Enter the Mind Palace: Reasoning and Planning for Long-term Active Embodied Question Answering

M. Fadhil Ginting[1,2], Dong-Ki Kim[2], Xiangyun Meng[2], Andrzej Reinke[2], Bandi Jai Krishna[2], Navid Kayhani[2], Oriana Peltzer[2], David D. Fan[2], Amirreza Shaban[2], Sung-Kyun Kim[2], Mykel J. Kochenderfer[1], Ali Agha[2], and Shayegan Omidshafiei[2]
[1]Stanford University, [2]Field AI

***Abstract*—As robots become increasingly capable of operating over extended periods—spanning days, weeks, and even months—they are expected to accumulate knowledge of their environments and leverage this experience to assist humans more effectively. This paper studies the problem of Long-term Active Embodied Question Answering (LA-EQA), a new task in which a robot must both recall past experiences and actively explore its environment to answer complex, temporally-grounded questions. Unlike traditional EQA settings, which typically focus either on understanding the present environment alone or on recalling a single past observation, LA-EQA challenges an agent to reason over past, present, and possible future states, deciding when to explore, when to consult its memory, and when to stop gathering observations and provide a final answer. Standard EQA approaches based on large models struggle in this setting due to limited context windows, absence of persistent memory, and an inability to combine memory recall with active exploration. To address this, we propose a structured memory system for robots, inspired by the mind palace method from cognitive science. Our method encodes episodic experiences as scene-graph-based world instances, forming a reasoning and planning algorithm that enables targeted memory retrieval and guided navigation. To balance the exploration-recall trade-off, we introduce value-of-information-based stopping criteria that determine when the agent has gathered sufficient information. We evaluate our method on real-world experiments and introduce a new benchmark that spans popular simulation environments and actual industrial sites. Our approach significantly outperforms state-of-the-art baselines, yielding substantial gains in both answer accuracy and exploration efficiency.**
Project website: mind-palace-laeqa.github.io.

I. Introduction

Humans naturally develop long-term situational awareness through repeated interactions with their environment, remembering routines, recognizing object placements, and anticipating future needs. For example, when making a shopping list for breakfast, one can recall household preferences and check available supplies to identify what needs to be bought. This type of memory retrieval and long-term temporal grounding is key to intelligent embodied behavior. Among tasks related to this, Embodied Question Answering (EQA) is particularly compelling, as it probes a robot's semantic understanding of its environment [9]. EQA approaches are typically framed either in active settings—where robots explore the environment from scratch to gather information [35]—or episodic settings—where robots answer questions using a single recorded trajectory [30]. While Vision-Language Models (VLMs) have improved performance [27, 20, 17], current approaches are limited to using only the robot's present observations or a single episodic memory, and do not generalize to using multiple past experiences or long-term knowledge. To address this gap, we introduce Long-term Active Embodied Question Answering (LA-EQA), where robots must both recall past experiences and actively explore their surroundings to answer complex questions (see Fig. 1). To our knowledge, this problem is largely unexplored, and no benchmark currently exists to evaluate it.

Performing LA-EQA with VLMs and LLMs using existing EQA approaches is challenging for two reasons. First, representing the robot's past observations accumulated over many deployments across days or months is difficult: a single run can generate thousands of images from diverse viewpoints, yet most questions only require a few relevant frames. Ingesting all this data directly is inefficient and often infeasible due to limited context windows. Second, retrieving relevant information from long-term memory and exploring relevant places in the environment creates a vast combined search space of past and new observations, where an uninformed search is computationally costly. These challenges raise a need for a new paradigm for long-term reasoning for embodied agents.

To address these challenges, we propose an approach for effective long-term memory representation and retrieval for embodied agents. Inspired by the mind palace technique [26]—where humans can effectively recall memories by associating them with spatial landmarks—we structure a robot's long-term observations into a series of spatial world instances. Each instance is represented by a hierarchical scene graph that spatially groups semantic observations. Spatiotemporal structure is captured by linking multiple episodic world instances over time, enabling reasoning and exploration using retrieval of relevant experiences based on spatial proximity and temporal context. Our method, titled Mind Palace Exploration, has three components: 1) Generation, converting long-term memory into multiple scene-graph world instances; 2) Reasoning and Planning, where the robot interleaves EQA reasoning to identify target objects and assess if sufficient information has been gathered; and 3) Stopping Criteria, using Value-of-Information to balance memory recall and active exploration.

