Beyond the Destination: A Novel Benchmark for Exploration-Aware Embodied Question Answering

Kaixuan Jiang[1] Yang Liu[1,4*] Weixing Chen[1] Jingzhou Luo[1] Ziliang Chen[2] Ling Pan[3]
Guanbin Li[1,2,4] Liang Lin[1,2,4]

[1]Sun Yat-sen University, China [2]Peng Cheng Laboratory [3] Hong Kong University of Science and Technology
[4] Guangdong Key Laboratory of Big Data Analysis and Processing

jiangkx3@mail2.sysu.edu.cn, liuy856@mail.sysu.edu.cn, chenwx228@mail2.sysu.edu.cn,
luojzh5@mail2.sysu.edu.cn, c.ziliang@yahoo.com, penny.ling.pan@gmail.com
liguanbin@mail.sysu.edu.cn, linliang@ieee.org

https://github.com/HCPLab-SYSU/EXPRESS-Bench

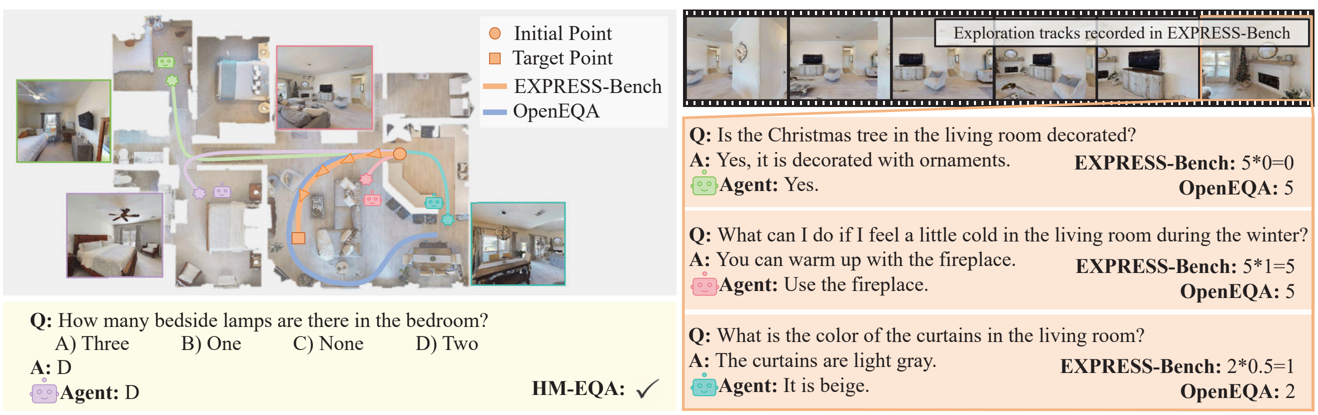

Figure 1. Comparison of our EXPRESS-Bench with other EQA datasets. The orange trajectory in the top-down map shows a complete exploration path from EXPRESS-Bench, with observation images at key waypoints (top-right). Data for this path is in the orange box. The blue trajectory simulates OpenEQA's episodic memory, passing near the target but not ending there. The yellow box simulates how multiple-choice data is generated in HM-EQA, lacking the exploration path. For each question, answers are based on visual observations at the endpoint, scored according to each dataset's evaluation method. Unlike HM-EQA and OpenEQA, which may give higher scores based on answer similarity, EXPRESS-Bench adjusts scores for incorrect or fabricated answers by grounding them in the agent's observations.

Abstract

Embodied Question Answering (EQA) is a challenging task in embodied intelligence that requires agents to dynamically explore 3D environments, actively gather visual information, and perform multi-step reasoning to answer questions. However, current EQA approaches suffer from critical limitations in exploration efficiency, dataset design, and evaluation metrics. Moreover, existing datasets often introduce biases or prior knowledge, leading to disembodied reasoning, while frontier-based exploration strategies struggle in cluttered environments and fail to ensure fine-grained exploration of task-relevant areas. To address these challenges, we construct the ***EXP****loration-awa****R****e* ***E****mbodied que****S****tion an****S****wering* ***Bench****mark (EXPRESS-Bench), the largest dataset designed specifically to evaluate both exploration and reasoning capabilities. EXPRESS-Bench consists of 777 exploration trajectories and 2,044 question-trajectory pairs. To improve exploration efficiency, we propose Fine-EQA, a hybrid exploration model that integrates frontier-based and goal-oriented navigation to guide agents toward task-relevant regions more effectively. Additionally, we introduce a novel evaluation metric, Exploration-Answer Consistency (EAC), which ensures faithful assessment by measuring the alignment between answer grounding and exploration reliability. Extensive experimental comparisons with state-of-the-art EQA models demonstrate the effectiveness of our EXPRESS-Bench in advancing embodied exploration and question reasoning.*

*Corresponding author

Table 1. Comparison to EQA benchmarks.

	Simulator	Dataset	Real Scenes	Exploration Track	Track Numbers	Target Point	Question Creation	Open Vocab
EQA-v1 [7]	House3D	SUNCG	✗	✓	-	✓	Rule-Based	✗
MP3D-EQA [37]	MINOS	MP3D	✓	✓	-	✓	Rule-Based	✗
MT-EQA [43]	House3D	SUNCG	✗	✓	-	✓	Rule-Based	✗
IQA [11]	AI2THOR	-	✗	✗	-	✗	Rule-Based	✗
VideoNavQA [4]	House3D	SUNCG	✗	✗	-	✗	Rule-Based	✗
K-EQA [35]	AI2Thor	-	✗	✗	-	✗	Rule-Based	✗
HM-EQA [31]	Habitat	HM3D	✓	✗	-	✗	VLMs	✗
S-EQA [8]	VirtualHome	-	✗	✗	-	✗	LLMs	✗
NoisyEQA [39]	-	-	-	✗	-	✗	VLMs	✓
CityEQA [45]	EmbodiedCity	-	✗	✗	-	✓	Manual	✓
OpenEQA [29]	Habitat	ScanNet/HM3D	✓	✓	152	✗	Manual	✓
EXPRESS-Bench (Ours)	Habitat	HM3D	✓	✓	777	✓	VLMs	✓

1. Introduction

Embodied Question Answering (EQA) is a pivotal challenge at the intersection of computer vision, natural language processing, and embodied intelligence. In this task, an embodied agent must navigate a 3D environment, actively gather visual information through exploration, and answer questions about the scene [7, 29]. Unlike traditional question-answering (QA) systems that rely on static images [3, 16] or pre-existing knowledge bases [2, 20], EQA demands sequential decision-making, where agents must dynamically explore their surroundings to acquire the necessary information before formulating an answer. Thus, traditional QA methods fail to generalize to EQA because they are not designed to handle the dynamic, multi-step reasoning and embodied navigation essential for EQA [34, 46]. This unique combination of perception, reasoning, and action makes EQA a compelling yet challenging problem with implications for real-world applications such as robotics, virtual assistants, and autonomous navigation [24].

