OpenEQA: Embodied Question Answering in the Era of Foundation Models

Arjun Majumdar[1*] Anurag Ajay[2*] Xiaohan Zhang[3*]
Pranav Putta[1] Sriram Yenamandra[1] Mikael Henaff[4] Sneha Silwal[4] Paul Mcvay[4]
Oleksandr Maksymets[4] Sergio Arnaud[4] Karmesh Yadav[4] Qiyang Li[5] Ben Newman[6]
Mohit Sharma[6] Vincent Berges[4] Shiqi Zhang[3] Pulkit Agrawal[2] Yonatan Bisk[4,6] Dhruv Batra[1,4]
Mrinal Kalakrishnan[4] Franziska Meier[4] Chris Paxton[4] Sasha Sax[4] Aravind Rajeswaran[4]

***Abstract*— We present a modern formulation of Embodied Question Answering (EQA) as the task of understanding an environment well enough to answer questions about it in natural language. An agent can achieve such an understanding by either drawing upon episodic memory, exemplified by agents on smart glasses, or by actively exploring the environment, as in the case of mobile robots. We accompany our formulation with OpenEQA – the first open-vocabulary benchmark dataset for EQA. OpenEQA contains over 1600 high-quality human generated questions drawn from over 180 real-world environments.**

In addition to the dataset, we provide an automatic LLM-powered evaluation protocol that has excellent correlation with human judgement. Using this dataset and evaluation protocol, we evaluate several state-of-the-art foundation models and find that they significantly lag behind human-level performance. Consequently, OpenEQA stands out as a straightforward, measurable, and practically relevant benchmark that poses a considerable challenge to current generation of foundation models. We hope this inspires and stimulates future research at the intersection of Embodied AI, conversational agents, and world models.

I. Introduction

AI agents are entering the physical world through devices like smartphones, smart glasses, and robots. To best assist non-expert users, Embodied AI (EAI) agents need a natural language interface and "common sense" grounded in human-centric perception and understanding. We posit that **Embodied Question Answering (EQA)** is both an essential capability for EAI agents and an intuitive way to probe an agent's understanding of the world. EQA involves answering a question posed in natural language, requiring visual understanding (see fig. 2).

EQA can be studied from an episodic memory (EM-EQA) perspective or through active exploration (A-EQA), depending on the EAI agent platform. EM-EQA is particularly relevant for devices like smart glasses which cannot move autonomously, but can process the episodic memory generated by human wearers to answer questions. This can enhance the memory of users, improve perceptual capabilities and understanding, and provide general world knowledge.

*Equal Contribution, [1]Georgia Tech, [2]MIT, [3]Binghamton University, [4]Met AI, [5]UC Berkeley, [6]CMU.
Correspondance to aravraj@meta.com
Work done at Fundamental AI Research (FAIR), Meta.

TABLE I: **OpenEQA vs existing benchmarks.** OpenEQA has multiple modalities, real scenes, active agents, and automated scoring.

	Modalities			2* Open Vocab	2* Real Scenes	2* EM (video)	2*A(ctive)	2* LLM Scoring
	RGB	Depth	Camera					
EQA-v1 [1]	✔	✔	✔	✖	✖	✖	✔	✖
MP3D-EQA [2]	✔	✔	✔	✖	✔	✖	✔	✖
MT-EQA [3]	✔	✔	✔	✖	✖	✖	✔	✖
IQA [4]	✔	✔	✔	✖	✖	✖	✔	✖
SQA3D [5]	✔	✔	✔	✖	✔	✖	✖	✖
ScanQA [6]	✔	✔	✔	✖	✔	✔	✖	✖
RoboVQA [7]	✔	✖	✖	✔	✔	✔	✖	✖
SEED-Bench [8]	✔	✖	✖	✖	✔	✔	✖	✖
MMBench [9]	✔	✖	✖	✔	✔	✖	✖	✔
OpenEQA (Ours)	✔	✔	✔	✔	✔	✔	✔	✔

The intersection of perception and language has long been a fertile ground for research in AI. While the broad problem of EQA [1], [3] and VQA [5], [6], [10] have been studied extensively, our approach and benchmark differ significantly along axis such as input modalities, scenes/scans of real-world spaces, and open-vocabulary questions and answers, as illustrated in table I. In particular, OpenEQA is the first open-vocabulary benchmark for EQA, and supports both the episodic-memory and active settings.