We introduce the first benchmark on LA-EQA and evaluate our approach against state-of-the-art baselines in EQA. In particular, the benchmark consists of diverse large-scale, high-fidelity simulation environments and real-world office and

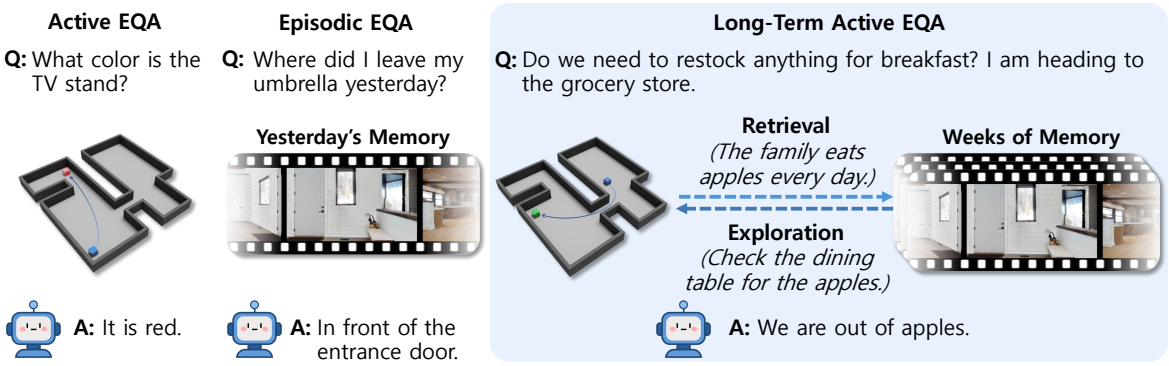

Fig. 1: Different EQA problem setups. We study a new problem of Long-term Active EQA that combines active exploration with long-term memory understanding over multiple episodes.

industrial sites across multiple days and months. Our approach outperforms baselines by 12–28% in answer correctness, achieves 16% higher exploration efficiency, and maintains a 12% correctness gain over the strongest baseline while using 77% fewer retrieved images, demonstrating both the effectiveness and efficiency of our approach. We further demonstrate the scalability and generalizability of our method in long-term settings across diverse environment types, including reasoning over memory from deployments spanning 2.4 km of robot trajectories collected over 6 months. We show the benefit of early memory retrieval stopping criteria in further reducing the number of past observation images while maintaining comparable performance. The real-world experiments demonstrate the feasibility of our approach in practical settings, where a legged robot deployed in a 1,000 m^2 office space uses past inspection memory to efficiently explore the environment and answer practical, day-to-day questions about the office.

II. Related Work

Embodied Question Answering (EQA) has been studied extensively from earlier works that employed learning-based models [9, 16, 10, 43, 48, 40] to more recent efforts leveraging foundation models [30]. Recent approaches generally fall into two settings: episodic-memory EQA, where the agent accesses a single episode of memory, such as in OpenEQA [30] and ReMEmbR [2], and active EQA [21, 12, 44], where the agent explores a novel environment to gather information for answering questions, such as in Explore-EQA [35], Efficient-EQA [7], and Graph-EQA [37]. We propose a new and more general problem of Long-term Active EQA, in which the agent must integrate information across multiple prior episodes and active exploration to answer the question.

Semantic scene representation is a critical component for embodied reasoning and planning. Various methods have been proposed to encode the semantics and contextual structure of the world, including dense 3D representations [32, 38], voxel maps [28], and scene graphs [3, 36, 42]. In our work, we opt for a scene graph approach [33], which has demonstrated effectiveness in EQA tasks [45, 37, 46], and can be integrated with scalable memory retrieval and planning. We extend the scene graph from a single environment snapshot to a series of episodic scene graphs labeled by macro-temporal intervals (e.g., hours, days), enabling the agent to reason over multiple world instances that capture how the environment evolves across long-term deployments.

Semantic-guided navigation focuses on reasoning and planning methods for robot navigation directed by semantic cues, which has a rich body of literature [1, 11] involving tasks specified by images [52, 31], object categories [47, 15], and natural language [8, 13, 14, 29]. Our work related to semantic-based planning to search objects [23] and gather information for EQA taks [7]. The problems are typically framed as either online planning, which builds representations incrementally during execution [24, 51, 6], or offline planning, which relies on pre-constructed maps of the environment [5, 18]. We address the challenge of leveraging multiple historical maps for online planning in long-term settings where the environment evolves over time. We propose a unified approach integrating offline memory retrieval with online exploration for LA-EQA.

III. Problem Formulation of LA-EQA

LA-EQA is a setting where an agent answers questions about the environment by actively exploring it and retrieving relevant information from long-term memory. The LA-EQA task is defined as tuple (Q, M, E, x_0, A^*), where Q is the question, $M = [m_1, \cdots, m_N]$ is a list of episodic memories, E is the current environment, x_0 is the initial robot pose, and A^* is the ground truth answer. The environment is dynamic: its visual appearance and object states can change over time. Each episodic memory $m_i = [m_{i,1}, \cdots, m_{i,L}]$ contains L tuples of past robot pose and image observations $m_{i,j} = (x_{i,j}, o_{i,j})$ collected within a specific macro-temporal interval (e.g., hours).

In LA-EQA, the agent follows policy $\pi(a_k \mid x_k, h_k, Q)$, mapping its state x_k at time step k, working memory h_k (history of action and observation since receiving Q), and the question to one of three possible actions: *retrieve*, *explore*, and *answer*. The *retrieve* action a^R recalls a past memory $m_{i,j}$ into h_k. The *explore* action a^E moves the robot to viewpoint w_i in E, storing the new observation o_k in h_k; w_i need not be near the robot and can be any obstacle-free space informed by prior experience. The *answer* action a^T generates an answer A in natural language based on h and terminates the task.

