Vision to Geometry: 3D Spatial Memory for Sequential Embodied MLLM Reasoning and Exploration

Zhongyi Cai*
Michigan State University
caizhon2@msu.edu

Yi Du*
University at Buffalo
yid@sairlab.org

Chen Wang
University at Buffalo
chenw@sairlab.org

Yu Kong
Michigan State University
yukong@msu.edu

Abstract

Embodied agents are expected to assist humans by actively exploring unknown environments and reasoning about spatial contexts. When deployed in real life, agents often face sequential tasks where each new task follows the completion of the previous one and may include infeasible objectives, such as searching for non-existent objects. However, most existing research focuses on isolated goals, overlooking the core challenge of sequential tasks: the ability to reuse spatial knowledge accumulated from previous explorations to guide subsequent reasoning and exploration. In this work, we investigate this underexplored yet practically significant embodied AI challenge. Specifically, we propose 3DSPMR, a 3D SPatial Memory Reasoning framework that utilizes Field-of-View (FoV) coverage as an explicit geometric prior. By integrating FoV-based constraints, 3DSPMR significantly enhances an agent's memory, reasoning, and exploration capabilities across sequential tasks. To facilitate research in this area, we further introduce SEER-Bench, a novel Sequential Embodied Exploration and Reasoning Benchmark that spans two foundational tasks: Embodied Question Answering (EQA) and Embodied Multi-modal Navigation (EMN). SEER-Bench uniquely incorporates both feasible and infeasible tasks to provide a rigorous and comprehensive evaluation of agent performance. Extensive experiments verify that 3DSPMR achieves substantial performance gains on both sequential EQA and EMN tasks[1].

1 Introduction

In real-world scenarios, embodied agents frequently encounter sequential tasks, where tasks follow one another and may even be infeasible, such as searching for non-existent objects. For feasible tasks, agents should operate efficiently; for infeasible ones, agents must detect the impossibility and communicate it clearly. However, mainstream research in Embodied Question Answering (EQA) [1, 2, 3, 4, 5, 6] and Embodied Multi-modal Navigation (EMN) [7, 8, 9, 10, 11, 12] remains largely confined to a single-task setting [13, 14, 15, 16], where the agent's memory is reset for each isolated goal. Only a few recent EMN studies [17, 18] have begun to consider the sequential setting, yet they still overlook infeasible tasks. To bridge this gap, we tackle the underexplored but critical problem of sequential embodied tasks, a crucial step toward developing more capable embodied agents. The core challenge in this problem lies in enabling agents to effectively reuse spatial information accumulated from prior tasks for both feasible and infeasible objectives.

*Equal contribution.
[1]Our code is publicly available at https://github.com/zhyczy/3DSPMR.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

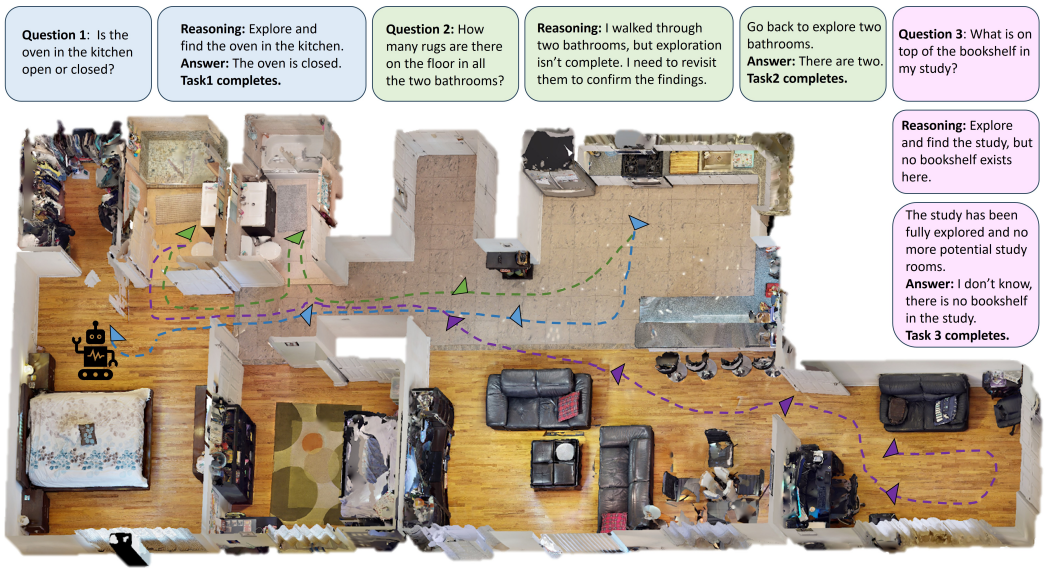

Figure 1: **Illustration of 3DSPMR completing sequential EQA tasks.** After navigating to the kitchen and answering the first question, the agent receives the second query. Using its spatial memory, it recognizes two previously passed but insufficiently explored bathrooms, revisits them, and answers the question after sufficient exploration. For the third question, the agent locates the study but cannot find a bookshelf; once the study is fully explored and no alternative candidate rooms remain, 3DSPMR correctly identifies the task as infeasible.

Multi-modal Large Language Models (MLLMs) [19, 20] are powerful general-purpose models that have been applied to complex embodied AI problems, yet they still struggle with spatial understanding and reasoning [21, 22]. Although recent studies [23, 24, 7] have employed memory-based reasoning frameworks to enhance embodied agents, these approaches still face three major challenges when applied to sequential embodied tasks. First, their memory designs either omit essential cues or store excessive redundant details from previously explored areas during earlier tasks, causing agents to struggle with retrieving relevant information for the new task. Second, MLLMs may overconfidently produce premature results or even hallucinations based on partial observations, due to incomplete exploration in earlier tasks. Finally, these approaches fail to identify infeasible tasks, leading agents to repeatedly or endlessly explore already well-explored areas.

To address these challenges, we propose 3D SPatial Memory Reasoning (3DSPMR), which introduces explicit geometric cues derived from the Field-of-View (FoV) coverage to simultaneously enhance memory, reasoning, and exploration. As illustrated in Fig. 2, FoV coverage represents the geometric region captured within the agent's viewing range. In this work, we perform ray casting from the agent's pose to identify visible cells in the top-down view. Those intersected by rays are marked as "covered" (yellow cells). By modeling FoV coverage, 3DSPMR provides a structured geometric representation of exploration progress.

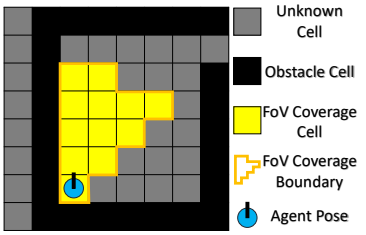

Figure 2: Top-down Projection of Agent FoV Coverage.

Specifically, our 3DSPMR framework enhances embodied agents across three critical dimensions. First, for memory, we propose a new FoV-based gating mechanism to decide when to save specific observations. As a result, the memory keeps representative data and filters spatial redundancy. Second, for model reasoning, we propose to leverage FoV coverage to verify whether all task-relevant spatial information has been sufficiently collected. Through grounded geometric examination, this post-inference mechanism rectifies the MLLM's predictions to mitigate common failure modes, such as premature task termination and the inability to recognize infeasible requests. Finally, for exploration, we propose utilizing FoV coverage to incentivize exploration of underexplored regions. Once triggered, the proposed Geometric and Semantic (Geo-Sem) module

fuses geometric coverage with task-relevant semantics to identify optimal frontier targets, thereby improving navigation goal selection.

To systematically evaluate agents in these realistic scenarios and address the lack of suitable datasets, we construct the Sequential Embodied Exploration and Reasoning Benchmark (SEER-Bench). Unlike prior datasets, SEER-Bench covers both EQA and EMN in a sequential setting, rigorously evaluating an agent's ability to reuse spatial memory across a mixture of feasible and infeasible objectives. Extensive experiments demonstrate that 3DSPMR substantially enhances both reasoning accuracy and exploration efficiency in previously unseen 3D environments.

Our contributions can be summarized as follows:

- We propose a compact and representative spatial memory system that retains essential geometric cues and filters redundancy, effectively bridging the 3D scene understanding gap for MLLMs in sequential tasks.
- We develop the 3DSPMR framework, which leverages explicit geometric priors to guide both reasoning and exploration. This enables agents to effectively execute sequential tasks and reliably identify infeasible objectives.
- We propose SEER-Bench, a Sequential Embodied Exploration and Reasoning Benchmark that covers EQA and EMN, including both feasible and infeasible tasks to systematically evaluate agents' performance.

2 Related Work

MLLMs for Embodied and Spatial Reasoning. Recent Multimodal Large Language Models (MLLMs) such as GPT-4o [19], Qwen2VL [25], and LLaVA-OneVision [26] have shown remarkable progress in general visual–language reasoning. These models have inspired growing efforts toward embodied and spatial reasoning, where agents interpret 3D environments and perform goal-directed behaviors [27, 28, 29]. However, most existing MLLMs rely on 2D image–text supervision and lack explicit 3D spatial awareness [21, 22]. This limitation constrains their ability to reason about geometry, occlusion, and room layouts, which are crucial for embodied perception and navigation. Moreover, their frame-wise visual processing hinders persistent memory formation, leading to fragmented and inconsistent spatial understanding in long-horizon tasks. As a solution, our proposed 3DSPMR approach leverages a unified spatial memory to distill complex 3D representations into text-form scene graphs, images, and a geometric map, providing a more accessible input format for MLLMs.

3D Representations and Memory in Embodied Agents. Structured spatial representations are central to effective reasoning and interaction within 3D environments. Scene-graph-based methods [30, 31, 32, 33, 34] encode object relationships and spatial hierarchies, offering interpretable relational abstractions. However, such symbolic graphs are typically sparse and lack the visual richness needed for perceptual grounding. Complementary approaches that store egocentric visual memories [23, 35, 36] allow retrieval of past observations but often lack geometric or relational awareness. Recent methods [37, 23, 38] have explored constructing 3D semantic maps and global spatial memories, which enhance navigation and question answering yet treat relational, visual, and geometric cues independently. In contrast, 3DSPMR integrates these complementary modalities into a unified 3D spatial memory, combining relational structure, visual appearance, and geometric coverage for coherent spatial reasoning.

Sequential Embodied Tasks. Traditional embodied research such as Embodied Question Answering (EQA) [39, 40] and Scene-Graph Navigation (SG-Nav) [7] evaluate isolated single-task scenarios where each episode is independent. Recent works have explored sequential and multi-task settings [18, 41, 42], but they typically assume all goals or questions are feasible. In realistic environments, some queries are inherently *infeasible*, such as asking for the color of a non-existent object or navigating to an unreachable area. Addressing such cases requires feasibility awareness in addition to spatial reasoning—an ability largely overlooked in prior benchmarks. Recent extensions of EQA, such as OpenEQA [13], GraphEQA [24], and exploration-aware EQA [15], have begun to tackle more diverse and dynamic settings but still do not explicitly incorporate infeasible task reasoning. Our SEER-Bench extends this paradigm by including both feasible and infeasible tasks in sequential EQA and EMN episodes. Therefore, SEER-Bench extends the embodied task formulation, offering a more realistic and comprehensive evaluation for real-world embodied agents.

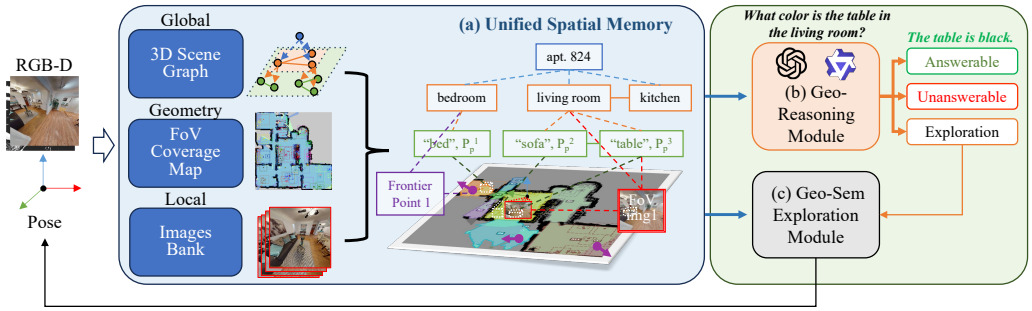

Figure 3: **Overview of 3DSPMR.** (a) Unified Spatial Memory: Raw observations are selectively stored to maintain an informative structured spatial representation. (b) Geo-Reasoning: A grounded inference module that ensures robust decision-making. (c) Geo-Sem Exploration: A strategic module that identifies optimal frontier targets.

3 Methodology

We introduce 3DSPMR, a framework designed to enhance memory retention and feasibility awareness in sequential embodied tasks. As exhibited in Fig. 3, 3DSPMR consists of three synergistic components: a unified spatial memory, a Geo-Reasoning module, and a Geo-Sem Exploration module.

3.1 Unified Spatial Memory

In sequential embodied tasks, agents must be able to consolidate and reuse historical information accumulated across multiple tasks, which necessitates a compact yet representative memory. However, existing memory architectures often exhibit two extremes: they either focus exclusively on the immediate task [3, 24], or suffer from excessive spatial redundancy [23, 43]. Consequently, when transitioning to new tasks, agents may either lack essential context or become encumbered by irrelevant information.

To address this issue, 3DSPMR maintains a unified spatial memory that continuously integrates representative information from all observations. These compact, persistent representations enable the agent to either directly facilitate reasoning, such as answering EQA queries from memory, or guide exploration by leveraging previously accumulated environmental cues.

As shown in Fig. 3(a), the unified spatial memory comprises three complementary modalities: global relational cues that provide topological context via a 3D scene graph, local visual cues that preserve high-fidelity visual attributes through an ego-centric image bank, and geometric cues derived from a Field-of-View (FoV) coverage map. Crucially, these geometric cues serve as a grounded anchor for the model to assess task feasibility and completion while seamlessly linking local and global representations. Furthermore, we employ a FoV-based gating mechanism to retain only non-redundant observations, significantly reducing the storage overhead typically incurred by resource-intensive visual data.

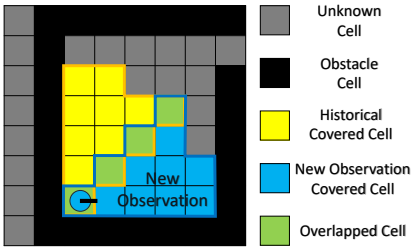

Figure 4: **Top-down Projection of FoV-based Gating.** After a 90° rotation, the agent obtains a new observation. Using the historical FoV coverage (yellow cells) and the coverage from the new observation (blue cells), we calculate the overlap (green cells) to decide whether to save this observation.