Despite its potential, current EQA methods face a fundamental limitation: models often generate answers without truly grounding them in exploration [29]. This issue stems from the reliance on manually created [7] or rule-based datasets [37], which are costly, inflexible, and prone to bias. As a result, agents frequently exploit spurious correlations rather than perform meaningful reasoning, undermining real-world reliability [23, 27]. Additionally, ambiguity in question design and scenario complexity often lead to non-unique answers, making evaluation unreliable [28]. More critically, some datasets embed prior knowledge, enabling models to generate answers without actual exploration—an issue known as unfaithful question answering [44]. For example, explicitly mentioning locations like "living room" allows guessing without interaction. Evaluation metrics further compound these challenges. Existing metrics fail to assess answer reliability and do not effectively detect model hallucinations—where responses seem plausible but are incorrect [5, 17]. This leads to biased assessments that overlook an agent's true exploration and reasoning abilities. Moreover, commonly used metrics, which primarily measure answer similarity, lack the granularity needed to evaluate exploration efficiency and quality, limiting their applicability to real-world scenarios [21]. These gaps highlight the urgent need for more comprehensive EQA benchmarks that assess both the exploration process and answer quality.

In addition to the limitations of benchmark datasets and evaluation metrics, existing EQA methods also suffer from a critical deficiency: inefficient exploration capabilities. Most current EQA models rely on frontier-based exploration [10, 14, 41, 42], where agents expand their search by identifying and navigating towards unexplored frontiers on a map. While effective in open spaces, this approach struggles in constrained environments such as narrow corridors or cluttered scenes, limiting the agent's capacity for comprehensive exploration [13, 30]. Furthermore, frontier-based strategies often lead to inefficient behaviors, such as redundant revisits to semantically important regions without acquiring new information. Crucially, existing methods lack mechanisms for fine-grained exploration of task-relevant areas, preventing agents from gathering the detailed environmental context needed for accurate reasoning [36]. Therefore, it is essential to integrate exploration efficiency with answer reliability to ensure meaningful reasoning.

To address these challenges, we introduce the exploration-aware EQA task, which emphasizes the need for agents to actively and rationally explore relevant environmental clues before answering. To comprehensively assess exploration abilities, we construct EXPRESS-Bench, a large-scale benchmark comprising 777 exploration trajectories and 2,044 question-trajectory pairs—offering superior coverage compared to OpenEQA (see Tab.1). To enhance exploration efficiency, we propose Fine-EQA, a hybrid method integrating frontier-based and goal-oriented exploration. Leveraging a global semantic map and functional region semantic map, Fine-EQA enables agents to efficiently navigate entire scenes while conducting fine-grained investigations of task-relevant areas. For faithful answer evaluation, we introduce the Exploration-Answer Consistency (EAC) metric, which explicitly measures the alignment between an agent's exploration process and its generated answers. EAC effectively detects ungrounded responses that appear correct but lack supporting exploration evidence, providing a more rigorous assessment of model performance. Our main contributions are:

- We introduce EXPRESS-Bench, a high-quality, large-scale EQA dataset designed to support active exploration while addressing key limitations of existing datasets, such

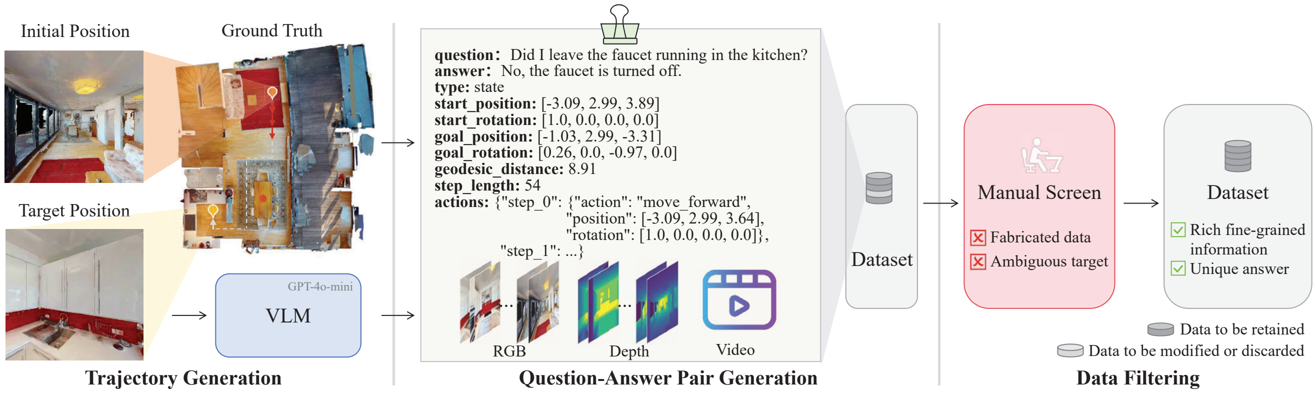

Figure 2. The construction process of EXPRESS-Bench.

as ambiguity and over-reliance on prior knowledge.

- We propose Fine-EQA, a two-stage exploration framework that enhances navigation efficiency and enables fine-grained exploration of task-relevant regions with flexible strategy switch, setting a new baseline for EQA.
- To ensure a faithful assessment of EQA, we introduce the EAC metric, which simultaneously evaluates answer grounding and exploration efficiency, providing a more rigorous measure of the model performance.