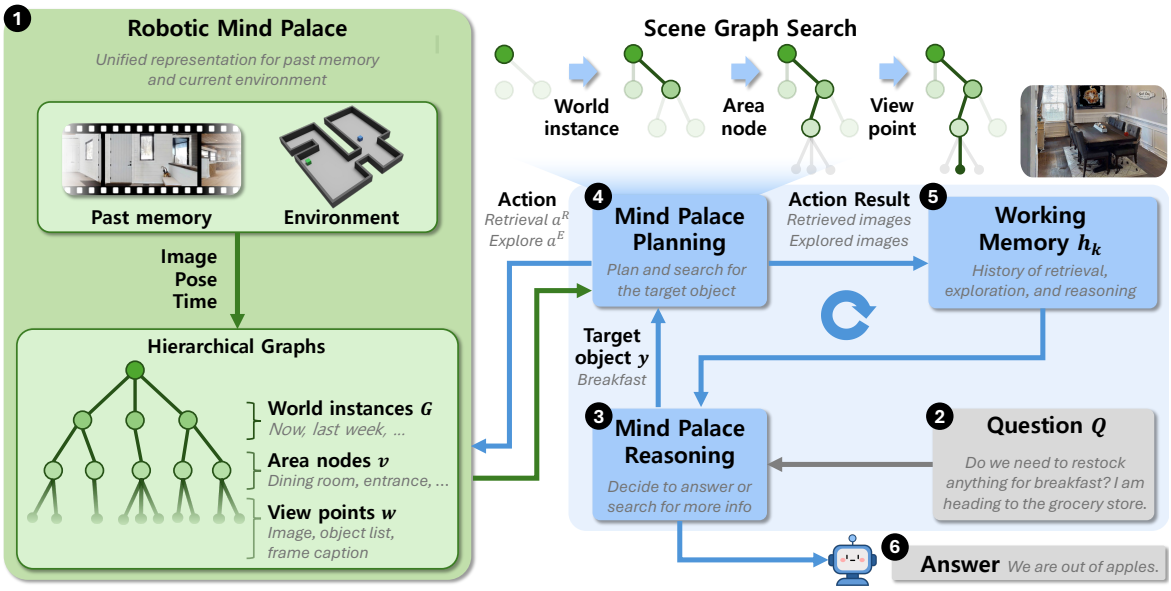

Fig. 2: **Mind Palace Exploration** builds a Robotic Mind Palace that unifies past memories and environment representation *(1)*. Given a question *(2)*, the agent alternates between reasoning over the question to identify a target object *(3)*, planning a search strategy through memory retrieval and exploration *(4)*, and updating its working memory *(5)*, until it is ready to answer the question *(6)*.

IV. Mind Palace Exploration for Solving Long-term Active EQA

Humans use the mind palace technique [26] to remember complex information by organizing it into a structured spatial memory, which enables efficient memory retrieval. We explore how this technique can be applied to long-term memory representation and reasoning in robots. Our approach consists of three key ideas (see Fig. 2). First, prior to the EQA task, we construct a long-term memory representation (referred to as the Robotic Mind Palace $\mathcal{M}$), which summarizes the robot's history of observations into multiple world instances of scene graphs $[G_0, G_1, \cdots, G_N]$. Then, during the LA-EQA scenario, the agent reasons over and explores these world instances in $\mathcal{M}$ to answer the question Q using a policy π. Additionally, we introduce early stopping criteria using the notion of *value of information* to avoid retrieving memory that is unlikely to improve the next exploration action a^E.

A. Mind Palace Generation

Mind Palace is a series of episodic world instances. The Mind Palace divides the long-term history of robot image observation and trajectories M into episodes m based on a macro-temporal term such as hours, times of day, and weeks. The chunking of the episodes comes naturally in robotics as a mobile robot in continuous operations needs to pause any activities while recharging the battery. Each episode becomes a world instance in the Mind Palace and is indexed by its macro-temporal label in texts, allowing an LLM-based agent to select relevant episodes to recall.

An Episodic world instance is represented as a hierarchical scene graph. Given the sequence of robot observation and trajectory within an episode m_i, we build a world representation as a hierarchical scene graph $G_i = (\mathcal{V}_i, \mathcal{E}_i)$, where $\mathcal{V}_i$ denotes the set of nodes and $\mathcal{E}_i$ denotes the edges connecting the nodes [36]. First, we sample dense viewpoints w from the past trajectory to form a set of viewpoint nodes. Each viewpoint node w_i is associated with the robot pose x, images, a list of detected objects in the image [49, 50], and frame captions. The list of objects and frame captions is used as an index for LLM-based agents for image retrieval selection. Then the viewpoint nodes w are clustered into area nodes $v \in \mathcal{V}_i$ based on the spatial and contextual similarity [4, 45, 22]. Each area node v is associated with the centroid of all the clustered viewpoints and the object list. The neighboring viewpoints w and areas v are connected with graph edges, and every w is connected to a v, forming a hierarchical scene graph G for each world instance.

The Robotic Mind Palace consists of a series of world instances representing the *past* long-term memory $[G_1, \cdots, G_N]$ and the *present* knowledge of the environment G_0. At the start of the LA-EQA task, we assume the robot has not explored the present environment yet, so world instance G_0 is only initialized with area nodes v because the state of the environment and object placement may have changed since the last mapping in G_1. We update G_0 as the robot explores the environment.