Global Relational and Local Visual Cues. Globally, we represent the environment as a hierarchical 3D scene graph G comprising *Apartment*, *Room*, and *Object* levels. We first build upon ConceptGraphs [30] to maintain object nodes, and then insert room nodes through morphological processing of the top-down projected occupancy map. Although this graph provides a highly interpretable and structured topological map, it inevitably omits fine-grained visual textures and high-fidelity appearance details. To compensate for this, we maintain a keyframe image bank I as a repository for high-resolution egocentric observations. The geometric

cues described below determine the timing and selection of image storage to ensure that I remains both compact and representative.

Geometric Cues. Unlike previous memory designs, we introduce FoV coverage (shown in Fig. 2) as a fundamental geometric cue. Based on this cue, we propose a geometric gating mechanism to maintain a compact image memory. We contend that visual observations should only be stored when they yield substantial geometric expansion, thereby enriching the memory with sufficiently novel spatial information. We quantify geometric novelty as the proportion of the current view that encompasses previously unobserved spatial regions. Specifically, given the current FoV coverage C_{new} (the blue boundary region in Fig. 4) and the accumulated historical coverage $C_{history}$ (the yellow boundary region in Fig. 4), the novelty score τ_{novel} is computed as:

$$\tau_{novel} = 1 - \frac{|C_{new} \cap C_{history}|}{|C_{new}|}. \tag{1}$$

At step t, if τ_{novel} exceeds a predefined threshold τ_0, the system archives this image in the image bank I_t and updates the coverage map via $C_t = C_{history} \cup C_{new}$. This union incorporates the newly observed cells (blue cells in Fig. 4). This strategy naturally leads to memory saturation once the environment is fully explored, effectively filtering redundant views while preserving a representative visual history for sequential tasks.

Beyond filtering, FoV coverage serves as a geometric bridge to establish bidirectional links between keyframe images and 3D graph nodes. Specifically, each saved image in I is associated with the object nodes within its visibility frustum. Conversely, each object node stores references to all images in which it is captured. This cross-modal indexing enables seamless transitions between global and local cues, synthesizing them into a unified memory architecture.

Finally, the FoV coverage provides a quantitative metric for exploration progress. By intersecting the historical FoV coverage $C_{history}$ with the whole area of a specific room S_{room} (*e.g.*, the area of Room 1 enclosed by the blue boundary in Fig. 5), we derive the coverage rate η to represent the explored proportion of this room:

$$\eta = \frac{|C_{history} \cap S_{room}|}{|S_{room}|}. \tag{2}$$

This metric provides grounded geometric evidence for the agent to assess each room's exploration status and serves as a termination criterion for task completion.

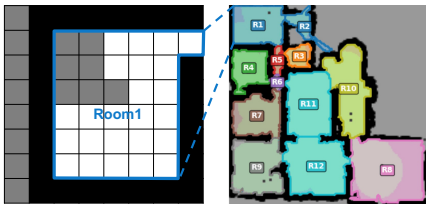

Figure 5: **Top-down Projection of Room Coverage.** The blue boundary defines the area of Room 1, while white cells mark covered regions $\mathrm{C_{history}}$. The room coverage ratio is calculated as the fraction of the total room area $\mathrm{S_{room}}$ (cells within the boundary) that has been covered by the agent.

Unified Spatial Memory. Together, these cues form the unified spatial memory at step t, defined as $\mathcal{M}_t = f_{\mathrm{3DSPMR}}(G_t, I_t, C_t)$, where the global 3D scene graph G_t encodes symbolic structure, the local image bank I_t provides perceptual grounding, and the FoV coverage map C_t maintains geometric context. At each step t, the agent acquires RGB-D observations from its current pose to perceive the local environment. These data are subsequently integrated into the unified spatial memory, as illustrated in Fig. 6. The system first computes the geometric novelty score τ_{novel} (Eq. 1) to decide whether to update C_t and to archive the observation frame in I_t. If the frame is stored, the scene graph is updated accordingly, and bidirectional links are established between the new image and the relevant object nodes. By synchronizing these heterogeneous cues, 3DSPMR bridges the gap in 3D scene understanding for MLLMs, providing a robust foundation for downstream reasoning.

3.2 Geo-Reasoning

Building on the unified spatial memory, we introduce a Geo-Reasoning module designed to utilize multi-modal cues for task execution and infeasibility detection. Although MLLMs exhibit remarkable perceptual and reasoning capabilities when processing text and image inputs, they are prone to reaching premature conclusions in partially explored environments and falling into repetitive loops when facing infeasible tasks. Unlike prior approaches [3, 15, 24] that rely on model-based confidence heuristics, which often suffer from calibration issues, our Geo-Reasoning module leverages model-agnostic geometric information (*i.e.*, FoV coverage) to provide a grounded, external verification.

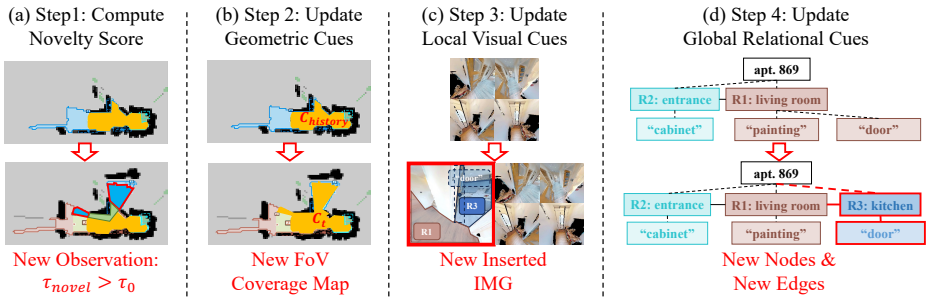

Figure 6: **Unified Spatial Memory Update.** The update process comprises four steps. (a) Step 1: compute the novelty score τ_{novel} of the new observation. If $\tau_{novel} > \tau_0$, update memory in later steps. (b) Step 2: update geometric cues by merging the historical coverage map with the new image's FoV coverage. (c) Step 3: update local visual cues by saving this image into the image bank. (d) Step 4: update global relational cues by updating the graph and establishing bidirectional links between the graph and the image.

Specifically, we introduce a Geometric Examination Mechanism to perform post-hoc validation of MLLM predictions. This mechanism leverages the cumulative coverage map C stored in the spatial memory to assess exploration completeness. If the coverage rate η (Eq. 2) of all task-relevant rooms exceeds a predefined threshold, the inference is deemed grounded and accepted. Conversely, if underexplored relevant regions persist, the mechanism rejects the inference and triggers further exploration. For accepted cases where a definitive answer is provided, the agent commits to it as the final prediction. Otherwise, the mechanism examines the coverage of irrelevant rooms to identify any unobserved yet potentially relevant areas. If no unexplored areas remain, the task is formally classified as infeasible. If unexplored regions persist, the system initiates Geo-Sem exploration to gather additional information. By bridging geometric occupancy with semantic reasoning, this mechanism effectively prevents premature task termination, detects infeasible tasks, and mitigates redundant revisiting of explored space. Details and visual illustrations of the reasoning process are provided in the Appendix.

3.3 Geo-Sem Exploration

In scenarios necessitating exploration, the agent must identify the most informative unexplored region to visit next. Prior works [13, 41] primarily leverage visual semantics from frontier images for navigation; however, such a paradigm ignores critical geometric considerations, which may result in inefficient trajectories, especially when semantic signals alone fail to distinguish between multiple candidates.

The proposed Geo-Sem Exploration module addresses these limitations by integrating geometric cues with visual semantics, enabling the agent to holistically determine the next target region. At each exploration step, the occupancy map identifies a set of frontiers $\mathcal{F}_t$, defined as the boundary regions between explored (yellow cells) and unexplored unknown space (gray cells). For each frontier node $i \in \mathcal{F}_t$ (the purple point shown in Fig. 7), Geo-Sem leverages an MLLM r to assign a semantic relevance score $v_t^{\text{sem}}(i)$ conditioned on the task query Q and the frontier's visual context in node i, defined as: $v_t^{\text{sem}}(i) = r(i, Q)$. This score quantifies the potential of frontier node i to yield information crucial for fulfilling the current task.

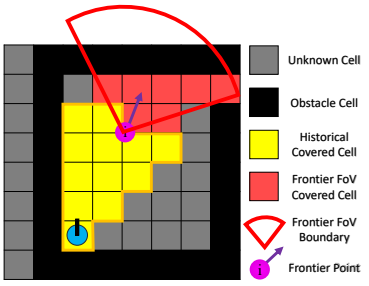

Figure 7: **Top-down Projection of Frontier Exploration FoV.** The red lines delineate the FoV projected ahead of a frontier node. Red cells indicate unknown cells within the FoV, whose area is denoted as U_t^{FoV} in Eq. 3.

Additionally, Geo-Sem utilizes two geometric heuristics to guide exploration. Most notably, frontier node i's FoV is used to evaluate the potential unobserved area beyond it. Based on this, Geo-Sem computes the exploration potential score $v_t^{\text{exp}}(i)$ as follows:

$$v_t^{\text{exp}}(i) = \frac{U_t^{\text{FoV}}(i)}{U_{\max}}, \tag{3}$$

where $U_t^{\mathrm{FoV}}(i)$ denotes potential explorable area (red cells in Fig. 7), and $U_{\max}$ denotes the largest possible explorable area (the area within the whole red boundary). By incentivizing the agent to prioritize regions with larger unexplored area, our strategy expedites the acquisition of a holistic environmental representation, effectively preventing suboptimal motion patterns such as repetitive oscillations or excessive detours. To favor proximal targets, Geo-Sem incorporates geometric proximity to the frontier as a secondary heuristic. This proximity score, $v_t^{\mathrm{dist}}(i)$, is formulated as: $v_t^{\mathrm{dist}}(i) = 1/(1+\|p_i - p_t\|/d_{\max})$, where p_t is the agent's current position, p_i is the frontier pos, and $d_{\max}$ serves as a distance normalization constant. This refinement enables the agent to select closer frontiers, particularly when the aforementioned heuristics yield comparable scores across multiple candidates.

Finally, these components are integrated into a unified composite score $V_t(i)$ as: $V_t(i) = w_{\mathrm{sem}}\, v_t^{\mathrm{sem}}(i) + w_{\mathrm{exp}}\, v_t^{\mathrm{exp}}(i) + w_{\mathrm{dist}}\, v_t^{\mathrm{dist}}(i)$, with weighting coefficients $(w_{\mathrm{sem}}, w_{\mathrm{exp}}, w_{\mathrm{dist}})$ normalized to sum to 1. The target frontier is selected by maximizing the composite score: $i^\star = \arg\max_{i \in \mathcal{F}_t} V_t(i)$.

Once the frontier $i^\star$ is selected, the agent navigates toward its location $p_{i^\star}$ using a shortest-path planner within the explored region. Such a multi-faceted integration enables the agent to comprehensively determine the next-best-step, significantly enhancing exploration efficiency through the synergy of semantic intent and geometric viability. In practice, we employ the pathfinder from Habitat-Sim [44] to compute such routes and to advance the agent by small steps at each simulation tick.

4 SEER Benchmark

Overview. Existing benchmarks [13, 3, 16] are mainly designed for single-task cases, which typically provide only one or several tasks per scenario and lack infeasible objectives, rendering them insufficient for a systematic evaluation of sequential embodied reasoning. To bridge this gap, we introduce the Sequential Embodied Exploration and Reasoning Benchmark (SEER-Bench). We define an exploration episode as the comprehensive process of exploring an environment and attempting all assigned tasks. By augmenting each scenario with additional annotations, SEER-Bench supports multi-task evaluation within a single episode and assesses an agent's ability to reuse spatial knowledge and detect infeasible objectives.

SEER-Bench incorporates two core components: Embodied Question Answering (EQA) and Embodied Multi-modal Navigation (EMN), across 60 scenarios developed using HM3DSem [45] within the Habitat simulator [44]. Each scenario comprises 5 to 10 sequential tasks, 40% of which are infeasible, providing a realistic foundation for long-horizon sequential embodied reasoning. Comprehensive details regarding annotations and data statistics are provided in the Appendix.

Evaluation Metrics. Let $\mathcal{E}$ denote the set of test episodes, where each episode $e \in \mathcal{E}$ comprises a sequence of N_e tasks. For EMN, the agent must navigate to a target goal for each individual task i. We define the success of task i using a binary indicator $S_i \in \{0, 1\}$. For each task i, let l_i denote the shortest path distance from the agent's initial position at the onset of the task to the goal, and let h_i represent the actual trajectory length traversed by the agent during that specific task.

Standard evaluation metrics, such as Success Rate ($SR = \frac{1}{N}\sum_{i=1}^{N} S_i$) and Success-weighted Path Length ($SPL = \frac{1}{N}\sum_{i=1}^{N} S_i \frac{l_i}{\max(h_i, l_i)}$) [46], where $N = \sum_{e \in \mathcal{E}} N_e$ denotes the total number of tasks, effectively measure the correctness and efficiency of isolated single tasks. However, they fail to capture task interdependencies or the efficiency of spatial knowledge reuse across an episode. While one could simply average the SR over tasks within an episode, such an approach often obscures critical failures within the episode. To address these limitations, we introduce two new metrics that impose strict requirements on long-horizon performance: Sequential Success Rate (SSR) for sequential correctness and Sequential SPL (SSPL) for sequential efficiency.

SSR measures the reliability of an agent across the entire episode. Unlike the average SR, which may award partial credit for completing individual steps, SSR serves as a strict binary metric for the entire episode. We define the success of an episode e using a binary indicator $SR_e \in \{0, 1\}$, which equals 1 if and only if every task within that episode is successfully completed ($S_i = 1$ for all $i \in \{1, \ldots, N_e\}$). Thus, SSR reflects the proportion of total episodes $\mathcal{E}$ that achieve full task completion: $\mathrm{SSR}_{\mathrm{EMN}} = \frac{1}{|\mathcal{E}|}\sum_{e \in \mathcal{E}} \mathbb{I}[\mathrm{SR}_e = 1] \times 100\%$.

SSPL evaluates how efficiently an agent utilizes its total step budget to complete an entire episode. Crucially, this metric serves as a proxy for knowledge reuse: in sequential tasks, maximizing efficiency requires the agent to recall previously explored areas to accelerate subsequent goals, rather than redundantly re-exploring the environment. Let H_e denote the total steps taken by the agent during episode e, and let L_e represent the step budget allocated for that episode. SSPL rewards agents that achieve complete sequential success ($SR_e = 1$) while conserving their step budget: $\text{SSPL}_{\text{EMN}} = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \left(\mathbb{I}[SR_e = 1] \times \frac{L_e}{\max(H_e, L_e)} \right) \times 100\%$.

For EQA, we also introduce SSR and SSPL to evaluate model performance over a full episode. Their formulas, along with the setting of the step budget L_e, are provided in the Appendix.

5 Experiment

Setup. Using SEER-Bench, we evaluated an agent's capacity for multi-step reasoning and active exploration across both the EQA and EMN tracks. For experimental setup, we constructed a testing set by sampling 5 tasks from the annotated tasks in each of the 60 scenarios. Specifically, for each scenario, we randomly selected three feasible and two infeasible tasks. These five tasks were then shuffled to create a sequential task chain, resulting in a total of 300 tasks in each track. The EQA track maintains a wide diversity of question types which include functional, counting, and spatial relationships, and incorporates both answerable and unanswerable queries to test the agent's discernment. The EMN track features diverse goal modalities, requiring the agent to navigate based on Object, Language, or Image targets.