2. Related Work

2.1. EQA Dataset

Compared to visual question answering (VQA) datasets, EQA datasets are constructed in a three-dimensional space, encompassing both static scene descriptions and dynamic interactions, making their development a significantly more challenging task. Early works [4, 7, 11, 35, 37, 43] adopted a template-based approach to accelerate data generation; however, this method often led to simplistic question formats and straightforward answers. With the advent of large models, [8, 31, 39] leveraged their capabilities to facilitate a more efficient dataset construction process that better aligns with research needs, thereby greatly enhancing data diversity and richness. [29] proposed an open-ended dataset through manual design, demonstrating innovation but primarily focusing on scenario-based memory questions while overlooking the critical role of active exploration in EQA tasks. [45] extended the task to city spaces, incorporating the complexities of urban environments. However, the complexity of 3D environments often leads to non-unique answers, making model evaluation challenging. Besides, existing datasets rarely support active exploration. Therefore, we propose an exploration-aware EQA benchmark that enables active exploration and ensures unique answers, providing high-quality evaluation and numerous EQA tasks.

2.2. Large Models for Embodied Agents

The strong reasoning and generalization capabilities of large models have driven their widespread adoption in embodied tasks, such as vision-language navigation [25, 26, 46] and embodied manipulation [18, 38, 40]. However, EQA demands more comprehensive reasoning, encompassing not only navigation and interaction but also question answering based on environmental information. This heightened requirement for semantic understanding makes it difficult to directly transfer methods from other tasks. In EQA, some studies use these models for target object detection [32] and subtask planning [45]. They are also employed to assign confidence scores to exploration directions [6, 31] or generate semantic labels [33] to update semantic maps, integrating a frontier-based exploration strategy. Yet, complex scenes often hinder exploration, causing inefficiencies or hallucinated responses due to missing critical information. To address this, we propose a two-stage EQA model that adaptively switches between frontier-based and goal-oriented exploration, enhancing efficiency and robustness.

3. EXPRESS-Bench

In this section, we describe how to construct the EXPRESS-Bench utilizing the multimodal large model GPT-4o-mini. Unlike existing EQA datasets, we not only record the generated questions and ideal responses, but also record the ground truth actions, as well as the state and information acquired by the agent after each operational step.

3.1. Simulator

HM3D is a comprehensive dataset featuring 3D reconstructions of 1,000 large-scale buildings from a wide range of real-world locations. We utilize the portion of the HM3D dataset that are rich in semantic information on the Habitat simulator to replicate real-world environments, and make extensive use of the Habitat interface to construct the data.

3.2. Dataset Generation Pipeline

In this section, we provide an overview of the question-answer pairs creation pipeline to create EXPRESS-Bench. The dataset construction process is illustrated in Fig.2.

Stage 1: Trajectory Generation. We randomly sample navigable initial and target positions within the scene. The shortest sequence of atomic actions ("move forward," "turn left," "turn right") and corresponding step count required to reach the goal from the starting position are computed as ground truth, along with the shortest geodesic distance. Given the large scale of the simulated environment, we constrain both the initial and target positions to the same floor and ensure the step length falls between 10 and 100. After each atomic action, the agent's state including coordinates, orientation and first-person visual observation is recorded. Based on the collected visual data, we further generate a trajectory video that visualizes the agent's complete exploration process from a first-person perspective.

Stage 2: Question-Answer Pair Generation. In EQA, the agent's final visual observation is key to answering the question. Therefore, we feed this observation into the multimodal large model, along with several example question-answer pairs. These questions are designed from a human perspective to simulate natural conversations in everyday home scenarios. Guided by the prompt, GPT-4o-mini generates both questions and corresponding answers. The answers are open-ended rather than simple yes/no or multiple-choice responses, which helps reduce data bias and prevents models from relying solely on their inherent common sense, placing higher demands on models' performance.

Stage 3: Data Filtering. Although large models perform well, their outputs are not always reliable and reasonable. In large environments with duplicate rooms or objects (e.g., multiple bedrooms), questions may have non-unique answers, necessitating manual data screening. First, we ensure all answers are relevant to the scene. Then, using trajectory videos and top-down maps from the Habitat simulator, we track the agent's movements to resolve cases where target regions are unclear or unreachable. Specifically, we retain only those questions where either no other regions of the same type as the target region, or where the target region is the closest to the initial position among similar regions, ensuring the uniqueness of the answer. When necessary, we also specify key regions of the questions and include contextual details, such as the location, attributes, and relationships of the target object. We invited several individuals to review the data. To ensure the dataset's high quality, a curator performed a final check to verify its overall consistency.

3.3. Dataset Statistics

Finally, the EXPRESS-Bench contains 777 trajectories, encompassing a total of 2,044 question-trajectory pairs. The EXPRESS-Bench predominantly consists of questions in the following seven categories: *state*, *knowledge*, *location*, *attribute*, *counting*, *existence*, and *object*. The distribution of these categories is shown in Fig.3. On average, the agent navigated the scene with a step size of 39.8 steps per question, covering a geodesic distance of 6.6 meters. More analysis of EXPRESS-Bench is in the Supplementary Material.

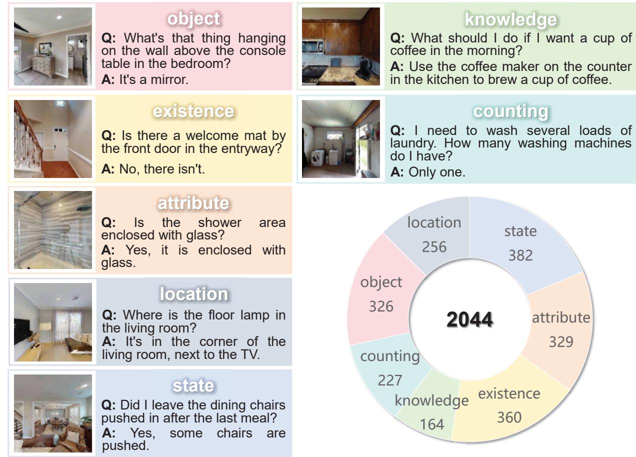

Figure 3. Overview of the EXPRESS-Bench statistics.