B. Mind Palace Reasoning and Planning

We perform reasoning and planning over the robotic mind palace to solve the LA-EQA task. This involves three interleaving steps: 1) reasoning over the question to determine what object or spatial concept y to search and when the agent can answer the question, 2) hierarchical planning over the Mind Palace to gather information, and 3) updating the information to the working memory h.

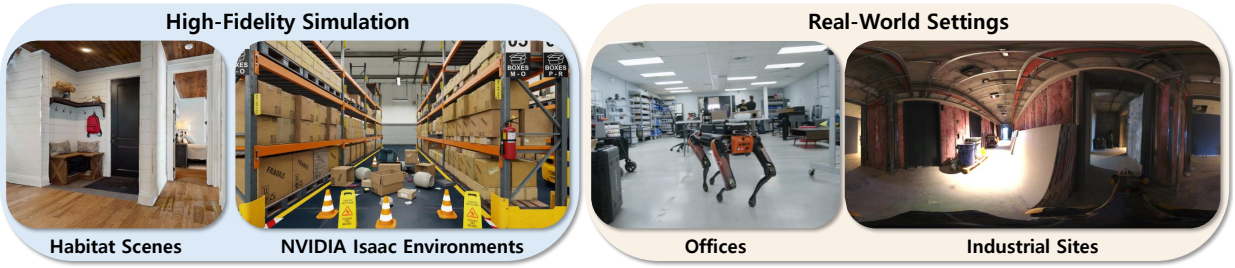

Fig. 3: **LA-EQA Benchmark:** simulated and real-world scenes spanning multiple days / months.

Reasoning over the LA-EQA: The first step in the reasoning process is to determine whether the robot has sufficient information to answer question Q using working memory h_k, which stores past actions, observations, and prior reasoning steps. The agent queries a VLM with h_k and Q. If the VLM responds it is possible to answer the question, the agent executes the *answer* action and provides an answer A with a VLM query. Otherwise, the agent queries and LLM to identify a target object or a spatial concept y, either a specific object explicitly stated in the question or an inferred cue (e.g., something to make a coffee), which becomes the next object goal for exploration.

Planning over episodic world instances G in the Mind Palace: Mind Palace planning begins by selecting a sequence of world instances G to locate y efficiently. We query an LLM with a two-step reasoning process because we observe direct query often yields inefficient plans. The first step asks the LLM to reason whether answering the question requires object search across multiple world instances or if it only concerns a specific instance. Based on this reasoning, the LLM selects a subset of $G \in \mathcal{M}$ and plans a sequence of G. We guide the sequential planning with a heuristic that suggests prioritizing past world instances over the present instance G_0 as using prior knowledge of y locations in the past can inform and improve object search efficiency in the present.

Planning over areas v in the scene graph: Given a world instance G_i, we plan a sequence of areas v to explore that maximizes the probability of finding y. This is framed as an object-goal navigation problem, and we adopt the planning formulation of object search over a scene graph [15]. We first query an LLM to output the probability of finding object y on each area $v \in G_i$, then use a forward search planner to find the best sequence of areas v to explore that minimizes the cost J to find y [25]. When exploring the present scene graph G_0, the cost is defined by the path length between the robot's current pose x_k and the centroid of each area. In contrast, when reasoning over past graphs, the agent can teleport to any area at a constant cost, regardless of the travel distance.

Exploring viewpoints w and replanning: Given an area v_i to search, we query the LLM to select viewpoints w based on the textual information in G_i. The object y may appear in frame captions but often is not mentioned, and relevant viewpoints must be inferred given the textual information [39]. The robot then explores the viewpoints by recalling images from the Mind Palace or navigating to the viewpoints in the environment using a robot-specific motion planner and taking the images. The retrieved or observed images are then stored in the working memory h_k. We repeat the planning over areas v and viewpoints w until the object y is detected in images by a VLM or until we reach the exploration limits. If the object is detected, we search for y in remaining world instances G and move to the *reasoning over the LA-EQA* step.

C. Early Stopping of Memory Retrieval for Navigation

This section examines how to reduce memory retrieval while maintaining exploration efficiency comparable to that of the unlimited memory retrieval case. In particular, we develop stopping criteria that decide when to halt past memory retrieval and proceed with exploration. Given a sequence of world instances that includes the present instance G_0 (e.g., $[G_1, G_2, G_0]$), we use an LLM to form a prediction set of areas $v \in G_0$, where the object y can be located with a probability above a threshold $P(y) \geq 1 - q$. Studies have shown that the LLM prediction and threshold $1 - q$ can be calibrated [35, 34, 41]. Using the prediction set, we define two possible conditions to immediately stop memory retrieval from past world instances $[G_1, G_2]$: 1) the prediction set contains only one area; 2) further memory retrieval will not improve the robot plan over the next sequence to explore v_i in the prediction set. We evaluate the possible improvements on the sequence using the notion of Value of Information (VoI) [19], which quantifies the expected utility gain from retrieving past memory, reducing the expected exploration cost J.

