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

Supplementary Material

1. Question Reasoning Module

State Judgment of Exploration. In EQA, the agent need to accumulate environmental information through dynamic interactions to achieve accurate responses. Crucially, this exploration process requires termination within reasonable constraints rather than continuing indefinitely. At each step, the agent performs a sufficiency evaluation of its acquired information to determine whether to end the exploration and proceed to the answer generating phase before reaching the predetermined maximum interaction threshold. Thus, we use the visual language model that systematically integrates real-time visual observations with textual query semantics to comprehensively analyze the relevance and information adequacy of the scene. Once the VLM determines that all essential information has been gathered and no further exploration is needed to answer the question, it signals the conclusion of the exploration phase. At this point, the exploration state is marked as “completed” and the agent transits to the QA phase.

Answer Generation. This process relies on the latest visual information obtained in the exploration phase and the understanding of the question. During the reasoning process, the VLM integrates the language information and the image features from the previous exploration to generate the answer that conforms to the question semantics and the actual situation of the scene.

2. Extra Experiments

2.1. Experimental Setup

- Maximum exploration limit. The agent’s total explorations within a scene are proportional to the scene size, while consecutive explorations within a task-relevant region are limited to three. Only in designated regions does the agent observe from four directions—front, back, left, and right—ensuring a comprehensive view.
- Maximum step length. The agent’s next exploration point must be within 3 meters of its current location, ensuring controlled movement within the scene.

2.2. Experimental Metrics

2.2.1. Formulas for Metric Calculation

C^* is the performance metric that ignores answer grounding(i.e., setting δ_i = 1):

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

Table 1. Performance comparison on the A-EQA subset of OpenEQA. Results marked with * are from the OpenEQA benchmark, where GPT-4V is evaluated on a random subset of 184 questions. In contrast, our Fine-EQA is evaluated on the full set of questions.

	$C'\uparrow$	$E'\uparrow$
OpenEQA w/ GPT-4V	$41.8_{\pm 3.2}$*	$7.5_{\pm 0.6}$*
Fine-EQA	**43.27**	**29.16**

The calculation formulas for metrics of reliability study are as follows:

$$ACE = \frac{1}{N}\sum_{i=1}^{N} ce_i \tag{2}$$

$$NPL = \frac{1}{N}\sum_{i=1}^{N}\frac{l_i}{max(p_i, l_i)} \tag{3}$$

$$WCE = \frac{1}{N}\sum_{i=1}^{N} ce_i \times \frac{l_i}{max(p_i, l_i)} \tag{4}$$

where ce_i represents the confidence of the VLM for the image, 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.

2.2.2. Experiments on Other Datasets

To further evaluate the performance of Fine-EQA, we conduct experiments on two additional datasets. For the OpenEQA[1], we focus specifically on the A-EQA subset, which assesses the agent’s ability to explore the environment and answer questions. We use the corresponding evaluation metrics C' and E' for performance measurement:

$$C' = \frac{1}{N}\sum_{i=1}^{N}\frac{\sigma_i^{'} - 1}{4} \times 100\% \tag{5}$$

$$E' = \frac{1}{N}\sum_{i=1}^{N}\frac{\sigma_i^{'} - 1}{4} \times \frac{l_i}{max(p_i, l_i)} \times 100\% \tag{6}$$

where $\sigma_i^{'}$ is determined by the LLMs based on the prompt from OpenEQA.

The results are presented in Tab.1. Fine-EQA outperforms the best-performing GPT-4V model from [1], particularly in terms of exploration efficiency. This is because the active exploration strategy in [1] relies entirely on the frontier-based method and fails to terminate exploration promptly after gathering the information necessary for the task.