3.4. Exploration-Answer Consistency Metric

Since the answers are open-vocabulary, directly assessing their correctness is not feasible. OpenEQA leverages GPT-4 to evaluate the model's output, assigning a score from 1 to 5. However, OpenEQA overlooks the grounding of the answers—meaning the model's response may be fabricated and unrelated to the environmental information observed by the agent. To address this limitation, we refine the correctness evaluation score, σ_i, which assesses the model's response based on the question, the correct answer and the visual observation. Additionally, we introduce a grounding evaluation score, δ_i, to assess the relevance of the answer to the environment with the support of the VLM:

$$\sigma_i = \varphi(Q_i, A_i^*, A_i, I_i) \tag{1}$$

$$\delta_i = \psi(Q_i, A_i, I_i) \tag{2}$$

where Q_i is the given question, A_i^* is the correct answer, A_i is the model's response, and I_i is the environment image from the agent's final first-person observation. The score of an answer is calculated as $\sigma_i * \delta_i$.

Due to the open-ended nature of answers, when the model's response aligns with the question and image but differs from the correct answer, incorporating the observed image into the evaluation of σ_i also enables the VLM to assign more precise scores. Guiding prompt for the model's scoring process is provided in the Supplementary Material.

The VLM assigns the σ_i score on a scale of 1 to 5, with higher scores indicating that the model's response is more accurate and closer to the correct answer. The δ_i score is assigned as 0, 0.5, or 1. A score of 1 signifies that the agent's observation is relevant to the question and the model's description of the environment is accurate, aligning with the agent's perception. A score of 0.5 denotes relevant observation but an incorrect model description. A score of 0

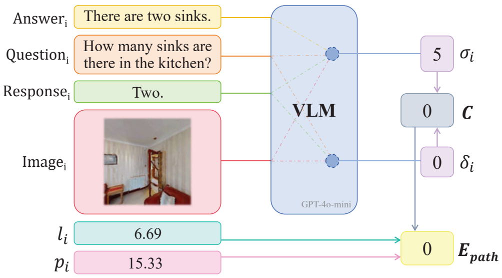

Figure 4. Exploration-Answer Consistency Metric.

implies irrelevant observation, with the model fabricating output unrelated to the environment, even if it matches the correct answer. The evaluation pipeline is shown in Fig.4.

Next, we calculate the overall correctness C of the output and the model's efficiency E_{path} in completing the task:

$$C = \frac{1}{N}\sum_{i=1}^{N}\frac{\sigma_i \times \delta_i}{5} \times 100\% \tag{3}$$

$$E_{path} = \frac{1}{N}\sum_{i=1}^{N}\frac{\sigma_i \times \delta_i}{5} \times \frac{l_i}{max(p_i, l_i)} \times 100\% \tag{4}$$

where N is the total number of questions, l_i represents the distance the agent navigate along the ground truth path that is sufficient to complete the task, and p_i is the actual distance the agent moves during the experiment.

Furthermore, we assess the agent's navigation performance by measuring the geodesic distance d_T from its final exploration position P_E^i to the target location P_T^i:

$$d_T = \frac{1}{N}\sum_{i=1}^{N} dis_g(P_E^i, P_T^i) \tag{5}$$

While the EQA task does not require the agent to reach the target as closely as navigation tasks, using this metric allows for a more thorough evaluation of the model.

4. Fine-EQA

We propose Fine-EQA, a flexible two-stage EQA framework that strategically integrates frontier-based exploration with goal-oriented exploration. Fine-EQA constructs dual semantic representations, global semantic maps and functional region semantic maps, to guide agents through complex environments with greater efficiency while ensuring thorough investigation of task-relevant areas.

4.1. Overview

Fine-EQA operates in two phases: exploration and question answering, as shown in Fig.5. The agent builds and maintains a comprehensive environmental representation using a 3D voxel map ($M_3 \in R^{d_x \times d_y \times d_z}$), which is projected onto a 2D planar map ($M_2 \in R^{d_x \times d_y}$) to track exploration states and spatial occupancy. Dimensions d_x, d_y, and d_z are set based on the scene size. During exploration, the agent alternates between frontier-based and goal-oriented strategies, with region-specific limits to avoid over-searching, while GPT-4o-mini decides when to stop and answer:

- **Frontier-Based Exploration (FBE)** identifies boundaries between known and unknown regions to expand environmental understanding.
- **Goal-Oriented Exploration (GOE)** targets regions with high task relevance based on semantic analysis.

4.2. Frontier-Based Exploration

FBE is fundamental to systematic environmental discovery, focusing on the boundaries between explored and unexplored regions. Inspired by [31], we enhance traditional frontier-based approaches by incorporating semantic understanding into the exploration process. We construct a global semantic map $M_{sem} \in R^{d_x \times d_y}$ by integrating both global semantic values v_g and local semantic values v_l. At each time step, the agent projects its current RGB observation onto the 2D map M_2 and applies farthest point sampling to identify navigable points that maximize spatial coverage. These sampled points p_{sample} are back-projected onto the original RGB image, where our VLM evaluates their exploration priority based on task relevance, establishing v_l. Concurrently, we extract the exploration decision confidence score of the VLM for the scene depicted in the RGB image, denoted as v_g. We compute the semantic value of p_{sample} through the weighted fusion of v_l and v_g and dynamically update M_{sem} using Gaussian smoothing:

$$M_{sem} \leftarrow (p_{sample}, v_l, v_g) \tag{6}$$

To identify frontier points, we analyze the exploration states surrounding each point in M_2 and cluster them into candidate frontier points $F = \{f_1, f_2, ..., f_n\}$. Each point f_i receives a composite weight w_i determined by:

$$w_i = \omega(v_{sem}^i, r_e^i, r_o^i, dis(f_i, p_{cur})) \tag{7}$$

where v_{sem}^i represents the semantic value, r_e^i and r_o^i are the unexplored rate and the unoccupied rate along the exploration direction, respectively. $dis(f_i, p_{cur})$ is the Euclidean distance to the agent's current position p_{cur}. ω incorporates an exponentially weighted enhancement of r_e^i, r_o^i, and v_{sem}^i near the frontier point, while simultaneously applying an exponential decay based on $dis(f_i, p_{cur})$. This encourages the agent to prioritize unexplored areas and minimize local redundancy, ultimately enhancing exploration efficiency.

