Beyond Episodic Evaluation: Memory Architectural Bottlenecks in Sequential Embodied Question Answering

Zikui Cai[1,*,†], Kaushal Janga[1,*], Tan Dat Dao[1,*], Seungjae Lee[1,*], Shivin Dass[2], Mingyo Seo[2], Kaiyu Yue[1], Mintong Kang[3], Nandhu Pillai[1], Monte Hoover[1], Aadi Palnitkar[1], Ruchit Rawal[1], Ruijie Zheng[1], Bo Li[3], Yuke Zhu[2], Roberto Martín-Martín[2], Tom Goldstein[1], and Furong Huang[1]

***Abstract*— Embodied question answering (EQA) is traditionally evaluated under an episodic formulation, where agents solve each task independently and reset internal state between episodes. However, real-world robots operate continuously and must accumulate, retain, and selectively reuse information acquired from prior interactions. Despite this practical requirement, the architectural mechanisms needed to support sequential memory in EQA remain underexplored. In this work, we investigate how different memory architectures behave when EQA agents are evaluated sequentially, with multiple questions answered in the same scene while memory is carried forward across queries. We find that simply preserving existing memory is often insufficient. Agents that retain only traversability information, such as 2D occupancy maps, remember where the robot has explored but not the visual-semantic evidence needed for later questions. Agents trained on short-horizon episodic data face a different challenge: when exposed to continuous, multi-query histories, their inherited context suffers from severe temporal mismatch, rather than forming a reusable scene representation. To overcome this architectural bottleneck, we highlight the necessity of structured, spatially grounded memory: architectures that map persistent visual observations onto metric 3D geometry preserve visual-semantic evidence in a coherent scene representation. Extensive experiments in simulated environments reveal that this form of memory breaks the accuracy-efficiency tradeoff in sequential settings, simultaneously achieving higher answer accuracy and lower navigation costs. We further validate these findings on a real-world mobile robot, demonstrating that spatially grounded visual memory is critical for enabling continuous, intelligent operation in physical environments.**

I. Introduction

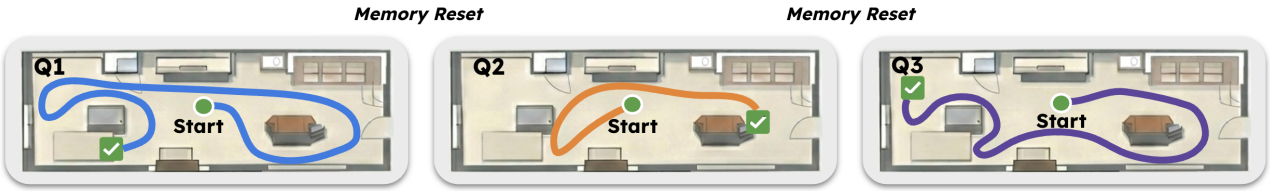
Embodied question answering (EQA) [1] is a core AI capability that requires a physical agent to perceive, reason, and interact within an environment to answer natural language queries grounded in the physical world [2, 3]. EQA serves as a foundational building block for a diverse range of real-world scenarios, ranging from domestic assistive robotics to industrial search-and-rescue and warehouse logistic, where agents must perform long-horizon tasks. Over the past several years, substantial progress has been made on EQA benchmarks and systems [4–11], enabling agents to achieve strong performance on a wide range of visually grounded questions in complex scenes. However, most existing evaluations adopt an episodic formulation, in which each question is treated as an independent task and the agent's internal state is reset between episodes [1, 4–6]. This episodic assumption is increasingly misaligned with real-world deployment, where robots operate continuously and must reuse knowledge acquired from prior interactions to perform subsequent tasks efficiently.

Concretely, an EQA episode begins when an agent receives a natural-language question such as "What color is the mug on the kitchen counter?" while located in a partially observed 3D environment. To answer, the agent must decide where to move, collect visual observations, determine when it has gathered sufficient evidence. Traditionally, this process has relied on modular robotic pipelines for active perception, localization, mapping, and collision-free navigation [1, 4, 12, 13]. While recent advances in LLMs and VLMs have significantly enhanced the agent's ability to ground language and reasoning over incomplete scene information, repeating this intensive exploration loop from scratch for every subsequent question is prohibitively inefficient. As embodied agents transition toward continuous, long-horizon operation, maintaining a persistent memory becomes an unavoidable necessity to bypass redundant re-exploration. Yet, a common implicit assumption in current systems is that simply leaving an agent's memory "on" across tasks will uniformly improve efficiency without degrading decision quality. In this work, we challenge this assumption and investigate a critical architectural bottleneck: how do different memory structures behave when transitioned from episodic to sequential EQA?

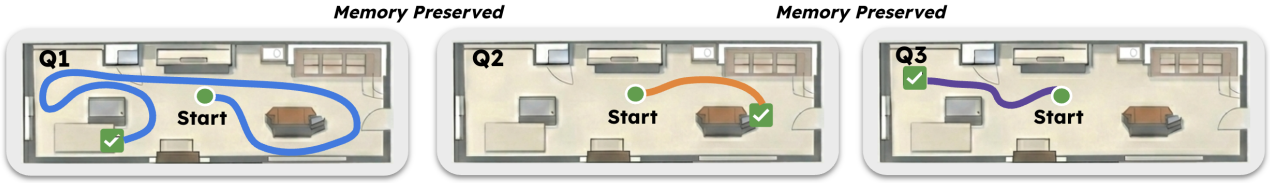
To systematically study this, we introduce the Sequential-EQA evaluation protocol, illustrated in Fig. 1. With minimal intervention, we convert widely used EQA benchmarks, such as OpenEQA [6], into a sequential evaluation setting by presenting multiple questions consecutively within the same environment and carrying the agent's internal state across questions, while leaving the underlying environments and model parameters unchanged. We evaluate existing embodied QA methods that employ different memory mechanisms, including short-term episodic buffers and persistent spatial–semantic representations. These methods were originally designed for independent episodes; we intentionally adapt them without additional optimization to reuse memory across tasks in order to study failure modes rather than optimized performance.