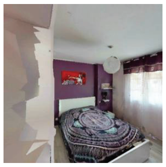
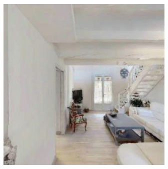
	Question: Are the cabinets in the kitchen white? **Answer:** Yes, the cabinets are white. **Response:** Yes, the cabinets in the kitchen are white. **OpenEQA:** 5 **EAC:** 5*1=5		**Question:** Is the artwork on the wall colorful in the living room? **Answer:** Yes, the artwork is very colorful. **Response:** Yes, the artwork on the wall is colorful. **OpenEQA:** 5 **EAC:** 5*0=0
	Question: Is there a lamp on the table near the living room wall? **Answer:** No, there is no lamp. **Response:** Yes, there is a lamp on the table near the living room wall. **OpenEQA:** 1 **EAC:** 1*0.5=0.5		**Question:** Did I close the curtains in the living room before left? **Answer:** No, the curtains are not closed. **Response:** Yes, the curtains in the living room are closed. **OpenEQA:** 1 **EAC:** 1*0=0

Figure 1. Comparison of the metrics proposed by OpenEQA and ours. The EAC metric combines σ and δ to jointly assess both the semantic validity and visual grounding of the response. By considering the grounding of the response, our metric offers a more reliable assessment of the model's performance.

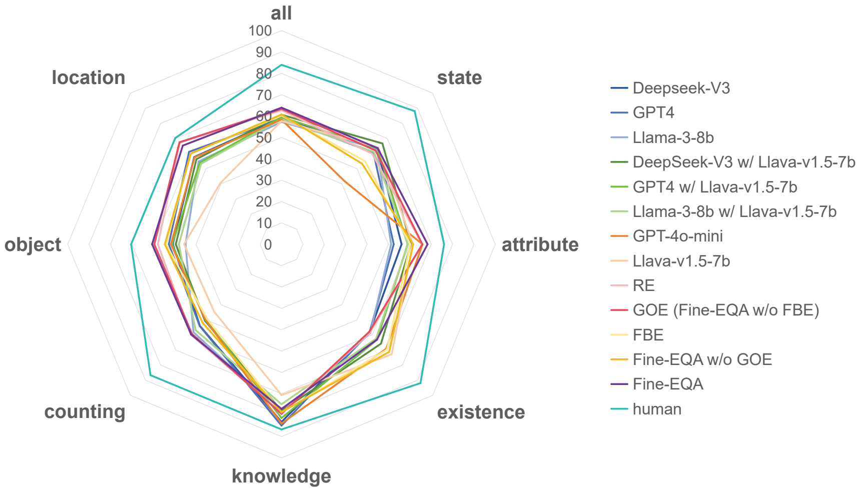


Figure 2. Performance of models in the C^* metric across different question types.

