SAFE-EQA: Semantic-Aware Efficient Exploration for Embodied Question Answering

Yijie Tang[1], Ke Xia[1], Jiazhao Zhang[2], Zhinan Yu[1], Zhiyuan Yu[3], Dezun Dong[1], Renjiao Yi[1], Chenyang Zhu[1†], and Kai Xu[4†]

[1] National University of Defense Technology
[2] CFCS, Peking University [3] Wuhan University
[4] Institute of AI for Industries, Chinese Academy of Sciences

Abstract. Embodied Question Answering (EQA) requires an agent to explore unseen environments, gather visual evidence, and answer natural-language questions. Efficient exploration requires both global planning and task-aware prioritization of informative regions. However, existing frontier-based methods typically rely on greedy, step-wise decisions, which often cause redundant motion and frequent Vision-Language Model (VLM) queries. We propose SAFE-EQA, a semantic-aware framework that improves exploration efficiency through globally coherent path planning and reduced VLM interaction. SAFE-EQA incrementally builds a landmark graph to capture scene topology and a query-conditioned semantic map to estimate task relevance of different regions. To support comprehensive, question-oriented exploration, we formulate planning as a Semantic-Aware Traveling Salesman Problem (TSP), which optimizes the visitation order over discovered landmarks. We further introduce an adaptive replanning mechanism that selectively interrupts local navigation and re-optimizes trajectories when higher-priority frontiers emerge. Extensive experiments on OpenEQA, EXPRESS-Bench and HM-EQA show that SAFE-EQA matches or outperforms state-of-the-art EQA methods while reducing VLM token consumption by 27.8–45.6%.

Keywords: Embodied Question Answering · Semantic-aware Path Planning · Travelling Salesman Problem

1 Introduction

Recent advances in embodied AI have shifted the focus from passive visual recognition to active perception, among which Embodied Question Answering (EQA) [11,35] is both fundamental and challenging. At the intersection of computer vision, natural language processing, and robotics, EQA requires an agent to autonomously navigate unseen environments, gather informative observations, and answer a natural-language query. Success therefore hinges on bridging high-level language understanding with goal-directed physical exploration to acquire the visual evidence needed for answering questions.

† Corresponding authors.
Homepage: `https://github.com/yjtang249/SAFE-EQA`.

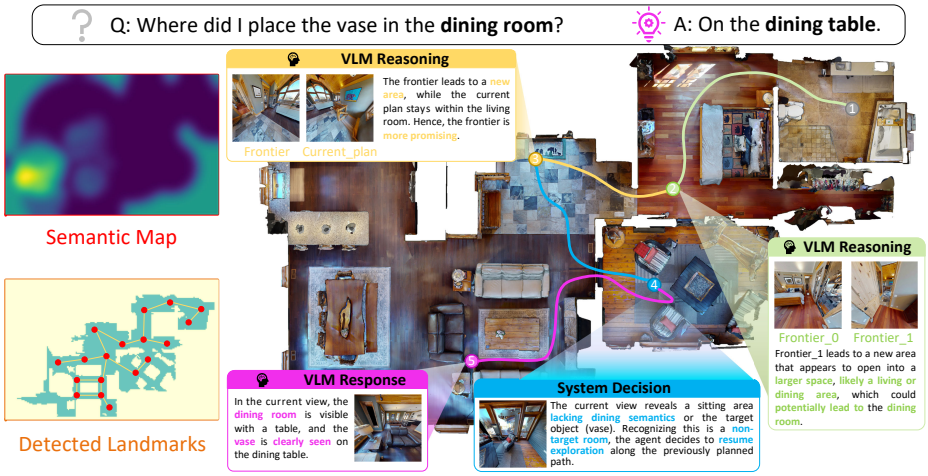

Fig. 1: SAFE-EQA performs question-driven exploration in unknown environments, combining semantic reasoning with navigable landmarks to guide the agent. Through globally coherent path planning, the agent efficiently prioritizes task-relevant regions, enabling accurate and resource-efficient open-ended question answering.

Consider an agent asked: "Is there a coffee mug on the kitchen table?" in an unknown house. A naive policy may exhaustively traverse all rooms or wander randomly. In contrast, an effective agent should leverage semantic commonsense—for example, inferring that a kitchen table is more likely to be near a kitchen or dining area than a bedroom—and prioritize exploration accordingly. Given the goal-oriented nature of EQA, agents must achieve both accurate evidence acquisition and high exploration efficiency. This requires coherent trajectories that globally prioritize task-relevant regions, reduce redundant motion, and avoid the latency caused by frequent step-wise decision making.

With the strong reasoning and decision-making capabilities of large Vision–Language Models (VLMs), recent EQA methods predominantly adopt VLMs as planners or decision makers [21, 35, 36]. Effective VLM-guided exploration typically relies on two key components: (i) a structured memory module for representing and retrieving scene knowledge, and (ii) a planning strategy for efficient navigation. While many works focus on advanced memory designs for efficient scene encoding [3,42,46], path-planning strategies remain comparatively underexplored. Most existing approaches retrieve task-relevant observations and query the VLM to select the next goal in a step-wise manner. However, this paradigm has two fundamental limitations. First, invoking the VLM at every step incurs substantial latency and token overhead, hindering practical online deployment. Second, many methods rely on greedy frontier-based planning, where the next target is selected from local frontier candidates. Such myopic planning lacks global topological awareness and long-horizon coordination, often leading to insufficient exploration (missing critical evidence), discontinuous trajectories (frequent detours and backtracking), and suboptimal global routes. As a result,

agents struggle to systematically cover informative regions and form coherent, efficient exploration plans.

To enable comprehensive and efficient goal-oriented exploration, we propose SAFE-EQA, a novel framework that reformulates EQA exploration as a *semantic-aware Traveling Salesman Problem (TSP)* over representative landmarks that approximately cover the navigable scene area (Fig. 1). Rather than planning solely over frontiers at explored–unexplored boundaries, SAFE-EQA incrementally constructs a 2D Voronoi graph from partial observations to extract topological landmarks online, allowing planning to jointly consider structural connectivity and exploration boundaries. Conditioned on the target region described in the question, we further construct a task-relevant semantic map [21,35] by querying the annotated scene graph with an LLM. We then solve a semantic-aware TSP to compute an efficient visitation order over discovered landmarks, balancing geometric distance and semantic relevance. Importantly, SAFE-EQA outputs multi-step local trajectories instead of single-step goals, substantially reducing VLM query frequency. To balance efficiency and accuracy, we additionally introduce an adaptive replanning mechanism: the current route is interrupted when a newly detected frontier has higher task priority, ensuring both local trajectory coherence and timely responses to critical visual cues.