The agent then selects the next location χ for exploration by treating these weights as probabilities:

$$\chi = \gamma(F, W),\ W = \{w_i / \sum_{j=1}^{N} w_j, i = 1, 2, ..., N\} \tag{8}$$

where γ represents the random sampling of points with different probabilities.

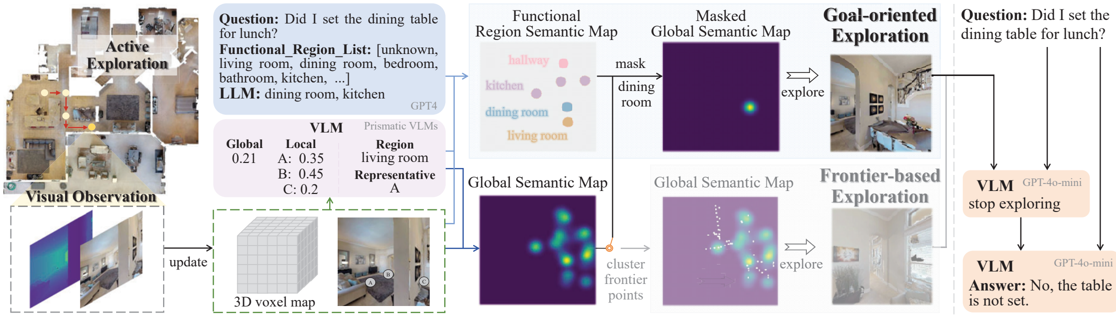

Figure 5. The Fine-EQA framework operates as follows: The agent initially performs coarse-grained exploration using a frontier-based strategy, then switches to goal-oriented fine-grained exploration once task-relevant regions are identified. A maximum exploration limit per region prevents excessive searching, prompting the agent to either return to frontier-based exploration or focus on the next most promising region. Throughout this process, the VLM continuously evaluates the relevance and completeness of the acquired information, guiding the agent's decision to either continue exploration or generate answers based on the most recent visual inputs, as detailed in the Appendix.

4.3. Goal-Oriented Exploration

The primary limitation of pure frontier-based exploration is its inability to comprehensively explore task-relevant regions, particularly in complex environments with occlusions and spatial constraints. Our goal-oriented exploration strategy addresses this limitation by explicitly modeling functional regions and their relevance to the given question.

Functional Region Semantic Mapping. We construct a functional region semantic map $M_{reg} \in R^{d_x \times d_y}$ to guide the agent toward areas with high task relevance. During observation, the VLM classifies the current scene into specific functional regions (e.g., kitchen, bedroom) and identify representative points q within these regions from p_{sample}. Formally, when a region Reg is identified with confidence exceeding a predefined threshold, we update M_{reg} around q:

$$M_{reg}(N(q)) = \text{ID}_{Reg} \tag{9}$$

where $N(q)$ is the neighborhood of point q and ID_{Reg} represents the semantic value associated with Reg. Points with identical semantic values in M_{reg} that exhibit spatial proximity are merged into coherent regions, creating a comprehensive functional decomposition of the environment.

Task-Relevant Region Prioritization. We use LLMs to analyze the question and rank regions by task relevance. This prioritization is essential because: 1) secondary regions often provide context or pathways to primary regions, and 2) exploring multiple regions strategically improves spatial awareness and navigation efficiency. It also enables explicitly modeling of scenarios with multiple task-related regions. When the agent identifies a higher-priority region, it shifts to goal-oriented exploration within that region.

Masked Semantic Mapping. To focus on high-priority regions, we apply masking operations ϕ to the global semantic map based on the functional region semantic map:

$$M_{masked} = \phi(M_{sem}, M_{reg}, r) \tag{10}$$

where r is the region with the current highest priority. This masked semantic map $M_{masked} \in R^{d_x \times d_y}$ exclusively contains semantic values for the prioritized region, effectively directing the agent's attention toward task-relevant areas. To prevent redundant exploration, previously visited points receive decreased semantic values within M_{masked}. With the masked semantic map, the agent selects the position $\chi = \text{argmax}_{x,y}(M_{masked})$ with the highest semantic value for further exploration. This adaptive switching between exploration strategies ensures efficient coverage of the environment and thorough investigation of task-relevant areas.

5. Experiments

We evaluate various models on our EXPRESS-Bench, reporting the accuracy of their responses. We also present the efficiency and navigation performance of agents equipped with exploration capabilities.

5.1. Baselines

We compare various models in a zero-shot setting, including 1) Blind LLMs, 2) Socratic Models, 3) Multi-Frame VLMs, 4) Exploring Agents, and 5) Human Performance.

Blind LLMs. The Blind LLMs refers to LLMs that produce answers based purely on questions, completely disregarding any contextual information from embodied scenarios, formulated as $A = \text{LLMs}(Q)$. Specifically, we employ DeepSeek-V3 [12, 22], GPT-4 [1], and LLaMA-3-8b [9].

Socratic Models. Using ground truth data G, we simulate an agent's exploration to obtain first-person visual observations, from which we randomly sample k frames. Each frame G_i is described by an image captioning model as $d_i = \text{Captioner}(Q, G_i)$, which are then concatenated into $D = [d_1, d_2, ..., d_k]$. We set $k = \lceil \frac{L}{10} \rceil$, where L denotes the step length of the ground truth. The LLMs then generate answers based on the question and the concatenated descriptions, i.e., $A = \text{LLMs}(Q, D)$. For caption generation, we employ GPT-4o-mini[15] and LLaVA-v1.5-7b, and for the LLMs, we use DeepSeek-V3, GPT-4 and LLaMA-3-8b.