Baselines. We benchmarked our approach against several representative methods across the EQA and EMN tracks to validate the efficacy of 3DSPMR. Specifically, we compared our approach with four state-of-the-art memory-enhanced active exploration methods: Explore-Mem [3], 3D-Mem [23], GraphEQA [24], and PredEQA [47]. To ensure a fair comparison, all baselines utilize Qwen2.5VL-72B [48] as their default backbone across both tracks. To isolate the influence of geometric information within our proposed 3DSPMR, we conducted an ablation study (w/o Geo) by removing the FoV-based geometric examination mechanism during reasoning and the geometric heuristics during exploration. For the EQA track, we additionally incorporated Blind LLMs, specifically Qwen3-14B [49] and GPT-5 [50], to serve as a lower bound by attempting to answer questions without active navigation. In the EMN track, we further evaluated two distinct architectures: Modular GOAT [36], a zero-shot modular baseline lacking persistent memory, and SenseAct-NN [17], a monolithic RL baseline. For a rigorous and robust assessment, we benchmarked the QA performance across multiple LLM evaluators, specifically GPT-5 [50], Gemini-3.6-Flash [51], and LLaMA-3.3-70B [52].

Table 1: SEER-Bench EQA Results.

	GPT-5				Gemini-3.6-Flash				LLaMA-3.3-70B			
Method	SSR ↑	SSPL ↑	SR ↑	SPL ↑	SSR ↑	SSPL ↑	SR ↑	SPL ↑	SSR ↑	SSPL ↑	SR ↑	SPL ↑
Blind LLMs												
Qwen3	0.7	\	26.0	\	0.3	\	26.2	\	0.7	\	28.9	\
GPT-5	1.8	\	26.6	\	1.7	\	27.5	\	1.5	\	28.3	\
Open-Sourced Qwen2.5VL Exploration												
Explore-Mem	0.3	0.2	22.1	8.5	0.2	0.1	22.1	8.6	0.2	0.1	23.2	8.8
GraphEQA	1.5	0.4	19.8	9.1	1.2	0.3	20.0	9.4	2.2	0.7	32.4	19.1
3D-Mem	1.2	0.6	31.2	19.5	2.7	1.5	33.1	21.0	3.2	1.9	35.2	21.9
PredEQA	2.3	2.1	33.5	22.7	4.2	3.4	35.7	25.1	5.7	4.2	40.3	28.2
3DSPMR w/o Geo	5.8	4.9	39.1	27.7	6.3	5.4	40.8	29.2	8.8	7.6	46.6	33.1
3DSPMR (Ours)	**14.5**	**12.9**	**48.8**	**41.3**	**12.8**	**11.4**	**49.1**	**41.6**	**16.2**	**14.4**	**51.9**	**44.8**

Discussion and Analysis. We present a quantitative analysis of SEER-Bench results, summarized in Table 1 (EQA) and Table 2 (EMN). 3DSPMR consistently outperforms all baselines across every metric on both tracks. On the EQA track, the results remain consistent across different LLM judges, proving that our method's superiority is robust to the choice of evaluator. Notably, our framework yields gains of approximately 10 to 20 percentage points on the rigorous SSR and SSPL metrics over 3D-Mem [23] across both EQA and EMN tracks.

Standard baselines struggle with the sequential demands of our benchmark. In particular, Explore-Mem [3] confines memory to the current task, precluding the reuse of historical context for subsequent

queries. Although 3D-Mem [23], GraphEQA [24], and PredEQA [47] maintain historical memory, they are still inferior to 3DSPMR in both task-level (SR, SPL) and episode-level (SSR, SSPL) metrics. This underperformance is due to their inability to both explicitly detect infeasible objectives and prevent premature task completion.

In contrast, the superior SR and SPL scores demonstrate that 3DSPMR is more resilient and efficient in challenging environments where infeasible queries are present. Furthermore, the significant gains in SSR and SSPL underscore the 3DSPMR's robust capability for spatial knowledge reuse. By leveraging a unified spatial memory alongside FoV-guided reasoning, 3DSPMR seamlessly retrieves prior observations and minimizes redundant exploration. As a result, our approach more effectively addresses the fundamental challenges of long-horizon, sequential embodied tasks.

Table 2: SEER-Bench EMN Results.

Method	SSR ↑	SSPL ↑	SR ↑	SPL ↑
GOAT-Bench Baselines				
Modular GOAT	0.0	0.0	18.0	10.6
SenseAct-NN	0.0	0.0	29.3	24.5
Open-Sourced Qwen2.5VL Exploration				
Explore-Mem	1.7	0.4	38.7	32.0
GraphEQA	1.7	0.8	35.0	20.3
3D-Mem	5.0	4.7	52.3	41.6
PredEQA	1.7	0.8	38.3	22.2
3DSPMR w/o Geo	5.0	3.3	47.3	42.7
3DSPMR (Ours)	**25.0**	**15.5**	**68.0**	**59.2**

Single Task. Furthermore, we evaluated 3DSPMR in the standard single-task scenario to verify its performance. Notably, this traditional setting can be viewed as a degenerated version of our proposed sequential setting. Our setting involves multiple tasks per episode with a mix of feasible and infeasible queries, while the traditional setting is limited to a single, always-feasible task per episode.

Table 3: Single-Task Experiments on HM-EQA.

Method	SR ↑	SPL ↑
Explore-Mem	30.6	24.5
3D-Mem	30.8	25.6
3DSPMR (Ours)	**35.8**	**27.9**

Specifically, we conducted EQA experiments on the HM-EQA dataset [3]. We compared 3DSPMR with two active exploration baselines, Explore-Mem [3] and 3D-Mem [23], using Qwen2.5VL-72B [48] as the backbone. Following established evaluation protocols [13], we reported SR and SPL. As shown in Table 3, 3DSPMR consistently outperforms all baselines even in these simplified single-task cases. These results highlight that our method excels not only in complex sequential scenarios but also generalizes robustly to standard paradigms.

Ablation Study. To investigate the individual and synergistic contributions of our geometric components, we conducted a fine-grained ablation study on the SEER-Bench EMN track. The "Geo" mechanism represents the unified integration of both FoV-Reasoning (grounded geometric examination) and FoV-Exploration (geometric heuristics). By systematically removing these modules, we quantified their impacts on the agent's ability to reuse spatial knowledge and detect task infeasibility in the long-horizon sequential setting. The experimental results are shown in Table 4.

Table 4: We conducted a detailed ablation study on EMN track. (1) w/o FoV-Reason, (2) w/o FoV-Explore, (3) w/o Geo (with neither FoV-Reason nor FoV-Explore), and (4) the full 3DSPMR, which integrates both modules and achieves peak performance across all metrics. Bold indicates the best performance.

Method	SSR ↑	SSPL ↑	SR ↑	SPL ↑
3DSPMR w/o FoV-Reason	8.3	5.2	56.3	49.6
3DSPMR w/o FoV-Explore	11.7	9.2	53.7	43.0
3DSPMR w/o Geo	5.0	3.3	47.3	42.7
3DSPMR	**25.0**	**15.5**	**68.0**	**59.2**

Notably, the "w/o Geo" method, which removes both FoV-based reasoning and exploration, shows the most substantial performance drop across every metric. Disabling the geometric examination mechanism in reasoning results in an SSR of 8.3% (a decrease of 16.7 percentage points). Removing the geometric exploration heuristics drops the SSPL from 15.5% to 9.2%. This loss leads to repeated exploration, a behavior where the agent lacks the geometric incentives to prioritize underexplored regions and instead moves redundantly within previously explored space. This validates that explicit geometric modeling is essential for maintaining spatial consistency and enabling efficient knowledge reuse across a continuous sequence of tasks. With both geometric reasoning and exploration, our full 3DSPMR method overcomes both premature termination and repeated exploration problems,

Table 5: SEER-Bench EQA Results over Long Horizons.

Method	GPT-5				Gemini-3.6-Flash				Llama-3.3-70B			
	SSR ↑	**SSPL ↑**	**SR ↑**	**SPL ↑**	**SSR ↑**	**SSPL ↑**	**SR ↑**	**SPL ↑**	**SSR ↑**	**SSPL ↑**	**SR ↑**	**SPL ↑**
Blind LLMs												
Qwen3-14B	0.0	\	28.5	\	0.0	\	27.9	\	0.0	\	30.6	\
GPT-5	0.0	\	23.1	\	0.0	\	24.0	\	0.0	\	25.2	\
Open-Sourced Qwen2.5VL Exploration												
Explore-Mem	0.0	0.0	16.9	5.0	0.0	0.0	15.0	5.0	0.0	0.0	17.1	5.2
GraphEQA	0.0	0.0	23.5	12.0	0.0	0.0	20.4	11.3	0.0	0.0	26.2	13.8
3D-Mem	1.7	0.4	35.8	27.8	0.4	0.2	35.0	27.6	5.4	1.3	41.5	31.8
PredEQA	0.0	0.0	32.3	24.1	0.0	0.0	29.2	22.1	0.0	0.0	37.1	27.8
3DSPMR (Ours)	**10.4**	**9.5**	**49.0**	**45.3**	**1.3**	**1.2**	**42.3**	**39.3**	**10.4**	**9.8**	**52.9**	**48.5**

therefore showing the best overall performance. Detailed analyses regarding hyperparameter settings are presented in the Appendix.

Long Horizon. To assess performance over a longer horizon, we conducted additional experiments using all 10 tasks per environment, rather than the 5 tasks used in the main experiments. We evaluated 3DSPMR and the baseline methods across 12 scenarios in the EQA track, adhering to the same evaluation protocol and metrics as in the main experiments.

As shown in Table 5, the results reveal two main observations. First, even under the more challenging setting with 10 tasks per environment, 3DSPMR still consistently outperforms all baseline methods across all evaluation settings. Second, in several evaluation cases, we observe that the gaps between the correctness metrics and their corresponding efficiency metrics become smaller as the task sequence grows longer, i.e., the gaps between SR and SPL and between SSR and SSPL. For example, while the gap between SR and SPL nearly reaches 10 percentage points under the 5-task setup, it generally stays within 5 percentage points under the 10-task setup. This trend suggests that memory reuse becomes more beneficial over longer task sequences. By retaining spatial information acquired from earlier tasks, 3DSPMR can answer subsequent queries and plan exploration with less repeated exploration of previously visited regions.

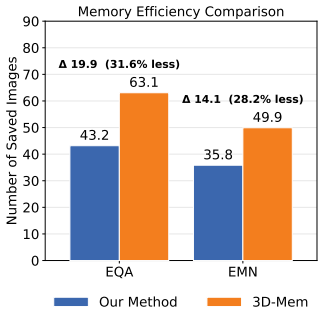

Figure 8: **Memory Efficiency Comparison.** On both tracks of SEER-Bench, 3DSPMR saves significantly fewer images than 3D-Mem in the memory.

Memory Efficiency. To validate the compactness of the unified spatial memory in 3DSPMR, we compared the memory efficiency with 3D-Mem [23]. Specifically, we averaged the number of saved images in each episode for experiments on SEER-Bench. As shown in Fig. 8, 3DSPMR reduces image memory usage by 31.6% on EQA and 28.2% on EMN. 3DSPMR stores substantially fewer images than 3D-Mem which retains observations based on object-clustering. The reduction arises from our FoV-driven gating mechanism, which filters redundant observations while retaining essential visual information. This is an important capability for avoiding memory saturation in long-horizon sequential tasks. Therefore, beyond reasoning and exploration improvements, our 3DSPMR features a highly efficient and compact memory system.

6 Conclusion

We propose 3DSPMR, an approach that augments MLLMs with a unified 3D spatial memory integrating explicit geometric FoV coverage, to address challenges in sequential embodied tasks. This geometric grounding effectively enhances spatial reasoning and optimizes exploration efficiency. Additionally, we introduce SEER-Bench, a rigorous framework to comprehensively evaluate embodied agents on sequential, feasibility-aware tasks. Experiments demonstrate that 3DSPMR achieves state-of-the-art performance in sequential embodied tasks, delivering superior correctness and efficiency. Future work will focus on extending this approach to dynamic environments and facilitating sim-to-real transfer.

7 Acknowledgement

This research was partially supported by the Army Research Office (ARO) grant W911NF-24-1-0385, the Office of Naval Research (ONR) grant N00014-23-1-2417, and DARPA under Award No. HR00112490426. The views, opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and should not be interpreted as representing the official policies or views, either expressed or implied, of ARO, ONR, or DARPA.

References

[1] Yifan Li, Yuhang Chen, Anh Dao, Lichi Li, Zhongyi Cai, Zhen Tan, Tianlong Chen, and Yu Kong. Industryeqa: Pushing the frontiers of embodied question answering in industrial scenarios. *arXiv preprint arXiv:2505.20640*, 2025.

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

[3] Allen Z Ren, Jaden Clark, Anushri Dixit, Masha Itkina, Anirudha Majumdar, and Dorsa Sadigh. Explore until confident: Efficient exploration for embodied question answering. *arXiv preprint arXiv:2403.15941*, 2024.

[4] Ronghao Dang, Yuqian Yuan, Wenqi Zhang, Yifei Xin, Boqiang Zhang, Long Li, Liuyi Wang, Qinyang Zeng, Xin Li, and Lidong Bing. Ecbench: Can multi-modal foundation models understand the egocentric world? a holistic embodied cognition benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24593–24602, 2025.

[5] Xiongkun Linghu, Jiangyong Huang, Xuesong Niu, Xiaojian Shawn Ma, Baoxiong Jia, and Siyuan Huang. Multi-modal situated reasoning in 3d scenes. *Advances in Neural Information Processing Systems*, 37:140903–140936, 2024.

[6] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6659–6668, 2019.

[7] Hang Yin, Xiuwei Xu, Zhenyu Wu, Jie Zhou, and Jiwen Lu. Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation. *Advances in neural information processing systems*, 37:5285–5307, 2024.

[8] Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171*, 2020.

[9] Jiazhao Zhang, Liu Dai, Fanpeng Meng, Qingnan Fan, Xuelin Chen, Kai Xu, and He Wang. 3d-aware object goal navigation via simultaneous exploration and identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6672–6682, 2023.

[10] Naoki Yokoyama, Ram Ramrakhya, Abhishek Das, Dhruv Batra, and Sehoon Ha. Hm3d-ovon: A dataset and benchmark for open-vocabulary object goal navigation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5543–5550. IEEE, 2024.

[11] Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. Instructnav: Zero-shot system for generic instruction navigation in unexplored environment. *arXiv preprint arXiv:2406.04882*, 2024.

[12] Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7641–7649, 2024.

[13] 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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16488–16498. IEEE, 2024.