Compared with prior discrete, step-wise exploration paradigms, our method achieves more accurate evidence acquisition by incorporating task semantics into global route planning, produces more coherent trajectories through topology-aware landmark ordering, and improves efficiency by reducing redundant motion and significantly lowering VLM query frequency (and thus token usage). We evaluate SAFE-EQA on OpenEQA [32] and EXPRESS-Bench [21]. SAFE-EQA matches state-of-the-art question-answering accuracy and exploration efficiency while reducing VLM token usage by 35.5–45.6% compared with existing VLM-guided baselines. In summary, our contributions are as follows:

- We propose SAFE-EQA, a semantic-aware global routing framework that reformulates EQA exploration as an online TSP over topological landmarks, moving beyond discrete step-wise goal selection to generate coherent multi-step trajectories.
- We develop an online planning pipeline that incrementally constructs a 2D Voronoi graph for landmark extraction and builds a question-conditioned semantic map via LLM queries, enabling semantic-aware route optimization and adaptive replanning with substantially fewer LLM/VLM interactions.
- Extensive experiments show that SAFE-EQA maintains state-of-the-art answering accuracy while reducing LLM/VLM token consumption by 35.5–45.6%, validating improved accuracy, trajectory coherence, and exploration efficiency in embodied question answering.

2 Related Works

2.1 Embodied Question Answering

Embodied Question Answering (EQA) requires an agent to actively explore an environment and gather visual evidence to answer a given question. Early studies [11, 12, 15, 39, 48] mainly relied on trained task-specific modules or heuristic policies, which limits their generalization to open-ended questions across diverse scenarios. With recent advances in vision-language models (VLMs) and large language models (LLMs), foundation models have shown strong spatial reasoning capabilities [1, 2, 13, 23, 24, 30, 37] and have been adopted in a broad range of embodied AI tasks [4, 8, 14, 18, 19, 31, 45]. Building on this progress, recent EQA methods increasingly incorporate VLMs/LLMs as high-level planners or decision-makers. Some approaches use question-aware semantic maps to guide exploration [10, 21, 35], while others emphasize efficient memory and representation design for information encoding and retrieval [3, 36, 42, 49].

2.2 Autonomous Scene Exploration

Autonomous scene exploration is a core capability in embodied AI and underpins a wide range of downstream tasks. In classical settings (e.g., 3D reconstruction and mapping), the primary objective is to maximize scanning coverage while maintaining exploration efficiency. Representative approaches include field-based methods [25], which use vector fields (e.g., gradient or tensor fields) to guide agents along smooth trajectories [26,33,41,43], and frontier-based methods, which improve coverage by selecting goals on the boundary between known and unknown regions [6, 17, 28, 44]. More recently, exploration has increasingly shifted toward goal-oriented settings, where agents are expected to discover task-relevant evidence rather than exhaustively scan the entire scene. In this context, LLMs/VLMs have been introduced to support online planning through semantic inference and spatial reasoning [5,7,9,27,40]. Benefiting from broad world knowledge and strong reasoning capabilities, these models can better connect partial observations to task objectives in both semantic and spatial dimensions, making them a promising direction for scalable, general-purpose embodied exploration.

3 Method

3.1 Overview

In Embodied Question Answering (EQA), an agent is initialized in an unseen environment and given an open-ended natural-language question $\mathbf{q}$. The goal is to actively explore the environment and gather sufficient visual evidence to answer $\mathbf{q}$ with high confidence.

As shown in Fig. 2, SAFE-EQA follows a two-stage framework. In the preprocessing stage, we parse the input question $\mathbf{q}$ to identify target regions and

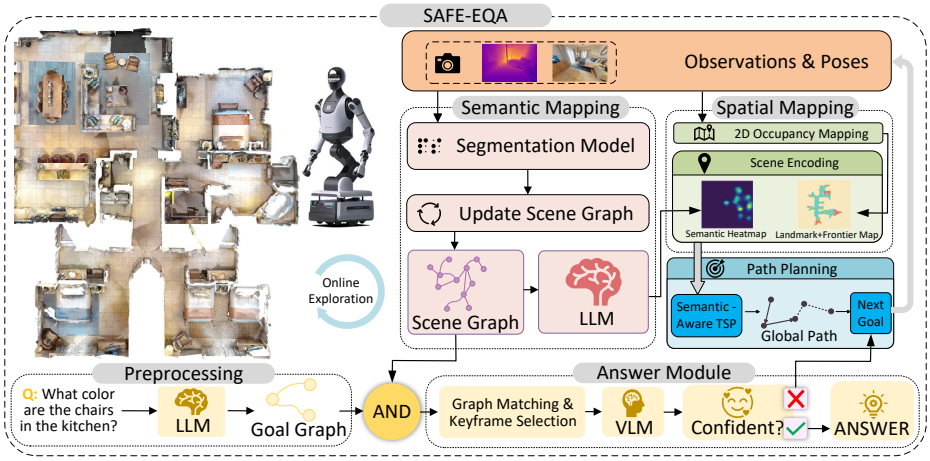

Fig. 2: Overall pipeline of SAFE-EQA. Given an input question $\mathbf{q}$, the *Preprocessing* stage constructs a *Goal Graph* for target region identification. During *Online Exploration*, a *Semantic Map* and a *Landmark Graph* are incrementally built from observations. The *Semantic-aware TSP*, guided by the Semantic Map, computes a visitation sequence over all landmarks, resulting in a globally coherent path that efficiently prioritizes task-relevant regions. The *Hierarchical Answer Module* retrieves key observations and invokes the VLM for answering only when sufficient evidence is available.

objects, and construct a *Goal Graph* [47] encoding their semantic relations. This graph provides high-level guidance for the subsequent exploration (Sec. 3.2).

In the second stage, online exploration (Secs. 3.3 and 3.4), SAFE-EQA incrementally builds a 3D *Scene Graph* from RGB-D observations and queries a large language model (LLM) to assess task relevance across regions, yielding a continuously updated 2D *Semantic Map*. In parallel, it constructs a *Landmark Graph* [7, 40] as a topological abstraction for global path planning. The planner then jointly considers geometric distance and semantic relevance to improve both exploration completeness and efficiency.

