maintenance and exploration costs. This expanded search space leads to an anticipated downward trend in the overall success rate. To evaluate robustness, we establish the optimal success rate of the variant lacking a decay mechanism in the single-room scenario as a fixed baseline (gray dashed line). With an appropriate decay rate, our approach consistently surpasses the fixed baseline. This confirms that Hit-Mem, powered by hierarchical temporal memory and multi-modal context-aware retrieval, reliably boosts task performance across varying spatial scales.

Case Visualization. Fig. 5 provides a visualization of a dynamic object relocation process. It illustrates the movement trajectories of both the ego-agent and the external agent, along with the ego-agent’s active visual observations. At t_1 , the external agent removes the task relevant target object (Book) and places it in a new location. Between t_1 and t_2 , the ego-agent partially observes and records the movement sequence of the external agent. Upon receiving the instruction to find the book, the ego-agent first queries its internal memory to retrieve the last known coordinates of the target. Arriving at this expected location at t_3 , the ego-agent experiences a conflict between its memory and current observations because the book is visually absent. To recover from this failure, the ego-agent infers a potential displacement vector utilizing the recorded trajectory points of the external agent. This geometric operation effectively filters the global search space. Subsequently, the system calculates a prioritized set of candidate carriers based on semantic class affinity. Guided by this ranked list, the agent navigates to the most probable carrier (a sofa) at t_4 . After a local visual inspection yields no results, the agent continues its sequential exploration. Ultimately at t_5 , the

agent successfully detects the target object on the next highest ranked carrier (an armchair) and proceeds with the remaining task execution.

5 Conclusion

In this paper, we propose HitMem, a novel hierarchical temporal memory framework designed for embodied agents performing long-term tasks in dynamic environments. HitMem unifies semantic, spatial, and temporal information into a lightweight topological graph while employing a temporal decay mechanism to dynamically modulate the activeness scores of memory nodes. Based on this memory representation, we further introduce a multi-modal context-aware retrieval mechanism for efficient target object localization. We specifically design a two-stage retrieval strategy that combines physical trajectory cues with semantic class affinity to reduce redundant exploration and improve search efficiency. Furthermore, we construct the Dyna-Thor benchmark to explicitly simulate unpredictable interferences from external agents. Extensive evaluations demonstrate that HitMem effectively adapts to dynamic environmental changes and scales well across different scenarios, while significantly reducing exploration costs during complex task execution.

Acknowledgements

This work is supported by the National Natural Science Foundation of China 62472412 and Project of Institute of Software, Chinese Academy of Sciences (ISCAS-JCMS-202402).

References

1. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Van Den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3674–3683 (2018)
2. 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 preprint arXiv:2212.06817 (2022)
3. Chen, B., Xia, F., Ichter, B., Rao, K., Gopalakrishnan, K., Ryoo, M.S., Stone, A., Kappler, D.: Open-vocabulary queryable scene representations for real world planning. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 11509–11522 (2023). <https://doi.org/10.1109/ICRA48891.2023.10161534>
4. Cui, X., Liu, Q., Liu, Z., Wang, H.: Frontier-enhanced topological memory with improved exploration awareness for embodied visual navigation. In: European Conference on Computer Vision. pp. 296–313. Springer (2024)