A. Our Contributions

a) Benchmark: We introduce a modern EQA formulation and a concrete evaluation benchmark (OpenEQA) containing **1600+ questions** across over **180+** real-world environments and photo-realistic simulations.

b) Evaluation: The open-vocabulary nature of our benchmark increases the complexity of evaluating answers generated by various models. We utilize LLMs [11], [12] to score answers and find, through a double blind study,a strong correlation between our LLM-Match metric and human preferences.

c) Baselines: We provide baseline results and implementations, including GPT-4V [13] and Socratic use of LLMs [11], [12] that leverage captioning models [14] or generated scene-graph representations [15].

II. Benchmark and Evaluation

This section presents the EM-EQA and A-EQA problem statements, how they are instantiated in OpenEQA, the dataset collection process, and the evaluation metrics.

Fig. 1: **Example questions and dataset statistics of OpenEQA.** The episode history H provides a human-like tour of a home. EQA agents must answer diverse, human-generated questions Q from 7 EQA categories, aiming match the ground answers A^*. Tours are collected from diverse environments including home and office locations (not shown above). Additional dataset examples are in appendix XIII. Dataset statistics (right) break down the question distribution by video source (top), question category (middle), and episodic memory vs active setting. Note that, by design, the HM3D questions are shared across the EM-EQA and A-EQA settings.

A. *Episodic-Memory Question Answering*

The EM-EQA task targets scenarios where EAI agents, leverage episodic memory to address queries without real-time exploration. Here, the task structure is denoted as (Q, H, A^*), with Q being an open-ended question, H representing the agent's historical observations, and A^* indicating the human-annotated correct response. The agent aims to produce an answer, A, based on its episodic memory that mirrors the accuracy of A^*. The function signature that is expected for the agent is described in algorithm 1 in appendix IV.

B. *Active Embodied Question Answering*

The A-EQA task addresses scenarios where an agent independently performs exploratory actions to respond to questions (e.g. *'Q: Do we have canned tomatoes at home? A: Yes, I found canned tomatoes in the pantry.'*). Our benchmark considers questions that require only navigation actions, in principle, this can be extended to mobile manipulators in order to incorporate actions like opening doors and drawers. The task is defined by (Q, S, A^*), where Q and A^* are the query and its correct answer and S symbolizes the simulation's initial state and responses are assessed in terms of accuracy and action efficiency. For the agent's function specifics, refer to Algorithm 1 in appendix IV.

C. *OpenEQA Dataset Collection and Validation*

To establish benchmarks for EM-EQA and A-EQA, we collect a human-generated dataset of (Q, H, A^*) using videos [16] and 3D scans of real-world environments [17], [18], [19], [20]. This dataset aims to reflect realistic inquiries users might pose to AI assistants in devices like smart glasses or robot helpers.

a) Data Sources: We collected episode histories H from two sources: ScanNet [16] and HM3D [17], [18]. For ScanNet, we utilized RGB-D data captured from human explorations, specifically selecting 90 validation scenes and 10 test scenes. In the case of HM3D, we defined a heuristic exploration policy to simulate human behavior and constructed episode histories for 87 validation scenes. For details, consult appendix II.

b) Question Generation: By having human annotators view H and generate questions about the environment, we identified seven primary EQA question categories that encapsulate typical user inquiries, ranging from object recognition to spatial and functional reasoning (illustrations of the question categories are provided in fig. 1). The final OpenEQA dataset is centered around these categories, ensuring a diverse range of questions. Additional details on the dataset collection and interface can be found in appendix II.

c) Dataset Validation: Each question-answer pair underwent rigorous validation by two annotators to ensure clarity, answerability, and correctness. The final dataset consists of 1636 validated questions following the statistics in fig. 1.

d) Dataset Splits: Validated (Q, A^*) pairs are applicable to both EM-EQA and A-EQA since S is recorded in addition to H for simulated scenes and A-EQA agents are initialized at the same start state S that was used to generate the episodic memory H for EM-EQA.

e) Additional Object Localization Answers: Recognizing the complexity of the *object localization* category questions, which may have multiple correct answers, we collected additional responses to capture a broader spectrum of plausible answers.