SAFE-EQA further employs a hierarchical answering module that performs coarse-to-fine graph matching and filtering to identify task-relevant evidence. This module removes irrelevant context and invokes the VLM only when necessary, improving inference efficiency and reducing token use (Sec. 3.5).

3.2 Preprocessing

In the preprocessing stage, we extract task-relevant semantic entities from the input question $\mathbf{q}$. Specifically, we use a large language model (LLM) to parse $\mathbf{q}$ and identify target objects and their relationships. Based on the parsed semantics, we construct a goal graph [47], denoted as $\mathbf{G}^{goal} = (\mathbf{V}^{goal}, \mathbf{E}^{goal})$, where $\mathbf{V}^{goal}$ contains target-object nodes and $\mathbf{E}^{goal}$ encodes inter-object relations.

Each node $\mathbf{v}_i^{goal} \in \mathbf{V}^{goal}$ is associated with a category label and an open-vocabulary feature $\mathbf{f}_i^{goal}$, extracted using a pretrained embedding model [29]. The

edge set $\mathbf{E}^{goal}$ represents spatial or semantic relations between objects, selected from a predefined relation vocabulary.

For questions involving abstract concepts (e.g., "to sit"), we introduce abstract labels into the goal graph to better identify regions that satisfy the intent of the query.

During online exploration, the goal graph is periodically matched with the scene graph to filter candidate target regions and guide the answering module.

3.3 Scene Construction and Exploration

To efficiently encode the information of the scanned scene during online exploration, we incrementally construct several key data structures, including a scene graph $\mathbf{G}^{scene}$, semantic map $\mathbf{S}$, landmark graph $\mathbf{L}$, frontier graph $\mathbf{F}$, and a Minimal Keyframe Set $\mathbf{K}$. Building upon them, we formulate the exploration process as a Traveling Salesman Problem (TSP) to ensure a comprehensive and efficient exploration strategy.

Scene Graph and Semantic Map. The scene graph $\mathbf{G}^{scene} = (\mathbf{V}^{scene}, \mathbf{E}^{scene})$ is a topological representation of the detected instances in the environment, where nodes $\mathbf{V}^{scene}$ represent individual instances, each associated with a label and feature, similar to the goal graph. The edges $\mathbf{E}^{scene}$ encode the spatial relationships between these instances. At each step $\mathbf{t}$, the agent captures $\mathbf{N}$ egocentric RGB-D images along with their corresponding camera poses. An open-vocabulary segmentation model [50] is used to generate 2D instance masks from the RGB images, which are then lifted into 3D through back-projection with the corresponding depth image and camera pose. Newly detected 3D instances are either added or merged with existing instances in $\mathbf{G}^{scene}$, ensuring that the scene graph remains up-to-date with the agent's surroundings.

To assign task priority to different regions, we partition $\mathbf{G}^{scene}$ into a set of connected subgraphs, denoted as $\mathbf{G}^{scene} = \{\mathbf{G}_i^{scene} \mid i = 1, 2, \ldots, m\}$. Each $\mathbf{G}_i^{scene}$ is then prompted to an LLM to obtain a task relevance score $\tau_i \in [0, 1]$, reflecting its relevance to the current question. The score is then semantically fused into the overall semantic map $\mathbf{S}$, applied to the unoccupied areas near the point cloud of $\mathbf{G}_i^{scene}$ within a distance threshold $\mathbf{r}$.

For each unoccupied point $p \in \mathcal{P}_{\mathbf{G}_i^{scene}}$, where $\mathcal{P}_{\mathbf{G}_i^{scene}}$ is the set of unoccupied points within the range $\mathbf{r}$ of $\mathbf{G}_i^{scene}$, the updated semantic value is given by:

$$w_t = w_{t-1} + 1, \quad \mathbf{S}(p)_t = \frac{w_{t-1}}{w_t} \cdot \mathbf{S}(p)_{t-1} + \frac{1}{w_t} \cdot \tau_i, \tag{1}$$

where $\mathbf{S}(p)_t$ represents the updated score, $\mathbf{S}(p)_{t-1}$ is the previous score, and w_t is the accumulated weight.

Frontier and Landmark Graph. In each step $\mathbf{t}$, the frontier graph $\mathbf{F}_t = \{\mathbf{f}_{t,i}\}$ records potential exploration directions at the boundary between explored and unexplored areas [44], where each frontier represents an aggregation of discrete regions.

However, relying solely on frontier-based exploration (FBE) [35, 46] can lead to incomplete or inefficient exploration due to the discontinuous nature of stepwise goal selection. To overcome this, we introduce the landmark graph $\mathbf{L}_t = \{\mathbf{l}_{t,i}\}$, which anchors the scanned scene by discretizing it into a set of representative points. The landmark graph is constructed incrementally by extracting Voronoi graph [7, 22, 40] over the navigable area already scanned, where each node represents a landmark, and edges define the navigable connections between them. As new regions are explored, the Voronoi graph is updated, and redundant or previously visited nodes are removed. This approach ensures that the landmarks represent key spatial points, while the complementary frontier and landmark graphs guide exploration, preserving both global consistency and semantic awareness of the agent's trajectory.

After each update of the landmark or frontier graph, we perform a distance-based clustering to assign each frontier to the nearest landmark. Specifically, for a given frontier $\mathbf{f}_{t,j}$, its corresponding landmark $f_l(\mathbf{f}_{t,i})$ is selected as:

$$f_l(\mathbf{f}_{t,i}) = \arg\min_{\mathbf{l}_{t,i} \in \mathbf{L}_t} d(\mathbf{f}_{t,j}, \mathbf{l}_{t,i}), \tag{2}$$

where $d(\mathbf{f}_{t,j}, \mathbf{l}_{t,i})$ represents the distance between frontier $\mathbf{f}_{t,j}$ and landmark $\mathbf{l}_{t,i}$. If a frontier does not have a landmark within a distance threshold $\mathbf{r}_f$, its corresponding landmark is set to *None*.

Minimal Keyframe Set. In order to reduce redundancy and maintain essential visual observations, we use a Minimal Keyframe Set $\mathbf{K}$, which consists of representative observations taken throughout the exploration. This set is constructed and updated incrementally by applying the co-visibility clustering based approach [46], ensuring that the selected keyframes provide a compact yet informative representation of the scanned scene.