[14] Yong Zhao, Kai Xu, Zhengqiu Zhu, Yue Hu, Zhiheng Zheng, Yingfeng Chen, Yatai Ji, Chen Gao, Yong Li, and Jincai Huang. Cityeqa: A hierarchical llm agent on embodied question answering benchmark in city space. *arXiv preprint arXiv:2502.12532*, 2025.

[15] 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. *arXiv preprint arXiv:2503.11117*, 2025.

[16] Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, et al. Robothor: An open simulation-to-real embodied ai platform. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3164–3174, 2020.

[17] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16373–16383, 2024.

[18] Xinshuai Song, Weixing Chen, Yang Liu, Weikai Chen, Guanbin Li, and Liang Lin. Towards long-horizon vision-language navigation: Platform, benchmark and method. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12078–12088. IEEE, 2025.

[19] Aaron Hurst et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[20] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[21] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025.

[22] Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. *arXiv preprint arXiv:2503.01773*, 2025.

[23] Yuncong Yang, Han Yang, Jiachen Zhou, Peihao Chen, Hongxin Zhang, Yilun Du, and Chuang Gan. 3d-mem: 3d scene memory for embodied exploration and reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17294–17303, 2025.

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

[25] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[26] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[27] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *arXiv preprint arXiv:2305.15021*, 2023.

[28] Dylan Goetting, Himanshu Gaurav Singh, and Antonio Loquercio. End-to-end navigation with vision-language models: Transforming spatial reasoning into question-answering. In *Proceedings of the International Conference on Neuro-symbolic Systems*, volume 288 of *Proceedings of Machine Learning Research*, pages 22–35. PMLR, 28–30 May 2025.

[29] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. Llava-plus: Learning to use tools for creating multimodal agents. In *European conference on computer vision*, pages 126–142. Springer, 2024.

[30] 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 *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028. IEEE, 2024.

[31] 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 (RSS)*, 2024.

[32] Nathan Hughes, Yun Chang, and Luca Carlone. Hydra: A real-time spatial perception system for 3d scene graph construction and optimization. *arXiv preprint arXiv:2201.13360*, 2022.

[33] Nathan Hughes, Yun Chang, Siyi Hu, Rajat Talak, Rumaia Abdulhai, Jared Strader, and Luca Carlone. Foundations of spatial perception for robotics: Hierarchical representations and real-time systems. *The International Journal of Robotics Research*, 43(10):1457–1505, 2024.

[34] 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. *arXiv preprint arXiv:2307.06135*, 2023.

[35] Yijun Yang, Tianyi Zhou, Kanxue Li, Dapeng Tao, Lusong Li, Li Shen, Xiaodong He, Jing Jiang, and Yuhui Shi. Embodied multi-modal agent trained by an llm from a parallel textworld. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[36] Matthew Chang, Theophile Gervet, Mukul Khanna, Sriram Yenamandra, Dhruv Shah, So Yeon Min, Kavit Shah, Chris Paxton, Saurabh Gupta, Dhruv Batra, et al. Goat: Go to any thing. *arXiv preprint arXiv:2311.06430*, 2023.

[37] R. Liu et al. Volumetric environment representation for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[38] Yi Du, Taimeng Fu, Zhuoqun Chen, Bowen Li, Shaoshu Su, Zhipeng Zhao, and Chen Wang. Vl-nav: Real-time vision-language navigation with spatial reasoning. *arXiv preprint arXiv:2502.00931*, 2025.

[39] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10, 2018.

[40] Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L Berg, and Dhruv Batra. Multi-target embodied question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6309–6318, 2019.

[41] Hang Yin, Xiuwei Xu, Linqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. Unigoal: Towards universal zero-shot goal-oriented navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[42] Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[43] Shuo Wang, Yongcai Wang, Wanting Li, Xudong Cai, Yucheng Wang, Maiyue Chen, Kaihui Wang, Zhizhong Su, Deying Li, and Zhaoxin Fan. Aux-think: Exploring reasoning strategies for data-efficient vision-language navigation. *arXiv preprint arXiv:2505.11886*, 2025.

[44] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in neural information processing systems*, 34:251–266, 2021.

[45] Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, et al. Habitat-matterport 3d semantics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4927–4936, 2023.

[46] 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.

[47] Bowen Yuan, Sisi You, and Bing-Kun Bao. Predict before you explore: Predictive planning with specialized memory for embodied question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 29610–29619, 2026.

[48] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[49] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[50] OpenAI. Gpt-5 technical overview. `https://openai.com/index/introducing-gpt-5/`, 2025.

[51] Google DeepMind. Gemini 3.6 flash model card. `https://deepmind.google/models/model-cards/gemini-3-6-flash/`, 2026.

[52] Meta. Llama 3.3 70B Instruct. `https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct`, 2024.

[53] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16901–16911, 2024.

[54] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, and et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

A Details of SEER-Bench

We propose a new Sequential Embodied Exploration and Reasoning Benchmark (SEER-Bench) to evaluate agent performance on sequential embodied tasks in section 4 of the main paper. Here, we supplement more details.

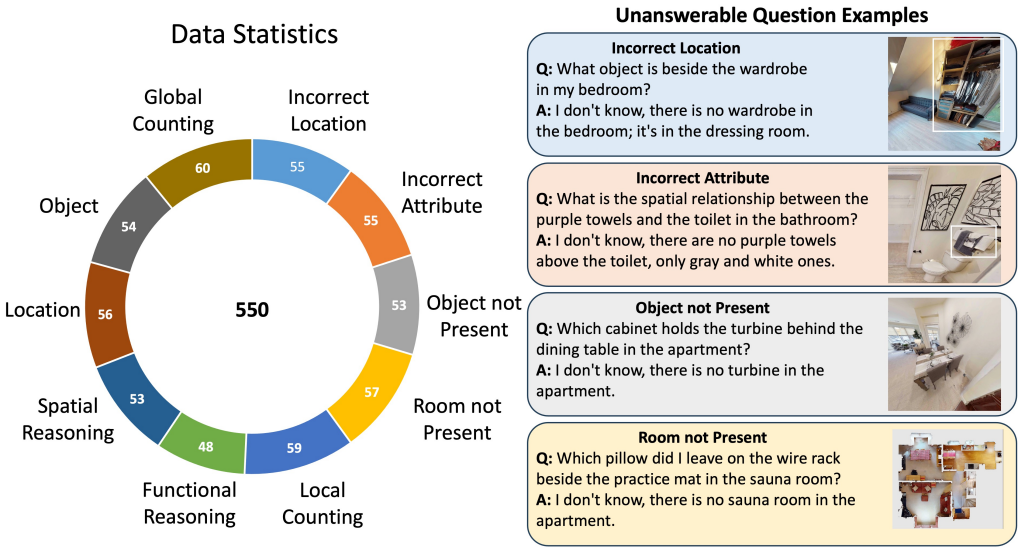

Figure 9: Data statistics and Unanswerable Question Examples of EQA in SEER-Bench.

A.1 Simulator

HM3DSem [45] is a large-scale dataset providing high-fidelity 3D reconstructions of 3,100 rooms with dense semantic annotations for objects, materials, and room structures. We adopted HM3DSem within the Habitat simulator [44] to construct photorealistic and semantically consistent embodied environments. Habitat supports high-performance egocentric rendering, efficient physics-based navigation, and modular interfaces for defining sensors, agents, and tasks. We leveraged these capabilities to create sequential task episodes that maintain consistent spatial layouts across tasks, enabling reproducible evaluation of spatial reasoning, memory reuse, and feasibility awareness in both Embodied Question Answering (EQA) and Embodied Multi-modal Navigation (EMN) settings. In this work, we did not include cross-floor tasks, treating each floor as a separate scenario instead.

A.2 Embodied Question Answering

As existing HM3DSem-related benchmarks [13, 15] do not provide sequential question–answer pairs for each scenario, we annotated 60 HM3DSem scenarios with 5 to 10 QA pairs for each scenario. The questions in each scenario span six normal categories: Object, Location, Spatial Reasoning, Functional Reasoning, Local Counting, and Global Counting. In EQA, infeasible tasks naturally correspond to unanswerable questions. To systematically capture such cases, we introduced four types of unanswerable questions: Incorrect Location (*e.g.*, erroneous room–object associations), Incorrect Attribute (*e.g.*, incorrect color or state of the mentioned object), Object not Present, and Room not Present. For each scenario, we constructed a QA set with at most one question from each answerable and unanswerable category. Data statistics and examples of unanswerable questions are provided in Fig. 9.

To support data annotation, we collected environmental information from each selected scenario. We implemented an autonomous exploration policy that, at each step, navigates the agent to the nearest frontier on the occupancy map, enabling systematic acquisition of ego-centric RGB observations. For each environment, the agent is initialized with two random feasible starting positions and camera orientations to increase trajectory diversity and improve spatial coverage.

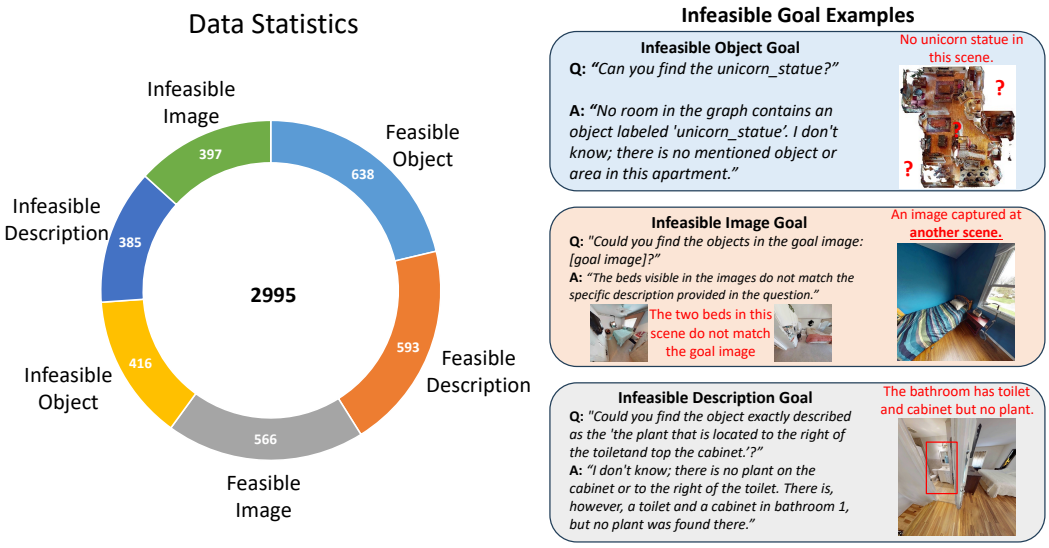

Figure 10: Data statistics of the EMN track in SEER-Bench and infeasible goal examples.

Using these collected frames, we employed GPT-5 [50] to generate most of the QA annotations, followed by manual filtering and correction to ensure correctness. For global counting questions, we obtained ground-truth answers by manually inspecting the layout figures or referencing the metadata. For unanswerable questions, GPT-5 is used to annotate cases of Incorrect Location and Incorrect Attribute. For the Object Not Present and Room Not Present categories, we designed two additional prompts that guide GPT-5 to generate questions involving objects (*e.g.*, carriage) or rooms (*e.g.*, sauna room) that are rarely found in typical residential environments. Furthermore, we recorded some unique objects and rooms appearing in each scenario and intentionally reused them in other environments where they do not exist, enabling controlled comparisons across scenarios.

A.3 Embodied Multi-modal Navigation

The Embodied Multi-modal Navigation (EMN) task in SEER builds upon the GOAT-Bench [17], extending it into a sequential, feasibility-aware formulation. Each navigation task requires an embodied agent to navigate toward a target object defined by one of three goal modalities: (1) an *object goal*, specified by the target object category; (2) a *language goal*, described by natural language instructions; or (3) an *image goal*, provided as a visual exemplar. Within an episode, the agent must complete each navigation task sequentially, reusing spatial knowledge accumulated from previous tasks for efficient planning.

Each task specifies a target using one of the three modalities: *object goal*, *language goal*, or *image goal*. To evaluate feasibility awareness, we introduce three corresponding types of infeasible goals:

- Object-Goal Infeasibility (Incorrect Location). Following the EQA setup, we designed target object categories that do *not* exist in the scene. The agent must determine that the goal is impossible after sufficient exploration and terminate appropriately.
- Language-Goal Infeasibility (Incorrect Description). We modified original GOAT-Bench language instructions by altering key attributes or spatial relations. For example, an instruction such as "the white surfboard near a plant and TV" may be changed to "the black surfboard near a bed and oven", producing a linguistically coherent but environmentally impossible goal.
- Image-Goal Infeasibility (Cross-Scene Images). We used an image query taken from a different HM3DSem scene than the current episode. For instance, an exemplar from Scene A is provided for a navigation task in Scene B, making the visual reference unmatchable within the environment.

Consistent with the configuration in the EQA track, we included 60 scenarios from HM3DSem and assigned 40% of the goals as infeasible, ensuring a balanced distribution between feasible and infeasible tasks. As summarized in Fig. 10, the EMN track contains a total of 2995 multi-modal goals. The right side of the figure illustrates example cases for each of the three infeasible goal types.

A.4 Evaluation Metrics

Evaluation Metrics for EQA. As traditional evaluation metrics Success Rate (SR) and Success-weighted Path Length (SPL) fail to measure model performance for an episode, we introduce two new evaluation metrics: Sequential Success Rate (SSR) for sequential correctness and Sequential SPL (SSPL) for sequential efficiency in Section 4 of the main paper. Here, we provide the SSR and SSPL formulas for Embodied Question Answering (EQA).

In an EQA task, the agent is required to infer an answer for each question. Consequently, the binary success indicator $S_i \in \{0, 1\}$ for task i in EMN is replaced by an LLM-based correctness score $\sigma_i \in \{1, \ldots, 5\}$ for the inferred answer. In this scale, 1 indicates an incorrect response, 5 signifies a fully correct response, and intermediate values represent varying degrees of semantic similarity to the ground truth. We adopt the correctness and efficiency metrics defined in OpenEQA [13] as SR and SPL. To adapt the SSR for EQA, we consider a task successful if its score σ_i is larger than 3. Based on this, we generalize $\text{SSR}_{\text{EMN}} = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \mathbb{I}[\text{SR}_e = 1] \times 100\%$ to account for the nuanced scores:

$$\text{SSR}_{\text{EQA}} = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \frac{\max(0, \sum_{i=1}^{N_e} (\sigma_i - 3))}{2 \times N_e} \times 100\%, \tag{4}$$

where $\mathcal{E}$ is the set of test episodes, $e \in \mathcal{E}$ is a single episode comprising a sequence of N_e tasks. Similarly, the SSPL for EQA is adapted from the standard EMN formulation, $\text{SSPL}_{\text{EMN}} = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \left(\mathbb{I}[SR_e = 1] \times \frac{L_e}{\max(H_e, L_e)} \right) \times 100\%$, by reformulating the success condition as follows:

$$\text{SSPL}_{\text{EQA}} = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \left(\frac{\max(0, \sum_{i=1}^{N_e} (\sigma_i - 3))}{2 \times N_e} \times \frac{L_e}{\max(H_e, L_e)} \right) \times 100\%. \tag{5}$$

Task Step Limit and Episode Step Limit. In this paragraph, we specify the task step limit, episode step limit, and step budget L_e allocated for an episode e. As discussed in the benchmark description, SEER-Bench includes *small*, *medium*, and *large* indoor scenes. For the EQA track, we set the task step limits for these three scene sizes to 50, 200, and 320 steps, respectively. The *episode step limit* is defined as twice the corresponding task step limit. These limits approximate the number of steps required to fully explore each scene size using a closest-frontier exploration policy, ensuring adequate spatial coverage while preventing unnecessarily long trajectories. For the EMN track, we extended the step-limit strategy used in GOAT-Bench [17]. Specifically, we used Habitat's scene-loading utilities to obtain the spatial extent of each environment, assigned a size-dependent local step limit, and again set the episode limit to twice the task budget. For both the EQA and EMN tracks, we set the expected optimal step budget L_e to a specified proportion (strictly less than 1) of the episode step limit. In both the EQA and EMN tracks, tasks that exceed the step limit are categorized as infeasible. If the episode step limit is exceeded, the current task and all remaining tasks in the episode are marked as failed.

