Embodied-RAG: General Non-parametric Embodied Memory for Retrieval and Generation

Quanting Xie[1,*], **So Yeon Min**[1,*], **Tianyi Zhang**[1], **Kedi Xu**[1], **Aarav Bajaj**[1], **Ruslan Salakhutdinov**[1], **Matthew Johnson-Roberson**[1], **Yonatan Bisk**[1]
[1]Carnegie Mellon University, Pittsburgh, PA 15213, USA
quantinx@andrew.cmu.edu, soyeonm@andrew.cmu.edu

Abstract: There is no limit to how much a robot might explore and learn, but all of that knowledge needs to be searchable and actionable. Within language research, retrieval augmented generation (RAG) has become the workhouse of large-scale non-parametric knowledge, however existing techniques do not directly transfer to the embodied domain, which is multimodal, data is highly correlated, and perception requires abstraction.

To address these challenges, we introduce Embodied-RAG, a framework that enhances the foundational model of an embodied agent with a non-parametric memory system capable of autonomously constructing hierarchical knowledge for both navigation and language generation. Embodied-RAG handles a full range of spatial and semantic resolutions across diverse environments and query types, whether for a specific object or a holistic description of ambiance. At its core, Embodied-RAG's memory is structured as a semantic forest, storing language descriptions at varying levels of detail. This hierarchical organization allows the system to efficiently generate context-sensitive outputs across different robotic platforms. We demonstrate that Embodied-RAG effectively bridges RAG to the robotics domain, successfully handling over 200 explanation and navigation queries across 19 environments, highlighting its promise for general-purpose non-parametric system for embodied agents.

Keywords: Embodied Memory and Retrieval, Object Goal Navigation, Semantic Navigation, Retrieval Augmented Generation

1 Introduction

Humans excel as generalist embodied agents in part due to our ability to build, abstract, and reason over rich memories. We seamlessly log experiences at appropriate levels of detail and retrieve information ranging from specific facts to holistic impressions, allowing us to respond to diverse requests across different contexts. In contrast, current embodied agents

In the language domain, foundation models combined with non-parametric memory mechanisms have achieved near human-level performance across various tasks. Retrieval-Augmented Generation (RAG)

However, applying RAG to embodied scenarios presents unique challenges due to key differences between textual data and embodied experiences. First, while RAG relies on existing documents, building memory from embodied experiences is itself a core research challenge. Current methods, such as dense point clouds or scene graphs, fail to capture the full range of experiences beyond object-level attributes, without relying on human-engineered schemas or exceeding memory budgets. Second, unlike documents, embodied experiences have inherent correlated structure — semantically similar objects are often spatially correlated and hierarchically organized so embodied experiences should not be treated as independent samples. Finally, embodied observations vary in

granularity and structure: outdoor scenes might be sparse, while indoor environments are cluttered, and repeated objects across frames can confuse LLMs, complicating retrieval.

To bridge this gap, we present Embodied-RAG. Embodied-RAG has two components, *Memory Construction* (Fig. 1(a)) and *Retrieval and Generation* (Fig. 1(b c)). During *Memory Construction*, the system autonomously builds a topological map for low-level navigation and a hierarchical *semantic forest* without relying on hand-crafted constraints or features. This forest is organized based on spatial correlations between hierarchical nodes, each containing language descriptions of observations, and can be expanded to handle temporal or multi-modal inputs. Root nodes represent global explanations, leaf nodes capture specific object arrangements, and intermediate nodes reflect various mid-level scales. Embodied-RAG allows retrieval at various levels of *abstraction* in the language query (explicit, implicit, global), matching it with the *spatial/semantic* resolution (local, intermediate, global) of the memory (Fig. 1(b)/(c)). In the *Retrieval and Generation* process, to mitigate perceptual hallucinations from semantic similarity searches, Embodied-RAG incorporates a robust reasoning component. This involves parallelized tree traversals scored by a language model, with retrieved results structured and used as context for generating explanations or navigational actions via an LLM.

