R-EQA: Retrieval-Augmented Generation for Embodied Question Answering

Hyobin Ong[1,2], Minsu Jang[1,2†]
[1]University of Science and Technology (UST), South Korea
[2]Electronics and Telecommunications Research Institute (ETRI), South Korea

Abstract

Embodied Question Answering (EQA) is a task where an agent explores its environment, gathers visual information and responds to natural language questions based on that information. The accuracy of the answer depends on which visual information is sampled for a given question. This study introduces R-EQA, a framework that uses Retrieval-Augmented Generation to evaluate the effectiveness of sampling semantically relevant visual information in the EQA setting. Experiments using the OpenEQA benchmark show that R-EQA achieves 10% higher performance compared to uniform sampling. This indicates that selective sampling of question-relevant information plays a critical role in improving answer quality in EQA.

1. Introduction

Embodied Question Answering (EQA) is a task in which an agent responds to natural language questions based on its perception and interaction within physical or simulated environments [6, 13]. An embodied agent collects history of observations (i.e. episodic memory) by observing and interacting with its environment. When a user question is given, the agent must selectively retrieve and utilize the most relevant information from episodic memory to generate an accurate response [7]. The need for question-relevant information selection has also been extensively studied in the field of natural language processing, particularly through Retrieval-Augmented Generation (RAG) approaches [10]. Recently, this paradigm has been extended beyond textual inputs to multimodal RAG frameworks [2, 4, 9, 11].

Prior studies on integrating videos into the RAG framework have demonstrated improved performance on video QA tasks when using RAG over uniform sampling [9]. This indicates that the effectiveness of RAG extends beyond text domains to video domains as well. In this work, we further evaluate the impact of applying RAG to the EQA task using the OpenEQA benchmark.

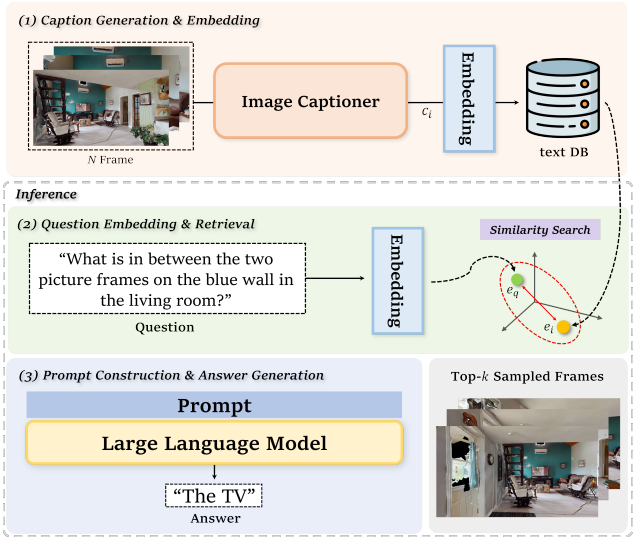

Figure 1. **An overview of R-EQA Framework.** Given a question, the system retrieves the top-k relevant frame image captions to generate a response. These frame captions are pre-processed and used during inference. The top-k sampled frames correspond to the retrieved frames.

2. Method

This section describes how the RAG framework is structured for episodic memory EQA and introduces the EQA agents evaluated in this study.

2.1. Overview

Figure 1 describes our method, which consists of three main components. Initially, given a sequence of N frames, we use a pretrained VLMs to generate textual captions $c = \{c_1, c_2, \ldots, c_N\}$, which are then embedded into a vector space $e_i = \text{Embed}(c_i)$. Second, given a natural language question q, we compute the similarity between its embedding e_q and each frame captions embedding e_i and retrieve the top-k most relevant frame captions. Finally, the retrieved captions $\{c_{i_j}\}_{j=1}^{K}$ are concatenated with the natural language question q to form the final prompt $P = \{c_{i_1}, \ldots, c_{i_K}, q\}$, which is then fed into a LLM to generate the answer $a = \text{LLM}(P)$. The model's response a is

subsequently evaluated using LLM-Match, as employed in OpenEQA:

$$C = \frac{1}{N} \sum_{i}^{N} \frac{\sigma_i - 1}{4} \times 100\%, \quad (1)$$

N deontes the number of questions and σ_i represents the score ranging from 1 to 5 provided by the LLM.

2.2. EQA Agents

We conducted our study using the following agents: Blind LLM, Uniform Sampling with frame captions, and R-EQA with frame captions. The predefined prompt w is constructed with in-context examples to align with the specific characteristics of each agent.

Blind LLM. Similar to OpenEQA, this agent serves as a baseline that uses a text-only LLM to assess how well EQA can be performed using commonsense knowledge alone. The agent operates as $a = LLM(w, q)$.

Uniform Sampling w/ Frame Captions. This agent performs uniform sampling over e_i to generate a response. The agent is defined as $a = LLM(w, c_{i_k}, q)$, where c_{i_k} denotes the k frame captions sampled via uniform sampling.

R-EQA w/ Frame Captions. This agent samples the top-k most relevant e_i to the question and generates a response. The agent is defined as $a = LLM(w, c_{i_k}, q)$, where c_{i_k} denotes the k frame captions sampled using retrieval.