5. Deitke, M., Han, W., Herrasti, A., Kembhavi, A., Kolve, E., Mottaghi, R., Salvador, J., Schwenk, D., VanderBilt, E., Wallingford, M., et al.: Robothor: An open simulation-to-real embodied ai platform. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 3164–3174 (2020)
6. Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K.M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R., et al.: Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. pp. 5021–5028. IEEE (2024)
7. Huang, W., Wang, C., Zhang, R., Li, Y., Wu, J., Fei-Fei, L.: Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973* (2023)
8. Hughes, N., Chang, Y., Carlone, L.: Hydra: A real-time spatial perception system for 3d scene graph construction and optimization. *arXiv preprint arXiv:2201.13360* (2022)
9. Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M.: Lerf: Language embedded radiance fields. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 19729–19739 (2023)
10. Kim, B., Kim, J., Kim, Y., Min, C., Choi, J.: Context-aware planning and environment-aware memory for instruction following embodied agents. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 10936–10946 (2023)
11. Kim, K., Yoon, K., Jeon, J., In, Y., Moon, J., Kim, D., Park, C.: Llm4sgg: Large language models for weakly supervised scene graph generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 28306–28316 (2024)
12. Kim, T., Min, C., Kim, B., Kim, J., Jeung, W., Choi, J.: Realfred: An embodied instruction following benchmark in photo-realistic environments. In: *European Conference on Computer Vision*. pp. 346–364. Springer (2024)
13. 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. *arXiv:2304.02643* (2023)
14. Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Gordon, D., Zhu, Y., Gupta, A., Farhadi, A.: AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv* (2017)
15. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: *ICML* (2022)
16. Lingelbach, M., Li, C., Hwang, M., Kurenkov, A., Lou, A., Martín-Martín, R., Zhang, R., Fei-Fei, L., Wu, J.: Task-driven graph attention for hierarchical relational object navigation. In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. pp. 886–893. IEEE (2023)
17. Liu, P., Guo, Z., Warke, M., Chintala, S., Paxton, C., Shafiullah, N.M.M., Pinto, L.: Dynamem: Online dynamic spatio-semantic memory for open world mobile manipulation. In: *2025 IEEE International Conference on Robotics and Automation (ICRA)*. pp. 13346–13355. IEEE (2025)
18. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499* (2023)
19. Liu, Y., Palmieri, L., Koch, S., Georgievski, I., Aiello, M.: Delta: Decomposed efficient long-term robot task planning using large language models. In: *2025 IEEE*

- International Conference on Robotics and Automation (ICRA). pp. 10995–11001. IEEE (2025)
20. Liu, Y., Chi, D., Wu, S., Zhang, Z., Hu, Y., Zhang, L., Zhang, Y., Wu, S., Cao, T., Huang, G., et al.: Spatialcot: Advancing spatial reasoning through coordinate alignment and chain-of-thought for embodied task planning. arXiv preprint arXiv:2501.10074 (2025)
 21. Maggio, D., Chang, Y., Hughes, N., Trang, M., Griffith, D., Dougherty, C., Cristofalo, E., Schmid, L., Carlone, L.: Clio: Real-time task-driven open-set 3d scene graphs. *IEEE Robotics and Automation Letters* **9**(10), 8921–8928 (2024)
 22. Mazur, K., Sucar, E., Davison, A.J.: Feature-realistic neural fusion for real-time, open set scene understanding. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 8201–8207. IEEE (2023)
 23. Park, J.S., O’Brien, J., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: Interactive simulacra of human behavior. In: Proceedings of the 36th annual acm symposium on user interface software and technology. pp. 1–22 (2023)
 24. Puig, X., Ra, K., Boben, M., Li, J., Wang, T., Fidler, S., Torralba, A.: Virtual-home: Simulating household activities via programs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8494–8502 (2018)
 25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021), <https://arxiv.org/abs/2103.00020>
 26. 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: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020), <https://arxiv.org/abs/1912.01734>
 27. Sun, W., Hou, S., Wang, Z., Yu, B., Liu, S., Yang, X., Liang, S., Gan, Y., Han, Y.: Dadu-e: Rethinking the role of large language model in robotic computing pipeline. arXiv preprint arXiv:2412.01663 (2024)
 28. Tang, Y., Wang, M., Deng, Y., Zheng, Z., Deng, J., Zuo, S., Yue, Y.: Openin: Open-vocabulary instance-oriented navigation in dynamic domestic environments. *IEEE Robotics and Automation Letters* **10**(9), 9256–9263 (2025)
 29. Wu, Z., Zhou, Y., Xu, X., Wang, Z., Yan, H.: Momanipvla: Transferring vision-language-action models for general mobile manipulation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1714–1723 (2025)
 30. Xia, F., R. Zamir, A., He, Z.Y., Sax, A., Malik, J., Savarese, S.: Gibson Env: real-world perception for embodied agents. In: Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on. IEEE (2018)
 31. Xie, Q., Min, S.Y., Ji, P., Yang, Y., Zhang, T., Xu, K., Bajaj, A., Salakhutdinov, R., Johnson-Roberson, M., Bisk, Y.: Embodied-rag: General non-parametric embodied memory for retrieval and generation. arXiv preprint arXiv:2409.18313 (2024)
 32. Yan, Z., Li, S., Wang, Z., Wu, L., Wang, H., Zhu, J., Chen, L., Liu, J.: Dynamic open-vocabulary 3d scene graphs for long-term language-guided mobile manipulation. *IEEE Robotics and Automation Letters* (2025)
 33. Yang, Y., Yang, H., Zhou, J., Chen, P., Zhang, H., Du, Y., Gan, C.: 3d-mem: 3d scene memory for embodied exploration and reasoning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17294–17303 (2025)
 34. 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., Chafflot, D.S., Batra, D., Mottaghi, R., Bisk, Y., Paxton, C.: Homerobot: Open-vocabulary mobile manipulation (2024), <https://arxiv.org/abs/2306.11565>