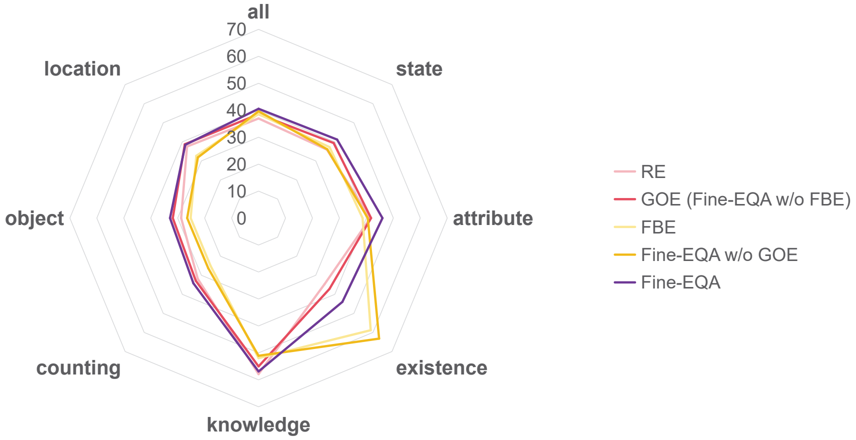
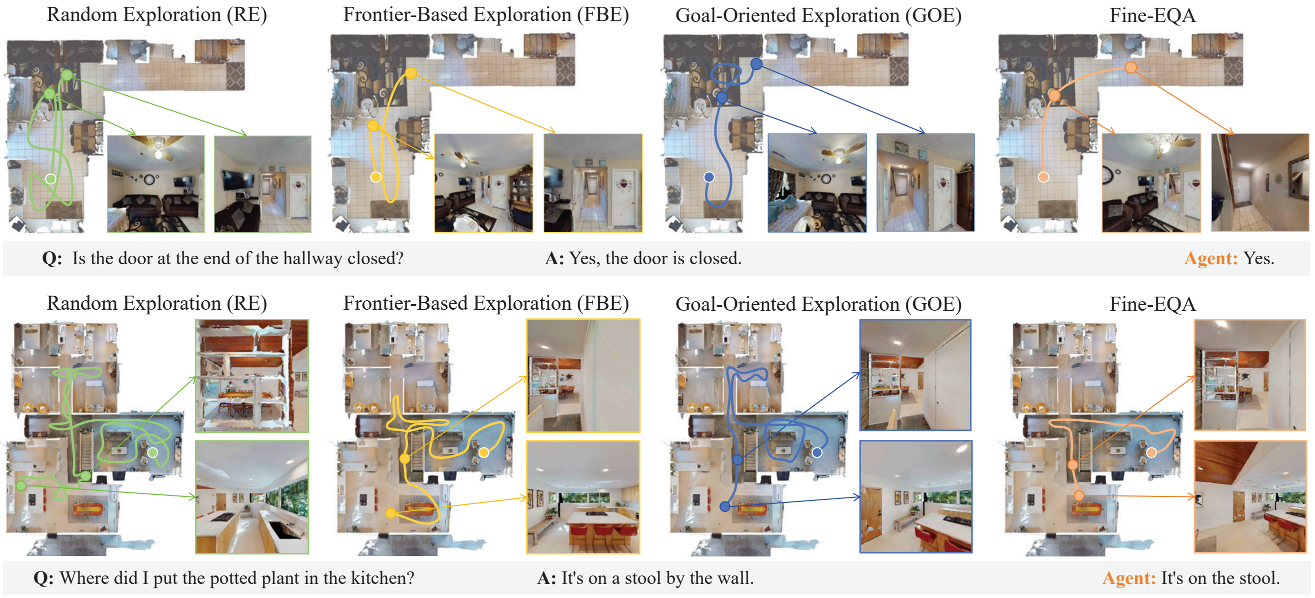


Figure 3. Performance of exploration-aware agents in the C metric across different question types.