D. LLM-Match: Evaluating Correctness of Answers

While the open-vocabulary nature makes EQA realistic, it poses a challenge for evaluation due to multiplicity of correct answers. One approach to evaluation is human trials, but it can be prohibitively slow and expensive. We use an LLM to evaluate the correctness of the answers produced by EQA agents. Specifically, we adapt the evaluation protocol introduced in MMBench [9] to the EQA task. Given a question Q_i, human annotated answer A_i^*, and model output A_i, the LLM is prompted to provide a score $\sigma_i \in \{1, \dots 5\}$. On this scale, 1 indicates an incorrect response, 5 is a correct response. We calculate an aggregate LLM-based **correctness** metric (LLM-Match) as:

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

LLM-Match is ilustrated in fig. 4 detailed in appendix III, and validated against human judgement in section V.

E. Evaluating Efficiency for A-EQA

In A-EQA, we also evaluate the **efficiency** of the agent, which measures how quickly the agent answered the question and favors agents that perform targeted exploration. We measure efficiency by weighting the correctness metric σ_i by the normalized length of the agent's path $l_i / \max(p_i, l_i)$, where p_i is the timesteps taken by the agent and l_i is the timesteps taken in a ground truth path that is sufficient for answering the question Q_i. Formally, our **efficiency** metric is defined as:

$$E = \frac{1}{N} \sum_{i}^{N} \frac{(\sigma_i - 1)}{4} \times \frac{l_i}{\max(p_i, l_i)} \times 100\%, \tag{2}$$

III. EQA Agents

This section outlines the EQA agents we evaluate, focusing on their use of foundation models (LLMs and VLMs) without additional tuning. We categorize agents as follows: (1) blind LLMs [11], [12], (2) Socractic LLMs w/ frame captions [21], (3) Socractic LLMs w/ scene-graph representations [15], and (4) VLMs that can directly process multiple frames (e.g. GPT-4V [13]); human performance is also studied for comparison. For a visual depiction of the different agents refer to fig. 5.

Blind LLMs. Rely solely on textual prompts and questions, ignoring visual context i.e. $A = \text{LLM}([\omega, Q])$ where ω is a generic prompt; and provide a baseline for knowledge or guess-based responses. We use GPT-4 and LLaMA-2-70B.

Socratic LLMs w/ Frame Captions. Utilize an image captioning model (e.g. LLaVA [21], [14]) to generate text captions of a subset of frames $s_1, \dots, s_k$ from the episode history H. This text captions augment the context of an LLM (GPT4 [11], LLaMA-2-70B [12]) as $A = \text{LLM}([\omega, z_1, \dots z_K, Q])$ to provide the final answer.

Socratic LLMs w/ Scene-Graph Captions. Construct an object-centric scene-graph representation G of H, by detecting objects in the scene, extracting its 3D locations using pose and depth information, and creating sematic descriptions by utilizing captioning models. We study two methods: ConceptGraph [15] and Sparse Voxel Map (SVM). Once a textual scene graph G is generated, we use it for EQA as $A = \text{LLM}([\omega, G, Q])$.

Multi-Frame VLMs. Process frames and textual queries simultaneously, i.e. $A = \text{MultiFrameVLM}([\omega, Q, H])$. We uniformly extract a subset of 50 frames from H and provide it to GPT-4V in addition to prompts for generating the answer; details in appendix V.

Human Agent. We run a study with human participants to establish human-level performance metrics on our benchmark. Each human annotator is provided with a video of the episode history H and asked to answer all of the questions Q for that scene.

Agents for A-EQA. We adapted all agents to the active exploration setting by generating their observational history with a simple task-agnostic baseline implemented through frontier exploration. Efficiency in exploration remains an area for future improvement.