3. Experiments and Results

Experiments Following the Episodic Memory EQA (EM-EQA) setup in OpenEQA [13], we evaluated each EQA agent using episode histories H collected from two sources: ScanNet [5] and HM3D [14]. We used Ferret 13B [16], LLaVA-v1.5 13B [12], and Qwen2.5-VL 7B [3] to generate image captions for frames in the episodic memory, and LLaMA 3.1 70B [8] to generate responses. Each image caption and question was embedded using Sentence-BERT [15], and similarity was computed using cosine similarity. We used $k = 10$ for uniform sampling and $k = 3$ for R-EQA. For evaluation, we used GPT-4 [1] for LLM-Match.

Results Table 1 presents the evaluation results of the agents described in Section 2.2 on the OpenEQA benchmark. Notably, while Uniform Sampling selects 10 frames evenly from episodic memory, the RAG-based method retrieves only 3 frames based on semantic similarity to the question, yet achieves better performance overall. This suggests that selecting fewer but more relevant frames can lead to improved answer quality. Across all settings, RAG-based sampling consistently outperforms uniform sampling, with the LLaMA-3.1 w/ Qwen2.5-VL achieving the best results. Even under the same image captioning conditions, R-EQA with LLaVA-v1.5 surpassed GPT-4 with uniform sampling by 0.3%, and R-EQA with Qwen2.5-VL outperformed its uniform counterpart by 2.4%. This indicates that the frame sampling strategy can have a greater impact on performance than model size. We found that the Blind LLM surprisingly showed competitive performance in Figure 2. This is partly because OpenEQA includes questions that can be answered using commonsense knowledge or simply by chance, e.g. *Is there a door that is open?* with the answer *yes*. Since the purpose of EQA benchmarks is to assess an agent's ability to acquire and process knowledge grounded in the physical environment, we argue that it is necessary to revisit these benchmarks to ensure they include only questions that genuinely require the acquisition of situated world knowledge.

Table 1. **LLM-Match scores on OpenEQA** Evaluation results of the EQA agent in EM-EQA using LLM-Match, broken down by data source (ScanNet, HM3D, ALL). **Bold** indicates the best results among the data sources in each EQA Agents. Method with * is reported from OpenEQA [13].

Methods	EM-EQA (LLM-Match scores) ScanNet	HM3D	ALL
Blind LLM			
LLaMA-3.1	31.1	38.0	34.0
Uni. w/ Frame Captions			
GPT-4 w/ LLaVA-v1.5*	45.4	40.0	43.6
LLaMA-3.1 w/ Ferret	42.1	35.2	38.6
LLaMA-3.1 w/ LLaVA-v1.5	37.9	35.5	36.6
LLaMA-3.1 w/ Qwen2.5-VL	40.3	31.7	36.0
R-EQA w/ Frame Captions			
LLaMA-3.1 w/ Ferret	42.7	38.3	40.4
LLaMA-3.1 w/ LLaVA-v1.5	46.5	41.2	43.9
LLaMA-3.1 w/ Qwen2.5-VL	**49.1**	**42.8**	**46.0**

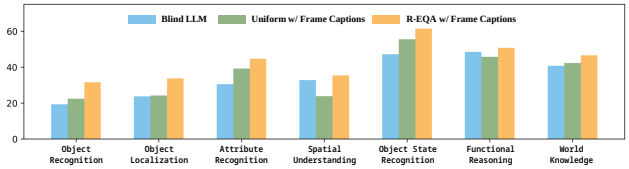

Figure 2. **Category-level performance on EM-EQA.** The results show performance across the 7 openEQA question categories. The scores for Uniform Sampling and R-EQA represent the average performance of the agents introduced in Section 2.2.

4. Conclusion and Future work

R-EQA achieved a 10% improvement over uniform sampling on the OpenEQA benchmark, validating the effectiveness of question-relevant frame sampling. As future work, we aim to investigate the integration of low-latency retrieval mechanisms to facilitate the real-world application of EQA.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2022-II220951, Development of Uncertainty-Aware Agents Learning by Asking Questions, 50%, No. RS-2024-00336738, Development of Complex Task Planning Technologies for Autonomous Agents, 50%).

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. 2

[2] Md Adnan Arefeen, Biplob Debnath, Md Yusuf Sarwar Uddin, and Srimat Chakradhar. irag: Advancing rag for videos with an incremental approach. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4341–4348, 2024. 1

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[4] 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. 1

[5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2

[6] 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. 1

[7] Samyak Datta, Sameer Dharur, Vincent Cartillier, Ruta Desai, Mukul Khanna, Dhruv Batra, and Devi Parikh. Episodic memory question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19119–19128, 2022. 1

[8] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2

[9] Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. Videorag: Retrieval-augmented generation over video corpus. *arXiv preprint arXiv:2501.05874*, 2025. 1

[10] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020. 1

[11] Weizhe Lin and Bill Byrne. Retrieval augmented visual question answering with outside knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11238–11254, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. 1

[12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2

[13] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Sasha Sax, and Aravind Rajeswaran. Openeqa: Embodied question answering in the era of foundation models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2

[14] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021. 2

[15] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 2

[16] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 2