Specifically, we compare three representative memory architectures: lightweight geometric memory, implicit memory in end-to-end VLA agents, and structured 3D spatial

[1]University of Maryland, College Park. [2]The University of Texas at Austin. [3]University of Illinois Urbana-Champaign. *Equal contribution. †Corresponding to: zikui@umd.edu.



(a) Episodic Evaluation



(b) Sequential Evaluation

Fig. 1: **Comparison between Episodic and Sequential Evaluation paradigms.** (a) In **episodic evaluation**, the agent's memory is cleared after every task; for each new query (e.g., Q2), the agent must repeat the full exploration process as if the environment were novel. (b) Conversely, **sequential evaluation** allows the agent to maintain a persistent memory across successive tasks. While the initial task (Q1) requires full exploration, the agent leverages its prior knowledge in subsequent tasks (Q2) to navigate directly to targets, eliminating redundant re-exploration and significantly improving efficiency.

memory. Our findings reveal that naively reusing memory exposes severe vulnerabilities in existing architectures. Agents relying on weak memory structures, such as 2D occupancy maps (e.g., ExploreEQA [14]), fail to retain the rich semantic information required for complex sequential reasoning. Conversely, agents trained end-to-end on short-horizon episodic data (e.g., vision-language-action (VLA) approach UniNavid [15]) suffer from severe out-of-distribution degradation when exposed to continuous, multi-query histories. For these architectures, memory reuse substantially reduces navigation cost but causes their accuracy to either stagnate or systematically degrade due to semantic interference and overwriting.

Using Sequential-EQA, we demonstrate the necessity of structured, spatially-grounded memory. We show that 3D-Mem [16], an architecture that maps persistent visual observations to rigid 3D geometry and leverages strong Vision-Language Models (VLMs), effectively accumulates environmental knowledge over time. Unlike reactive baselines, 3D-Mem uniquely breaks the accuracy-efficiency tradeoff in sequential settings. By anchoring visual memory to spatial coordinates, it avoids catastrophic forgetting and semantic noise, simultaneously achieving higher answer accuracy and drastically lower navigation costs on subsequent queries.

To ensure our findings are not merely simulation artifacts, we reproduce our evaluation on a real mobile robot operating in physical indoor environments. The real-world deployment confirms that spatially-grounded visual memory is critical for enabling continuous, intelligent operation under realistic sensing and actuation noise.

In summary, our main contributions are:

- We introduce Sequential-EQA, an evaluation protocol that transitions EQA from isolated episodes to continuous multi-query operation, measuring memory reuse efficacy without modifying underlying models or environments.
- We identify a critical architectural bottleneck: memory persistence does not imply knowledge accumulation, demonstrated across four architectures spanning VLM-agent and VLA paradigms.
- We show that spatially grounded 3D memory is necessary to break the accuracy–efficiency tradeoff, simultaneously achieving +33.3% accuracy gain and +53.3% navigation cost reduction in sequential settings.
- We validate on a physical quadruped robot across indoor and outdoor environments, confirming that identified bottlenecks generalize beyond simulation.

II. Related Work

Embodied Question Answering (EQA) has emerged as a fundamental task for intelligent agents, requiring the integration of visual perception, spatial reasoning, and navigation capabilities. Early work by Das et al. [1] introduced the first EQA benchmark using House3D environments, establishing the foundational paradigm of agents navigating to answer questions about their surroundings. This was followed by several benchmarks that expanded the scope and complexity of embodied reasoning tasks [4, 6, 8, 14, 17, 18]. The evolution of EQA benchmarks has primarily focused on increasing dataset scale, improving question naturalness, and expanding environmental diversity. However, by evaluating agents on isolated episodes, existing benchmarks fail to capture essential capabilities for practical deployment, where agents operate continuously, and must learn from accumulated expierence Recent work has begun to address some aspects of continual learning in embodied AI. GOAT-Bench [7] introduces lifelong navigation memory, while PartnR [19] incorporates task tracking across multi-agent scenarios. Long-horizon EQA has also been studied directly: Enter the Mind Palace [20] evaluates agents that must reason and plan over extended active question-answering interactions. Our work is complementary rather than claiming that Sequential-EQA is the first long-horizon

EQA setting. The distinction is that we focus on a diagnostic evaluation protocol for existing episodic EQA agents: we keep the environments, questions, and model weights fixed, then vary only whether memory persists across multiple questions in the same scene. This lets us isolate whether a memory architecture itself supports reusable knowledge accumulation, rather than measuring the performance of a method specifically designed or trained for long-horizon operation.