V. Long-term Active EQA Benchmark

Existing EQA datasets [30, 35, 2, 21] focus on scene understanding over short time spans (i.e., the same day), limiting their ability to capture long-term evolution of a scene (e.g., days and months). To address this, we curate the first LA-EQA dataset and benchmark, consisting of 3 simulated and 2 real-world scenes (see Fig. 3). For each simulation scene, we generate 5–10 scene variations over multiple days, reflecting changes caused by common routines. For real-world scenes, we collected 11 trajectories (30–60 mins) in an industrial site and an office environment over a 6-month period.

Question types: We categorize the questions based on their required temporal reasoning to capture different aspects of long-term scene understanding. **1) Past questions** pertain to a specific event observed in a single past trajectory. **2) Present questions** require only exploration of the current environment. **3) Multi-past questions** involve synthesizing information

Methods	Answer	Expl. Eff.	Mem. (#)
Mind Palace (Ours)	**65.0%**	**0.45**	22.86
Mind Palace w/ stopping	61.8%	0.42	**15.73**
Multi-Frame VLMs [30]	52.9%	-	100
Socratic LLMs [30]	44.3%	-	0
ReMEmbR [2]	46.1%	-	0
Active EQA w/ Frames	43.7%	0.29	100
Active Socratic EQA	36.8%	0.19	0

TABLE I: LA-EQA results over answer correctness, exploration and retrieval efficiency.

from multiple past trajectories (e.g., "What do we usually eat for breakfast?"). **4) Past-present questions** require reasoning over both historical memory and the current scene (e.g., "Are we missing anything we usually have for breakfast?"). **5) Past-present-future questions** involve predicting future outcomes based on both past and present observations (e.g., "When do you think we will run out of apples for breakfast?").

We curated 150 questions, which uniformly cover the question types. The questions were generated by seven people to ensure the diversity of the questions. The dataset consists of past trajectories and observations, simulation environments, ground truth answers, and exploration solutions.

VI. Experiments and Discussion

To evaluate our method, we answer: **Q1)** Does Mind Palace Exploration outperform other EQA methods across question types and memory lengths in long-term active EQA? **Q2)** Does the early stopping criteria reduce the amount of memory retrieved without sacrificing performance? **Q3)** Can Mind Palace Exploration be practically deployed in real-world settings?

Methods: We compare our approach against the following baselines: **1) Multi-Frame VLMs** process the question with images and robot poses through a VLM to output the answer. This method is the strongest approach in the OpenEQA benchmark. **2) Socratic LLMs w/ Frame and Scene Graph Captions** use image and scene-graph captions and robot poses to answer the question. **3) ReMEmbR [2]** is a state-of-the-art method in episodic EQA by building a queryable vector database representation of the robot pose, observation time, and image caption embedding and retrieving relevant entries in the database using an LLM. We use the open-source code of the method. **4) Active EQA Agent w/ Frames as the Memory** has the same information as Multi-Frame VLMs, but it lets the agent explore the environment by providing a list of viewpoints that the robot can visit. This approach is similar to the state-of-the-art method of using long-context VLMs with topological graphs [8] applied to the LA-EQA setting. **5) Active Socratic EQA Agent w/ Captions as the Memory** uses the same past memory information as Socratic LLMs w/ Frame and Scene Graph Captions, but it lets the agent explore viewpoints and analyze explored images with VLMs. All approaches use the GPT-4o as the language and vision model [20] and have the same maximum budget of image retrieval and exploration budgets on all active methods.

Metrics: We evaluate all the agents using three metrics: **1) Answer correctness** is compared to the human-annotated answer and judged by an LLM-based scoring [30]. **2) Exploration efficiency** measures the path length of robot exploration compared to the oracle path length weighted by *answer correctness*. **3) Memory retrieval efficiency** measures the number of past images retrieved to answer the question.

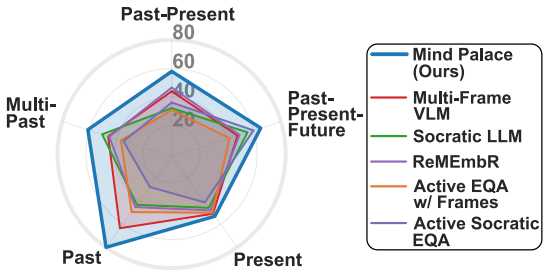

Fig. 4: Performance over temporal reasoning question types.

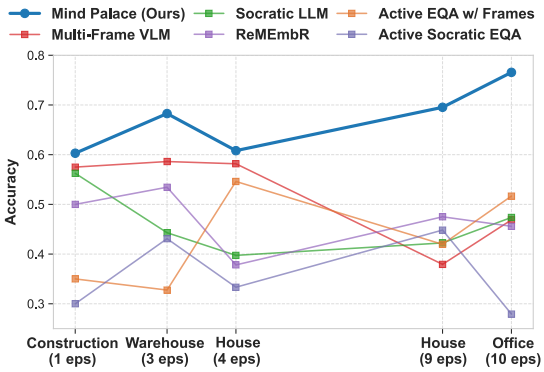

Fig. 5: Performance of five different environments in the LA-EQA benchmark with a varying number of episodes.