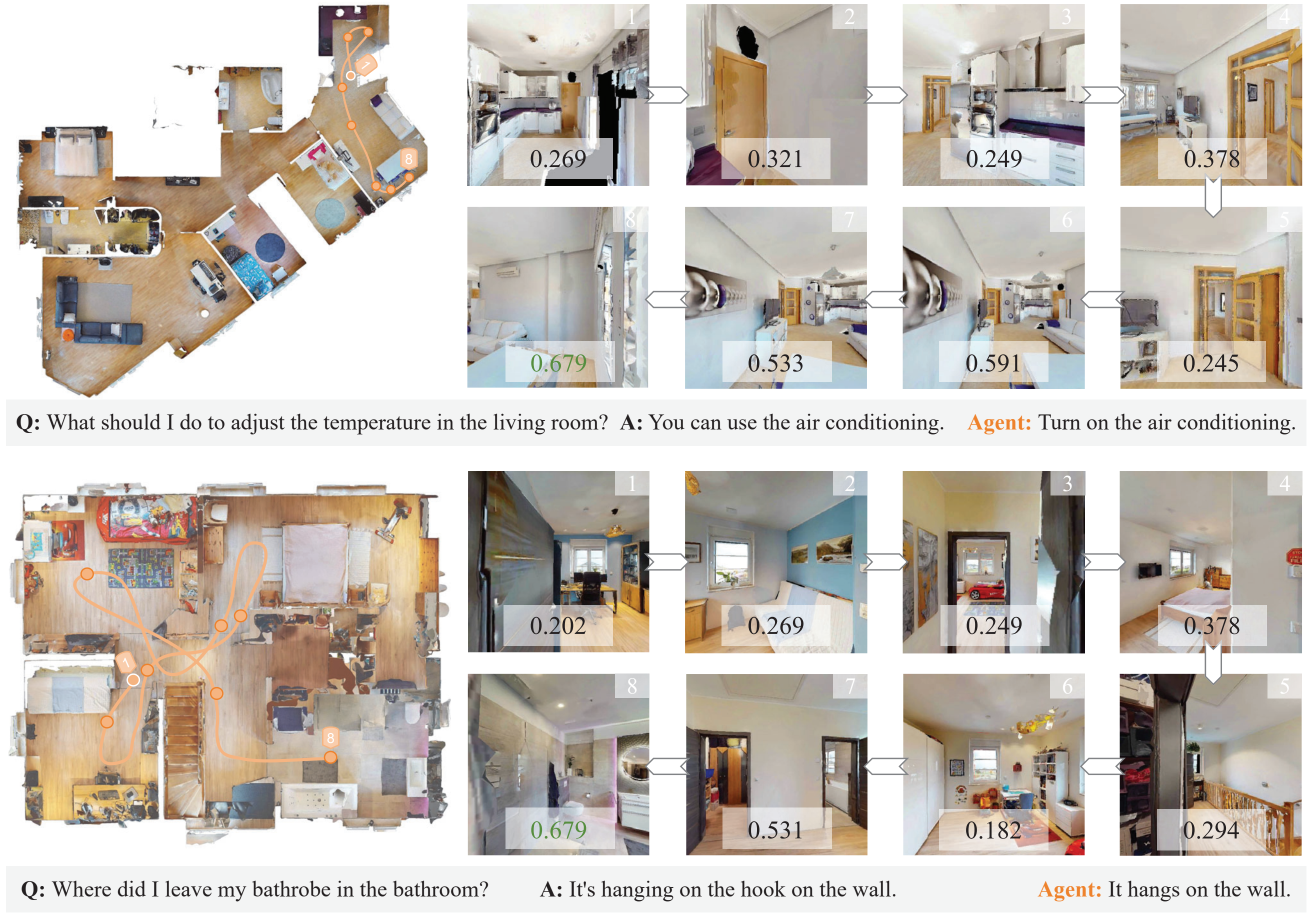

Random Exploration (RE)
Frontier-Based Exploration (FBE)
Goal-Oriented Exploration (GOE)
Fine-EQA

Q: Is the door at the end of the hallway closed? **A:** Yes, the door is closed. **Agent:** Yes.

Random Exploration (RE)
Frontier-Based Exploration (FBE)
Goal-Oriented Exploration (GOE)
Fine-EQA

Q: Where did I put the potted plant in the kitchen? **A:** It's on a stool by the wall. **Agent:** It's on the stool.