Memory Architectures in Embodied AI. Memory plays a central role in embodied intelligence. Prior work has explored a variety of representations, for building and retaining information, including semantic and voxel maps [21], scene graphs [22], and other structured memory representations [23, 24]. These representations enable localization, object search, and intra-episode reasoning by integrating observations over time. Recent systems further combine structured spatial memory with large Vision–Language Models to improve semantic grounding and exploration strategies [14, 16, 25, 26]. However, most existing memory mechanisms are designed to support decision-making within a single episode. Their performance is rarely evaluated under conditions where memory must persist across multiple, temporally linked queries. Consequently, it remains unclear whether these architectures encode environmental knowledge in a form that supports long-horizon reuse, or whether they primarily function as short-term navigation aids. Our work isolates this question by transitioning episodically trained agents to sequential evaluation without modifying their training procedures.

III. Sequential Evaluation Protocol

Traditional episodic evaluation treats each question as independent: the agent is reset and no internal memory, maps, or perceptual representations carry over. This simplifies evaluation but ignores whether agents can accumulate knowledge from previous interactions. We instead evaluate sequences of questions in the same environment continuously without resetting internal state, exposing whether memory persistence actually improves later answers.

A. Problem Formulation: Episodic vs. Sequential

We define the EQA task within an environment $\mathcal{E}$ containing an ordered set of N questions $Q = \{q_1, q_2, \ldots, q_N\}$. The agent is characterized by a policy π and an internal state (memory) $m \in \mathcal{M}$, which may include spatial maps, latent embeddings, or navigation history.

Episodic Evaluation: In the standard episodic protocol, each question $q_i \in Q$ is treated as a disjoint Markov decision process (MDP). The agent's memory is re-initialized to a null state $m_0 = \emptyset$ at the start of every question. The trajectory τ_i for question q_i is generated as:

$$\tau_i \sim \pi(q_i, m_0). \tag{1}$$

This ensures the agent ignores any environmental knowledge gained in previous trials.

Sequential Evaluation: In the sequential protocol, the memory is accumulated over the set of questions Q. If m_i is the memory at the terminal state for question q_i, then the trajectory for question q_{i+1} is generated as

$$\tau_{i+1} \sim \pi(q_{i+1}, m_i). \tag{2}$$

This allows the policy to leverage accumulated environmental context.

B. Sequentializing an Existing Benchmark

To isolate memory reuse, we group questions from the same environment $\mathcal{E}$ into a sequence $\mathcal{S}$. Given the original dataset $\mathcal{D}_{\text{episodic}} = \{(q_j, \mathcal{E}_j)\}_{j=1}^{M}$, the sequential version is:

$$\mathcal{D}_{\text{seq}} = \{(\mathcal{S}_k, \mathcal{E}_k)\}_{k=1}^{K}. \tag{3}$$

Each $\mathcal{S}_k = (q_{k,1}, \ldots, q_{k,N_k})$ contains all selected questions from scene $\mathcal{E}_k$; thus N_k varies by scene. Sequence order is randomly shuffled with a fixed seed and held constant for all agents and episodic/sequential comparisons. We will release the generated sequences, split code, and evaluation configuration for reproducibility.

C. Evaluation Metrics

We quantify the efficacy of memory retention through both task success and navigational efficiency. Let $y_{k,i}$ be the success indicator (binary or real-valued score) for the i-th question in sequence $\mathcal{S}_k$, and $L_{k,i}$ be the geodesic path length (in meters) traversed during that question.

Accuracy metrics. Success Rate (SR) is the mean accuracy under episodic resets:

$$SR = \frac{1}{\sum_{k=1}^{K} N_k} \sum_{k=1}^{K} \sum_{i=1}^{N_k} y_{k,i} \quad \text{s.t.} \quad m_{0,k,i} = \emptyset. \tag{4}$$

Success Rate with Memory (SR_{mem}) is the mean accuracy under state persistence:

$$SR_{\text{mem}} = \frac{1}{\sum_{k=1}^{K} N_k} \sum_{k=1}^{K} \sum_{i=1}^{N_k} y_{k,i} \quad \text{s.t.} \quad m_{0,k,i} = m_{T_{k,i-1},k,i-1}. \tag{5}$$

Memory Advantage (MA) is the absolute performance gain of using sequential memory:

$$MA = SR_{\text{mem}} - SR. \tag{6}$$

Efficiency metrics. Path Length (PL) is the total navigation cost required without information reuse:

$$PL = \sum_{k=1}^{K} \sum_{i=1}^{N_k} L_{k,i} \quad \text{s.t.} \quad m_{0,k,i} = \emptyset. \tag{7}$$

Path Length with Memory (PL_{mem}) is the total cost when leveraging prior exploration:

$$PL_{\text{mem}} = \sum_{k=1}^{K} \sum_{i=1}^{N_k} L_{k,i} \quad \text{s.t.} \quad m_{0,k,i} = m_{T_{k,i-1},k,i-1}. \tag{8}$$

Step Advantage (SA) is the normalized efficiency gain, expressed as a percentage:

$$SA = \left(\frac{PL - PL_{\text{mem}}}{PL} \right) \times 100\%. \tag{9}$$

IV. Memory Representation Paradigms in EQA

Existing EQA systems differ in how they connect perception, memory, and action. **VLM-agent methods** keep a pretrained VLM frozen and collect structured observations for inference, using memories such as 2D maps [14], dense semantic libraries [26], or 3D reconstructions [16]. This makes memory design modular, but also means that retrieval quality and representation structure strongly determine performance. **VLA methods** instead fine-tune the VLM to output actions or waypoints end-to-end, with memory implicit in short-horizon hidden states. This tight perception-action coupling can be effective within the training distribution, but limits memory to the horizon seen during training, which becomes vulnerable under sequential evaluation.