To evaluate the performance of Embodied-RAG, we developed an Embodied-RAG Benchmark, which consists of queries that require multimodal outputs (navigational waypoints and text responses) and reasoning (implicit questions and global summaries). Across over 200 benchmark tasks, we compared Embodied-RAG with two other non-parametric memory baselines: Semantic Match and vanilla RAG. We found that our method serves as an initial step toward solving the problems mentioned above in applying non-parametric memory to embodied agents, showing superior performance against these baselines on the Embodied-RAG Benchmark in the following aspects: (1) More robust against object detection errors on explicit queries (direct object retrieval) since it leverages hierarchical spatial relevancy—for example, recognizing that a toothbrush is more likely found in a bathroom; (2) Improved reasoning on implicit queries (indirect object retrieval), achieving a 220% improvement over Semantic Match and a 30% relative improvement over RAG; (3) Generating more accurate global summarization and trend analysis within the environment, where Semantic Match is unsupported and RAG shows poor quality.

The key contributions and implications of this paper include:

- **Method** We introduce the system of Embodied-RAG. This method addresses problems of naively apply non-parametric memories like RAG to embodied setting.
- **Task** We introduce the general task of *Embodied-RAG benchmark*, formulating semantic navigation and question answering under a single paradigm (Table 1, Figure 1).
- **Implications** Our results and discussion provide a basis for rethinking approaches to generalist robot agents based on non-parameteric memories.

2 Task: Embodied-RAG Benchmark

The Embodied-RAG benchmark contains queries from the cross-product of {explicit, implicit, global} questions with potential {navigational action, language} generation outputs. A task consists of:

- **Query**: The content can be explicit (e.g. a particular object instance), implicit (e.g. looking for adequacy, instruction with more pragmatic understanding required), or global. The request might pertain to a location or general vibe.
- **Experience**: The experience is a sequence of egocentric visual perception and odometry, occurring in indoor, outdoor, or mixed environments.
- **Output**: The expected output can be both navigation actions with language descriptions (Fig 2 top, Fig. 1 c-1), or language explanations (Fig 2 bottom, Fig. 1 c-2).

Table 1: Comparison of related tasks and datasets.

Task	Dataset	Scope of Query			Output Format		Experience	
		Explicit	Implicit	Global	Navigational	Free-form	Indoor	Outdoor
Semantic Navigation		✓	✗	✗	✓	✗	✓	✗
		✓	✗	✗	✓	✗	✓	✗
		✓	✗	✗	✓	✗	✓	✗
Embodied QA		✓	✗	✗	✗	✓	✓	✗
		✓	✗	✗	✗	✓	✓	✗
		✓	✗	✗	✗	✓	✓	✗
VideoQA		✓	✗	✗	✗	✗	✓	✓
		✓	✗	✗	✗	✗	✓	✓
Embodied-RAG		✓	✓	✓	✓	✓	✓	✓

Example tasks are shown in Fig. 2, with instances of explicit, implicit, and global queries in Fig. 1. Spatially, the queries range from specific regions small enough to contain certain objects to global regions encompassing the entire scene. Linguistically, global queries are closer to retrieval-augmented generation tasks, while explicit/implicit ones are more retrieval-focused. *Explicit* and *implicit* queries are *navigational* tasks that expect navigation actions and text descriptions of the retrieved location. *Global* queries are *explanation* tasks requiring text generation at a more holistic level; there are no global navigation tasks since they pertain to large areas, sometimes the entire environment. who carefully went through the simulated or real environment.