B Implementation

B.1 Geo-Reasoning Details

We introduce the Geo-Reasoning module in Section 3.2 of the main paper. Here, we provide a detailed description and a visual illustration of the reasoning pipeline. As shown in Fig. 11, Geo-Reasoning operates in two distinct phases: first, Reasoning Modules provide an initial inference; second, a Geometric Examination Mechanism performs post-reasoning validation. We utilize MLLMs in a training-free, hierarchical manner to interpret the multi-modal cues stored in memory.

For the EQA task, the MLLM reasoning process is executed through four specialized stages: Question Analysis decomposes the query into hierarchical semantic requirements; Global Reasoning prunes the 3D scene graph G to isolate task-relevant sub-graphs; Local Reasoning then retrieves the associated keyframes I indexed by these sub-graphs, further narrowing the visual search space to extract precise perceptual evidence; and Question Answering formulates the final response by synthesizing symbolic and visual evidence. This pipeline effectively distills the complex, multi-modal spatial memory into task-level decisions supported by structured evidence. These initial conclusions are then subjected to the Geometric Examination to ensure their geometric grounding. Depending on the extent of spatial exploration, this examination either validates or rectifies the initial outputs.

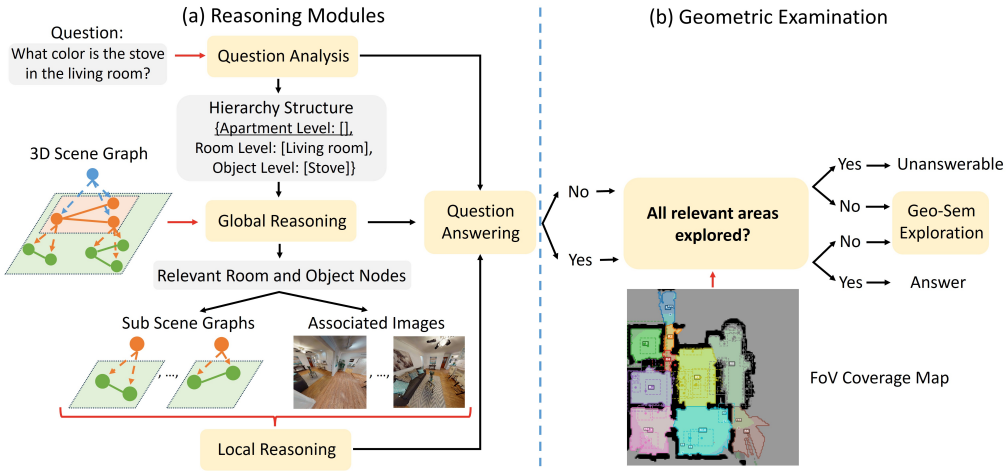

Figure 11: **Illustration of Geo-Reasoning in 3DSPMR for EQA tasks.** (a) Phase 1: Reasoning Modules leverage the graph and images in spatial memory to generate an initial inference. (b) Phase 2: Geometric Examination mechanism utilizes FoV coverage to assess task completion. If the relevant spatial regions are fully explored, the system either confirms the inferred answer or flags task infeasibility; otherwise, it triggers Geo-Sem Exploration to determine the next optimal exploration direction.

B.2 Hyperparameters

FoV-Based Image Gating. The unified spatial memory in 3DSPMR leverages a gating mechanism based on Field-of-View (FoV) to determine stored observation frames. In this mechanism, the geometric novelty score threshold τ_0 controls the minimum proportion of novel spatial coverage required for an observation to be committed to the image bank. An observation is only stored if the calculated novelty score (representing the ratio of newly observed cells to the total cells in the current view) exceeds the default threshold of 0.25 in EQA and 0.5 in EMN.

Geo-Sem Exploration Weights. In our Geo-Sem Exploration, the agent identifies the optimal frontier target by maximizing a composite score that integrates semantic relevance, geometric exploration potential, and proximity. We assigned a weight w_{sem} of 0.5 to semantic relevance to prioritize task-pertinent visual cues, a weight w_{dist} of 0.4 to geometric proximity to favor closer targets, and another weight w_{exp} of 0.1 to unexplored area potential to promote environmental coverage. To prevent trajectory oscillations, the agent only switches to a new frontier target if its composite score exceeds the current target's score by a switching threshold of 0.1.

B.3 Backbones and System Components

We used Qwen2.5VL-72B [48] as our default open-source MLLM backbone. To construct and maintain object-level nodes in the 3D scene graph, we build upon ConceptGraphs [30], which loads the following perception models: YOLO-World [53] for open-vocabulary object detection, SAM (Segment Anything Model) [54] for instance segmentation, CLIP [55] for visual embedding and category grounding. These components collectively provide reliable open-vocabulary detection, segmentation, and feature embedding, enabling robust and consistent object-node creation and updates throughout exploration.

B.4 Computing Resources

In this work, we used 8 GPUs for embodied agent reasoning. Specifically, we employed RTX A6000 or RTX 6000 Ada in all the experiments. With a single GPU, it on average took about 3 hours for one scenario with 5 sequential embodied tasks.

C More Experimental Results On Benchmark Datasets

Experimental Results on More Datasets. In Section 5 of the main paper, we benchmark various methods on our proposed SEER-Bench for both EQA and EMN tasks. Here, we supplement more experiments on two more existing active EQA datasets, HM-EQA [3] and OpenEQA [13]. Since these datasets are not designed for sequential tasks and do not contain infeasible tasks, we adapt these datasets to our sequential setting.

We identified scenarios containing multiple single tasks previously tested in isolation and merged them into sequential episodes. Following the SEER-Bench annotation protocol, we introduced unanswerable questions and randomly selected two infeasible tasks to append to these merged sequences. Through this process, we curated an additional 48 scenarios totaling 240 tasks from the HM-EQA dataset. Correspondingly, we extracted 12 scenarios comprising 60 tasks from OpenEQA for our testing suite.

We compared our proposed 3DSPMR method with two state-of-the-art memory-enhanced active exploration baselines: Explore-Mem [3] and 3D-Mem [23] with Qwen2.5VL-72B [48]. The comparative results on the adapted HM-EQA and OpenEQA datasets are presented in Tables 6 and 7, respectively. The results demonstrate that our 3DSPMR approach consistently surpasses the baseline methods across all evaluated tasks and metrics. Most notably, performance on the rigorous SSPL metric nearly quadruples that of the baselines. The superior performance of 3DSPMR is consistent with the observations on SEER-Bench, demonstrating its robust adaptability across various question formats. Taken together, these results highlight 3DSPMR as an efficient and effective framework for sequential tasks, particularly excelling in complex real-world cases that demand reasoning over both feasible and infeasible objectives.

Table 6: Experiments on adapted HM-EQA dataset.

Method	SSR ↑	SSPL ↑	SR ↑	SPL ↑
Blind LLMs				
Qwen3	0	\	20.7	\
GPT-5	0.8	\	19.9	\
Open-Sourced Qwen2.5VL Exploration				
Explore-Mem	0.8	0.2	22.0	9.2
3D-Mem	3.5	2.1	29.2	18.0
3DSPMR (Ours)	**10.8**	**9.7**	**46.2**	**38.8**

Table 7: Experiments on adapted OpenEQA dataset.

Method	SSR ↑	SSPL ↑	SR ↑	SPL ↑
Blind LLMs				
Qwen3	0	\	25.0	\
GPT-5	1.7	\	34.5	\
Open-Sourced Qwen2.5VL Exploration				
Explore-Mem	2.4	0.6	27.1	13.5
3D-Mem	0.8	0.4	27.1	18.4
3DSPMR (Ours)	**3.3**	**2.8**	**37.1**	**33.7**

Experimental Results with Different MLLMs. To assess how the choice of foundational MLLM affects 3DSPMR, we evaluated our method on SEER-Bench across 12 randomly sampled EQA scenarios, adhering to the same experimental setup and protocol as the main experiment. Specifically, we compared a closed-source model (GPT-5 [50]) against the open-source model used in our primary experiments (Qwen2.5VL-72B [48]). As Table 8 shows, the performance of 3DSPMR correlates directly with the capability of the underlying backbone, with GPT-5 yielding superior results. This trend highlights that our framework continuously benefits from advancements in MLLM reasoning capabilities.

Table 8: EQA Experiments using different MLLMs.

	GPT-5				Gemini-3.6-Flash				Llama-3.3-70B			
MLLM Base	**SSR ↑**	**SSPL ↑**	**SR ↑**	**SPL ↑**	**SSR ↑**	**SSPL ↑**	**SR ↑**	**SPL ↑**	**SSR ↑**	**SSPL ↑**	**SR ↑**	**SPL ↑**
Qwen2.5VL	5.0	4.6	43.3	34.2	5.8	5.3	44.2	35.6	8.3	7.9	46.3	35.9
GPT-5	**32.5**	**23.0**	**58.8**	**45.7**	**28.3**	**18.8**	**56.7**	**44.3**	**30.8**	**21.0**	**58.8**	**44.1**

D Additional Experiments and Analyses

D.1 Hyperparameter Analysis

In this section, we analyzed the impact of two important hyperparameters in our 3DSPMR approach.

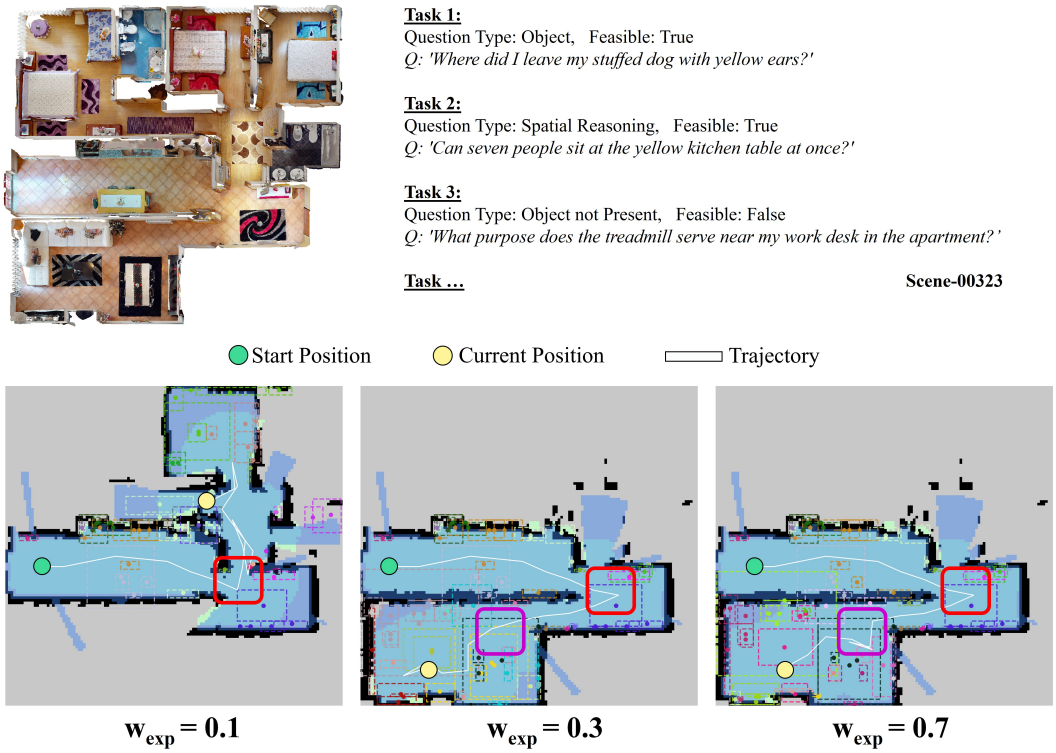

Figure 12: Agent Performance Analysis with Different w_{exp}. Increasing w_{exp} encourages the agent to prioritize regions containing extensive unexplored areas. When $w_{\text{exp}} \in \{0.3, 0.7\}$, the agent decides to explore the region with larger unexplored areas (illustrated within the red bounding boxes). When $w_{\text{exp}} = 0.7$, the agent is reluctant to fully explore this room and prefers to explore other underexplored rooms (illustrated within the purple bounding boxes).

FoV-Based Image Gating Threshold. We first investigated the influence of the Geometric Novelty Score Threshold τ_0 on the gating mechanism across 60 scenes from the SEER-Bench EMN track. In addition to the four metrics for success and efficiency (SSR, SSPL, SR, and SPL) in SEER-Bench, we introduced Average Number of Images (ANI) as a novel metric to evaluate memory footprint and storage efficiency. Specifically, we conducted a series of experiments using a set of threshold values $\tau_0 \in \{0.1, 0.25, 0.5, 0.8\}$. Experiment results shown in Table 9 reveal a clear trend: higher novelty thresholds τ_0 consistently reduce the ANI saved in our unified spatial memory.

Table 9: Experiments on the SEER-Bench EMN track with different Geometric Novelty Score Thresholds (τ_0).

Threshold	SSR ↑	SSPL ↑	SR ↑	SPL ↑	ANI ↓
0.1	18.3	11.4	65.3	57.1	52.3
0.25	18.3	12.8	62.0	54.3	38.4
0.5	**25.0**	**15.5**	**68.0**	**59.2**	25.2
0.8	21.7	12.5	66.0	56.8	**17.0**

Crucially, we observed that performance peaks at a moderate threshold, whereas extreme values in either direction lead to failure. Excess redundancy at $\tau_0 = 0.1$ (ANI = 52.3) limits sequential performance to 18.3% SSR, as a dense image bank likely introduces noise or exceeds the efficient reasoning capacity of the MLLM. Conversely, extreme sparsity at $\tau_0 = 0.8$ (ANI = 17.0) causes the SSR to drop to 21.7% because the memory lacks the critical spatial cues required to resolve task inter-dependencies and reuse knowledge across the episode. We identify $\tau_0 = 0.5$ as the optimal configuration, achieving peak sequential success (25.0% SSR, 15.5% SSPL, 68.0% SR, 59.2% SPL) with a representative memory density (ANI = 25.2) that balances storage efficiency with informative environmental representation.