Force-A-Guess when Agents Abstain. Agents are forced to make a guess rather than abstainining to answer to ensure that each agent has at least an informed random chance to answer correctly a question; this avoids automatic failure for models that are overly conservative when full context to answer a question is missing. Details of this protocol are in appendix VII and an analysis of the effects of this procedure are in appendix VIII.

IV. Experimental Results on OpenEQA

We present evaluation results of agents described in section III in table II, our main observations are the following:

1) Humans achieve excellent performance on the benchmark (>85%), confirming the validity of the benchmark and evaluation metrics.

2) Multi-frame VLMs outperform other agents. Suggesting that a tight integration of perception and language may significantly benefit EQA.

3) We find that blind LLMs are surprisingly strong baselines, suggesting a large degree of regularity in the world given that some questions can be "guessed" without explicit visual context of a specific environment. We note that early works in VQA [22] found similar results.

4) All agents with access to perceptual information in the form of frame captions or scene-graphs outperform blind LLMs.

5) When comparing the performance of agents in EM-EQA and A-EQA, we generally observe lower scores in A-EQA, underscoring the challenging nature of the A-EQA benchmark and the importance of efficient exploration in interactive settings.

TABLE II: **LLM-Match and efficiency scores on OpenEQA.** EM-EQA results are broken down by data source. A-EQA results include both correctness (Eq. 1) efficiency (Eq. 2) scores.

# method	EM-EQA			A-EQA	
	ScanNet eq. (1)	HM3D eq. (1)	ALL eq. (1)	HM3D eq. (1)	HM3D eq. (2)
Blind LLMs					
1 GPT-4	32.5±1.2	35.5±1.7	33.5±1.0	35.5±1.7	-
2 LLaMA-2	27.9±1.2	29.0±1.7	28.3±1.0	29.0±1.7	-
Socratic LLMs w/ Frame Captions					
3 GPT-4 w/ LLaVA-1.5	45.4±1.3	40.0±1.8	43.6±1.1	38.1±1.8	7.0±0.4
4 LLaMA-2 w/ LLaVA-1.5	39.6±1.3	31.1±1.8	36.8±1.1	30.9±1.8	5.9±0.4
Socratic LLMs w/ Scene-Graph Captions					
5 GPT-4 w/ CG	37.8±1.3	34.0±1.7	36.5±1.0	34.4±1.8	6.5±0.4
6 LLaMA-2 w/ CG	31.0±1.2	24.2±1.6	28.7±1.0	23.9±1.6	4.3±0.3
7 GPT-4 w/ SVM	40.9±1.3	35.0±1.8	38.9±1.0	34.2±1.8	6.4±0.4
8 LLaMA-2 w/ SVM	36.0±1.3	30.9±1.8	34.3±1.0	29.9±1.7	5.5±0.4
Multi-Frame VLMs					
9 GPT-4V*	51.3±2.5	46.6±3.1	49.6±2.0	41.8±3.2	7.5±0.6
Human Agent	87.7±0.7	85.1±1.1	86.8±0.6	85.1±1.1	-

Figure 10 breaks down performance on EM-EQA by the question categories described in section II-C. Among all the categories, functional reasoning is the easiest for EQA agents, followed by object state recognition and world knowledge. EQA agents suffered the most on object localization and spatial understanding questions. To our surprise, agents that use scene-graph representations are no better than frame-captioning agents, even on spatial reasoning questions. This suggests that more work is needed to incorporate space understanding and geometry into large models.

V. Analysis and Discussions

Human Alignment and Robustness of LLM-Match. In section II-D we introduced the LLM-Match metric. We now test it along two axis: (1) How closely aligned is the LLM-Match metric with human evaluators? (2) How sensitive is the LLM-Match metric towards specific choice of prompts and the LLM?