3 Related Works

Nonparametric Methods Outside the Embodied Domain In the text and multimodal domain, RAG

Parametric Use of Foundation Models in the Embodied Domain Current approaches in embodied AI often rely on the *parametric* use of foundation models to perceive environments and plan

Existing Methods of Semantic Memory and Retrieval Several methods have been proposed for storing and querying semantic memory in spatial environments, but they remain limited and task-specific compared to the potential of foundation models. Approaches like

Other approaches, such as OCTREE maps

Semantic Navigation and Question Answering Tasks like ObjectNav

4 Method: Embodied Retrieval and Generation

4.1 Memory Construction

The memory construction process of Embodied-RAG consists of two parts: a topological map and a semantic forest.

Topological map We employ a topological graph composed of nodes with the following attributes:

- Position Information: Allocentric coordinates (x, y, z) and the yaw angle θ.
- Image Path: Each node contains a path to an associated ego-centric image.
- Captions: Generated by a vision-language model, these captions provide object-level natural language textual descriptions of the image.

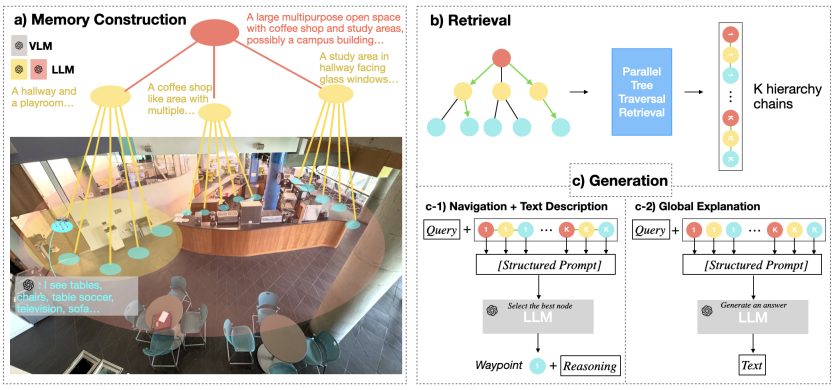

Figure 1: **Embodied-RAG method overview**. (a) Memory is constructed by hierarchically organizing the nodes of the topological map into a semantic forest. (b) The memory in (a) can be retrieved for a query, with parallelized tree traversals. (c) Navigation actions with text outputs, or global explanations can be generated for the query, with using the retrieval results as LLM contexts.

The nodes form a topological map (blue nodes in Fig. 1), eliminating the need for specific control parameters like velocity and yaw, which often vary across different drive systems. This abstraction enables compatibility with any local planner, regardless of the robot's embodiment. Furthermore, the topological structure is far more memory-efficient than traditional metric maps

Semantic Forest We use a separate tree structure, referred to as a semantic forest, to capture meaning at various spatial resolutions. The nodes of this forest are those of the topological map, with the non-leaf nodes capturing larger spaces at a thinner density of semantic specificity. First, we create the forest through hierarchical clustering. Since spatially approximate leaf nodes exhibit semantic correlations, we employ an agglomerative clustering mechanism

This iterative process continues until a root node is formed, stopping when no further relevance is found based on a threshold set by the algorithm. Once we have a complete forest with one or more root nodes, each non-leaf node receives a language description. We achieve this by prompting a large language model (LLM, e.g., GPT-4) to generate a abstraction that encompasses the descriptions of its direct child nodes (see website for the prompting). This process is conducted bottom-up, starting from the leaf nodes and moving up to the parent nodes. We parallelized this process across all nodes at the same hierarchy.

4.2 Retrieval

We run the following *process*, which takes a single tree as input and outputs a single leaf node. Starting by visiting the root node, we run BFS with LLM selection; we ask *LLM_Selector* to choose the best child node of the currently visited node based on compatibility with the given query. For example, if the query is "find me a place that is bright and quiet but has some presence of people," we prompt the LLM to select the best description among the children of the currently visited node. We then visit the selected best child node and iterate this process until we reach a leaf node. Once we obtain k leaf nodes ($\frac{k}{N}$ nodes from each tree) by running this process $\frac{k}{N}$ times for each of the N trees, we obtain the "chain" from the selected node to the root node. The $\frac{k}{N}$ processes are parallelized across the N trees. The set of these best k chains is the retrieval output, containing semantics at all scales for any specific location that corresponds to the leaf scale. Embodied-RAG unifies the retrieval process to handle explicit, implicit, and global queries, producing both explanations and navigational actions as outputs. Note, these hierarchies and corresponding trees allow for querying automatically created semantic regions, something particularly useful for outdoor navigation where walls and structures cannot be used to determine function.

