Supplementary Material for
FAST-EQA: Efficient Embodied Question Answering with Global and Local Region Relevancy

1. Evaluation Metrics

The evaluation metrics used in the results Table 2 are calculated as follows, where N_{total} is the total number of questions:

Success rate (SR) for multiple-choice questions:

$$SR = \frac{Correct}{N_{total}} \times 100\%$$

where $Correct$ is the number of questions answered correctly. For multiple-choice questions, we ask the model to output the letter corresponding to the choice.

Normalized steps (Steps) from MemoryEQA [5]:

$$Steps = \frac{1}{N_{total}} \sum_{i=1}^{N_{total}} \frac{q_i}{\sqrt{S_i * \gamma_s}},$$

where q_i is the number of steps taken for question i, S_i is the total room size, γ_s is the ratio between max steps and room size.

LLM Score from Fine-EQA [2]:

$$\mathit{LLM\ Score} = \frac{1}{N_{total}} \sum_{i=1}^{N_{total}} \frac{\sigma_i}{5} \times 100\%.$$

where σ_i is the raw score given by the LLM from 1 to 5. We use the same prompts as [4] for the LLM scoring procedure.

LLM-Match from OpenEQA [4]:

$$\mathit{LLM\text{-}Match} = \frac{1}{N_{total}} \sum_{i=1}^{N_{total}} \frac{\sigma_i - 1}{4} \times 100\%.$$

where σ_i is the raw score given by the LLM from 1 to 5. We use the same prompts as OpenEQA [4] for the LLM scoring procedure.

Path Efficiency (E_{path}) (from OpenEQA [4] and Fine-EQA [2]):

$$E = \frac{1}{N_{total}} \sum_{i=1}^{N_{total}} \delta_i \times \frac{l_i}{\max(p_i, l_i)} \times 100\%,$$

where δ_i is the normalized score given by the LLM which is equal to $\frac{\sigma_i - 1}{4}$ for OpenEQA and $\frac{\sigma_i}{5}$ for Fine-EQA. l_i is the geodesic distance taken in the ground-truth trajectory and p_i is the distance traveled by the agent.

	SR	Steps (↓)
FAST-EQA$_{k=1}$	69.0	0.66
FAST-EQA$_{k=2}$	74.0	0.63
FAST-EQA$_{k=3}$	76.0	0.65
FAST-EQA$_{k=4}$	76.0	0.64
FAST-EQA$_{k=5}$	80.0	0.63
FAST-EQA$_{k=6}$	79.0	0.64

Table 1. Tuning the size k of our bounded visual memory on a fixed subset of HM-EQA

2. Additional Ablation Results

We conduct an ablation study on the size of our bounded visual memory, formed by the top-k relevant image observations per target. This visual memory is used for both determining the stopping condition and final question answering. We vary the value of parameter k and measure the question-answering success rate and normalized steps taken in Table 1. We observe that the success rate decreases when only one relevant image is retrieved, likely due to the lack of multi-view angles required for answering some questions. As we increase k, we see that performance increases, plateaus at $k = 3$, and then achieves a peak at $k = 5$. We evaluate FAST-EQA on all the datasets for both $k = 3$ as well as 5. The results are reported in Table 2, where we see that increasing the value of k from 3 to 5 does not impact overall performance in any significant way. For HM-EQA and MT-HM3D, the success rate (SR) increases for $k = 5$, whereas for EXPRESS-Bench, it stays the same. However, A-EQA presents contrary results. When we increase visual memory size, the LLM-Match score decreases. Since $k = 3$ gives us the best overall results across all the benchmarks and reduces memory usage, we report results with $k = 3$ in the main paper.

Table 2. Comparison across EQA benchmarks against SOTA baseline methods. * indicates that the result is from a reproduced experiment reported by others. † indicates results are on full A-EQA split.

	HM-EQA		MT-HM3D		EXPRESS-Bench		A-EQA (184)	
	SR	Steps (↓)	SR	Steps (↓)	LLM Score	E_{path} (↑)	LLM-Match	E_{path} (↑)
GPT-4V (OpenEQA)	–	–	–	–	–	–	41.8	7.5
Explore-EQA	58.4	0.52	36.2*	0.64	–	–	46.9*	23.4
Graph-EQA	63.5	**0.20**	45.63*	0.45	–	–	30.1*†	–
Memory-EQA	63.4	0.40	**55.1**	**0.41**	–	–	36.8†	–
Fine-EQA	56.0	0.54	–	–	63.95	25.58	43.3†	29.2
3D-Mem	–	–	–	–	–	–	**52.6**	**42.0**
FAST-EQA ($k = 3$)	**69.2** ±0.7	0.65 ±0.01	50.5 ±0.3	0.52 ±0.01	**68.7** ±0.5	**29.25** ±0.55	49.0 ±1.7	27.70 ±1.70
FAST-EQA ($k = 5$)	**71.0** ±0.7	0.65 ±0.01	52.36 ±0.3	0.52 ±0.01	**68.8** ±0.5	**27.41** ±0.55	46.73 ±1.7	31.93 ±1.70