A. Evaluated Agents

We select four representative agents to span these paradigm boundaries; their key structural differences in memory, retrieval, and adaptation are summarized in Table I.

VLM-agent with geometric memory. ExploreEQA [14] maintains a 2D occupancy map encoding traversable space and frontier scores. A frozen VLM scores candidate frontiers for semantic relevance given the current question, guiding exploration toward likely answer locations. Because the map stores only geometric occupancy rather than object-level appearance or semantic attributes, it effectively remembers where the agent has been but not what was observed there.

VLM-agent with dense episodic semantic memory. MemoryEQA [26] builds an explicit semantic library in which each entry binds an RGB frame to its spatial pose, a natural language description, and a feature embedding. At inference time, an entropy-based retrieval strategy selects the most relevant entries for the current query. While semantically rich, entries are stored as independent pose-tagged events rather than fused into a global spatial structure, which limits coherent reasoning across viewpoints.

VLM-agent with structured 3D spatial-semantic memory. 3D-Mem [16] constructs a persistent metric reconstruction of the environment, binding visual embeddings directly to 3D coordinates. Observations from different viewpoints are fused into a spatially consistent representation, preserving both layout and object identity. Retrieval operates over structured 3D geometry, enabling the agent to reason about spatial relationships explicitly and accumulate knowledge compositionally across queries.

VLA with implicit latent memory. UniNavid [15] fine-tunes a transformer-based VLM to directly predict navigation waypoints from a short window of sequential RGB frames and a language query. Memory is implicit in the model's attention state across this window. Because training always begins from a fresh episode, the model has no mechanism to consolidate observations across query boundaries, making it inherently episodic by design.

B. Memory Adaptation Protocol

To isolate the contribution of memory architecture from the confounding effects of sequential training, we adopt a minimal

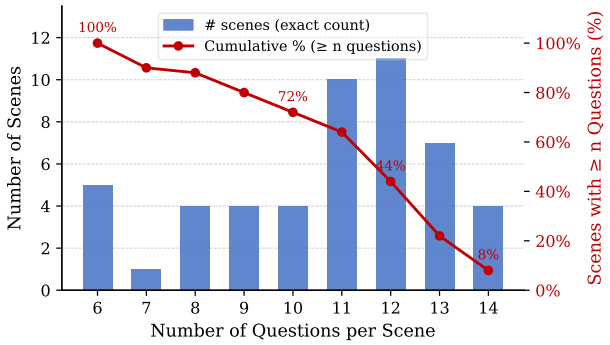


Fig. 2: Distribution of questions per scene in the sequentialized OpenEQA benchmark. Bars show the number of scenes containing exactly N_k questions; the overlaid line shows the cumulative percentage of scenes with at least N_k questions. We evaluate all generated scene-level sequences; aggregate metrics pool all questions, while query-index analyses average over scenes with $N_k \geq i$.

persistence protocol that transitions each agent from episodic to sequential evaluation without any modification to model weights or training procedures. This design choice is deliberate: our goal is not to find the best-performing sequential agent after task-specific optimization, but rather to diagnose which architectural properties intrinsically support knowledge accumulation across queries. The protocol applies uniformly to all agents. At the boundary between consecutive questions q_i and q_{i+1} within a scene, we perform three operations: (1) **state inheritance**: The agent's terminal memory state, $m_{T_i,i}$, is carried forward as the initial state for the next query, $m_{0,i+1}$, with no truncation or summarization. (2) **query update**: The task-specific input is replaced with q_{i+1}, and any goal-directed planners are re-initialized accordingly. (3) **frozen weights**: All network parameters θ remain strictly unchanged throughout the sequence, and no uncertainty recalibration or auxiliary adaptation is applied. This minimal intervention ensures that any performance difference between episodic and sequential conditions is attributable solely to the structure of the inherited representation, not to any learned sequential behavior. Agents that benefit from this protocol do so because their memory architecture is compositional by design; agents that fail do so because their representations were never structured to support cross-query reuse.

V. Experimental Results and Analysis

A. Experimental Setup

Dataset. We construct our sequential benchmark by regrouping questions from the OpenEQA [6] dataset by scene and evaluate the aforementioned 4 different types of representative agents. OpenEQA provides real indoor environments paired with free-form questions that require spatial and semantic reasoning. Under our sequential protocol, all selected questions belonging to the same scene are concatenated into one ordered sequence, and the agent's internal state is preserved across them without any resets. We evaluate all resulting scene-level sequences rather than truncating every scene to a fixed horizon: aggregate results pool all questions

TABLE I: Comparison of memory paradigms in embodied EQA. We categorize each method by its architectural paradigm, the specific content stored, and the primary retrieval mechanism employed.

Method	Paradigm	What is Stored	Retrieval Mechanism
Explore-EQA	VLM-Agent	Occupancy + frontier scores	Frontier ranking by VLM
MemoryEQA	VLM-Agent	RGB + pose + language description + embedding	Entropy-based adaptive retrieval
3D-Mem	VLM-Agent	3D point cloud + visual embeddings	Spatially indexed 3D lookup
UniNavid	VLA	Transformer hidden state	Implicit attention

TABLE II: Performance comparison on Sequential-EQA benchmark. Results show success rates under episodic (SR %) and sequential (SR_{mem} %) conditions, memory advantage (MA %), path lengths (PL), and step advantage (SA %). ↑ indicates higher is better, ↓ indicates lower is better. Standard errors shown in Fig. 3.