To answer the question on **human alignment**, we designed an experiment to measure the agreement between LLM-Match metric and human evaluators. We uniformly sampled a subset of 300 questions from the dataset. To ensure coverage of the answer distributions, we sampled responses from blind LLaMA-2, GPT-4V, and human annotated answers. In a double blind study, we asked 4 human evaluators to score the 300 responses using an evaluation prompt similar to the one used by LLM-Match. We found a **Spearman's $\rho = 0.909$ between human and LLM evaluation** (bootstrap CI=(0.883,0.928), N=9999), indicating excellent agreement with human judgement. For reference, human evaluators correlated with each other in $\rho \in [0.91, 0.93]$. Essentially, LLM-Match agrees with human evaluation nearly as much as human subjects do with one another.

To answer the question of **LLM-Match robustness**, we designed an experiment to test sensitivity under small perturbations to the prompt (see appendix XI). Table VIII in appendix XI shows that changing the LLM's role in the system prompt does not significantly change results, the scores have a tight correlation with a Spearman's $\rho > 0.95$. Similarly, Table IX in appendix XI shows analogous results $\rho > 0.95$ for changing the description of a '5' from 'perfect match' to 'contains correct answer', 'similar to a reasonable person', or 'reasonable professional'. Sensitivity to seed and temperature has negligible impact as well. Finally, we vary the LLM used for scoring and find that GPT4 has excellent agreement with human judgement, but GPT-3.5 and LLaMA-2 have significantly lower correlation (ρ<0.7). Thus, **for now, we recommend using GPT4 for LLM-Match.**

Force-A-Guess. As discussed in section III, when studying Socractic LLMs augmented with perceptual information, we found that agents often abstained from answering (e.g. *'Not enough information to answer the question.'*). Our LLM-Match metric does not give preferential scoring for abstention. Thus, we defaulted to the answer from the blind LLM powering an agent when it abstained. In appendix VIII, we provide statistics on the frequency of abstencion, and study performance without defaulting to a blind LLM. In general, we find that GPT-4-based Socratic agents abstain frequently (up to 55%), and thus, rely heavily on the blind LLM-based score correction that we apply in our benchmark evaluations. By contrast, GPT-4V and LLaMA-2 based models do not abstain as often (up to 12%), and thus the differences between the two variants is minimal.

VI. Conclusion

We introduce OpenEQA, the first realistic benchmark to study open-vocabulary EQA in both episodic memory and active settings. OpenEQA includes challenging, human-generated, open-vocabulary questions that require understanding an environment and answering question in natural language. Our benchmark is primarily enabled by (1) videos and scans of real-world indoor environments and (2) LLMs that can be used for scoring open-ended answers in an efficient and reliable manner, as we demonstrated through our analyses. We use OpenEQA to benchmark various state-of-the-art foundation models and their combinations. This includes approaches that leverage image captions, scene-graphs, and multi-frame VLMs. Ultimately, we find a large gap between the best models (GPT-4V at 49.6%) and human-level performance (at 86.8%). In particular, for questions that require spatial understanding, the aforementioned agents perform similarly to blind LLMs, suggesting that further improvement on perception and semantic grounding is necessary before EQA agents are ready for real-world domains. In an era where LLMs are smashing hard QA tasks (e.g. SAT math exams), OpenEQA stands out as a straightforward, quantifiable, and practically relevant benchmark that poses considerable challenge to the current generation of foundation models. We thus believe OpenEQA is well positioned to serve as barometer for tracking future progress in multimodal learning and scene understanding.

REFERENCES

[1] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, "Embodied Question Answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] E. Wijmans, S. Datta, O. Maksymets, A. Das, G. Gkioxari, S. Lee, I. Essa, D. Parikh, and D. Batra, "Embodied Question Answering in Photorealistic Environments with Point Cloud Perception," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[3] L. Yu, X. Chen, G. Gkioxari, M. Bansal, T. L. Berg, and D. Batra, "Multi-target embodied question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[4] D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, and A. Farhadi, "Iqa: Visual question answering in interactive environments," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4089–4098.

[5] X. Ma, S. Yong, Z. Zheng, Q. Li, Y. Liang, S.-C. Zhu, and S. Huang, "Sqa3d: Situated question answering in 3d scenes," in *International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=IDJx97BC38

[6] D. Azuma, T. Miyanishi, S. Kurita, and M. Kawanabe, "Scanqa: 3d question answering for spatial scene understanding," 2022.