λ	SR
0.0	69.0
0.5	76.0
0.6	72.0
0.7	78.0
1.0	72.0

Table 3. Tuning the scoring parameter λ for relevant memory retrieval on a fixed subset of HM-EQA

3. Parameter Tuning

When retrieving relevant memory, we use a tunable parameter λ to weigh Prismatic and CLIP scores. This combination scoring allows the agent to retrieve observations that align with both the focused target goal and the question-answering goal. We vary the value of parameter λ and measure the question-answering success rate and normalized steps taken in 3. Based on this, we choose $\lambda = 0.7$ in our final experiments.

4. Prompts

FAST-EQA employs a variety of prompts when interacting with VLMs or LLMs across different stages of the pipeline. As illustrated in Figure 1, the prompt shown is issued at the beginning of each episode to identify the relevant regions and visual targets corresponding to the given question. We provide illustrative few-shot examples to clarify what is meant by relevant regions and visual targets for different types of questions. The model's response is returned in the form of a JSON string, which is subsequently parsed to generate a structured list of relevant regions and visual targets.

To determine the current region R_t, FAST-EQA queries Prismatic-VLM [3] with the current observation o_t using the prompt shown in Figure 2 (a). For the stopping condition, FAST-EQA queries GPT-4o [1] with the prompt in Figure 2 (b) which includes the question Q and the most relevant images retrieved from visual memory, asking whether sufficient information is available to answer the question. An exception is made for questions involving counting or object existence, where GPT-4o is instructed to continue exploring, as such tasks often require a more exhaustive examination of the scene.

For final question answering, FAST-EQA uses the prompt in Figure 2 (c) for multi-choice QA and (d) for open vocabulary QA. The prompt includes instruction to think step-by-step to encourage chain-of-thought reasoning.

References

[1] 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. 2

[2] Kaixuan Jiang, Yang Liu, Weixing Chen, Jingzhou Luo, Ziliang Chen, Ling Pan, Guanbin Li, and Liang Lin. Beyond the destination: A novel benchmark for exploration-aware embodied question answering. In *arXiv preprint arXiv:2503.11117*, 2025. 1

[3] 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. 2

[4] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Sil-

Initial GPT-4o prompt for extracting Relevant Regions and Targets:

You are an agent tasked with exploring an environment to answer a question.

First, given the question, output the detailed visual object goal(s) you need to observe in order to answer the question as a list of strings. There can be multiple. If no target object is mentioned, fill in 'object'.

Please also give the relevant rooms you need to explore to see the goal object. Choose ONLY from this list: [hallway, dining room, living room, kitchen, bedroom, bathroom, entryway, storage room, gym, home office, laundry room, garage, porch].