Method	SR ↑	SR_{mem} ↑	MA ↑	PL ↓	PL_{mem} ↓	SA ↑
ExploreEQA	43.8	46.5	2.7	84.3	84.3	0.0
MemoryEQA	61.0	62.4	1.4	43.6	43.9	-0.5
3D-Mem	25.5	58.8	33.3	5.6	2.6	53.3
UniNavid	36.4	37.3	0.9	12.2	12.4	-1.5

from all sequences, while per-query-index plots include, at index i, only scenes whose sequence length satisfies $N_k \geq i$. Figure 2 shows the distribution of sequence lengths across scenes in the resulting benchmark. The majority of scenes contain between 10 and 13 questions, with over 60% of scenes providing sequences of at least 10 queries. The exact shuffled sequence files, split-generation code, and evaluation hyperparameters will be released with the benchmark to make the random ordering reproducible and to support future controlled comparisons.

Foundation model and compute resources. We use Qwen3 VL 8B-Instruct [27] VLM with FB8 quantization, all experiments are run on A5000 GPUs.

B. Sequential Memory Does Not Uniformly Improve Performance

A natural expectation is that preserving internal state should reduce redundant exploration while accumulating useful environmental knowledge. However, Table II shows that this assumption only partly holds. Sequential evaluation separates two effects that are coupled in standard episodic testing: whether memory helps the agent answer better, measured by MA, and whether it helps the agent move less, measured by SA. The results lead to four main takeaways.

Accuracy and efficiency. Table II shows that memory persistence alone does not reliably improve answer quality. ExploreEQA, UniNavid, and MemoryEQA all obtain near-zero memory advantage (MA < 3%), but for different architectural reasons. ExploreEQA retains traversed geometry but not object-level semantic evidence, so it remembers where the agent has been rather than what was observed. UniNavid inherits recurrent context across query boundaries, although it was trained on short episodic horizons; this makes sequential context an out-of-distribution input rather than a reusable scene model. MemoryEQA stores richer RGB observations, poses, descriptions, and embeddings, but these entries accumulate as independent events rather than a globally aligned scene representation, increasing retrieval noise as sequences grow.

In contrast, 3D-Mem improves both answer accuracy and navigation efficiency, with +33.3% MA and +53.3% SA. This improvement is not merely trajectory compression: the agent answers more questions correctly while navigating less. By anchoring observations to metric 3D geometry, each query enriches a coherent scene representation rather than adding unstructured history.

Figure 3 visualizes the same pattern. Across most methods, memory persistence changes navigation cost more readily than answer quality. Only 3D-Mem moves toward the desirable region of higher accuracy and lower navigation cost, indicating that spatially grounded memory turns prior observations into reusable evidence rather than simply shortening paths.

Key Takeaways from Sequential Evaluation

Persistence ≠ Accumulation. Merely retaining state across sequential queries does not guarantee successful environmental knowledge accumulation.

Spatial Grounding is Critical. Only spatially grounded memory consistently translates accumulated observations into simultaneous gains in answer accuracy and navigation efficiency.

Decoupling of Efficiency and Success via Premature Termination. Shorter navigation paths do not imply better reasoning; unstructured memory can bias agents to stop early in familiar regions without finding correct evidence.

Structured Memory Supports Cross-query Accumulation. Structuring observations geometrically allows the agent's representation to become exponentially more useful over later queries.

Dynamics across query position. Aggregate metrics show average gains, but they do not reveal whether memory becomes more useful as a sequence progresses. We therefore analyze accuracy and navigation time by query position i.

Figure 4 averages each position over scenes with $N_k \geq i$, so later points use the subset of longer sequences. Navigation time generally decreases under memory reuse, since even weak memory can reduce repeated exploration or bias the agent toward familiar regions. However, lower navigation cost is not sufficient evidence of knowledge accumulation: it may

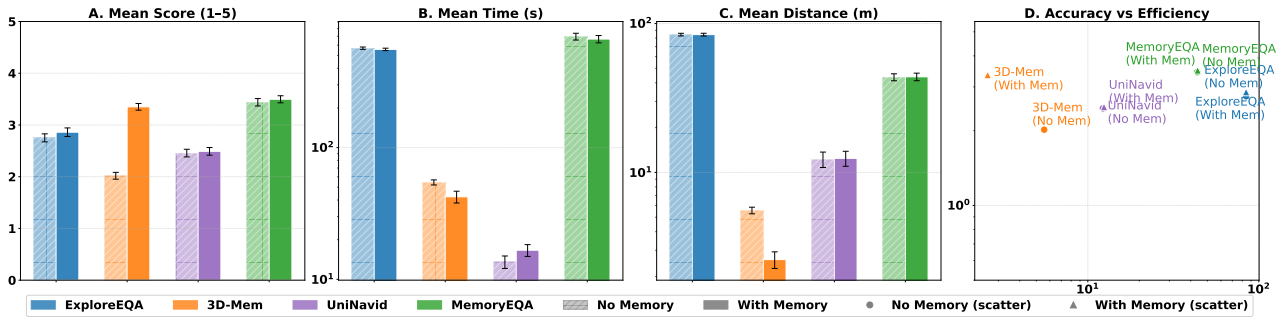


Fig. 3: **Accuracy–efficiency tradeoff under sequential memory reuse.** From left to right: mean answer score (1–5; ↑), mean navigation time (seconds; ↓), mean navigation distance (meters; ↓), and a joint efficiency-vs-accuracy scatter plot. Light bars denote episodic (no memory) evaluation; dark bars denote sequential (with memory) evaluation; error bars indicate standard error.