Multi-Frame VLMs. We randomly select k frames from the ground truth, which are subsequently fed into the

Table 2. Performance comparison on EXPRESS-Bench.

	$C\uparrow$	$C^*\uparrow$	$E_{path}\uparrow$	$d_T\downarrow$
Human Agent	-	83.99	-	-
Blind LLMs				
DeepSeek-V3	-	59.15	-	-
GPT4	-	58.96	-	-
LLaMA-3-8b	-	57.25	-	-
Socratic Models				
DeepSeek-V3 w/ GPT-4o-mini	-	62.60	-	-
GPT4 w/ GPT-4o-mini	-	62.56	-	-
LLaMA-3-8b w/ GPT-4o-mini	-	59.95	-	-
DeepSeek-V3 w/ LLaVA-v1.5-7b	-	60.63	-	-
GPT4 w/ LLaVA-v1.5-7b	-	59.53	-	-
LLaMA-3-8b w/ LLaVA-v1.5-7b	-	58.59	-	-
Multi-Frame VLMs				
GPT-4o-mini	-	58.37	-	-
LLaVA-v1.5-7b	-	57.66	-	-
Exploring Agents				
RE	36.95	62.75	12.06	7.26
FBE	38.60	60.24	14.55	6.64
GOE	38.54	63.34	12.74	6.46
Fine-EQA (Ours)	**40.55**	**63.95**	**16.22**	**6.43**

VLMs, i.e., $A = \text{VLMs}(Q, G_1, G_2, ..., G_k)$. We employ GPT-4o-mini and LLaVA-v1.5-7b.

Exploring Agents. We use various exploration strategies S to guide the agent, i.e. $\chi = \pi(S)$. The final answer is generated based on the visual information I at the termination point, given by $A = \text{VLMs}(Q, I)$.

(1) Random Exploration (RE). The agent randomly explores the scene by selecting a movement distance and direction, navigating to the nearest point. To avoid excessive exploration, GPT-4o-mini decides when to stop.

(2) Frontier-Based Exploration (FBE). We use a pure frontier exploration method that ignores semantic information, with GPT-4o-mini determining when to stop.

(3) Goal-Oriented Exploration (GOE). The agent starts with random exploration and switches to goal-oriented exploration upon reaching a task-relevant region. Similar to Fine-EQA, once the maximum exploration limit for a given region is reached, the agent transitions to either random exploration or explores a suboptimal region.

(4) Fine-EQA (Ours). For LLMs, we use GPT-4 to identify task-related functional regions. For VLMs, we employ Prismatic VLMs (Prism-DINOSigLIP 7B) [19] to assign semantic values and evaluate functional regions. GPT-4o-mini is used to decide when to stop exploration and generate responses.

Human Agent. Five participants review the ground truth and provide answers, denoted as $A = \text{Human}(Q, G)$. Their scores are then averaged.

5.2. Comparison with State-of-the-art Methods

Besides the metrics in Section 3.4, we also compute C^*, which is C without answer grounding (i.e., setting $\delta_i = 1$). Since models like GPT-4 cannot perform exploration, we only assess their performance using C^*, as shown in Tab.2.

Non-Exploratory Agents. Among LLMs without vision, DeepSeek-V3 performs best. When paired with VLMs for reasoning, their performance is improved, with DeepSeek-V3-based socratic agents yielding the highest results. However, since EXPRESS-Bench focuses on active exploration, with questions typically related to the last frame, gains from passively viewing video frames remain limited. Additionally, Multi-Frame VLMs performs worse than the Socratic models, indicating that critical information extraction is essential for exploration-aware EQA than simply using numerous images. This supports the reliability of EXPRESS-Bench in evaluating active exploration capabilities.

Active Exploration Agents. Agents with active exploration demonstrate enhanced environmental perception and outperform nearly all non-exploratory models. RE explores freely until finding the answer or reaching the maximum exploration limit, leading to a high C^* value. However, its C value drops significantly, while its d_T and E_{path} metrics perform worse than other methods. This indicates overestimated performance due to model hallucinations. In contrast, FBE excels in unknown environments, and GOE in focused target exploration, leading to better E_{path} and d_T, with FBE slightly ahead in C. By strategically integrating both strategies, our Fine-EQA surpasses all baselines across these metrics, demonstrating the effectiveness of our exploration strategy. In terms of navigation performance, Fine-EQA performs well on d_T, indicating it efficiently reaches target regions and enhance visual perception. Though EQA does not require exact positioning, results suggest that agents closer to the target tend to provide more accurate answers, which aligns with intuition.

Human Agent. Although Fine-EQA performs well, all methods exhibit a significant gap in C^* scores compared to human performance. This result validates the effectiveness of EXPRESS-Bench while highlighting the limitations of current approaches in complex environments, emphasizing the need for further improvements of EQA methods.

To further assess Fine-EQA, we test it on two additional datasets and the results are listed in the Appendix. Furthermore, we analyze the ambiguity rate in questions, finding 35.6% in HM-EQA, 62.3% in Open-EQA, and only 9.7% in EXPRESS-Bench. The lack of full exploration trajectories or clear target locations in HM-EQA and Open-EQA hinders evaluation reliability. In contrast, EXPRESS-Bench offers clearer, more interpretable assessments.

Table 3. Ablation study of model modules.

	$C\uparrow$	$C^*\uparrow$	$E_{path}\uparrow$	$d_T\downarrow$
Fine-EQA w/o FBE	38.54	63.34	12.74	6.46
Fine-EQA w/o GOE	39.63	60.74	14.64	6.54
Fine-EQA	**40.55**	**63.95**	**16.22**	**6.43**

5.3. Ablation Studies

As shown in Tab.3, the ablation of both FBE and GOE leads to a decline in performance, affecting both response accuracy and exploration efficiency. Specifically, the FBE module has a greater impact, highlighting its key role in information acquisition. We also isolate the impact of GPT-4o-mini on Fine-EQA, with results provided in the Appendix.