Geo-Sem Exploration Weights. Then, we investigated the influence of Geo-Sem Exploration weights. The core advantage of our Geo-Sem Exploration module stems from its synergistic combina-

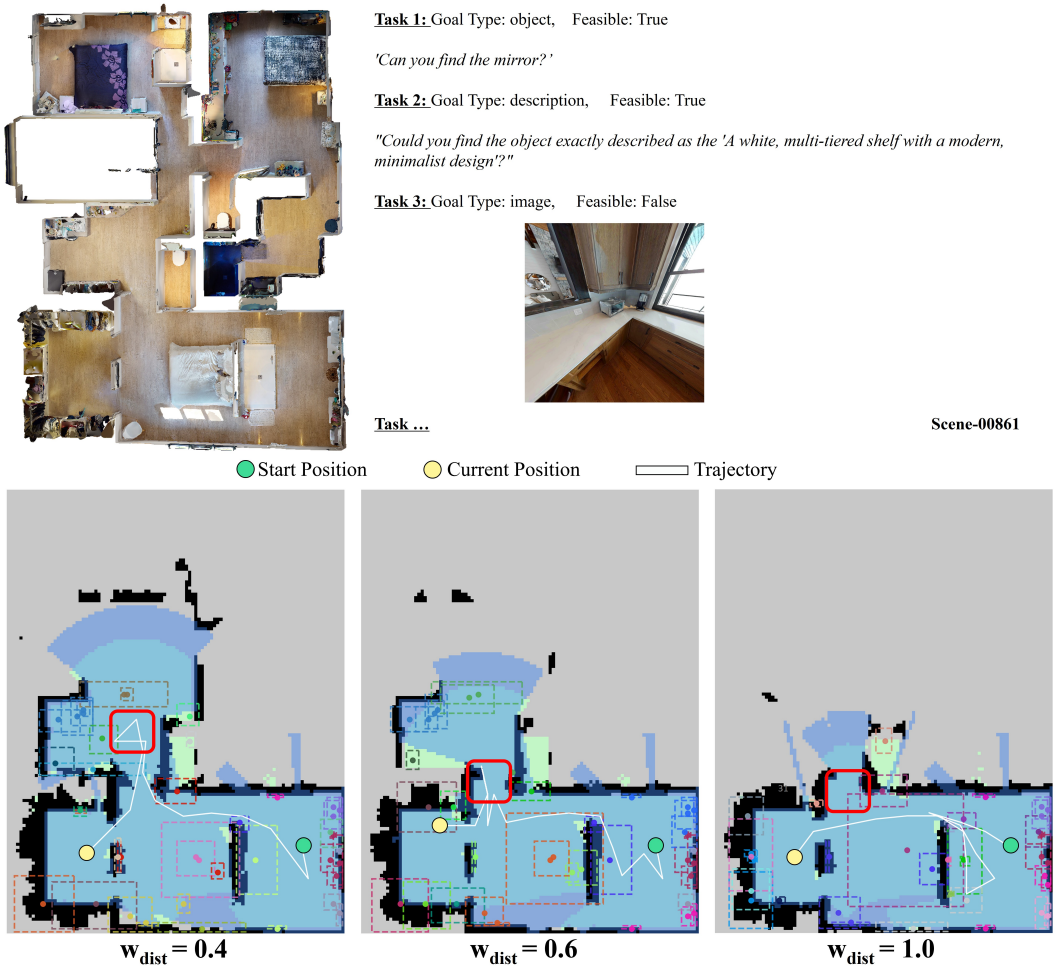

Figure 13: Agent Performance Analysis with Different w_{dist}. Increasing w_{dist} encourages the agent to prioritize proximal unexplored regions. In the highlighted red boxes, it could be observed that as w_{dist} increases, the agent becomes less inclined to navigate toward far-off regions, preferring to exhaust the closest unexplored spaces first.

tion of semantic guidance, geometric exploration potential, and path proximity. Relying on any single heuristic often leads to inefficient navigation, such as repetitive oscillations or excessive detours, but our weighted scoring dynamically balances these complementary signals to optimize the agent's trajectory. Therefore, we dynamically adjusted two geometric weights w_{exp} and w_{dist} within a single episode to visualize and analyze their influence on the agent's exploration strategy.

The FoV heuristic is weighted by w_{exp}, which encourages the agent to prioritize regions containing extensive unexplored space. We evaluated different values of $w_{\text{exp}} \in \{0.1, 0.3, 0.7\}$ and visualized their resulting trajectories within a representative scenario from the SEER-Bench EQA track. As illustrated in Fig. 12, assigning an excessively large weight w_{exp} overwhelms the MLLM's semantic scoring and makes the agent reluctant to fully explore one room, causing the agent to blindly prioritize large, unobserved spaces while ignoring smaller, highly task-relevant regions. The proximity heuristic is modulated by w_{dist} to bias the agent's selection toward proximal unexplored regions. Similarly, we evaluated different values of $w_{\text{dist}} \in \{0.4, 0.6, 1.0\}$ and visualized their resulting trajectories within a representative scenario from the SEER-Bench EMN track. As illustrated in Fig. 13, an overly large proximity weight w_{dist} downgrades the sophisticated multi-modal module into a naive, closest-frontier exploration strategy. This greediness not only strips away the agent's semantic foresight but also prevents it from navigating towards larger, more informative unexplored areas.

By employing the optimally calibrated weights, our system ensures that task-oriented semantic intent remains the primary driver of exploration, while the geometric heuristics effectively break ties to mitigate repetitive visiting, minimize excessive detours, and maximize overall trajectory efficiency.

D.2 Map Prior

For service robots operating in the same environment over extended periods, pre-exploring the environment is a natural and practical assumption. To investigate how map priors impact sequential tasks, we evaluate 3DSPMR under a pre-explored setting. Specifically, the unified spatial memory is initialized with a pre-explored room-level map containing room segmentation, connectivity, and spatial layouts. While this room-level structure remains fixed, object nodes are updated online to accommodate dynamic changes in object locations and states. We evaluate this setup across 12 EQA scenarios in SEER-Bench, adhering to the same protocol and metrics as the main experiments.

As shown in Table 10, incorporating map priors consistently boosts both task accuracy and navigation efficiency, as the agent directly exploits prior knowledge of room layouts. These results highlight 3DSPMR's effectiveness in pre-explored environments, while underscoring its broad compatibility across both map-free exploration and map-prior deployment settings.

Table 10: Effect of the map prior on SEER-Bench EQA Track.

	GPT-5				Gemini-3.6-Flash				Llama-3.3-70B			
Setting	**SSR** ↑	**SSPL** ↑	**SR** ↑	**SPL** ↑	**SSR** ↑	**SSPL** ↑	**SR** ↑	**SPL** ↑	**SSR** ↑	**SSPL** ↑	**SR** ↑	**SPL** ↑
w/o Map Prior	7.5	7.1	38.8	29.8	5.0	4.6	35.4	26.7	7.5	7.2	43.8	33.6
w/ Map Prior	**11.7**	**10.7**	**47.9**	**40.0**	**6.7**	**6.2**	**44.6**	**36.7**	**16.7**	**15.7**	**54.6**	**45.3**

E Limitations

While our 3DSPMR method organizes complex and extensive 3D information gathered during the long-horizon task execution into a compact and representative memory system, which facilitates MLLM's spatial understanding and reasoning about the surrounding scenario, it does not alter the model's intrinsic reasoning abilities. Consequently, a limitation is that the overall framework's performance is constrained by the capabilities of the foundational MLLM. As demonstrated in Table 8, 3DSPMR paired with the closed-source GPT-5 outperforms the open-source Qwen2.5VL on EQA tasks, showcasing this dependency. Another limitation is that our 3DSPMR method does not take dynamic environments into consideration. The premise in this work is that the surrounding environment is static, such as a residential apartment. This premise does not hold when applied to outdoor, city-level environments. In future work, we plan to (1) develop MLLM's internal spatial reasoning ability, and (2) integrate temporal update mechanisms to extend our framework into dynamic environments.

F Broader Impacts

In this paper, we propose a novel geometric memory system for long-horizon sequential embodied tasks. Extensive experiments validate the effectiveness of 3DSPMR in both EMN and EQA scenarios. Beyond navigation, our memory framework is highly extensible to other Vision-Language-Action (VLA) agents. For instance, in robotic manipulation, 3DSPMR can actively verify object existence and explicitly alert users to infeasible commands caused by missing targets. By preventing embodied agents from endlessly attempting impossible tasks, our reasoning pipeline significantly enhances operational safety, conserves computational and physical resources, and fosters greater human-robot trust in real-world deployments.

G Examples and Visualizations

G.1 Trajectory Comparison Between 3D-Mem and 3DSPMR on Sequential Tasks

Figures 14 and 15 provide qualitative comparisons of exploration behavior in sequential tasks. In the rightmost columns, the blue regions denote areas covered by FoV keyframes stored in memory, while the green regions correspond to frontier regions explored during the episode (the overlapped area looks like light blue). When trajectories are normalized to the same number of steps, 3DSPMR consistently covers a larger portion of previously unseen space, whereas 3D-Mem frequently revisits already-explored regions and exhibits oscillatory back-and-forth movements. These results highlight the advantages of our unified spatial memory and geometric reasoning in enabling efficient, non-redundant exploration across sequential tasks.

G.2 Comparison of Feasible and Infeasible Tasks

In real-world scenarios, certain tasks within an episode may be inherently infeasible, *e.g.*, searching for a non-existent object or answering a question that refers to a room that does not exist. In this section, we examined how the proposed 3DSPMR behaves on infeasible tasks, providing fine-grained comparisons against their corresponding feasible counterparts through several concrete examples. Specifically, we presented three pairs of feasible and infeasible tasks from the EQA track of SEER-Bench in Figures 16, 17, and 18, assuming a single task in each episode for clarity.

Feasible vs. Infeasible Task: Incorrect Attribute. In Fig. 16, we selected a single environment and constructed a comparison by modifying the object attribute in the question to an incorrect description. Under the same environment and initial position, 3DSPMR successfully locates the referenced object and answers correctly when the attribute is accurate. When the attribute is incorrect, 3DSPMR identifies the mismatch, determines that the task is infeasible, and terminates further exploration.

Feasible vs. Infeasible Task: Object not Present. In Fig. 17, we selected two different environments, one containing a Christmas tree and the other not. The proposed 3DSPMR approach adapts to both settings and responds correctly to the same question in each case. When a Christmas Tree is present, 3DSPMR locates it and answers accurately; when it is absent, 3DSPMR identifies the missing object, determines that the task is infeasible, and terminates further exploration.

Feasible vs. Infeasible Task: Room not Present. In Fig. 18, we selected two different environments, one containing bedrooms and the other not. We applied the same question to both settings and ensured that the referenced object (a lamp, in this example) exists in both environments. For infeasible tasks caused by the absence of the target room, 3DSPMR still adapts appropriately to each case. In the feasible environment, 3DSPMR locates the referenced lamp within the actual bedroom and answers correctly. In the infeasible environment, although it detects lamps, it realizes that they are located in the living room rather than the target bedroom. After exploring all potentially relevant areas, 3DSPMR concludes that the bedroom is absent, identifies the task as infeasible, and halts further exploration.

H Prompts for Data Annotation

In this section, we presented the prompts used for the Multi-Modal Large Language Models (MLLMs). We introduced the annotation procedure for the EQA track of SEER-Bench, conducted using GPT-5 [50]. For most sequential Question–Answering (QA) annotations, we employed the prompts illustrated in Fig. 19 and Fig. 20. To enrich the dataset with objects and rooms that do not typically appear in HM3DSem [45], we further used GPT-5 to generate two categories of unanswerable questions (Object Not Present and Room Not Present) using the prompts shown in Fig. 21 and Fig. 22.

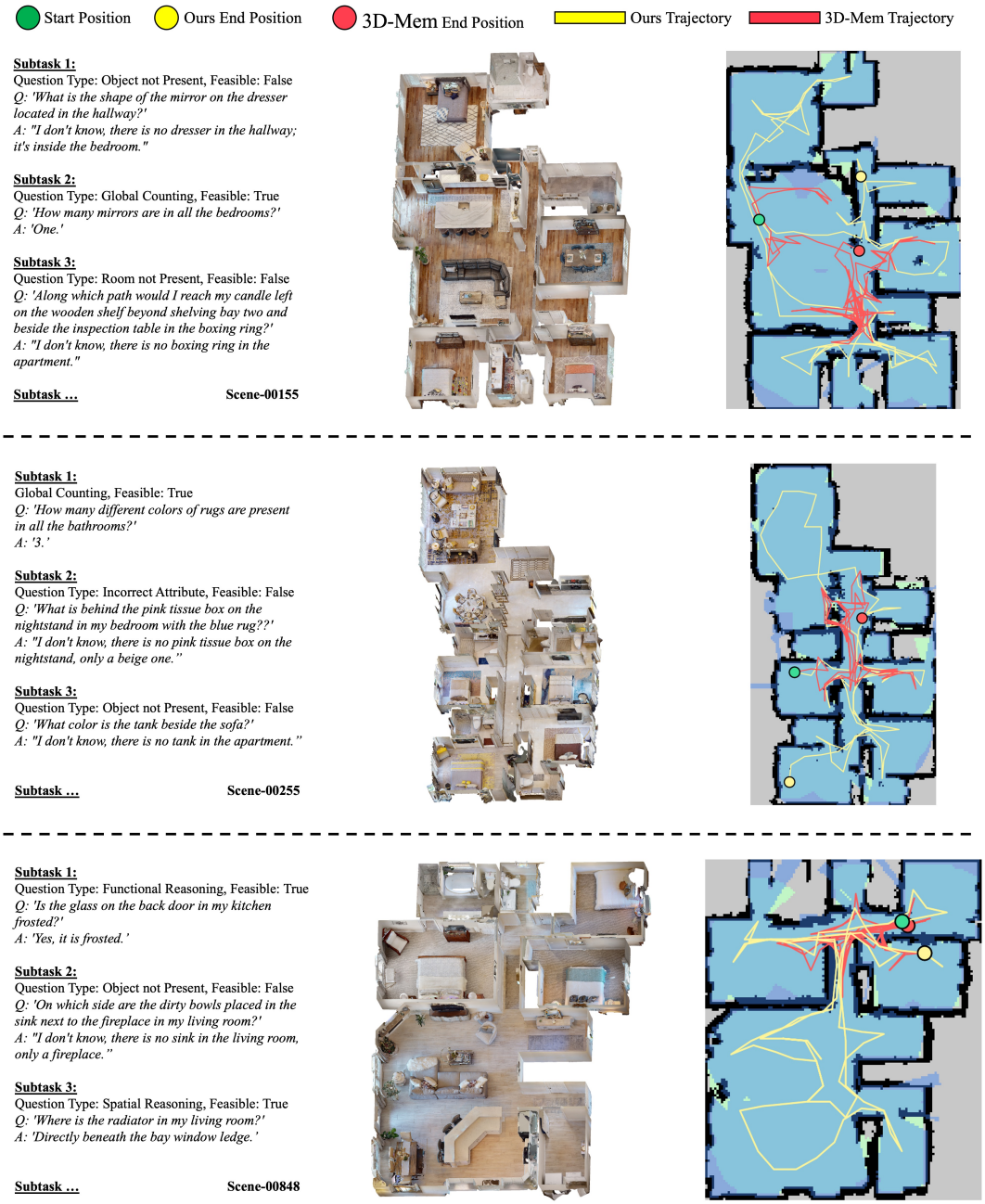

Figure 14: **Comparative Navigation Trajectories for Sequential EQA Tasks within One Episode.** The figure presents the results for three distinct scenes, with each row corresponding to a different scene and its sequential tasks (Left). The middle column shows the 3D map of the scene. The right column compares the navigation trajectories (normalized to the same step numbers) generated by our Proposed Method (yellow line) and the 3D-Mem Baseline (red line).