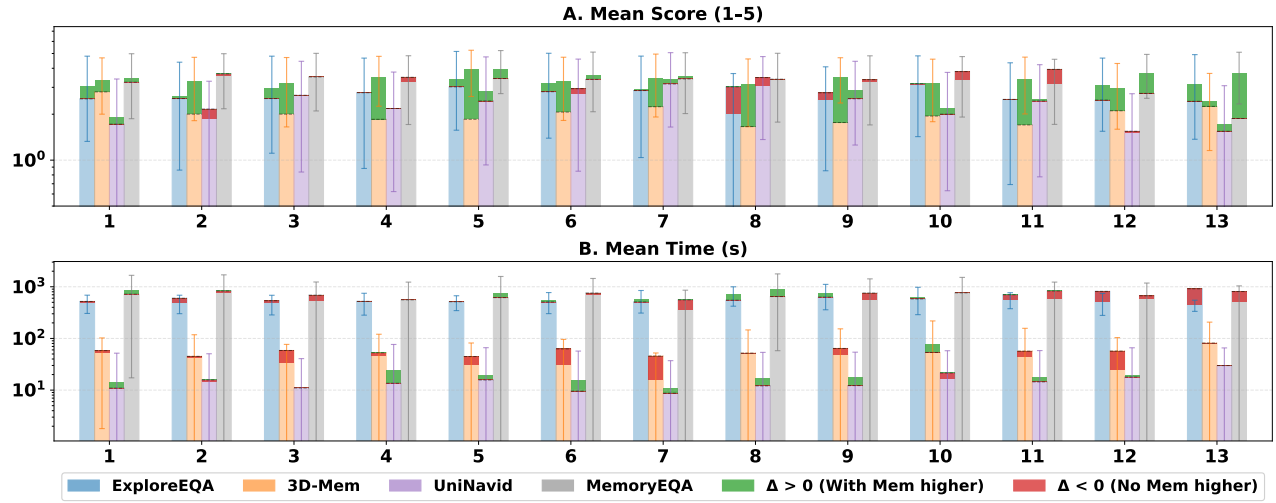


Fig. 4: **Per-query-index accuracy and navigation time under sequential memory reuse.** Panel A shows mean answer score (1–5; ↑, log scale); Panel B shows mean navigation time (seconds; ↓, log scale). The x-axis is query position i within a scene sequence; each position averages over all sequences with $N_k \geq i$, not only sequences of length 13. Δ denotes sequential minus episodic performance and is color-coded by sign: green for $\Delta > 0$ and red for $\Delta < 0$. Because metric direction differs across panels, green indicates improvement only for Panel A, whereas red/negative Δ in Panel B indicates an improvement because lower time is better.

also reflect premature stopping or over-reliance on incomplete history.

Accuracy reveals this distinction. ExploreEQA, UniNavid, and MemoryEQA show no stable improvement across query positions; gains and losses alternate as history accumulates. This suggests that their retained state does not monotonically improve scene understanding. Only 3D-Mem shows sustained positive accuracy Δ across query positions, consistent with a representation that becomes more complete over time. Taken together, the aggregate and per-query analyses show that useful sequential memory must be both persistent and spatially compositional.

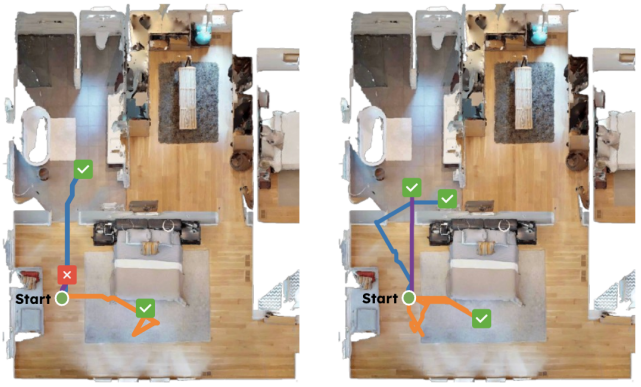
C. Qualitative Analysis

Figure 5 provides a concrete illustration of the behavioral difference between episodic and sequential evaluation on a representative scene from the benchmark, using 3D-Mem as the focal architecture. The left panel shows three episodic trajectories — one per query, each beginning from the same start position with no inherited state. The agent re-explores large portions of the apartment for every question, accumulating substantial redundant path length across the sequence. Despite this exploration, it answers only 1 of 3 questions correctly: when asked "What color is the smoke detector?", it returns an answer about a coat closet — a symptom of incomplete observation coverage within the allocated episode budget.

The right panel shows the same three queries under sequential evaluation. With a persistent 3D map accumulated from prior queries, the agent navigates directly to previously observed regions rather than re-exploring from scratch. Path length drops sharply across Q2 and Q3, consistent with the aggregate SA of 53.3% reported in Table II. Critically, accuracy improves alongside efficiency: all three questions are answered correctly, including the smoke detector query, which the agent now answers correctly by drawing on observations already fused into its 3D representation from the first query's exploration.

This example illustrates the core mechanism behind 3D-Mem's sequential advantage. The improvement is not simply that the agent travels less — it is that prior exploration leaves a spatially consistent record that later queries can retrieve from

(a) **3D-Mem** trajectories on a representative MP3D scene. Under episodic evaluation, the agent re-explores from scratch for each query, answering only 1/3 correctly. Persistent 3D memory enables more direct navigation and correct answers for all 3 questions.