35. Yin, H., Xu, X., Wu, Z., Zhou, J., Lu, J.: Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation. *Advances in neural information processing systems* **37**, 5285–5307 (2024)
36. Zhang, H., Zhang, Z., Wang, Z., Zhang, Z., Fang, L., Zhou, Q., Gan, C.: Ella: Embodied social agents with lifelong memory. *arXiv preprint arXiv:2506.24019* (2025)
37. Zhang, J., Gireesh, N., Wang, J., Fang, X., Xu, C., Chen, W., Dai, L., Wang, H.: Gamma: Graspability-aware mobile manipulation policy learning based on online grasping pose fusion. In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. pp. 1399–1405. IEEE (2024)
38. Zhang, J., Wang, K., Wang, S., Li, M., Liu, H., Wei, S., Wang, Z., Zhang, Z., Wang, H.: Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. *Robotics: Science and Systems* (2025)
39. Zhang, J., Dong, R., Ma, K.: Clip-fo3d: Learning free open-world 3d scene representations from 2d dense clip. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 2048–2059 (2023)
40. Zhang, L., Liu, Y., Zhang, Z., Aghaei, M., Hu, Y., Gu, H., Alomrani, M.A., Bravo, D.G.A., Karimi, R., Hamidizadeh, A., et al.: Mem2ego: Empowering vision-language models with global-to-ego memory for long-horizon embodied navigation. *arXiv preprint arXiv:2502.14254* (2025)
41. Zhang, Y., Huang, X., Ma, J., Li, Z., Luo, Z., Xie, Y., Qin, Y., Luo, T., Li, Y., Liu, S., et al.: Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514* (2023)
42. Zheng, Y., Chen, X., Zheng, Y., Gu, S., Yang, R., Jin, B., Li, P., Zhong, C., Wang, Z., Liu, L., et al.: Gaussiangrasper: 3d language gaussian splatting for open-vocabulary robotic grasping. *IEEE Robotics and Automation Letters* (2024)
43. Zhi, P., Zhang, Z., Zhao, Y., Han, M., Zhang, Z., Li, Z., Jiao, Z., Jia, B., Huang, S.: Closed-loop open-vocabulary mobile manipulation with gpt-4v. In: *2025 IEEE International Conference on Robotics and Automation (ICRA)*. pp. 4761–4767. IEEE (2025)
44. 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)
45. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: *Conference on Robot Learning*. pp. 2165–2183. PMLR (2023)