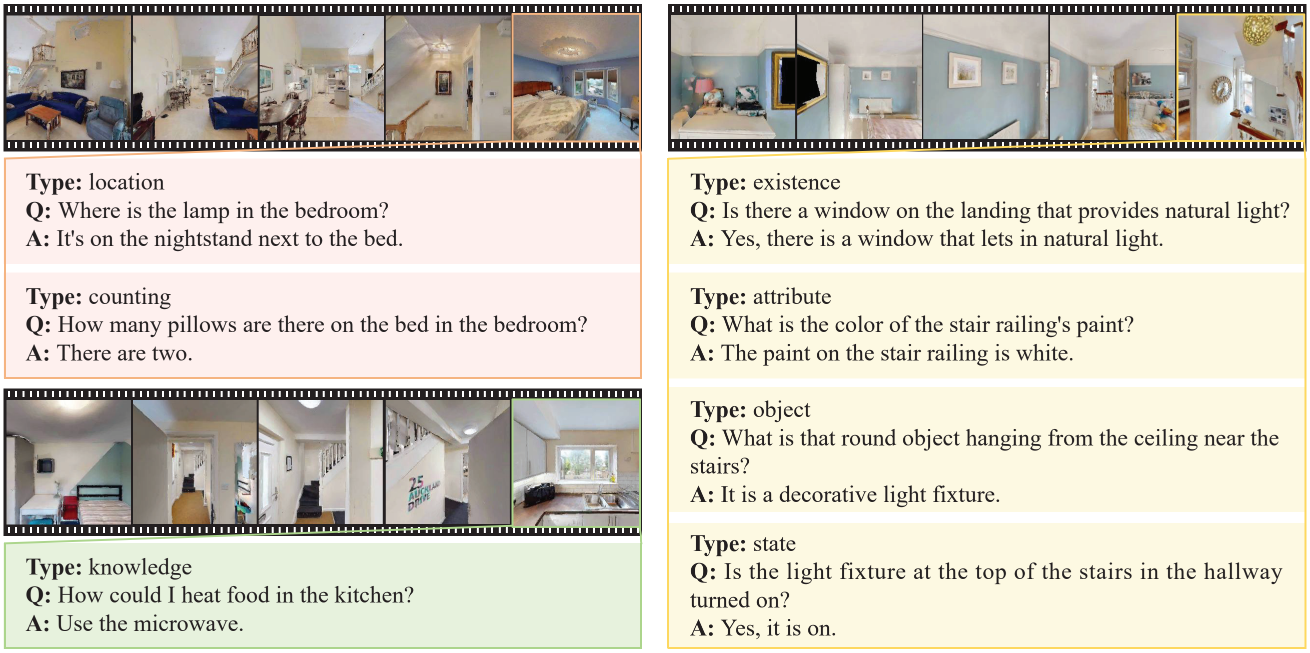
Figure 4. Visualization of the agent's exploration trajectory under different strategies. While the agent correctly answers the question using all approaches, our Fine-EQA achieves the highest exploration efficiency.



Q: What should I do to adjust the temperature in the living room? **A:** You can use the air conditioning. **Agent:** Turn on the air conditioning.

Q: Where did I leave my bathrobe in the bathroom? **A:** It's hanging on the hook on the wall. **Agent:** It hangs on the wall.

Figure 5. Confidence of visual observations at waypoints during the agent's exploration. The highest confidence is indicated in green in the final frame, demonstrating the reliability of our question reasoning module.



Type: location
Q: Where is the lamp in the bedroom?
A: It's on the nightstand next to the bed.

Type: counting
Q: How many pillows are there on the bed in the bedroom?
A: There are two.

Type: knowledge
Q: How could I heat food in the kitchen?
A: Use the microwave.

Type: existence
Q: Is there a window on the landing that provides natural light?
A: Yes, there is a window that lets in natural light.

Type: attribute
Q: What is the color of the stair railing's paint?
A: The paint on the stair railing is white.

Type: object
Q: What is that round object hanging from the ceiling near the stairs?
A: It is a decorative light fixture.

Type: state
Q: Is the light fixture at the top of the stairs in the hallway turned on?
A: Yes, it is on.

Figure 6. Examples of different question types from EXPRESS-Bench.

Table 2. Performance comparison on HM-EQA.

	Accuracy(%)↑	Path Length(m)↓
Explore-EQA	50.4	93.687
Fine-EQA	**56.0**	**54.267**

Table 3. Ablation study of VLMs.

	C↑	C^*↑	E_{path}↑	d_T↓
Janus-Pro-7B	39.88	63.35	20.86	6.18
Qwen-Vl-Plus	40.02	63.06	17.57	6.41
GPT-4o-mini	40.55	63.95	16.22	6.43

We also evaluated the performance of Explore-EQA[2] and Fine-EQA on the multiple-choice dataset HM-EQA[2], using answer accuracy and the length of the agent's navigation path as metrics. As shown in Tab.2, Fine-EQA achieves substantial improvements over Explore-EQA in both metrics.

2.2.3. Ablation Study of VLMs

To isolate the impact of GPT-4o-mini on Fine-EQA's performance, we replace it with other VLMs and conducted evaluations on the EXPRESS-Bench. As shown in Tab.3, Fine-EQA built using different VLMs exhibit varying performance across different metrics, but consistently outperform other models. We observed that Fine-EQA models with higher C scores tend to engage in more extensive exploration within the environment, which is reflected in their lower E_{path} scores.

