arXiv:2409.18313v4 [cs.RO] 8 Oct 2024

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], and Yonatan Bisk[1]

***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.

***Index Terms*—Autonomous Agents, Autonomous Vehicle Navigation, AI-Enabled Robotics**

I. 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 [1]–[4] lack such versatile memory capabilities, limiting their ability to operate effectively in unbounded and complex real-world environments. While existing methods such as semantic mapping [1], [2] and scene graphs [5], [6] attempt to capture spatial and contextual relationships, they largely fall short of the dynamic and flexible memory, retrieval, and generative abilities exhibited by humans.

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) [7]–[9] has been widely adopted in the field of Natural Language Processing (NLP) as a non-parametric memory mechanism over large document corpora, enhancing the accuracy and relevance of responses generated by Large Language Models (LLMs). Similarly, the continuous stream of experiences gathered by embodied agents forms vast databases that exceed the context window limitations of LLMs. To address this, approaches like RAG are essential for enabling human-like embodied agents to operate effectively in large, dynamic environments. By integrating non-parametric memory, foundation models within robots can store and retrieve a diverse range of experiences, enhancing their ability to navigate and respond in real-world scenarios.

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. 2(a)) and *Retrieval and Generation* (Fig. 2(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. 2(b)/(c), Fig. 3). In the *Retrieval and Generation* process, to mitigate perceptual hallucinations from

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.

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

[*] Equal contribution.

https://quanting-xie.github.io/Embodied-RAG-web/

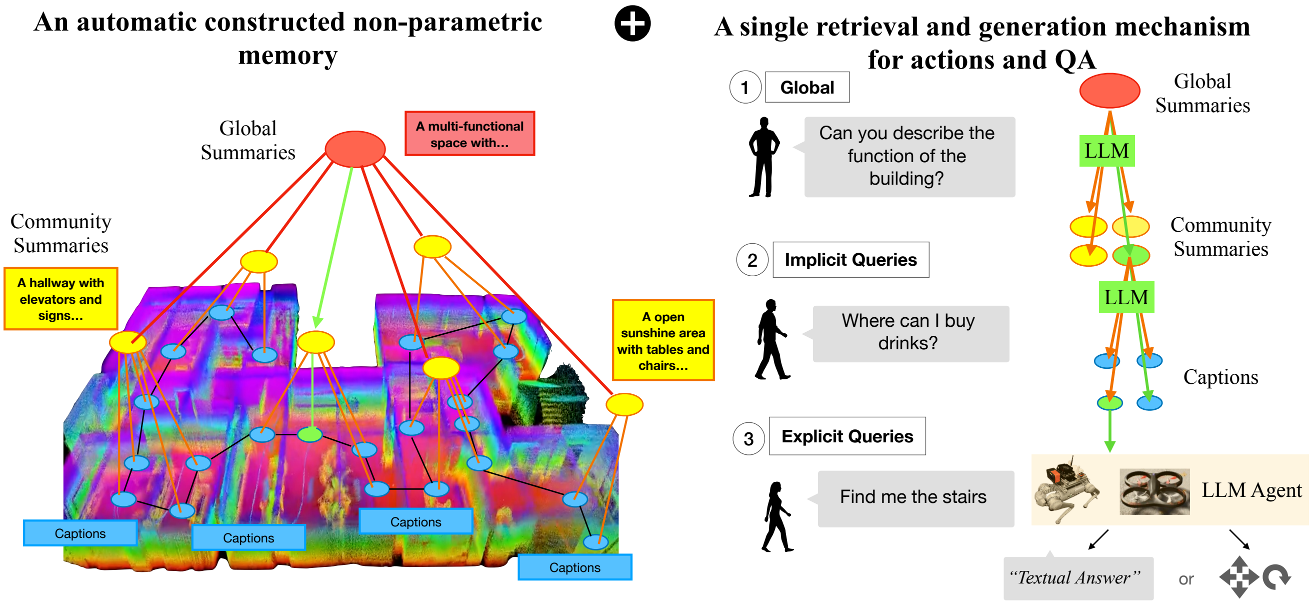
An automatic constructed non-parametric memory
Global Summaries
A multi-functional space with...
Community Summaries
A hallway with elevators and signs...
A open sunshine area with tables and chairs...
Captions
Captions
Captions
Captions
A single retrieval and generation mechanism for actions and QA
1 Global
Can you describe the function of the building?
2 Implicit Queries
Where can I buy drinks?
3 Explicit Queries
Find me the stairs
Global Summaries
LLM
Community Summaries
LLM
Captions
LLM Agent
"Textual Answer"
or

Fig. 1: **Overview**: Our goal is for robots to navigate and communicate effectively in any environment where humans are present. We introduce Embodied-RAG, a framework for automatically building hierarchical spatial memory and providing both explanations and navigation across multiple levels of query abstraction. Embodied-RAG supports robotic operations regardless of the query's abstraction level, the platform, or the environment.

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.

Furthermore, our experiments demonstrate that this pipeline is versatile and applicable across various practical forms of embodiment (drones, locobots, quadrupeds) and can be seamlessly integrate with existing low-level autonomous navigation pipelines. This highlights Embodied-RAG's potential as a general system capable of task-, environment-, and platform-agnostic operation, enabling robots to effectively navigate and communicate in any environment where humans are present. 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 I, Figure 1).
- **Implications** Our results and discussion provide a basis for rethinking approaches to generalist robot agents based on non-parameteric memories.

II. 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 4 top, Fig. 2 c-1), or language explanations (Fig 4 bottom, Fig. 2 c-2).

Example tasks are shown in Fig. 4, with instances of explicit, implicit, and global queries in Fig. 1. Spatially, the