Exploration Based on Semantic-aware TSP. Inspired by prior work in autonomous scene exploration [41,43], we treat the extracted landmarks as a coarse topological coverage of the scene for globally coherent path routing. Specifically, we formulate exploration as an online TSP over the landmark graph. Given the current landmark set $\{\mathbf{l}_{t,i}\}$, our goal is to determine their optimal visitation order. This order is represented by a permutation $\mathbf{\Pi} = (\pi_0, \pi_1, \ldots, \pi_{n-1})$, where π_i denotes the i-th landmark in the sequence.

Due to the goal-oriented nature of our exploration task, solving the TSP purely based on geometric distance cannot lead to optimal solution. Instead, we propose semantic-aware TSP, which incorporates both the geodesic distances between landmarks and their corresponding task relevance scores, derived from the semantic map $\mathbf{S}$. This enables the agent to prioritize regions with higher task relevance, while still striving to minimize the overall travel cost, ensuring both efficient and task-aware exploration. Specifically, our semantic-aware TSP optimization begins with determining a reference point $\mathbf{p}_r$, which can either be the agent's current position or a manually selected point, as described in Sec. 3.4. The landmark which will be visit firstly, π_0, is set as the closest landmark to $\mathbf{p}_r$.

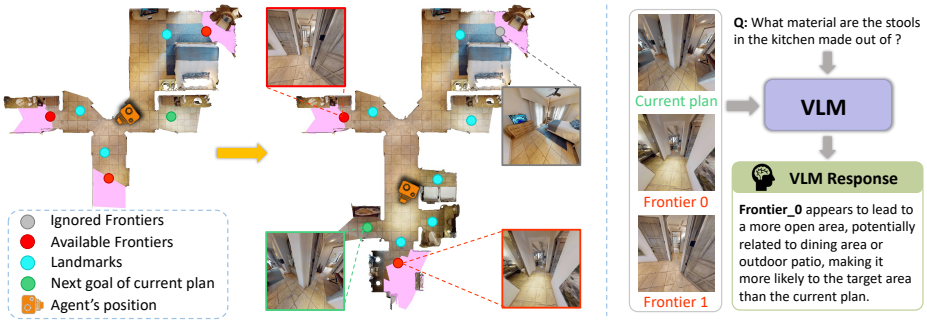

Fig. 3: Priority assessment for new frontiers. When a new frontier is detected, the agent captures an image of it. The image, along with those of other selected frontiers and the next goal in current path, are combined and fed to the VLM to assess whether any new frontier should be prioritized for exploration.

The semantic-aware TSP is thus formulated as:

$$\mathbf{s}_i = \frac{\mathbf{S}(\pi_i)}{\sum_j \mathbf{S}(\pi_j)}, \quad \lambda_i = 1 - w_{sem} \cdot \mathbf{s}_i \tag{3}$$

$$\mathcal{TSP}(\mathbf{p}_r, \mathbf{\Pi}) = \min_{\mathbf{\Pi}} \sum_{i=1}^{n-1} [\lambda_i \cdot \text{dist}(\pi_{i-1}, \pi_i)] \tag{4}$$

where $\mathbf{s}_i$ represents the normalized semantic score of landmark π_i over all landmarks, and λ_i is defined as the *discount factor* applied to each landmark, extracted with a predefined weight w_{sem}. The geodesic distance between consecutive landmarks is represented as $\text{dist}(\pi_{i-1}, \pi_i)$. The final solution to the TSP balances the routing cost and the semantic relevance of each landmark, ensuring both efficiency and task-awareness in the exploration path.

3.4 Online TSP Solving

Building upon the formulation of Semantic-aware TSP in Eqs. (3) and (4), we now focus on the online TSP solving, which is triggered whenever a new landmark is detected or a frontier with higher priority is identified during exploration. In such cases, the visitation sequence over all landmarks should be updated to ensure that the agent's exploration path remains optimal and task-aware.

Reference Point Selection. As mentioned above, the reference point $\mathbf{p}_r$ must be determined before solving the TSP each time, as it directly determines the first landmark to visit (i.e. π_0).

Suppose TSP solving is triggered at step $\mathbf{t}$. If new landmarks are detected, $\mathbf{p}_r$ is set to the agent's current position, $\mathbf{p}_t$. If new frontiers are detected, a Priority Assessment is performed. As illustrated in Fig. 3, we firstly filter the frontier graph $\mathbf{F}_t = \{\mathbf{f}_{t,i}\}$ to obtain a subset of frontiers close to $\mathbf{p}_t$, and with an associated landmark (i.e. $f_l(\mathbf{f}_{t,i}) \neq \text{None}$). This subset is denoted as $\mathbf{F'}_t = \{\mathbf{f'}_{t,i}\}$.

Subsequently, the agent captures the observation of each new frontier in $\mathbf{F'}_t$, and retrieve the observations of all selected frontiers, denoted as $\{\mathbf{o}_{t,i}\}$. The position of each $\mathbf{f'}_{t,i}$ is then marked in $\mathbf{o}_{t,i}$, and these observations, along with the observation of the next landmark in the existing visitation order (denoted as $\mathbf{o}^n$) are combined. This combined input is fed to a VLM to determine whether any of the selected frontiers should explore firstly than the current next goal. The optimal observation selection is formulated as follows (*the prompt is given in the supplementary*):

$$\mathbf{o}' = VLM(\{\mathbf{o}_{t,i}\}, \mathbf{o}^n) \tag{5}$$

If any of the frontiers is considered to have a higher priority (i.e., $\mathbf{o}' \in \{\mathbf{o}_{t,i}\}$), its corresponding landmark $f_l(\mathbf{f}_{t,i})$ is set as the reference point, and TSP solving is triggered.

TSP Solving. Once the reference point is determined, The TSP is solved to update the exploration path. To maintain continuity and avoid drastic path changes, we recalculate the TSP only for a local set of k landmarks, while keeping the remaining landmarks in their original order in the visiting queue.

The local landmark set consists of the k landmarks with the highest priority scores, selected through the following procedure for each landmark $\mathbf{l}_i$:

1. Compute the geodesic distance $\mathbf{d}_i$ from $\mathbf{l}_i$ to the reference point using A*.
2. Adjust the semantic score $\mathbf{S}(\mathbf{l}_i)$ by applying a sigmoid function followed with a scaling factor λ_a, yielding $s_i = \lambda_a \cdot \sigma(\mathbf{S}(\mathbf{l}_i))$.
3. Calculate the priority score of $\mathbf{l}_i$ as:

$$pr_i = \frac{s_i^\alpha}{(d_i + \epsilon)^\beta}, \tag{6}$$

where α and β control the influence of the semantic score and distance, and ϵ avoids division by zero.