2.2.4. Comparison of Metrics

By incorporating answer grounding, our metric provides a more accurate evaluation of the model's performance. Fig.1 compares our metric with that of OpenEQA using several examples.

2.3. Performance of Different Problem Types

We categorize the dataset based on question types and evaluate the models' performance across these categories.

Fig.2 presents the C^* scores of all models across different question types. It is evident that human performance significantly surpasses that of all other models. Overall, the performance gap between models and humans is smallest in the knowledge category, while it is more pronounced in the state, existence, and counting categories. Among the models, Fine-EQA demonstrates strong performance, ranking either the best or second-best in most categories, except for knowledge and existence questions.

Additionally, Fig.3 illustrates the performance of agents with exploration capabilities in terms of the C metric across various question types. After accounting for the grounding of the responses, all agents experience a notable decline in performance. While Fine-EQA generally performs well in most categories, its performance on existence-type questions is relatively weaker.

2.4. Exploration and Answering Effectiveness

We present visualizations of the exploration paths from different agents across additional examples in Fig.4. Fine-EQA consistently demonstrated the highest performance.

Fig.5 also presents the confidence scores assigned by VLMs for the agent's visual observations in two trajectory examples.

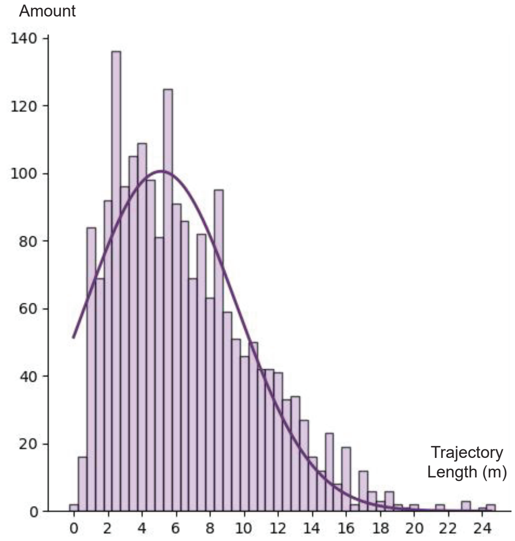


Figure 7. Distribution of trajectory lengths.

3. More examples of EXPRESS-Bench

Fig.6 presents data from seven distinct question types across three tracks. A single trajectory can generate multiple data of different types, all derived from the final frame of the trajectory videos.

We also analyze the distribution of trajectory lengths in the dataset, as shown in Fig.7.

4. Prompt Used

We present our data generation prompt ($prompt_1$ 8), the scoring evaluation prompt ($prompt_2$ 9), the prompt for determining whether the agent should terminate exploration ($prompt_3$ 10), and the prompt for answering questions ($prompt_4$ 11). The design of $prompt_4$ is inspired by [1].

References

[1] 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. 1, 5

[2] Allen Z Ren, Jaden Clark, Anushri Dixit, Masha Itkina, Anirudha Majumdar, and Dorsa Sadigh. Explore until confident: Efficient exploration for embodied question answering. *arXiv preprint arXiv:2403.15941*, 2024. 4

You are an expert at generating embodied question answering datasets.
The input is an image. You need to generate questions and corresponding answers based on the image to expand the embodied question answering task dataset. The questions are asked from the perspective of the owner of the house. The robot tasked with these questions needs to navigate around the house, exploring the environment until it captures an observation similar to the input image to gather the information required to answer the question.
Below are examples of different types of questions. Refer to these examples to generate appropriate questions and answers. Your output does not need to cover every type of question; you can generate multiple questions of the same type. The structure of the questions can differ from the examples, but they should fit the respective types. Make sure the output is formatted the same as the examples. No additional explanatory text is needed. The input image is assumed to be your first-person observation, so your output should avoid phrases like "in the image".

type: state
question: Did I turn off the light in the living room before I went out?
answer: No, the living room light is still on.

type: knowledge
question: What could I do to cool down in the living room?
answer: Turn on the air conditioning.

type: location
question: I forgot where I leave my mug in the dining room. Do you see it?
answer: It's on the corner of the dining table.

type: attribute
question: Is the floor in the bedroom wooden?
answer: Yes, it is made of wood.

type: counting
question: There will be five guests visiting in the afternoon. Are the back cushions on the sofa in the living room enough for them?
answer: No. There are only three back cushions on the living room sofa.

type: existence
question: Is there a dressing mirror in the bedroom for me to see how my clothes look?
answer: No.

type: object
question: What is the gray object I put on the bed in my bedroom?
answer: It's a hat.