A. Q1) Mind Palace Exploration Outperforms other EQA Approaches

Mind Palace Exploration outperforms baselines in all metrics. As shown in Table I, our approach considerably outperforms all methods across the metrics, highlighting the gap in the current approaches in the long-term EQA setting.

Efficient past image retrieval is the key to multi-episodic world understanding. Our approach significantly outperforms the others that require specific information from past memory, represented by *past* and *multi-past* question types in Fig. 4. This is largely because images convey richer visual contexts than captions, enabling more accurate answers about objects. In the LA-EQA setting, multi-frame VLMs struggle as the maximum context length of the state-of-the-art VLMs is not comparable to the sheer amount of past observations in the memory. Our image retrieval approach is critical for efficient image analysis, as EQA questions typically need only several question-related images across multiple episodic memories. The results in Table I show that our approach only needs 77.14% fewer images compared to VLM-based methods, with much higher answer correctness.

Leveraging long-term memory improves active exploration efficiency. Our method achieves higher exploration efficiency than other active EQA agents (Table I), particularly on *past-present* questions (Fig. 4) in which past information

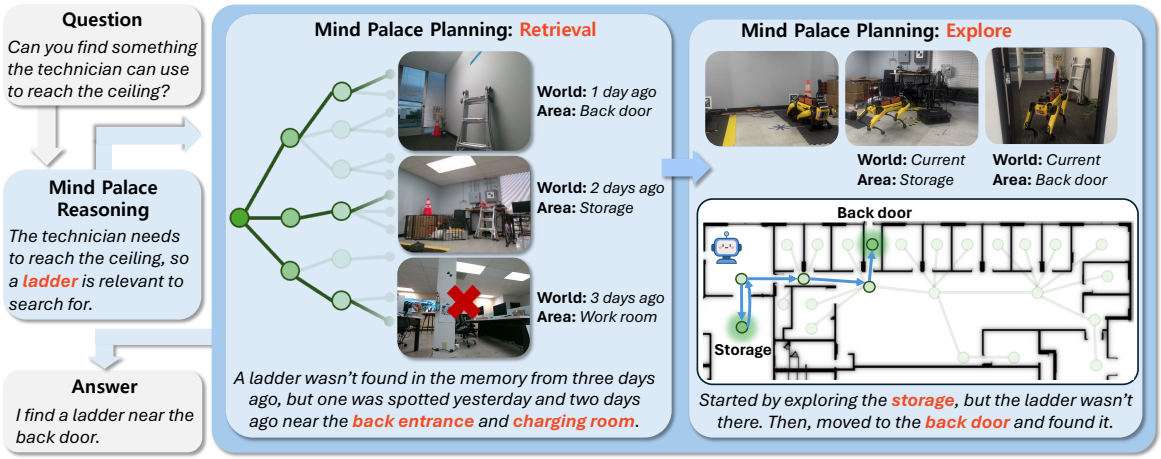

Fig. 6: **Long-term Active EQA Hardware Experiments.** The robot retrieve relevant past information from the past memory and navigate around the office to answer the question.

can benefit present exploration. Our approach often recalls past memories to locate objects of interest so it can more accurately predict the probabilities of the object placements across areas v in the present environment.

Mind Palace Exploration is a scalable approach for LA-EQA. We evaluate scalability by plotting answer accuracy across different environments with varying numbers of past episodic memories (Fig. 5). Our approach shows increasing performance gains over other methods with the number of past episodes in the memory. Given the same image retrieval limits, multi-frame VLM performance considerably drops as the images have less coverage across all the memories. ReMEmbR performs steadily, highlighting the value of retrieval-based approaches in long-term EQA problems.

Our approach generalizes to diverse environments beyond the standard house setting. To test our approach further beyond standard EQA home environments benchmarks, we evaluated Mind Palace Exploration in larger real-world construction sites, a large office, and a simulated warehouse (Fig. 5), where it consistently outperforms others, highlighting its flexibility. Building a structured memory representation for efficient exploration and retrieval becomes more critical as the environment size increases across many episodes.

B. Q2) Benefits of Early Memory Retrieval Stopping

Early memory retrieval stopping reduces the number of memories retrieved without sacrificing performance. As shown in Table I, early stopping reduces the amount of image retrieval from the past memory while maintaining comparable answer accuracy. The early stopping reduces the number of past world instances that the agent retrieves if there is no new observation that will change the agent's next exploration action. Examples in the experiment that we observe include when the robot predicts the possible areas where the object of interest is on the second floor, the robot will stop retrieving past world instances and move to the second floor. The stopping criteria are beneficial to even further improve the memory retrieval efficiency in Mind Palace Exploration.

C. Q3) Real-world Hardware Experiments

We demonstrate the efficacy of Mind Palace Exploration in real-world LA-EQA use cases in an office space spanning over 1,000 m^2 with 27 different areas, using a legged robot as an office assistant. The robot accesses 10 past episodes of past runs, inspecting the office for the past four days and six monthly inspections from October 2024 to March 2025. All the Mind Palace memory storage and planning, other than the GPT4-o query, is performed on the robot. A user sends the question to the robot remotely through a computer, and the robot reports back the answer once it finishes the task. We select 7 questions from the LA-EQA benchmark that require active exploration (Fig. 6).