Recalculating the TSP only for the selected local landmarks, this approach minimizes the difference between consecutive planned paths, ensuring smoother transitions and prioritizing task-relevant regions with high semantic scores. Algorithm 1 summarizes the TSP solving process.

3.5 Hierarchical Answering Module

To avoid querying the VLM for answer prediction at every step, we propose a hierarchical answering approach. This method only invokes the VLM when a task-relevant region is identified, significantly reducing token consumption and improving exploration efficiency.

The first stage aims to coarsely localize potentially relevant regions. To achieve this, we perform graph matching [47] between the goal graph $\mathbf{G}^{goal}$ and the scene graph $\mathbf{G}^{scene}$, extracting the overlapped nodes by computing pairwise feature similarities between $\{\mathbf{f}_i^{scene}\}$ and $\{\mathbf{f}_i^{goal}\}$. These overlapped nodes are referred to as the *target anchor nodes*. To preserve more scene context, we then

Algorithm 1: Solve Semantic-aware TSP

Input: Landmarks' visitation sequence $\mathbf{\Pi}^{t-1} = \{\pi_0^{t-1}, \ldots, \pi_{n-1}^{t-1}\}$, reference point $\mathbf{p}_r$

Compute priority scores: $\{pr_i = \mathrm{Pri}(\pi_i^{t-1})\}$;
Select top-k landmarks to form the local landmark set: $\mathbf{L}_s = \arg\max_{\mathcal{S} \subseteq \mathbf{\Pi}^{t-1}, |\mathcal{S}|=k} \sum_{\pi \in \mathcal{S}} \mathrm{Pri}(\pi)$;
foreach $\mathbf{l}_i \in \mathbf{L}_s$ **do**
  $\mathbf{\Pi}^{t-1} \leftarrow \mathbf{\Pi}^{t-1} - \{\mathbf{l}_i\}$;
Solve semantic-aware TSP on $\mathbf{L}_s$: $\mathbf{\Pi}_s = \mathcal{TSP}(\mathbf{p}_r, \mathbf{L}_s)$;
Concatenate the two sequences: $\mathbf{\Pi}^t \leftarrow \mathbf{\Pi}_s \,\|\, \mathbf{\Pi}^{t-1}$;
return *Updated visitation sequence* $\mathbf{\Pi}^t = \{\pi_0^t, \ldots, \pi_{n-1}^t\}$

extract the k_n-neighbors of the target anchor nodes in $\mathbf{G}^{scene}$, forming a subset of the scene graph, denoted as $\mathbf{G}^{scene\prime}$.

In the second stage, a filtering-and-answering approach [46] is employed. First, the labels of the nodes in $\mathbf{G}^{scene\prime}$ are fed to the LLM along with the question $\mathbf{q}$ to identify the potential target instances $\mathbf{T}^{scene}$ in the scene. If $\mathbf{T}^{scene} \neq \emptyset$, the corresponding visual observations of these target instances are retrieved from the Minimal Keyframe Set $\mathbf{K}$ and fed to the VLM for answer prediction. The exploration ends if a confident answer is provided; otherwise, the agent continues exploring according to the planned path.

To prevent redundant queries, target anchor nodes that have been queried are marked as *"queried"* and ignored in the first stage, until their relationship with other nodes is updated.

4 Experiments

In this section, we conduct extensive experiments to evaluate our method against state-of-the-art methods on a set of publicly available Embodied Question Answering benchmarks. We first describe the experimental setup and implementation details in Sec. 4.1. We then present quantitative and qualitative results in Sec. 4.2 and Sec. 4.3, respectively. Finally, we provide ablation studies to analyze the contribution of our key design choices in Sec. 4.4.

4.1 Experimental Setup

Implementation Details. We use GPT-4o [20] as the VLM for frontier priority assessment and answer prediction, and Qwen3-30B-A3B-Instruct-2507 [38] as the LLM for preprocessing, semantic mapping, and target-object prefiltering. For scene reconstruction, we follow prior work [35] and adopt a TSDF-based volumetric representation with a voxel size of 10 cm, from which we extract the 2D occupancy and semantic maps. We build the Voronoi graph over the 2D free space and update it incrementally, ignoring newly detected vertices that fall within 1.2 m of existing ones. To further reduce redundancy, we track the

Methods	OpenEQA		EXPRESS-Bench		HM-EQA†
	LLM-Match (↑)	LLM-Match SPL (↑)	LLM-Match (↑)	LLM-Match SPL (↑)	LLM-Match (↑)
Explore-EQA [35]	46.9	23.4	-	-	-
CG w/ Frontier Snapshots [16]	47.2	33.3	-	-	-
MemoryEQA* [49]	45.3	28.4	52.7	33.6	42.8
Fine-EQA [21]	43.8	26.6	51.4	28.1	44.2
3D-Mem [46]	52.6	42.0	**67.5**	50.3	59.0
SAFE-EQA (Ours)	**56.4**	**53.9**	63.3	**53.5**	**59.2**

Table 1: Comparison on answering accuracy and exploration efficiency. "CG" represents ConceptGraphs, we take the results of this method reported in 3D-Mem [46].

agent's trajectory and discard newly proposed landmarks within 1.2 m of any previously visited position. For semantic mapping, we use OpenSEED [50] as the 2D foundation model to generate instance masks from RGB images, and lift them into 3D to update the scene graph following ConceptGraph [16]. We set the hyperparameters to $k = 5$, $w_{\text{sem}} = 0.8$, and $\alpha = \beta = 1$ in our experiments.

Datasets. We evaluate on three benchmarks built on HM3D scenes [34]: OpenEQA [32], EXPRESS-Bench [21], and HM-EQA [35]. For OpenEQA, we use the A-EQA split with 184 open-ended questions. For EXPRESS-Bench, which covers 174 scenes and each scene contains one or more questions, we select the first question per scene to form a 174-question subset due to resource constraints. HM-EQA is originally a multiple-choice benchmark; to better match real-world open-ended EQA, we remove the answer choices and use the content of the correct option as the ground-truth answer, denoting this adapted dataset as HM-EQA†. Together, these benchmarks cover diverse tasks, such as object localization, attribute recognition, functional reasoning, and spatial understanding.