Figure 8. $prompt_1$ for data generation.

You are an AI assistant who will help me to evaluate the response given the question, the correct answer and the scene observed by the robot.
The input includes the Question, the Answer, the Response given by the model and the Image of the environmen. You need to evaluate the alignment between the Response and the Image, as well as between the Response and the Answer, and assign a score for each.
First, assess whether the Response depends on the observed environment Image and assign one of three possible scores [0, 0.5, 1]. If the target object referenced in the Question or the Answer is present in the Image and is described accurately, assign a score of 1. If the object is present but inaccurately described, assign a score of 0.5. If the object does not exist in the Image, meaning the answer is entirely unrelated to the Image and fabricated, assign a score of 0.
Additionally, compare the model's Response with the Answer and Image, assigning a score scale from 1 to 5 based on its accuracy.
Here are some examples illustrating the degree to which response align with the correct answer, accompanied by an explanation of the score provided in parentheses.

Question: There will be 4 guests. Are there enough chairs around the dining table?
Answer: Yes, there are 6 tables.
Response: Yes.
Your mark: 5 (Correct answer. Giving a specific number is not necessary for this question.)

Question: What color is the sofa in the living room?
Answer: It is light beige.
Response: White.
Your mark: 4(The output is close to the answer but deviates.)

Question: Are the curtains in the living room closed?
Answer: No, the curtains are partially open.
Response: Yes, the curtains are closed.
Your mark: 3(The output is close to the answer but deviates because the curtain is not completely closed.)

Question: Can you tell me where the light switch for the basement is?
Answer: It is on the wall near the entrance door.
Response: The light switch on the wall near the door.
Your mark: 5(The output is completely correct.)

Question: What could I do if I get cold in the living room?
Answer: You can use the blanket on the couch next to the window.
Response: You can turn on the fireplace.
Your mark: 5(The response is inconsistent with the answer but consistent with common sense, and a fireplace can be observed in the image.)

Question: Are there any plants in the living room?
Answer: Yes, there is a plant near the sofa.
Response: No.
Your mark: 1(The output is the opposite of the answer.)

Question: What is the blue item on the bed in the nursery?
Answer: It's a baby blanket.
Response: It's a coat.
Your mark: 2(Object identification error.)

Your output should consist of exactly two fractions, separated by a comma. No further elaboration is necessary. Please provide the output that fulfills these criteria given the input.

Figure 9. $prompt_2$ for scoring evaluation.

You are an intelligent assistant tasked with determining whether the given image contains sufficient information to answer the provided question.
The input consists of QUESTION and IMAGE. The QUESTION is what you need to evaluate, while the IMAGE represents the currently observed environment.
Respond only with “yes” or “no” without attempting to answer the question itself.

Figure 10. $prompt_3$ for determining whether the agent should terminate exploration.

You are an intelligent question answering agent. I will ask you questions about an indoor space and you must provide an answer.
You will be shown a image that have been collected. Given a user query, you must output ‘text’ to answer to the question asked by the user. No explanatory text is required.
If the query and the image do not provide enough information to properly answer, provide an appropriate guess. Avoid stating uncertainty about answering a question. Below are several examples.

Q: What machine is on top of the stove?
A: The microwave.
Explanation: Stoves are typically found in kitchens and near microwaves.

Q: What piece of furniture is in the middle of the bedroom?
A: It is a bed.
Explanation: Bedrooms almost always contain a bed.

Q: Is the door open or closed?
A: The door is open.
Explanation: The door can be in either state, so we just randomly pick one.

Figure 11. $prompt_4$ for question answering.