Mind Palace Exploration enables efficient exploration for practical real-world tasks. By consolidating knowledge of past object placements, the robot can efficiently locate relevant objects, saving an average of 3–10 room searches across the seven evaluated questions compared to a robot without memory access. The questions reflect realistic office scenarios (e.g., searching for tools, tracking missing packages, or identifying vacant desks unused for days) demonstrating the practical utility of LA-EQA. The robot can answer all the questions given that the information is available in its past memory and the current environment.

VII. Conclusion

We present the problem of LA-EQA, a new task that requires robots to combine long-term environment understanding with active exploration. We propose Mind Palace Exploration to address LA-EQA by representing long-term memory and the present environment with a robotic mind palace, enabling reasoning and planning over the Mind Palace. We introduce the first benchmark for long-term active EQA, spanning days of simulation environments and months of real-world data, to foster future research in long-term reasoning. Our approach significantly outperforms state-of-the-art EQA baselines, highlighting the need for a new paradigm for LA-EQA.

References

[1] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018.

[2] Abrar Anwar, John Welsh, Joydeep Biswas, Soha Pouya, and Yan Chang. Remembr: Building and reasoning over long-horizon spatio-temporal memory for robot navigation. *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[3] Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *International Conference on Computer Vision (ICCV)*, 2019.

[4] Yun Chang, Nathan Hughes, Aaron Ray, and Luca Carlone. Hydra-multi: Collaborative online construction of 3d scene graphs with multi-robot teams. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023.

[5] Boyuan Chen, Fei Xia, Brian Ichter, Kanishka Rao, Keerthana Gopalakrishnan, Michael S. Ryoo, Austin Stone, and Daniel Kappler. Open-vocabulary queryable scene representations for real world planning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[6] Junting Chen, Guohao Li, Suryansh Kumar, Bernard Ghanem, and Fisher Yu. How to not train your dragon: Training-free embodied object goal navigation with semantic frontiers. In *Robotics: Science and Systems*, 2023.

[7] Kai Cheng, Zhengyuan Li, Xingpeng Sun, Byung-Cheol Min, Amrit Singh Bedi, and Aniket Bera. Efficienteqa: An efficient approach for open vocabulary embodied question answering, 2024. arXiv preprint arXiv:2410.20263.

[8] Hao-Tien Lewis Chiang, Zhuo Xu, Zipeng Fu, Mithun George Jacob, Tingnan Zhang, Tsang-Wei Edward Lee, Wenhao Yu, Connor Schenck, David Rendleman, Dhruv Shah, et al. Mobility vla: Multimodal instruction navigation with long-context vlms and topological graphs. *Conference on Robot Learning (CoRL)*, 2024.

[9] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[10] Abhishek Das, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Neural modular control for embodied question answering. In *Conference on Robot Learning (CoRL)*, 2018.

[11] Matt Deitke, Dhruv Batra, Yonatan Bisk, Tommaso Campari, Angel X Chang, Devendra Singh Chaplot, Changan Chen, Claudia Pérez D'Arpino, Kiana Ehsani, Ali Farhadi, et al. Retrospectives on the embodied ai workshop. *arXiv preprint arXiv:2210.06849*, 2022.

[12] Vishnu Sashank Dorbala, Prasoon Goyal, Robinson Piramuthu, Michael Johnston, Reza Ghanadhan, and Dinesh Manocha. S-eqa: Tackling situational queries in embodied question answering. *arXiv preprint arXiv:2405.04732*, 2024.

[13] Yao Fu, Dong-Ki Kim, Jaekyeom Kim, Sungryull Sohn, Lajanugen Logeswaran, Kyunghoon Bae, and Honglak Lee. Autoguide: Automated generation and selection of context-aware guidelines for large language model agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 119919–119948, 2024.

[14] Muhammad Fadhil Ginting, Dong-Ki Kim, Sung-Kyun Kim, Bandi Jai Krishna, Mykel J Kochenderfer, Shayegan Omidshafiei, and Ali-akbar Agha-mohammadi. Saycomply: Grounding field robotic tasks in operational compliance through retrieval-based language models. *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[15] Muhammad Fadhil Ginting, Sung-Kyun Kim, David D. Fan, Matteo Palieri, Mykel J. Kochenderfer, and Ali akbar Agha-mohammadi. SEEK: Semantic reasoning for object goal navigation in real world inspection tasks. In *Proc. of Robotics: Science and Systems*, 2024.

[16] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. Iqa: Visual question answering in interactive environments. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[17] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[18] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[19] Ronald A Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1): 22–26, 2007.

[20] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[21] Kaixuan Jiang, Yang Liu, Weixing Chen, Jingzhou Luo, Ziliang Chen, Ling Pan, Guanbin Li, and Liang Lin. Beyond the destination: A novel benchmark for exploration-aware embodied question answering, 2025. arXiv preprint arXiv:2503.11117.

[22] Navid Kayhani, Brenda McCabe, and Bharath Sankaran. Semantic-aware quality assessment of building elements

using graph neural networks. *Automation in Construction*, 155:105054, 2023.