[7] P. Sermanet, T. Ding, J. Zhao, F. Xia, D. Dwibedi, K. Gopalakrishnan, C. Chan, G. Dulac-Arnold, S. Maddineni, N. J. Joshi, *et al.*, "Robovqa: Multimodal long-horizon reasoning for robotics," *arXiv preprint arXiv:2311.00899*, 2023.

[8] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan, "Seed-bench: Benchmarking multimodal llms with generative comprehension," 2023.

[9] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, K. Chen, and D. Lin, "Mmbench: Is your multi-modal model an all-around player?" 2023.

[10] D. Z. Chen, A. X. Chang, and M. Nießner, "Scanrefer: 3d object localization in rgb-d scans using natural language," *16th European Conference on Computer Vision (ECCV)*, 2020.

[11] OpenAI, "Gpt-4 technical report," *ArXiv*, vol. abs/2303.08774, 2023. [Online]. Available: https://arxiv.org/abs/2303.08774

[12] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[13] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang, "The dawn of lmms: Preliminary explorations with gpt-4v (ision)," *arXiv preprint arXiv:2309.17421*, vol. 9, 2023.

[14] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, 2023.

[15] Q. Gu, A. Kuwajerwala, S. Morin, K. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa, C. Gan, C. de Melo, J. Tenenbaum, A. Torralba, F. Shkurti, and L. Paull, "Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning," *arXiv*, 2023.

[16] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[17] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. M. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang, M. Savva, Y. Zhao, and D. Batra, "Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. [Online]. Available: https://openreview.net/forum?id=-v4OuqNs5P

[18] K. Yadav, R. Ramrakhya, S. K. Ramakrishnan, T. Gervet, J. Turner, A. Gokaslan, N. Maestre, A. X. Chang, D. Batra, M. Savva, *et al.*, "Habitat-matterport 3d semantics dataset," *arXiv preprint arXiv:2210.05633*, 2022. [Online]. Available: https://arxiv.org/abs/2210.05633

[19] Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, D. Parikh, and D. Batra, "Habitat: A Platform for Embodied AI Research," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[20] A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. Chaplot, O. Maksymets, A. Gokaslan, V. Vondrus, S. Dharur, F. Meier, W. Galuba, A. Chang, Z. Kira, V. Koltun, J. Malik, M. Savva, and D. Batra, "Habitat 2.0: Training home assistants to rearrange their habitat," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[21] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," 2023.

[22] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Batra, and D. Parikh, "Vqa: Visual question answering," 2016.

[23] Y. Hong, H. Zhen, P. Chen, S. Zheng, Y. Du, Z. Chen, and C. Gan, "3d-llm: Injecting the 3d world into large language models," 2023.

[24] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[25] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[26] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.

[27] P. Gao, J. Han, R. Zhang, Z. Lin, S. Geng, A. Zhou, W. Zhang, P. Lu, C. He, X. Yue, *et al.*, "Llama-adapter v2: Parameter-efficient visual instruction model," *arXiv preprint arXiv:2304.15010*, 2023.

[28] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[29] D. Ko, J. S. Lee, W. Kang, B. Roh, and H. J. Kim, "Large language models are temporal and causal reasoners for video question answering," *arXiv preprint arXiv:2310.15747*, 2023.

[30] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," 2017.

[31] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "Ok-vqa: A visual question answering benchmark requiring external knowledge," 2019.

[32] D. Schwenk, A. Khandelwal, C. Clark, K. Marino, and R. Mottaghi, "A-okvqa: A benchmark for visual question answering using world knowledge," 2022.

[33] Y. Zhong, J. Xiao, W. Ji, Y. Li, W. Deng, and T.-S. Chua, "Video question answering: Datasets, algorithms and challenges," *arXiv preprint arXiv:2203.01225*, 2022.

[34] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao, "Activitynet-qa: A dataset for understanding complex web videos via question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 9127–9134.

[35] J. Lei, L. Yu, M. Bansal, and T. L. Berg, "Tvqa: Localized, compositional video question answering," *arXiv preprint arXiv:1809.01696*, 2018.