Metrics. Our evaluation considers three aspects: answering accuracy, exploration efficiency, and querying overhead. Following OpenEQA, we adopt *LLM-Match* to measure answering correctness, where an LLM assigns a score $\theta_i \in \{1, \ldots, 5\}$ to the predicted answer that is further mapped to a score from 0 to 100. We also report *LLM-Match SPL*, which weights θ_i by the exploration path length to reflect efficiency. In addition, we measure the number of tokens consumed by both VLM and LLM queries as an indicator of the overall resource overhead.

Baseline Methods. We compare SAFE-EQA with several recent VLM-guided exploration methods for open-ended, active Embodied Question Answering to evaluate the effectiveness of our approach. In particular, we consider Explore-EQA [35], ConceptGraph with frontier snapshots [16], and other state-of-the-art baselines including Fine-EQA [21], MemoryEQA [49], and 3D-Mem [46]. Most of these methods rely on frequent VLM calls for stepwise decision making or next-goal selection. Notably, the original MemoryEQA only supports EQA tasks formulated as multiple-choice questions. We modify its prompts to handle open-vocabulary questions, and denote this adapted version as MemoryEQA*.

4.2 Quantitative Results

Results of Accuracy and Efficiency. We report LLM-Match and LLM-Match SPL in Tab. 1, following the same evaluation protocol as 3D-Mem. SAFE-EQA matches or achieves state-of-the-art question-answering accuracy on both datasets, performing comparably to 3D-Mem and outperforming other baselines by over 20%. Benefiting from the compact yet comprehensive visual memory proposed by 3D-Mem, our method can retrieve informative past observations immediately once graph matching identifies a potential target region. Moreover, our globally coherent path-planning strategy yields superior exploration efficiency, reflected by the highest LLM-Match SPL. Unlike other methods that query the VLM for stepwise next-goal selection, SAFE-EQA plans multi-step trajectories by solving a semantic-aware TSP, ensuring exploration that is both consistent, efficient and task-oriented.

Results of VLM/LLM Cost. To demonstrate SAFE-EQA's reduced resource overhead, we report the token consumption of LLM and VLM queries in Tab. 2. The reduction mainly stems from two factors. First, our TSP-based planner produces multi-step trajectories, and the VLM is invoked for goal comparison only when new frontiers are detected, far less frequently than the stepwise next-goal selection used by most baselines. Second, our hierarchical answering module triggers VLM-based answer prediction only when graph matching identifies a potential target region, rather than at every step.

Methods	OpenEQA	EXPRESS-Bench	HM-EQA
3D-Mem	19488.7 (+84.1%)	23270.1 (+55.2%)	22461.7 (+38.5%)
SAFE-EQA	**10585.8**	**14991.5**	**16214.1**

Table 2: Token consumption per scene (↓). Percentages in parentheses denote the increase relative to SAFE-EQA.

Token Overhead Analysis. We conduct a breakdown experiment to analyze the contribution of the key modules to overall token overhead. Specifically, we consider the LLM-based modules *Preprocessing* and *Semantic Mapping*, as well as the VLM-based modules *Priority Assessment* and *Answering*. The analysis in Fig. 4 shows each module's token-consumption percentage across exploration trajectories of varying lengths: *Short* (< 5 m), *Middle* (5–25 m), and *Long* (> 25 m). Notably, as the explored area grows, token usage for *Semantic Mapping* and *Priority Assessment* increases proportionally, reflecting their more frequent invocation in larger environments. This underscores their crucial role in accurate task-priority identification and efficient path planning.

Runtime Analysis. We further analyze the runtime cost of the key modules on the A-EQA-41 split of OpenEQA [32]. As shown in Tab. 3, VLM-based modules dominate the per-call latency, with Answer prediction taking 11.88 s and Priority Assessment taking 4.32 s on average. Nevertheless, they are invoked only occasionally (0.24 and 0.18 times per step), while lightweight graph and planning modules are called more frequently. In particular, solving the semantic-aware TSP takes only 0.31 s per call. This scheduling is consistent with our multi-step planning and hierarchical answering design, which limits expensive VLM reasoning to critical decision points and keeps the overall exploration practical.

Modules	Avg. Time	Avg. Frequency
Scene graph querying	2.93	0.56
Landmark graph updating	2.09	0.72
Priority Assessment	4.32	0.18
Answer prediction	11.88	0.24
Solving TSP	0.31	0.23

Table 3: Runtime analysis. The average time cost (s) and invoking frequency (per step) of the key modules.

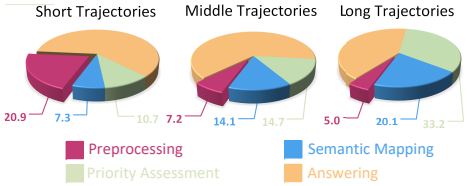

Fig. 4: Token overhead breakdown. Token usage of each module is shown for exploration paths of varying lengths.

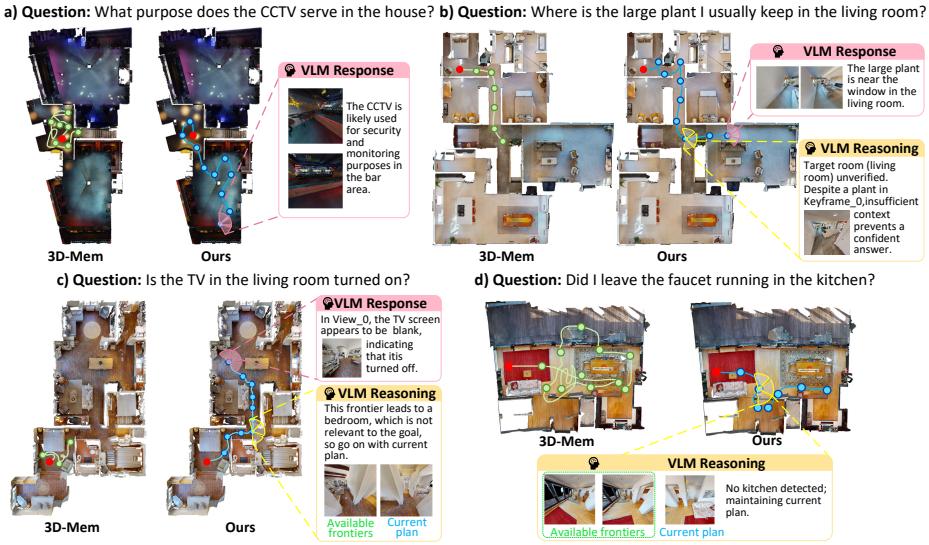

Fig. 5: Comparison of exploration trajectories. The starting point of each trajectory is marked as a red circle.