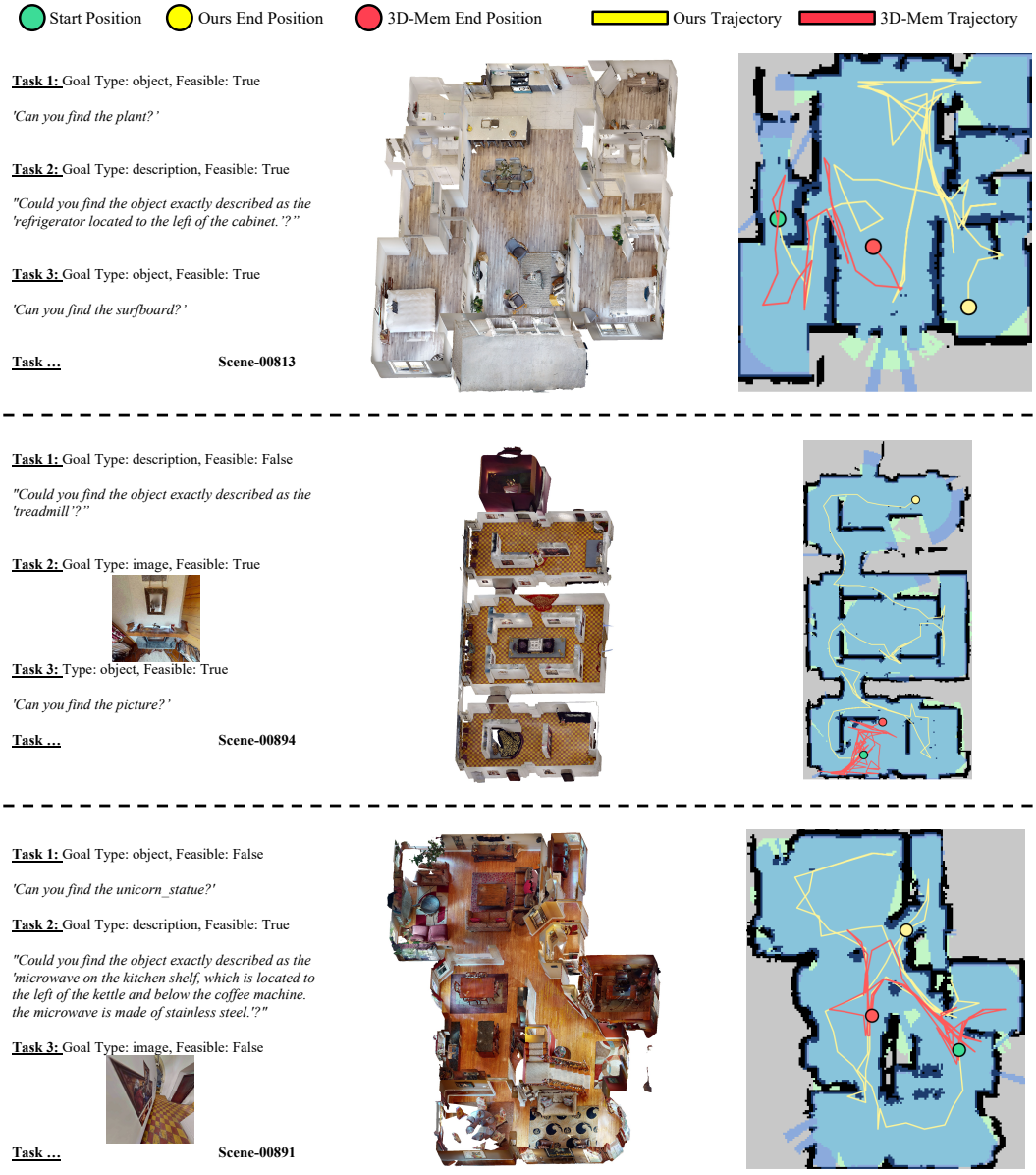

Figure 15: **Comparative Navigation Trajectories for Sequential EMN Tasks within One Episode.** The figure presents the results for three distinct scenes, with each row corresponding to a different scene and its sequential tasks (Left). The middle column shows the 3D map of the scene. The right column compares the navigation trajectories (normalized to the same step numbers) generated by our Proposed Method (yellow line) and the 3D-Mem Baseline (red line).

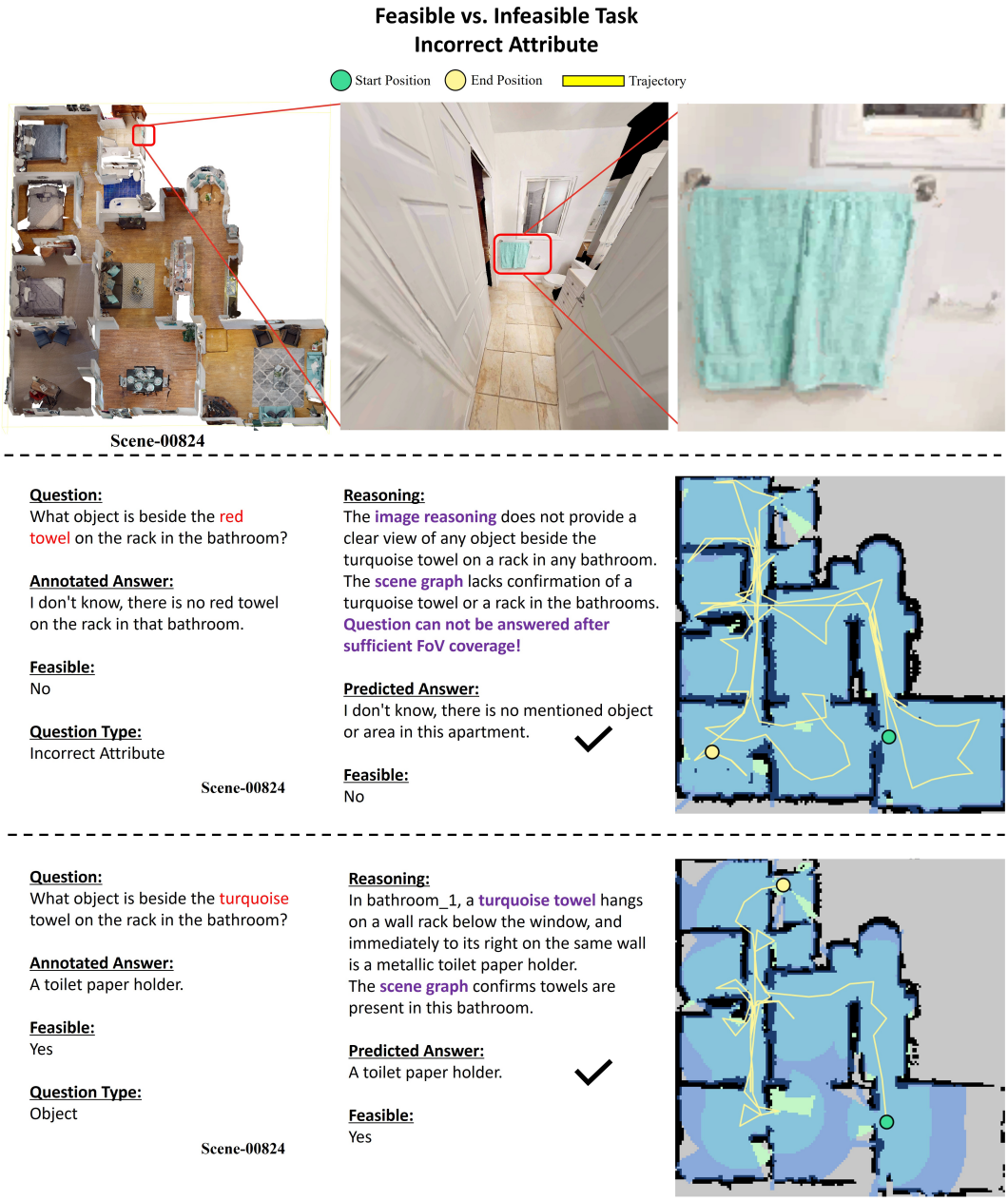

Figure 16: **Feasible vs. Infeasible Task: Incorrect Attribute.** This figure illustrates 3DSPMR's reasoning and exploration process on two EQA tasks: one feasible and one infeasible due to an incorrect object attribute. In the feasible case, 3DSPMR locates the referenced object and answers correctly. In the infeasible case, it detects the attribute mismatch, determines the task is infeasible, and halts further exploration.

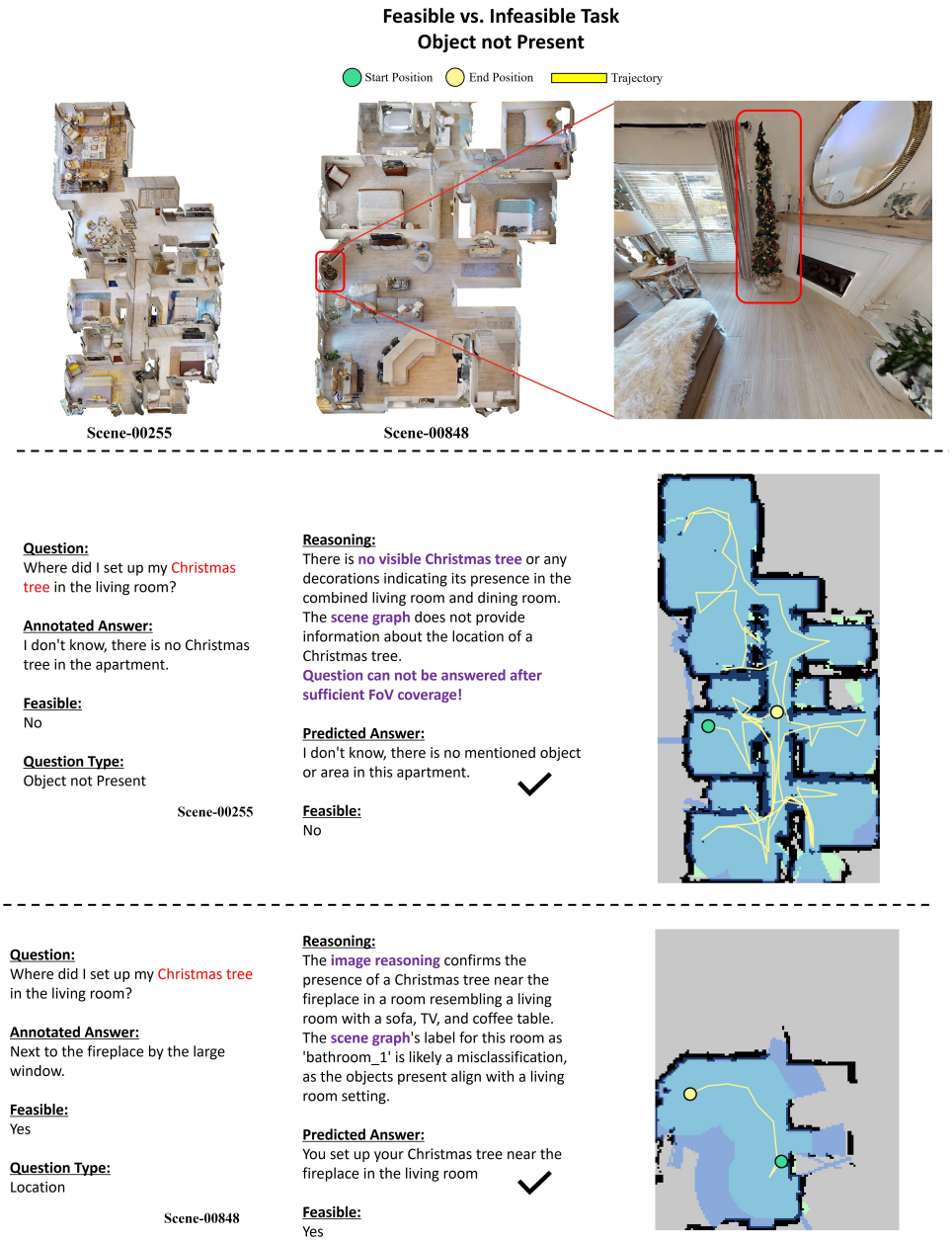

Figure 17: **Feasible vs. Infeasible Task: Object not Present.** This figure illustrates 3DSPMR's reasoning and exploration process on two EQA tasks: one feasible and one infeasible due to the absence of the mentioned object *Christmas Tree*. In the feasible case, 3DSPMR locates the *Christmas Tree* and answers correctly. In the infeasible case, it identifies the *Christmas Tree*'s absence, determines that the task is infeasible, and terminates further exploration.