[36] B. Wu, S. Yu, Z. Chen, J. B. Tenenbaum, and C. Gan, "STAR: A benchmark for situated reasoning in real-world videos," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. [Online]. Available: https://openreview.net/forum?id=EfgNF5-ZAjM

[37] Z. Hou, L. Ji, D. Gao, W. Zhong, K. Yan, C. Li, W.-K. Chan, C.-W. Ngo, N. Duan, and M. Z. Shou, "Groundnlq@ ego4d natural language queries challenge 2023," *arXiv preprint arXiv:2306.15255*, 2023.

[38] J. Xiao, X. Shang, A. Yao, and T.-S. Chua, "Next-qa: Next phase of question-answering to explaining temporal actions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9777–9786.

[39] A. Das, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, "Neural Modular Control for Embodied Question Answering," in *Proceedings of the Conference on Robot Learning (CoRL)*, 2018.

[40] Z. Yin, J. Wang, J. Cao, Z. Shi, D. Liu, M. Li, L. Sheng, L. Bai, X. Huang, Z. Wang, J. Shao, and W. Ouyang, "Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark," 2023.

[41] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[42] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

APPENDIX

APPENDIX I
RELATED WORK

The intersection of perception and language [6], [23], [8], [24], [25], [26], [27], [28], [29] has long been a fertile ground for AI research. Prior works studying perception and language have proposed Visual Question Answering (VQA) benchmarks, such as VQA-v1 [22], VQA-v2 [30], OK-VQA [31] and A-OKVQA [32], that focus on answering questions from a single image. Later works extended question answering tasks to videos [33], [34], [35] and 3D scenes [5], [10], [6], [23]. These include benchmark such as VideoQA [33], SQA3D [5] and ScanRefer [10]. While conceptually similar to our EM-EQA setting, these prior benchmarks focused on singular and narrow themes such as situated reasoning, object localization, object recognition, activity recognition, temporal window localization, and future forecasting [36], [37], [8], [34], [35], [38], [5], [10]. Another closely related line of work is prior benchmarks on Embodied QA [1], [2], [3], [39] and is conceptually similar to our A-EQA setting. They focus on leveraging RGB-D to accomplish navigation tasks in simulation [2], in which the agent must seek out multiple target locations or objects sampled from a closed vocabulary set [3] Our work takes inspiration from such prior works [1] and modernizes it to be relevant in the current era of foundation models. To our knowledge, our benchmark is the only one that incorporates all elements of a real-world use case for EQA: (1) The study of both episodic memory and active settings to accommodate for a wide variety of embodied agents like smartphones and mobile robots, (2) High quality real-world datasets with broad and non-templated questions, and (3) Embracing open-vocabulary interactions with users. In addition, our baselines use modern foundation models trained on vast internet data, enabling world knowledge beyond the reach of methods trained solely on simulated interactions.

LLMs have been used to scale the size of benchmarks either with their use for question and answer generation [8] or during evaluation time [9], [40]. Evaluation of open-vocabulary answers remains an open problem in AI. While the gold-standard remains human evaluations, they are time-consuming and expensive. An automatic evaluation process is desirable for benchmarking, quick iteration of research ideas, and model selection. We setup such a process by taking inspiration from recent works that study if LLMs can be used as an evaluation proxy in place of human raters [9]. Through a randomized control trial, we found a high correlation between human ratings and GPT-4, as evidenced by a Spearman correlation coefficient of 0.909.

APPENDIX II
OPENEQA BENCHMARK DETAILS

This section provides further details on the construction of the OpenEQA benchmark (section II-C). Specifically, we describe the process for generating human-like episode histories H for EM-EQA (appendix A), the interface for collecting question-answer pairs (Q, A^*) (appendix B), and the interface used to validate the dataset (appendix C).

A. Generating Episode Histories H

Episode histories H provide agents with observations of the environment, and are used for the EM-EQA split of OpenEQA in both ScanNet and HM3D environments. The ScanNet dataset was originally collected by people who were asked to scan indoor environments with an RGB-D camera. We use the initial 30 seconds (or 600 frames) of these human trajectories from ScanNet as EM-EQA episode histories H.