4.3 Generation

We pass the retrieved k best chains as part of a context, for the LLM to generate navigation and text description (Fig. 2 top) or global explanations (Fig. 2 bottom). Given the query and the k chains, we prompt the LLM to "select" a waypoint with a reasoning, or to "explain" (prompt in our project website).

Navigation We select a waypoint (a leaf node of the semantic forest) and use a planner to generate navigational actions—sequences of (torque, velocity) pairs— to reach the waypoint. To select this waypoint, we ask the LLM to choose the best single leaf node, togehter with textual reasoning, using the query and the chain as input. Again, including the entire chain as input ensures that a waypoint can be generated for implicit navigation tasks as well.

Text Answers As depicted in Figure 1 (c), we concatenate the k chains as part of the prompt to the LLM. We ask it to generate an answer to the query based on the k retrieved chains. The spatial scale of attention in each node of the chain facilitate the LLM to generate responses at any semantic scale (explicit, implicit, general) based on the retrieved result.

5 Experiments

Task To assess the efficacy of our approach and ensure statistical robustness, we collected data across 19 diverse environments, including both indoor and outdoor settings. These environments span simulated settings (AirSim

Embodiment For the real-world robotic configuration, we utilized a Unitree Go2, equipped with three Realsense cameras to capture a 180-degree field of view. Positional data was acquired using the Go2's integrated lidar and SLAM algorithms. For simulations, we use the default drone setup with a 210-degree panoramic view for and APIs for drone manipulation and positional data acquisition for AirSim. For Habitat, we use the default locobot setup. To construct the experience, human annotators teleoperated and mapped each environment. However, our methodology is adaptable to any frontier-based exploration with minimal modifications

Evaluation Before evaluation, users familiarized themselves with the collected dataset to understand the environment. The four human annotators who generated the queries cross-evaluated, with each query receiving three evaluations, excluding the one from its author. For navigation output, participants chose binary success or fail, and we calculated the Success Rate (SR) from the average across evaluators and tasks. For text output, participants rated the relevance and correctness of the response on a Likert scale of 1 to 5.

Baselines We benchmarked it against two baseline methodologies. The first baseline, *Semantic Match*, follows existing methods by computing cosine similarities between the query and the semantic embeddings of captions from nodes in the topological map, which are also leaf nodes of the semantic forest

6 Results

Table 2: Comparison of Methods on different Embodied-RAG Benchmarks.

Env.	Explicit			Implicit			Global		
	Embodied-RAG	RAG	Sem.	Embodied-RAG	RAG	Sem.	Embodied-RAG	RAG	Sem.
Small	0.955	0.955	0.955	1.000	0.818	0.364	4.88	3.67	-
Large	0.977	0.947	0.895	0.914	0.695	0.426	4.86	2.43	-
Total	**0.969**	0.949	0.877	**0.926**	0.706	0.410	**4.87**	2.68	-

6.1 Quantitative Result

We first present *quantitative* results that demonstrate the effectiveness of our approach in Table 2. As outlined in Section 2, we categorize the Embodied-RAG benchmark queries into three major types: explicit retrieval, implicit retrieval, and global retrieval. Additionally, we classify environments as either small or large based on the number of topological nodes mapped. Our results indicate that Embodied-RAG consistently outperforms RAG and Semantic Match across all tasks and environments. Crucially, all approaches yield expected strong results for explicit queries where a single node is being extracted. RAG's multi-hypothesis approach outperforms Semantic Similarity, and the hierarchy of Embodied-RAG provides a small further boost. The story changes dramatically as we move to implicit queries where the lack of structure causes RAG and and Semantic search performance to drop dramatically, while Embodied-RAG maintains robustness even in large environments. A similar result is seen in the likert scale for Global questions. Note, Semantic Match cannot be applied for Global as it lacks summarizing and reasoning.