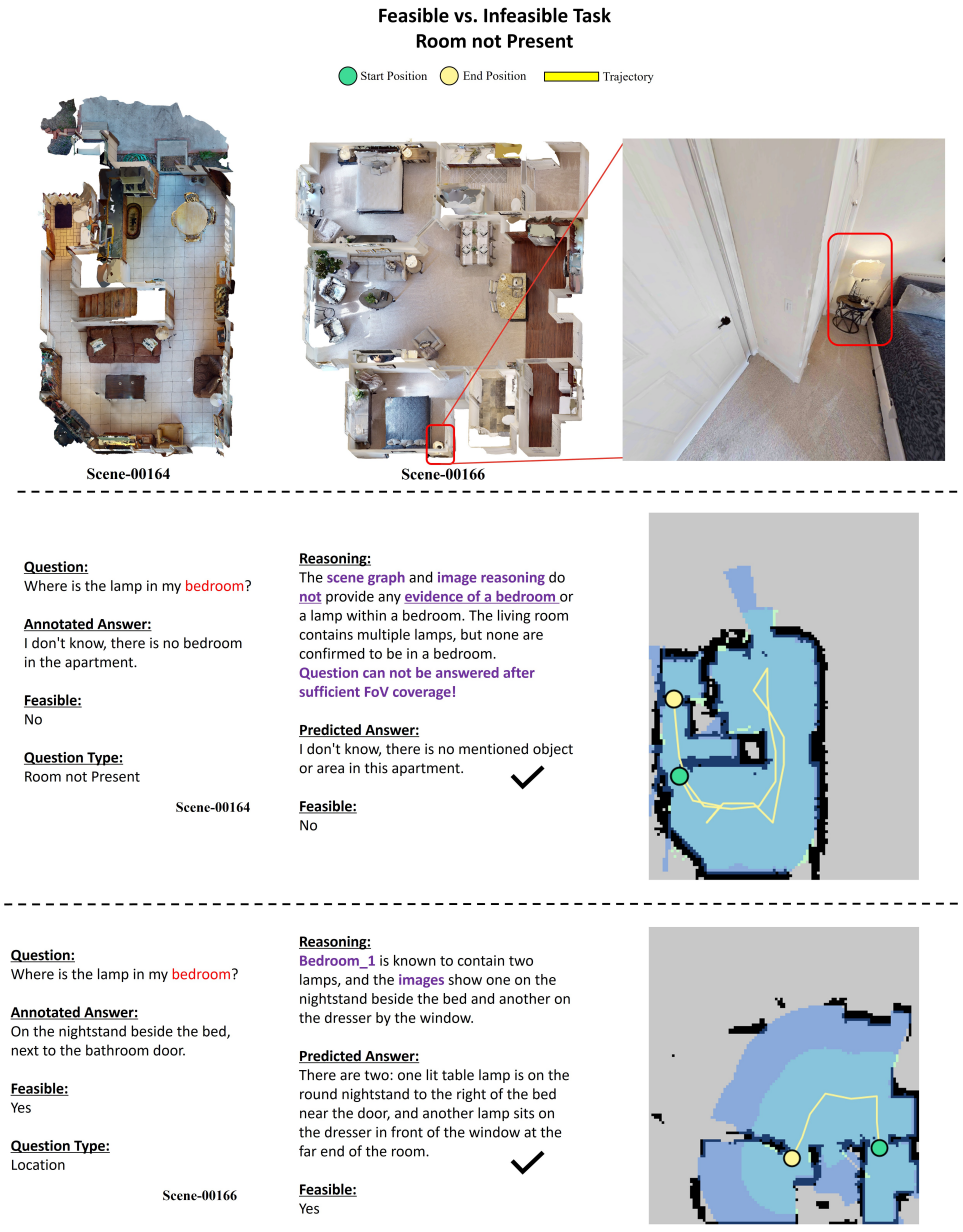

Figure 18: **Feasible vs. Infeasible Task: Room not Present.** This figure illustrates 3DSPMR's reasoning and exploration process on two EQA tasks: one feasible and one infeasible due to the absence of the mentioned room *bedroom*. In the feasible case, 3DSPMR locates the referenced lamp within the actual *bedroom* and answers correctly. In the infeasible case, although it detects lamps, it finds that they are located in the living room rather than the targeted *bedroom*. After exploring all potentially relevant areas, 3DSPMR concludes that the *bedroom* is absent, identifies the task as infeasible, and halts further exploration.

System Persona:
You are an **EQA-Generator**, an AI expert in computer vision and embodied robotics. Your task is to generate high-quality, spatially grounded **question–answer (QA)** pairs from a sequence of visual frames and a room layout.

Primary Objective:
Given a **room layout** and **video frame sequence** of a home environment, generate **12 diverse QA pairs** in **valid JSON array** format.

User's Perspective: Frame questions as if a homeowner ("I," "my") is asking their home robot for help.

Question Categories:
Category Coverage (total=12):
- Object Recognition: 1
- Object Attributes: 1
- Object Status: 1
- Object Location: 1
- Spatial Reasoning: 2
- Functional Reasoning: 2
- Local Counting: 1 (cover local counting in each room)
- Unanswerable: 3 (cover at least two distinct subtypes)
Generate questions from the following categories.

-Object Recognition: Questions about object identification.
Example Question: "What is the gray object I put on the bed in my bedroom?"
Example Answer: "It's a hat."

-Object Attributes: Questions about object attributes (color, material).
Example Question: "Is the floor in the bedroom wooden?"
Example Answer: "Yes, it is made of wood."

-Object Status: Questions about object status (on/off, open/closed).
Example Question: "Is the light on in the kitchen?"
Example Answer: "No, the kitchen light is off."

-Object Location: Questions about an object's location or its spatial relationship to other objects.
Example Question: "Where did I leave my coffee mug?"
Example Answer: "It's on the small table next to the sofa in the living room."

-Spatial Reasoning: Questions requiring inference about space, capacity.
Example Question: "Is there enough space on the bookshelf for another book?"
Example Answer: "Yes, there is room on the top shelf."

-Functional Reasoning: Questions based on what an object can be used for (its function or affordance).
Example Question: "Where can I hang up my coat?"
Example Answer: "On the hook by the front door."

-Local Counting: Questions that count the object number in one single room.
Example Question: "How many couch pillows are there in the living room?"
Example Answer: "Five."

Figure 19: **Prompt for QA annotation I.**

-Unanswerable: Plausible-sounding questions that are impossible to answer based on the visual evidence.
Subtypes:
Object Not Present:
Example Question: "What color is the rocket?"
Example Answer: "I don't know, there is no rocket."

Incorrect Location:
Example Question: “What is beside the microwave in the bedroom?"
Example Answer: "I don't know, there is no microwave in the bedroom."

Incorrect Attribute:
Example Question: “What is to the left of the blue pillow on the couch in the living room?"
Example Answer: "I don't know, there are no blue pillows on the couch, only red ones."

Room Not Present :
Example Question: “How are the chairs set up in the meeting room?"
Example Answer: "I don't know, there is no meeting room in this apartment.“

Hard Constraints:
1) Embodiment: Each question must require navigation/exploration or multi-step spatial reasoning (not solvable by a single static close-up with no context).
Do not ask questions that don't require spatial or agent-centric reasoning (e.g., "What is the color of the wall?", "What brand is the laptop?", "What is the distance between the Earth and the Moon?").
2) Visual Grounding: Use the layout to anchor rooms/adjacency. For answerable items, the answer MUST be verifiable from the provided frames.
3) Evidence Frames:
- "frame_index" must be an integer or a array of provided GLOBAL indices.
- Only use indices that were explicitly listed with the frames in this request.
4) Unambiguous Wording: Avoid vague pronouns and subjective adjectives; anchor locations (e.g., "in the living room").
NO Ambiguous Questions: The question must have one clear, verifiable answer (e.g., AVOID "What's that interesting thing in the corner?").
NO Navigation Question in Spatial Reasoning, (e.g., AVOID "How do I get from the entry hallway console to the dining table?")
5) Answer Format:
- Counting answers must be numeric strings (e.g., "3").
- Unanswerable answers must start with: "I don't know, ..." and state the reason (not present / wrong location / wrong attribute).
6) User's Perspective: homeowner voice ("I", "my"). No agent-relative pose questions (e.g., "to my left").

Subtype Rules:
- If the chosen type has subtypes, set "subtype" accordingly; otherwise repeat the type.

Output Specification
Your output must be a single, valid JSON array. Each object in the array represents one QA pair and must contain exactly four keys:
frame_index: Integer or Array of Integers. The frame number(s) where the answer can be verified.
question: String. The embodied question.
answer: String. A concise and direct answer.
question_type: String. Must be one of: Object Recognition, Object Attributes, Object Status, Object Location, Spatial Reasoning, Functional Reasoning, Local Counting, Unanswerable
subtype: String. If the question type has subtypes, provide the subtype. Otherwise, use question_type.

Now, analyze the provided frames and generate 12 embodied question-answer pairs in the specified JSON format. Do not include any other text or explanations in your response.

Figure 20: **Prompt for QA annotation II.**

You are EQA-Generator, an expert in Embodied Question Answering (EQA) for computer vision and robotics. Your task is to generate high-quality, embodied question–answer (QA) pairs that are unanswerable because the referenced object does not exist in the apartment.

Each QA object must contain exactly four keys, in this order:
{ "question": "...",
  "answer": "...",
  "question_type": "Unanswerable",
  "subtype": "Object Not Present in the indoor environment"}

Voice & Perspective
- Write from the homeowner's perspective, using "I" or "my." Example: "Where did I leave ...", "Can my robot check ...".
- Do not use agent-relative expressions such as "to my left/right" or "in front of you."

Core Unanswerability
- Every answer must begin exactly with:
  "I don't know, there is no {object} in the apartment." where {object} is identical (case-exact) to the noun in the question.
- Each question must be unanswerable solely because the referenced object does not exist in any indoor living area.

Forbidden Question Patterns
1. Hypothetical / Imaginative — forbidden examples:
   "If there were a {object}, how would it behave in the apartment?"
2. Formulaic / Geometric — forbidden example:
   "In which corner ... in the apartment?"
3. Binary / Weakly Embodied — forbidden examples:
   "Can the robot find the submarine in the apartment?"
4. Avoid mentioning specific rooms as the question should refer to the whole apartment.
   Use natural, perception-based phrasing starting with What, Where, Which, or How.

Object Selection Principles
Choose objects that cannot realistically exist in a normal apartment but are physically possible in the real world.

ALLOWED OBJECTS (You could extend more)
Vehicles & Machinery: rocket, tank, submarine, helicopter, forklift, airplane, excavator, drone, tractor, hovercraft
Industrial / Construction: crane, turbine, generator, conveyor belt, robotic arm, 3D printer, oil pump
Animals: lion, elephant, horse, cow, dolphin, owl, tiger, penguin, snake, camel, parrot
Scientific / Technical: microscope, centrifuge, medical robot, oxygen tank, X-ray machine, control panel
Entertainment / Sports: treadmill, surfboard, trampoline, billiard table, vending machine, arcade cabinet, ATM, server rack

DISALLOWED OBJECTS
Natural phenomena (rainbow, volcano, tornado, river, waterfall)
Mythical / imaginary (dragon, unicorn, phoenix, mermaid, alien)
Ordinary household items (cup, chair, knife, laptop, book, bed)
Large structures (bridge, castle, stadium, lighthouse, dam, tower, hangar)
Abstract / non-physical (happiness, time, gravity, memory)

Hard Constraints
- Each question implies embodied reasoning or perception-based exploration.
- Use concrete nouns and clear spatial anchors ("behind the sofa," "near the balcony door").
- No Yes/No questions.
- No unsafe, branded, or personal content.
- Every question must be unique.
- The main object should differ whenever possible.
- If an object repeats, its phrasing or relation must differ.

Figure 21: **Annotation prompt for generating unanswerable QA pairs where the referenced object is not present.**

You are EQA-Generator, an expert in computer vision and robotics. Your sole task is to create high-quality, embodied question–answer (QA) pairs where each question mentions a closed/indoor room that would not realistically exist inside a normal apartment. Every QA must be unanswerable for the same reason: the referenced room does not exist in the apartment.

Each QA object must contain exactly four keys, in this order:
{ "question": "...",
"answer": "...",
"question_type": "Unanswerable",
"subtype": " Room Not Present in the indoor environment "}

Voice & Perspective
- Write from the homeowner's perspective using "I"/"my" (e.g., "Where did I leave…", "Can my robot check…").
- Do NOT use agent-relative poses (no "to my left/right", no "in front of you").

Core Unanswerability
- Each question must place a COMMON apartment object in, on, or relative to an impossible room
- The answer MUST start EXACTLY with: "I don't know, there is no {room_name} in the apartment." where {room_name} is a case-exact string that matches the room name used in the question.

Rooms (ALLOWED — use these names verbatim)
Medical/Clinical: operating room, MRI suite, X-ray room, dental clinic
Transportation: aircraft hangar, bicycle repair bay, motorcycle workshop bay, racing simulator cockpit, flight simulator cockpit, drone operations desk, portable car lift bay
Sports and recreation: badminton court, table tennis room, basketball court, boxing ring, ice rink, swimming pool lane, climbing gym, bowling alley, curling sheet, squash court, gymnastics practice room, trampoline room, sauna room
Industrial: foundry, assembly line, brewing room, walk-in freezer, server data hall, woodworking room, photo darkroom, printing room, pottery kiln room, sewing room,
Security: surveillance monitoring room, weapons vault, shooting range, home vault, display room Entertainment: home DJ booth, casino room, arcade room
Hobbyist: hydroponics grow room, seedling nursery room Creative: sculpture casting room, paint spray booth room

Household Object Pool (examples only; no brands)
mug, cup, bowl, plate, spoon, fork, knife, cutting board, pan, pot, kettle, sponge, soap dispenser, dish rack, microwave, toaster, trash bin, broom, mop, vacuum, laundry basket, detergent bottle, towel,
bedsheet, pillow, blanket, lamp, light switch, power strip, phone charger, laptop, tablet, headphones, remote control, keychain, wallet, book, notebook, pen, scissors, tape, umbrella, shoe rack, sneakers,
coat, hanger, mirror, plant pot, picture frame, candle, tissue box, shampoo bottle, conditioner bottle, toothpaste,

Hard Constraints (apply to EVERY question)
1) Embodiment: The question must imply navigation/exploration or multi-step spatial reasoning
(e.g., multiple landmarks, containers, or path constraints). Avoid trivial single close-ups.
2) Unambiguous wording:
- Use specific nouns; avoid vague terms like "that thing" and subjective adjectives like "interesting".
- Anchor with clear locations/relations (e.g., "on the metal counter beside the sink in the operating room").
3) NO yes/no questions. Each question must seek a concrete fact (identity, attribute, relation, function, count, capacity).
4) Safety/Content: No brands, private data, or unsafe/harmful instructions.
5) Room consistency: Mention EXACTLY ONE room (from ALLOWED). Repeat the SAME string in the answer template .
6) Object plausibility: Use objects common to apartments (see pool). They may be referenced within the impossible room.
7) Diversity: Vary objects, prepositions, surfaces, containers, landmarks, paths, and phrasing. Avoid near-duplicates and repeated (object + relation + room) triples.

Figure 22: **Annotation prompt for generating unanswerable QA pairs where the referenced room is not present.**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We make our main claims in the abstract, where we point out that sequential embodied tasks are an underexplored problem, and we propose a novel 3DSPMR method to address this along with a novel benchmark. We also list our three main contributions in the introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We analyze the limitation of our proposed 3DSPMR method in Section E of the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We propose a new 3DSPMR method, which is introduced in Section 3 of the main paper.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the full SEER-Bench dataset and the corresponding codebase via the attached link.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We include experimental setup in the main paper. Implementation details and hyperparameters are included in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported because it would be too computationally expensive.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute resources are included in Appendix B.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: No ethics concern involved.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Included in Section F of the Appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks. No model is released.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We add proper citations for employed simulator and 3D scenarios.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce a new data benchmark for sequential embodied tasks (SEER-Bench). The details are described in the Appendix and we provide the dataset in the attached link.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work involves neither crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are used for proofreading and QA annotation.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.