(b) **UniNaviD** trajectories on another MP3D scene. The agent answers 2/3 questions correctly episodically; sequential state inheritance corrects Q2, yielding 3/3. Unlike 3D-Mem, navigation efficiency changes minimally between settings.

Fig. 5: Qualitative comparison of agent trajectories under episodic (left) and sequential (right) evaluation for three consecutive questions (Q1: blue, Q2: orange, Q3: purple).

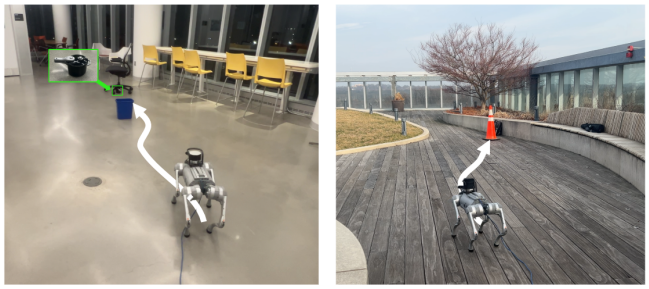
directly. Efficiency and accuracy improve together because they share the same cause: a representation that accumulates observations compositionally rather than discarding them at episode boundaries.

VI. Real-Robot Validation

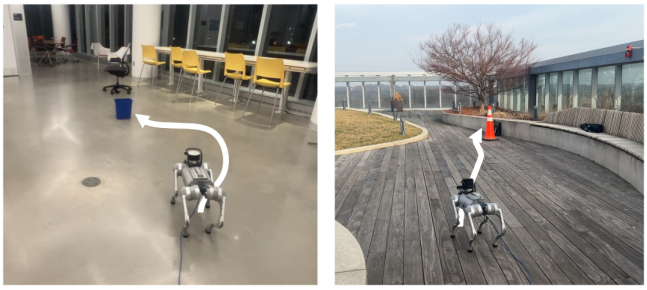
To verify that our simulated findings reflect physical deployment conditions, we evaluate all four methods on a real mobile robot in diverse environments. These experiments test whether the simulated bottlenecks — distributional shift under sequential context and lack of structured spatial memory — persist under sensing noise, actuation error, and environmental variability.

A. Experimental Setup

We deploy a Unitree Go2 quadruped with an Intel RealSense D435i depth camera and onboard LiDAR L2. ExploreEQA, MemoryEQA, 3D-Mem and UniNavid are all evaluated on both an episodic and a sequential trial per environment. All trials consist of five questions, spanning object identification, counting, and localization. Additionally, every trial resets the robot to the same start location before each query; however, the sequential trials preserved state across the five spatially dependent queries. The environments tested are a furnished



(a) Question 1: Indoor ("What color is the pot next to the black chair?") and Outdoor ("What's the color of the cone?")



(b) Question 2: Indoor ("How many yellow chairs are in this room?") and Outdoor ("What's the object behind the cone?")

Fig. 6: **Real-world deployment of the Unitree Go2.** In both environments, UniNavid fails to leverage previous knowledge. In the indoor setting (left), the agent re-navigates to the chair for the second question as if the first encounter never occurred. Similarly, in the outdoor setting (right), the agent repeats its navigation pattern for the second query. In both cases, the agent exhibits erratic rotation and fails to answer correctly, confirming that the **sequential memory reuse bottleneck** generalizes to physical deployment.

indoor lab space, an open lobby, and a long hallway. UniNavid was also tested on an outdoor rooftop terrace.

B. Results and Analysis

Figure 6 shows the UniNavid in both indoor and outdoor environments. UniNavid achieved 15% accuracy in both episodic and sequential settings across all scenes. Failure severity is similar in both indoor and outdoor settings. 3D-Mem stayed consistent with its performance in simulation, improving its success rate from 20% to 40% when switching from episodic to sequential. ExploreEQA actually had a decrease in accuracy from 33% to 26%; however, that could be due to the small test size. Finally, MemoryEQA improved from 40% to 47% accuracy when maintaining memory.

These results indicate that the simulation trend is not an artifact of idealized dynamics. Physical sensing noise and actuation error amplify weak memory architectures: noisy observations reduce evidence quality, and drift accumulates over longer trajectories. Without a structured map, the agent has no stable reference for deciding whether a later query can be answered from prior evidence. Structured spatial memory is therefore critical for deployable continuous EQA agents.

VII. Conclusion

We introduced Sequential-EQA, an evaluation protocol that transitions embodied question answering from isolated

episodes to continuous, multi-query evaluation. Our analysis reveals a critical architectural gap: memory persistence does not guarantee knowledge accumulation. While most paradigms fail to translate retained state into better performance due to semantic interference or a lack of spatial grounding, structured 3D spatial memory successfully breaks this bottleneck. By anchoring visual-semantic evidence to metric geometry, it eliminates redundant exploration and prevents forgetting. Our findings suggest that advancing sequential agents requires building better spatial memory structures rather than simply expanding memory capacity. By exposing these hidden failure modes in memory reuse, Sequential-EQA provides the community with a framework to bridge the gap between simulation and physical deployment. Ultimately, resolving this memory-efficiency bottleneck is the key to unlocking truly continuous, lifelong embodied assistants capable of operating reliably in human environments over days and weeks, rather than isolated, minutes-long episodes. We hope this work motivates the next generation of memory architectures that can operate continuously and reliably in the physical world.

References

[1] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, "Embodied question answering," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[2] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3674–3683.