6.2 Qualitative Result

We further conduct a *qualitative* comparison on the reasoning generated by Embodied-RAG and the baseline models.

Implicit Query: Find where I can buy some drinks? From the figure, we see that Embodied-RAG correctly identifies a food service area, while the baselines provide incorrect answers. For RAG and direct semantic match, the most relevant results retrieved are those with direct semantic associations, such as a refrigerator or water fountain. However, there is a clear mismatch between the user's intention and the retrieved objects. The goal is to 'buy' water, which typically requires a counter or vending machine for the transaction, rather than simply grabbing it from a refrigerator or drinking from a water fountain. Embodied-RAG performs multi-step reasoning from the top of the tree to the bottom, and retrieves more diverse and plausible locations. It successfully identifies counters as the most appropriate locations for the user's intention.

Global Query: As illustrated in Figure 2, Embodied-RAG demonstrates a comprehensive understanding of the environment by accurately describing it as a suburban neighborhood intertwined with a park. This holistic perception is attributed to Embodied-RAG's hierarchical organization of information, Figure 2, enables it to pseudo-attend to every node in the map. In contrast, RAG retrieves only the most similar nodes, resulting in a fragmented view characterized by redundant items and a failure to integrate observations into a cohesive environmental context. This limitation of RAG, where subareas are treated as independent rather than interconnected components of the whole, aligns with findings from previous work

7 Conclusions

We present Embodied-RAG, a system capable of capturing spatial memory at any spatial and semantic resolution in both indoor and outdoor environments, and retrieving and generating responses for navigation and explanation requests. Additionally, we introduce the task of Embodied-RAG benchmark, unifying semantic navigation and question answering. Our findings demonstrate that Embodied-RAG can robustly handle implicit and global queries, as well as ambiguously phrased requests from human annotators. Our results indicate that Embodied-RAG shows potential as the basis for incorporating large non-parameteric memories into robotics foundation models. We are excited for future extensions to manipulation and dynamic environments that enable robotics tasks out of reach for current memory/context constrained approaches.

Acknowledgment

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR00112490375 and partially supported by funding from Lockheed Martin Corporation.