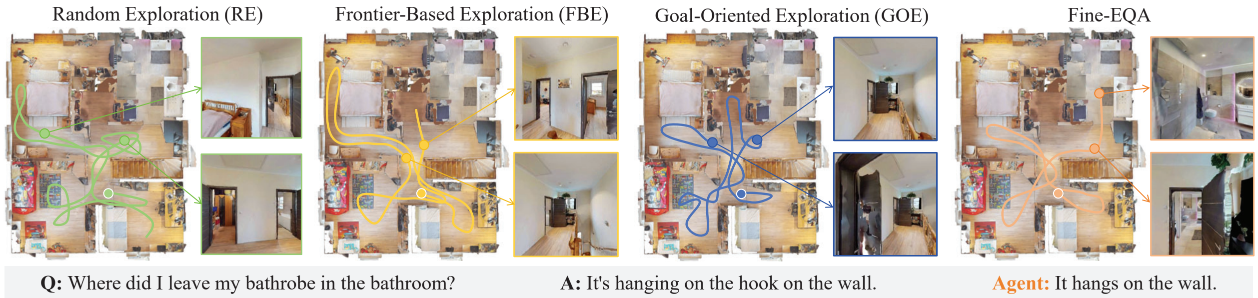

Figure 6. Exploration trajectories of agents employing different strategies within the scene. Only our Fine-EQA effectively guides the agent to observe question-relevant regions and generate the correct answer.

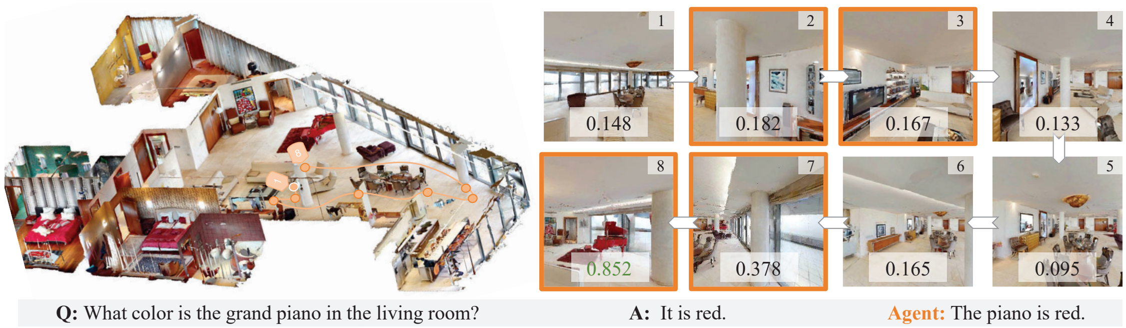

Figure 7. Confidence of visual observations at waypoints during the agent's exploration. The highest confidence is indicated in green in the final frame, demonstrating the faithfulness of our Fine-EQA by maintaining exploration-answer consistency.

5.4. Effectiveness of Exploration and Answering

We perform both qualitative and quantitative studies to evaluate the effectiveness of our exploration strategy and the faithfulness of our question reasoning module.

Exploration Effectiveness. In Fig.6, we randomly select a scene and visualize the agent's exploration trajectory under different exploration strategies. RE disregards both scene and instruction information, selecting the next location arbitrarily. This leads to significant redundancy and inefficiency. FBE and GOE demonstrate improved performance over RE. However, they still struggle to effectively explore the scene, failing to cover relevant regions or gather sufficient contextual information. Fine-EQA further enhances exploration efficiency by integrating scene context and the regional cues embedded in the instructions. While rapidly expanding the unknown areas, it strategically directs the agent toward task-relevant regions for a more thorough and efficient exploration. As shown in Fig.6, our Fine-EQA agent, starting in the hallway, identified the relevant bathroom ahead. It then navigated strategically towards the bathroom, ultimately reaching the correct answer.

Answer Faithfulness. Moreover, we utilize VLMs to assess the faithfulness of our question reasoning module. Specifically, we use Prismatic VLMs to determine whether the environmental observations collected during the agent's exploration support an accurate response to the question. From Fig.7, the observation image captured at the agent's termination point has the highest confidence. We select the four images with the highest confidence, highlighted with orange borders, as inputs for LLaVA and GPT-4o-mini. Both models agreed that the final image provides the most informative basis for answering the question.

Table 4. Quantitative comparison of model performance.

	$NPL\uparrow$	$ACE\uparrow$	$WCE\uparrow$
RE	0.23	0.55	0.16
FBE	0.28	0.58	0.19
GOE	0.27	0.58	0.18
Fine-EQA	**0.34**	**0.59**	**0.22**

We also compute the average confidence (ACE), normalized path length (NPL), and the normalized path length-weighted confidence (WCE) of the agent's observations at the termination location under various exploration strategies. The detailed formulas of these metrics are provided in the Supplementary Material. According to the quantitative results in Tab.4, Fine-EQA demonstrates the best performance, verifying the faithfulness of our Fine-EQA by concurrently improving the agent's exploration efficiency and maintaining exploration-answer consistency.

6. Conclusion

In this paper, we introduce EXPRESS-Bench, the largest dataset for evaluating both exploration and reasoning in EQA. We also propose Fine-EQA, a hybrid exploration model that improves exploration efficiency by combining frontier-based and goal-oriented navigation. Additionally, we introduce the Exploration-Answer Consistency (EAC) metric to better assess exploration and reasoning alignment. Extensive experiments demonstrate the effectiveness of EXPRESS-Bench in advancing exploration-aware EQA.