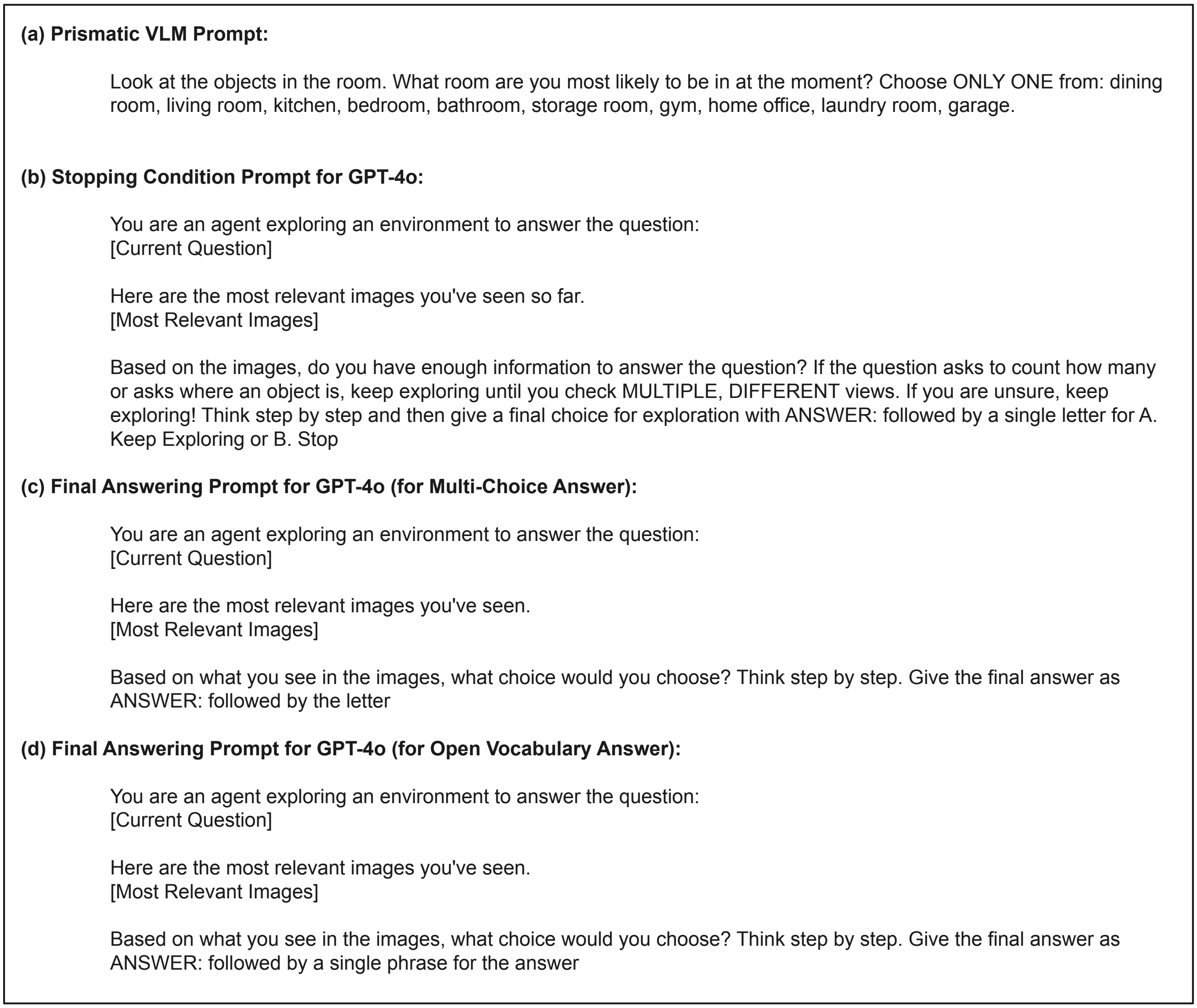
If no room is mentioned in the question, make a guess as to which rooms are relevant. Provide both lists as one VALID JSON dict and output nothing else!

```
    For example:
    Question: What color is the kettle on the counter in the kitchen?
    Output:
    {"visual_goals":["kettle on the counter"], "rooms":["kitchen"]}

    Question: Where did I leave my blue water bottle? I can't find it.
    Output:
    {"visual_goals":["blue water bottle"], "rooms":["kitchen", "living room", "dining room", "bedroom"]}

    Question: What is on the black nightstand?
    Output:
    {"visual_goals":["object on black nightstand"] "rooms":"bedroom"]

    Question: What is the white object next to the potted plant on the table?
    Output:
    {"visual_goals":["white object next to potted plant on table"], "rooms":["living room", "dining room", "bedroom", "kitchen"]}

    Question: Is the living room coffee table the same color as the dining table?
    Output:
    {"visual_goals":["living room coffee table", "dining table"], "rooms":["living room", "dining room"]}

    Question: [Current Question here]
    Output:
```

Figure 1. Prompt for extracting relevant regions and targets

wal, Paul McVay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent-Pierre Berges, Shiqi Zhang, Pulkit Agrawal, Dhruv Batra, Yonatan Bisk, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Alexander Sax, and Aravind Rajeswaran. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[5] Mingliang Zhai, Zhi Gao, Yuwei Wu, and Yunde Jia. Memory-centric embodied question answer. In *arXiv preprint arXiv:2505.13948*, 2025. 1



(a) Prismatic VLM Prompt:

Look at the objects in the room. What room are you most likely to be in at the moment? Choose ONLY ONE from: dining room, living room, kitchen, bedroom, bathroom, storage room, gym, home office, laundry room, garage.

(b) Stopping Condition Prompt for GPT-4o:

You are an agent exploring an environment to answer the question:
[Current Question]

Here are the most relevant images you've seen so far.
[Most Relevant Images]

Based on the images, do you have enough information to answer the question? If the question asks to count how many or asks where an object is, keep exploring until you check MULTIPLE, DIFFERENT views. If you are unsure, keep exploring! Think step by step and then give a final choice for exploration with ANSWER: followed by a single letter for A. Keep Exploring or B. Stop

(c) Final Answering Prompt for GPT-4o (for Multi-Choice Answer):

You are an agent exploring an environment to answer the question:
[Current Question]

Here are the most relevant images you've seen.
[Most Relevant Images]

Based on what you see in the images, what choice would you choose? Think step by step. Give the final answer as ANSWER: followed by the letter

(d) Final Answering Prompt for GPT-4o (for Open Vocabulary Answer):

You are an agent exploring an environment to answer the question:
[Current Question]

Here are the most relevant images you've seen.
[Most Relevant Images]

Based on what you see in the images, what choice would you choose? Think step by step. Give the final answer as ANSWER: followed by a single phrase for the answer

Figure 2. Prompts for (a) determining current region R_t, (b) stopping condition, (c) final answering for MCQA, and (d) final answering for Open Vocabulary Questions