[3] J. Thomason, M. Murray, M. Cakmak, and L. Zettlemoyer, "Vision-and-dialog navigation," in *Conference on Robot Learning*, 2020, pp. 394–406.

[4] E. Wijmans, S. Datta, O. Maksymets, A. Das, G. Gkioxari, S. Lee, I. Essa, D. Parikh, and D. Batra, "Embodied question answering in photorealistic environments with point cloud perception," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[5] K. Jiang, Y. Liu, W. Chen, J. Luo, Z. Chen, L. Pan, G. Li, and L. Lin, "Beyond the destination: A novel benchmark for exploration-aware embodied question answering," in *IEEE/CVF International Conference on Computer Vision*, 2025.

[6] A. Majumdar, A. Ajay, X. Zhang, P. Putta, S. Yenamandra, M. Henaff, S. Silwal, P. Mcvay, O. Maksymets, S. Arnaud, K. Yadav, Q. Li, B. Newman, M. Sharma, V. Berges, S. Zhang, P. Agrawal, Y. Bisk, D. Batra, M. Kalakrishnan, F. Meier, C. Paxton, A. Sax, and A. Rajeswaran, "Openeqa: Embodied question answering in the era of foundation models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[7] M. Khanna, R. Ramrakhya, G. Chhablani, S. Yenamandra, T. Gervet, M. Chang, Z. Kira, D. S. Chaplot, D. Batra, and R. Mottaghi, "Goat-bench: A benchmark for multi-modal lifelong navigation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 373–16 383.

[8] L. Yu, X. Chen, G. Gkioxari, M. Bansal, T. L. Berg, and D. Batra, "Multi-target embodied question answering," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6309–6318.

[9] T.-C. Chi, M. Shen, M. Eric, S. Kim, and D. Hakkani-Tur, "Just ask: An interactive learning framework for vision and language navigation," in *AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 2459–2466.

[10] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[11] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox, "Alfred: A benchmark for interpreting grounded instructions for everyday tasks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 740–10 749.

[12] B. Yamauchi, "A frontier-based approach for autonomous exploration," in *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97.'Towards New Computational Principles for Robotics and Automation'*, IEEE, 1997, pp. 146–151.

[13] S. Thrun et al., "Robotic mapping: A survey," *Exploring artificial intelligence in the new millennium*, vol. 1, no. 1-35, p. 1, 2002.

[14] A. Z. Ren, J. Clark, A. Dixit, M. Itkina, A. Majumdar, and D. Sadigh, "Explore until confident: Efficient exploration for embodied question answering," *arXiv preprint arXiv:2403.15941*, 2024.

[15] J. Zhang, K. Wang, S. Wang, M. Li, H. Liu, S. Wei, Z. Wang, Z. Zhang, and H. Wang, "Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks," *arXiv preprint arXiv:2412.06224*, 2024.

[16] Y. Yang, H. Yang, J. Zhou, P. Chen, H. Zhang, Y. Du, and C. Gan, "3d-mem: 3d scene memory for embodied exploration and reasoning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 17 294–17 303.

[17] D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, and A. Farhadi, "Iqa: Visual question answering in interactive environments," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[18] Y. Li, Y. Chen, A. Dao, L. Li, Z. Cai, Z. Tan, T. Chen, and Y. Kong, "Industryeqa: Pushing the frontiers of embodied question answering in industrial scenarios," *arXiv preprint arXiv:2505.20640*, 2025.

[19] M. Chang, G. Chhablani, A. Clegg, M. D. Cote, R. Desai, M. Hlavac, V. Karashchuk, J. Krantz, R. Mottaghi, P. Parashar, S. Patki, I. Prasad, X. Puig, A. Rai, R. Ramrakhya, D. Tran, J. Truong, J. M. Turner, E. Undersander, and T.-Y. Yang, "Partnr: A benchmark for planning and reasoning in embodied multi-agent tasks," in *International Conference on Learning Representations*, 2025.

[20] M. F. Ginting, D.-K. Kim, X. Meng, A. M. Reinke, B. J. Krishna, N. Kayhani, O. Peltzer, D. Fan, A. Shaban, S.-K. Kim, M. Kochenderfer, A.-a. Agha-mohammadi, and S. Omidshafiei, "Enter the mind palace: Reasoning and planning for long-term active embodied question answering," in *Proceedings of Machine Learning Research*, vol. 305, 2025, pp. 5072–5106.

[21] N. M. M. Shafiullah, C. Paxton, L. Pinto, S. Chintala, and A. Szlam, "Clip-fields: Weakly supervised semantic fields for robotic memory," *arXiv preprint arXiv:2210.05663*, 2022.

[22] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, and A. Hauptmann, "A comprehensive survey of scene graphs: Generation and application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 1–26, 2021.

[23] S. Gupta, J. Davidson, S. Levine, R. Sukthankar, and J. Malik, "Cognitive mapping and planning for visual navigation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2616–2625.

[24] D. S. Chaplot, D. P. Gandhi, A. Gupta, and R. R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4247–4258, 2020.

[25] N. Yokoyama, S. Ha, D. Batra, J. Wang, and B. Bucher, "Vlfm: Vision-language frontier maps for zero-shot semantic navigation," in *IEEE International Conference on Robotics and Automation*, 2024, pp. 42–48.

[26] M. Zhai, Z. Gao, Y. Wu, and Y. Jia, "Memory-centric embodied question answer," *arXiv preprint arXiv:2505.13948*, 2025.

[27] Qwen Team, "Qwen3-VL technical report," *arXiv preprint arXiv:2511.21631*, 2025.