4.3 Qualitative Results

We compare exploration trajectories produced by our method and 3D-Mem in Fig. 5. When the start and target regions are in different rooms and the target is not directly visible, 3D-Mem may get stuck and oscillate around nearby frontiers. In (a) and (c), for example, it follows a zig-zag path and fails to move to other rooms, whereas our method thoroughly explores the current room before discovering the next-room entrance.

The yellow boxes in (c) and (d) show that our frontier-triggered priority assessment enables the VLM to choose the best next goal by comparing planned landmarks with newly detected frontiers. In (b), even when a target-like object appears in an irrelevant region, our method keeps exploring because the answering prompt enforces scope-consistent matching, as detailed in the appendix.

	LLM-Match SPL (↑)	Token number (↓)
Distance-only	45.7	14376.9
Semantic-aware	**53.9**	**10585.8**

Table 4: Ablation study on path planning strategy. The Distance-only TSP is solved based on geodesic distance.

	Trajectory length (↓)	Token number (↓)
wo. PA	39.4	**35921.3**
w. PA	**32.7**	38753.2

Table 5: Ablation study on priority assessment. The trajectory length is measured in meters.

4.4 Ablation Study

Effect of Semantic-aware TSP solving. To assess the effect of our path planning strategy, we compare the proposed Semantic-aware TSP with the traditional TSP, which determines the landmark visitation sequence solely by distance. As shown in Tab. 4, incorporating semantic relevance enables the agent to prioritize task-relevant regions while maintaining a globally coherent path. Compared to the vanilla TSP, our Semantic-aware TSP improves exploration efficiency by 17.9% and reduces query overhead by 26.4%. These results demonstrate that integrating semantic information into the path planning significantly enhances both the completeness and task-awareness of the exploration strategy.

Effect of Priority Assessment. To evaluate our Priority Assessment (PA) mechanism, we test questions with long exploration trajectories (> 25,m) from EXPRESS-Bench, with results shown in Tab. 5. Removing this VLM-based module slightly reduces the number of VLM queries during exploration, but the gain is limited. In contrast, it increases the average trajectory length by 20.5%, reducing exploration efficiency. These results show that PA improves exploration efficiency with a manageable query overhead.

Hyperparameter Ablations. We further conduct ablation studies to examine the sensitivity of SAFE-EQA to several key hyperparameters, with details provided in the Appendix.

5 Conclusion

We present SAFE-EQA, a semantic-aware exploration framework for Embodied Question Answering. SAFE-EQA builds a 3D Scene Graph, a Landmark Graph, and a query-conditioned Semantic Map online to support globally coherent and task-relevant exploration. By formulating exploration as a semantic-aware Traveling Salesman Problem (TSP), SAFE-EQA plans multi-step trajectories that prioritize task-relevant regions and reduce redundant motion. Its frontier-triggered Priority Assessment and hierarchical answering modules further limit unnecessary VLM queries, improving efficiency and resource utilization. Experiments on Open-EQA, EXPRESS-Bench, and HM-EQA show that SAFE-EQA achieves strong question-answering accuracy and exploration efficiency while reducing token consumption compared with SOTA methods. Overall, SAFE-EQA provides initial evidence that combining semantic reasoning with global path planning is a promising direction for efficient, task-oriented EQA.

6 Acknowledgements

This work is supported in part by the NSFC (62132021), the Major Program of Xiangjiang Laboratory (23XJ01009), NSFC (62572477, 62522219, 62372457, 62325211) and the Research Fund of Jiangsu Key Laboratory of AI for Industries (E6420016G8).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
3. Anwar, A., Welsh, J., Biswas, J., Pouya, S., Chang, Y.: Remembr: Building and reasoning over long-horizon spatio-temporal memory for robot navigation. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 2838–2845. IEEE (2025)
4. Cai, W., Ponomarenko, I., Yuan, J., Li, X., Yang, W., Dong, H., Zhao, B.: Spatialbot: Precise spatial understanding with vision language models. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 9490–9498. IEEE (2025)
5. Cai, W., Huang, S., Cheng, G., Long, Y., Gao, P., Sun, C., Dong, H.: Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 5228–5234. IEEE (2024)
6. Cao, C., Zhu, H., Choset, H., Zhang, J.: Tare: A hierarchical framework for efficiently exploring complex 3d environments. In: Robotics: Science and Systems. vol. 5, p. 2 (2021)
7. Cao, Y., Zhang, J., Yu, Z., Liu, S., Qin, Z., Zou, Q., Du, B., Xu, K.: Cognav: Cognitive process modeling for object goal navigation with llms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9550–9560 (2025)
8. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14455–14465 (2024)
9. Chen, J., Lin, B., Xu, R., Chai, Z., Liang, X., Wong, K.Y.: Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 9796–9810 (2024)
10. Cheng, K., Li, Z., Sun, X., Min, B.C., Bedi, A.S., Bera, A.: Efficienteqa: An efficient approach to open-vocabulary embodied question answering. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 20357–20364. IEEE (2025)
11. Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., Batra, D.: Embodied question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–10 (2018)

12. Das, A., Gkioxari, G., Lee, S., Parikh, D., Batra, D.: Neural modular control for embodied question answering. In: Conference on robot learning. pp. 53–62. PMLR (2018)
13. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al.: Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378 (2023)
14. Gao, J., Sarkar, B., Xia, F., Xiao, T., Wu, J., Ichter, B., Majumdar, A., Sadigh, D.: Physically grounded vision-language models for robotic manipulation. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 12462–12469. IEEE (2024)
15. Gordon, D., Kembhavi, A., Rastegari, M., Redmon, J., Fox, D., Farhadi, A.: Iqa: Visual question answering in interactive environments. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4089–4098 (2018)
16. Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K.M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R., et al.: Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 5021–5028. IEEE (2024)
17. Holz, D., Basilico, N., Amigoni, F., Behnke, S.: Evaluating the efficiency of frontier-based exploration strategies. In: ISR 2010 (41st International Symposium on Robotics) and ROBOTIK 2010 (6th German Conference on Robotics). pp. 1–8. VDE (2010)
18. Hong, Y., Lin, C., Du, Y., Chen, Z., Tenenbaum, J.B., Gan, C.: 3d concept learning and reasoning from multi-view images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9202–9212 (2023)
19. Hu, Y., Lin, F., Zhang, T., Yi, L., Gao, Y.: Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. arXiv preprint arXiv:2311.17842 (2023)
20. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
21. Jiang, K., Liu, Y., Chen, W., Luo, J., Chen, Z., Pan, L., Li, G., Lin, L.: Beyond the destination: A novel benchmark for exploration-aware embodied question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9091–9101 (2025)
22. Kalra, N., Ferguson, D., Stentz, A.: Incremental reconstruction of generalized voronoi diagrams on grids. Robotics and Autonomous Systems **57**(2), 123–128 (2009)
23. Kamath, A., Hessel, J., Chang, K.W.: What's "up" with vision-language models? investigating their struggle with spatial reasoning. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 9161–9175 (2023)
24. Karamcheti, S., Nair, S., Balakrishna, A., Liang, P., Kollar, T., Sadigh, D.: Prismatic vlms: Investigating the design space of visually-conditioned language models. In: Forty-first International Conference on Machine Learning (2024)
25. Khatib, O.: Real-time obstacle avoidance for manipulators and mobile robots. The international journal of robotics research **5**(1), 90–98 (1986)
26. Koren, Y., Borenstein, J., et al.: Potential field methods and their inherent limitations for mobile robot navigation. In: Icra. vol. 2, pp. 1398–1404 (1991)