Acknowledgement

This work is supported in part by the National Key R&D Program of China under Grant No.2021ZD0111601, in part by the National Natural Science Foundation of China under Grant No. 62436009, No. 62002395 and No. 62322608, in part by the open research fund of Pengcheng Laboratory under Grant 2025KF1B0050, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant NO. 2025A1515011874 and No.2023A1515011530. We also thank the National Supercomputer Center in Guangzhou for computational support.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 6

[2] Ali Mohamed Nabil Allam and Mohamed Hassan Haggag. The question answering systems: A survey. *International Journal of Research and Reviews in Information Sciences (IJRRIS)*, 2(3), 2012. 2

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2

[4] Cătălina Cangea, Eugene Belilovsky, Pietro Liò, and Aaron Courville. Videonavqa: Bridging the gap between visual and embodied question answering. *arXiv preprint arXiv:1908.04950*, 2019. 2, 3

[5] Weixing Chen, Yang Liu, Binglin Chen, Jiandong Su, Yongsen Zheng, and Liang Lin. Cross-modal causal relation alignment for video question grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2

[6] Kai Cheng, Zhengyuan Li, Xingpeng Sun, Byung-Cheol Min, Amrit Singh Bedi, and Aniket Bera. Efficienteqa: An efficient approach for open vocabulary embodied question answering. *arXiv preprint arXiv:2410.20263*, 2024. 3

[7] 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. 2, 3

[8] Vishnu Sashank Dorbala, Prasoon Goyal, Robinson Piramuthu, Michael Johnston, Reza Ghanadhan, and Dinesh Manocha. S-eqa: Tackling situational queries in embodied question answering. *arXiv preprint arXiv:2405.04732*, 2024. 2, 3

[9] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 6

[10] Luigi Freda and Giuseppe Oriolo. Frontier-based probabilistic strategies for sensor-based exploration. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pages 3881–3887. IEEE, 2005. 2

[11] 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 conference on computer vision and pattern recognition*, pages 4089–4098, 2018. 2, 3

[12] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 6

[13] Megha Gupta, Thomas Rühr, Michael Beetz, and Gaurav S Sukhatme. Interactive environment exploration in clutter. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5265–5272. IEEE, 2013. 2

[14] Dirk Holz, Nicola Basilico, Francesco Amigoni, and Sven Behnke. Evaluating the efficiency of frontier-based exploration strategies. In *ISR 2010 (41st International Symposium on Robotics) and ROBOTIK 2010 (6th German Conference on Robotics)*, pages 1–8. VDE, 2010. 2

[15] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6

[16] Md Farhan Ishmam, Md Sakib Hossain Shovon, Muhammad Firoz Mridha, and Nilanjan Dey. From image to language: A critical analysis of visual question answering (vqa) approaches, challenges, and opportunities. *Information Fusion*, 106:102270, 2024. 2

[17] Chaoya Jiang, Hongrui Jia, Mengfan Dong, Wei Ye, Haiyang Xu, Ming Yan, Ji Zhang, and Shikun Zhang. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 525–534, 2024. 2

[18] Ivan Kapelyukh, Yifei Ren, Ignacio Alzugaray, and Edward Johns. Dream2real: Zero-shot 3d object rearrangement with vision-language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4796–4803. IEEE, 2024. 3

[19] Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *Forty-first International Conference on Machine Learning*, 2024. 7

[20] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019. 2

[21] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024. 2

[22] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 6

[23] Yang Liu, Guanbin Li, and Liang Lin. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11624–11641, 2023. 2

[24] Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *arXiv preprint arXiv:2407.06886*, 2024. 2

[25] 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. 3

[26] Yuxing Long, Xiaoqi Li, Wenzhe Cai, and Hao Dong. Discuss before moving: Visual language navigation via multi-expert discussions. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 17380–17387. IEEE, 2024. 3

[27] Haonan Luo, Guosheng Lin, Yazhou Yao, Fayao Liu, Zichuan Liu, and Zhenmin Tang. Depth and video segmentation based visual attention for embodied question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6807–6819, 2022. 2

[28] Jingzhou Luo, Yang Liu, Weixing Chen, Zhen Li, Yaowei Wang, Guanbin Li, and Liang Lin. Dspnet: Dual-vision scene perception for robust 3d question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2

[29] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16488–16498, 2024. 2, 3

[30] Farzad Niroui, Kaicheng Zhang, Zendai Kashino, and Goldie Nejat. Deep reinforcement learning robot for search and rescue applications: Exploration in unknown cluttered environments. *IEEE Robotics and Automation Letters*, 4(2):610–617, 2019. 2

[31] 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. 2, 3, 5

[32] Koya Sakamoto, Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Map-based modular approach for zero-shot embodied question answering. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10013–10019. IEEE, 2024. 3

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

[34] Xinshuai Song, Weixing Chen, Yang Liu, Weikai Chen, Guanbin Li, and Liang Lin. Towards long-horizon vision-language navigation: Platform, benchmark and method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2

[35] Sinan Tan, Mengmeng Ge, Di Guo, Huaping Liu, and Fuchun Sun. Knowledge-based embodied question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11948–11960, 2023. 2, 3

[36] Rui Wang, Jie Zhang, Ming Lyu, Cheng Yan, and Yaowei Chen. An improved frontier-based robot exploration strategy combined with deep reinforcement learning. *Robotics and Autonomous Systems*, 181:104783, 2024. 2

[37] 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. 2, 3

[38] Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots*, 47(8):1087–1102, 2023. 3

[39] Tao Wu, Chuhao Zhou, Yen Heng Wong, Lin Gu, and Jianfei Yang. Noisyeqa: Benchmarking embodied question answering against noisy queries. *arXiv preprint arXiv:2412.10726*, 2024. 2, 3

[40] Jinxuan Xu, Shiyu Jin, Yutian Lei, Yuqian Zhang, and Liangjun Zhang. Rt-grasp: Reasoning tuning robotic grasping via multi-modal large language model. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7323–7330. IEEE, 2024. 3

[41] Brian Yamauchi. A frontier-based approach for autonomous exploration. In *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97.'Towards New Computational Principles for Robotics and Automation'*, pages 146–151. IEEE, 1997. 2

[42] Brian Yamauchi. Frontier-based exploration using multiple robots. In *Proceedings of the second international conference on Autonomous agents*, pages 47–53, 1998. 2

[43] 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. 2, 3

[44] Kun Zhang, Jiali Zeng, Fandong Meng, Yuanzhuo Wang, Shiqi Sun, Long Bai, Huawei Shen, and Jie Zhou. Tree-of-reasoning question decomposition for complex question answering with large language models. In *Proceedings of the AAAI Conference on artificial intelligence*, pages 19560–19568, 2024. 2

[45] 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. 2, 3

[46] 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*, pages 13624–13634, 2024. 2, 3