References

Appendix

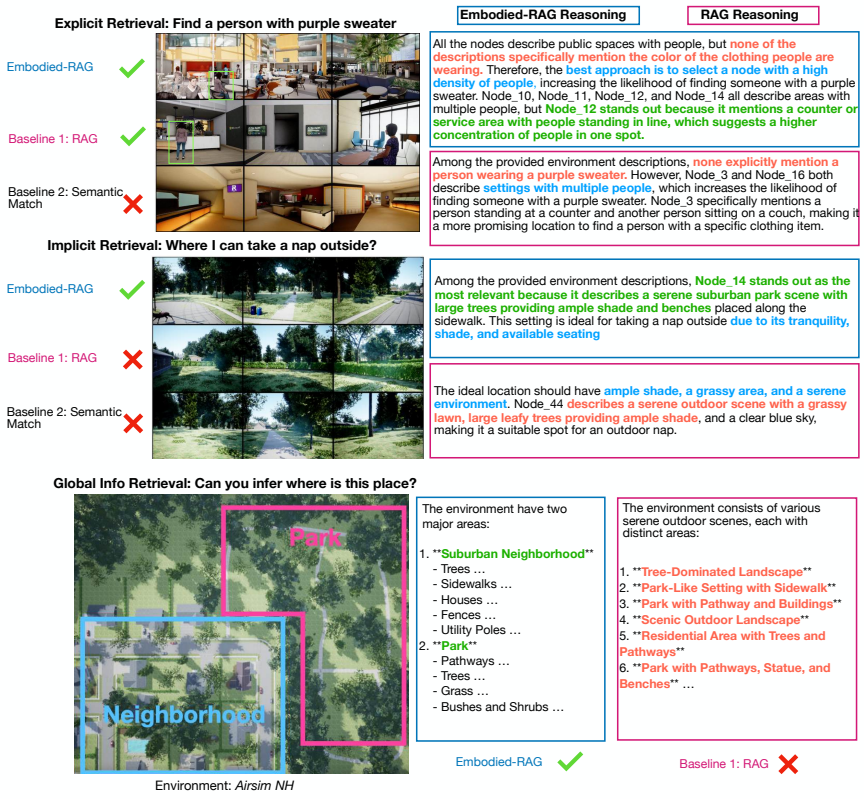

Figure 2: Example reasoning of Embodied-RAG and RAG for generation tasks are highlighted in blue and pink boxes, respectively.

A Computational Efficiency

Both memory construction and retrieval have a computational complexity of $O(\log N)$, where N represents the number of nodes in the environment. This choice allows us to efficiently scale to larger environments, as the time complexity only increases logarithmically with the number of nodes. Additionally, when performing the k retrievals, we execute them in parallel to minimize the overall time cost. In our real-life experiments, the time costs are demonstrated in the supplementary video, which is 8x fast-forwarded. On average, a single retrieval takes around 20 seconds in most of our environments, and the travel time depends on the speed of the specific embodiment in use.

B Ablation

We investigate the impact of $k \in \{1, \text{GPT4 Token Limit}\}$ on Embodied-RAG and RAG in Figure 3. A total of 15 experiments were conducted for each k in each environment. We observe that with larger k, both RAG and Embodied-RAG show improved performance, but this improvement plateaus at higher values. RAG still fails to capture the larger holistic resolution with just more object-level nodes and cannot adequately solve the implicit/general queries, further justifying our hierarchy and tree selection approach.

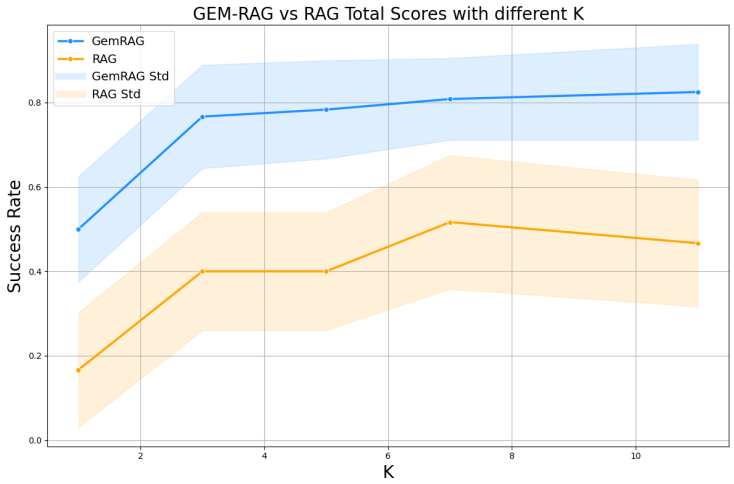

Figure 3: Effect of total number of K searches or K retrievals

C Limitations and Future work

We primarily focused on semantic forests rather than a topological map. Therefore, we may not be robust in obstacle avoidance involving dynamic objects and people. Furthermore, Embodied-RAG currently struggles with requests that require precise counting of objects at a small scale (e.g., "How many chairs are there around the red table?"). This limitation arises because the agglomerative clustering of the semantic forest does not consider multi-view consistency. Future work could incorporate multi-view consistency in the hierarchies of the semantic forest with a learned or pre-trained mechanism to cluster with positional information (e.g. utilizing a LLM).