27. Kuang, Y., Lin, H., Jiang, M.: Openfmnav: Towards open-set zero-shot object navigation via vision-language foundation models. In: Findings of the Association for Computational Linguistics: NAACL 2024. pp. 338–351 (2024)
28. Kulich, M., Faigl, J., Přeučil, L.: On distance utility in the exploration task. In: 2011 ieee international conference on robotics and automation. pp. 4455–4460. IEEE (2011)
29. Lee, S., Shakir, A., Koenig, D., Lipp, J.: Open source strikes bread - new fluffy embeddings model (2024), https://www.mixedbread.ai/blog/mxbai-embed-large-v1
30. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
31. Ma, X., Yong, S., Zheng, Z., Li, Q., Liang, Y., Zhu, S.C., Huang, S.: Sqa3d: Situated question answering in 3d scenes. arXiv preprint arXiv:2210.07474 (2022)
32. Majumdar, A., Ajay, A., Zhang, X., Putta, P., Yenamandra, S., Henaff, M., Silwal, S., Mcvay, P., Maksymets, O., Arnaud, S., et al.: Openeqa: Embodied question answering in the era of foundation models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16488–16498 (2024)
33. Ok, K., Ansari, S., Gallagher, B., Sica, W., Dellaert, F., Stilman, M.: Path planning with uncertainty: Voronoi uncertainty fields. In: 2013 IEEE International Conference on Robotics and Automation. pp. 4596–4601. IEEE (2013)
34. Ramakrishnan, S.K., Gokaslan, A., Wijmans, E., Maksymets, O., Clegg, A., Turner, J., Undersander, E., Galuba, W., Westbury, A., Chang, A.X., et al.: Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. arXiv preprint arXiv:2109.08238 (2021)
35. Ren, A.Z., Clark, J., Dixit, A., Itkina, M., Majumdar, A., Sadigh, D.: Explore until confident: Efficient exploration for embodied question answering. arXiv preprint arXiv:2403.15941 (2024)
36. Saxena, S., Buchanan, B., Paxton, C., Liu, P., Chen, B., Vaskevicius, N., Palmieri, L., Francis, J., Kroemer, O.: Grapheqa: Using 3d semantic scene graphs for real-time embodied question answering. In: 9th Annual Conference on Robot Learning
37. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: A family of highly capable multimodal models. arxiv 2023. arXiv preprint arXiv:2312.11805 (2024)
38. Team, Q.: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388
39. Wijmans, E., Datta, S., Maksymets, O., Das, A., Gkioxari, G., Lee, S., Essa, I., Parikh, D., Batra, D.: Embodied question answering in photorealistic environments with point cloud perception. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6659–6668 (2019)
40. Wu, P., Mu, Y., Wu, B., Hou, Y., Ma, J., Zhang, S., Liu, C.: Voronav: Voronoi-based zero-shot object navigation with large language model. In: International Conference on Machine Learning. pp. 53757–53775. PMLR (2024)
41. Xi, Y., Zhu, C., Duan, Y., Yi, R., Zheng, L., He, H., Xu, K.: Thp: Tensor-field-driven hierarchical path planning for autonomous scene exploration with depth sensors. Computational Visual Media **10**(6), 1121–1135 (2024)
42. Xie, Q., Min, S.Y., Ji, P., Yang, Y., Zhang, T., Xu, K., Bajaj, A., Salakhutdinov, R., Johnson-Roberson, M., Bisk, Y.: Embodied-rag: General non-parametric embodied memory for retrieval and generation. arXiv e-prints pp. arXiv–2409 (2024)
43. Xu, K., Zheng, L., Yan, Z., Yan, G., Zhang, E., Niessner, M., Deussen, O., Cohen-Or, D., Huang, H.: Autonomous reconstruction of unknown indoor scenes guided

by time-varying tensor fields. ACM Transactions on Graphics (TOG) **36**(6), 1–15 (2017)
44. Yamauchi, B.: A frontier-based approach for autonomous exploration. In: Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97.'Towards New Computational Principles for Robotics and Automation'. pp. 146–151. IEEE (1997)
45. Yang, Y., Liu, J., Zhang, Z., Zhou, S., Tan, R., Yang, J., Du, Y., Gan, C.: Mindjourney: Test-time scaling with world models for spatial reasoning. arXiv preprint arXiv:2507.12508 (2025)
46. Yang, Y., Yang, H., Zhou, J., Chen, P., Zhang, H., Du, Y., Gan, C.: 3d-mem: 3d scene memory for embodied exploration and reasoning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17294–17303 (2025)
47. Yin, H., Xu, X., Zhao, L., Wang, Z., Zhou, J., Lu, J.: Unigoal: Towards universal zero-shot goal-oriented navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19057–19066 (2025)
48. Yu, L., Chen, X., Gkioxari, G., Bansal, M., Berg, T.L., Batra, D.: Multi-target embodied question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6309–6318 (2019)
49. Zhai, M., Gao, Z., Wu, Y., Jia, Y.: Memory-centric embodied question answering. arXiv preprint arXiv:2505.13948 (2025)
50. Zhang, H., Li, F., Zou, X., Liu, S., Li, C., Yang, J., Zhang, L.: A simple framework for open-vocabulary segmentation and detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1020–1031 (2023)