[23] Mukul Khanna, Ram Ramrakhya, Gunjan Chhablani, Sriram Yenamandra, Theophile Gervet, Matthew Chang, Zsolt Kira, Devendra Singh Chaplot, Dhruv Batra, and Roozbeh Mottaghi. Goat-bench: A benchmark for multi-modal lifelong navigation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[24] S. Kim, Amanda Bouman, Gautam Salhotra, David Fan, Kyohei Otsu, Joel Burdick, and A. Agha-mohammadi. PLGRIM: Hierarchical value learning for large-scale exploration in unknown environments. In *International Conference on Automated Planning and Scheduling*, 2021.

[25] Mykel J Kochenderfer, Tim A Wheeler, and Kyle H Wray. *Algorithms for Decision Making*. MIT Press, 2022.

[26] Eric LG Legge, Christopher R Madan, Enoch T Ng, and Jeremy B Caplan. Building a memory palace in minutes: Equivalent memory performance using virtual versus conventional environments with the method of loci. *Acta Psychologica*, 141(3):380–390, 2012.

[27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[28] Peiqi Liu, Zhanqiu Guo, Mohit Warke, Soumith Chintala, Chris Paxton, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Dynamem: Online dynamic spatio-semantic memory for open world mobile manipulation, 2024. arXiv preprint arXiv:2411.04999.

[29] Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. Instructnav: Zero-shot system for generic instruction navigation in unexplored environment. *Conference on Robot Learning (CoRL)*, 2024.

[30] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul McVay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[31] Lina Mezghan, Sainbayar Sukhbaatar, Thibaut Lavril, Oleksandr Maksymets, Dhruv Batra, Piotr Bojanowski, and Karteek Alahari. Memory-augmented reinforcement learning for image-goal navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022.

[32] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[33] Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. *Conference on Robot Learning (CoRL)*, 2023.

[34] Allen Z Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, et al. Robots that ask for help: Uncertainty alignment for large language model planners. *Conference on Robot Learning (CoRL)*, 2023.

[35] Allen Z Ren, Jaden Clark, Anushri Dixit, Masha Itkina, Anirudha Majumdar, and Dorsa Sadigh. Explore until confident: Efficient exploration for embodied question answering. *Robotics: Science and Systems*, 2024.

[36] Antoni Rosinol, Andrew Violette, Marcus Abate, Nathan Hughes, Yun Chang, Jingnan Shi, Arjun Gupta, and Luca Carlone. Kimera: From SLAM to spatial perception with 3d dynamic scene graphs. *The International Journal of Robotics Research*, 40(12-14):1510–1546, 2021.

[37] Saumya Saxena, Blake Buchanan, Chris Paxton, Bingqing Chen, Narunas Vaskevicius, Luigi Palmieri, Jonathan Francis, and Oliver Kroemer. Grapheqa: Using 3d semantic scene graphs for real-time embodied question answering, 2024. arXiv preprint arXiv:2412.14480.

[38] Nur Muhammad Mahi Shafiullah, Chris Paxton, Lerrel Pinto, Soumith Chintala, and Arthur Szlam. Clip-fields: Weakly supervised semantic fields for robotic memory. In *ICRA Workshop on Pretraining for Robotics (PT4R)*, 2023.

[39] Dhruv Shah, Michael Robert Equi, Błażej Osiński, Fei Xia, Brian Ichter, and Sergey Levine. Navigation with large language models: Semantic guesswork as a heuristic for planning. In *Conference on Robot Learning (CoRL)*, 2023.

[40] Jesse Thomason, Daniel Gordon, and Yonatan Bisk. Shifting the baseline: Single modality performance on visual navigation & qa. In *North American Association for Computational Linguistics (NAACL)*, 2019.

[41] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[42] Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. In *Robotics: Science and Systems*, 2024.

[43] Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. Embodied question answering in photorealistic environments with point cloud perception. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[44] Tao Wu, Chuhao Zhou, Yen Heng Wong, Lin Gu, and Jianfei Yang. Noisyeqa: Benchmarking embodied question answering against noisy queries. *arXiv preprint*

arXiv:2412.10726, 2024.

[45] Quanting Xie, So Yeon Min, Tianyi Zhang, Kedi Xu, Aarav Bajaj, Ruslan Salakhutdinov, Matthew Johnson-Roberson, and Yonatan Bisk. Embodied-rag: General non-parametric embodied memory for retrieval and generation, 2024. arXiv preprint arXiv:2403.00000.

[46] Yuncong Yang, Han Yang, Jiachen Zhou, Peihao Chen, Hongxin Zhang, Yilun Du, and Chuang Gan. Snapmem: Snapshot-based 3d scene memory for embodied exploration and reasoning. *arXiv preprint arXiv:2411.17735*, 2024.

[47] Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[48] Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L. Berg, and Dhruv Batra. Multi-target embodied question answering. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[49] Mike Zhang, Kaixian Qu, Vaishakh Patil, Cesar Cadena, and Marco Hutter. Tag map: A text-based map for spatial reasoning and navigation with large language models. *Conference on Robot Learning (CoRL)*, 2024.

[50] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[51] Zirui Zhao, Wee Sun Lee, and David Hsu. Large language models as commonsense knowledge for large-scale task planning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[52] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2017.