HM3D consists of scanned 3D environments, but does not come with pre-collected environment tours. Thus, we generate episode histories H using a two-step, semi-automated process. First, we generate a shortest-path trajectory from a starting location x_{src} to a destination x_{dst} in the environment. We select locations such that the geodesic distance between x_{src} and x_{dst} is $> 10m$ and the path curves (enforced by the criteria that $thegeodesicpathdistance \geq 1.1 \times Euclideanpathdistance$). Under these constraints, the paths typically traverse multiple rooms in the environment. To collect an episode history H, an agent travels along the path, while scanning the scene every $1m$ by rotating up to $180°$. These scans are intended to mimic human-like exploration behavior. After collecting the trajectories, we manually inspect each trajectory to ensure they properly explore the scene; we exclude trajectories with extended periods closely facing walls. This process results in one episode history H for 63 different HM3D validation environments.

B. Collecting Question-Answer Pairs (Q, A^*)

We use a Google Form to collect question-answer pairs (Q, A^*) annotations from 8 different AI researchers. Specifically, the annotators watch a video of a given episode history H, and generate questions for the 7 categories listed in section II-C. In the form, each category is described and one to two *good* and *bad* example questions are provided.

C. Interface for Dataset Validation

After the initial collection of question-answer pairs Q, A^*, we ask two independent people to validate each question. Specifically, the validators are shown an episode history H and a corresponding question Q on a simple HTML page. They are asked to provide an answer or mark the question as invalid (i.e. ambiguous or unanswerable). For the subset of *object localization* questions, we ask the validators to provide two answers for each questions because referring expressions often have multiple valid options (e.g. an object may be both *'left of the sink'* and *'right of the stove'*). We remove any question marked invalid by either validator.

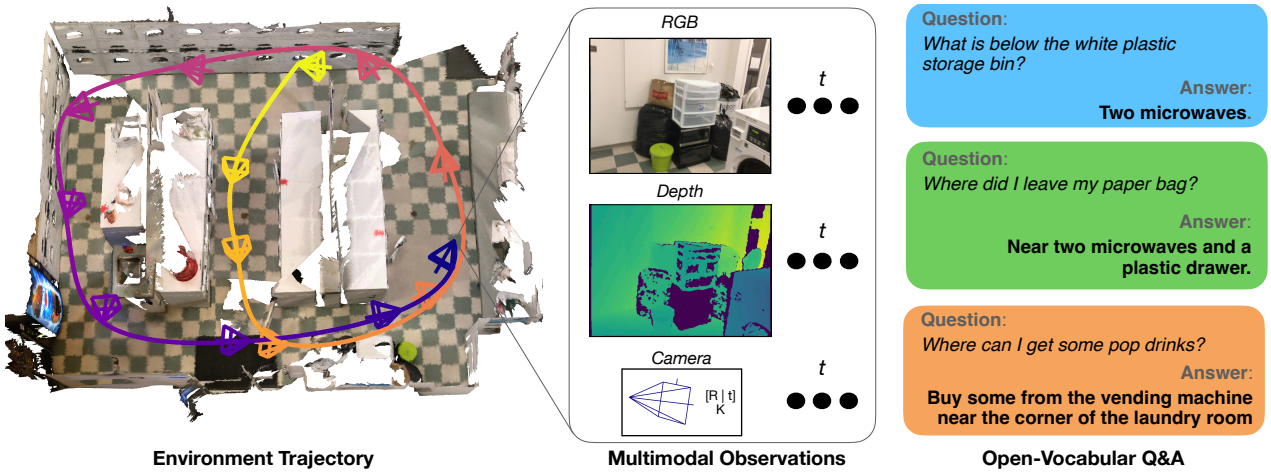


Fig. 2: **Illustration of an episode history along with questions and answers from our OpenEQA benchmark**, which contains 1600+ untemplated questions that test aspects of attribute recognition, spatial understanding, functional reasoning, and world knowledge. In episodic-memory EQA (EM-EQA), agents parse a sequence of historical sensory observations, and in active EQA (A-EQA), agents must explore real-world scanned environments to gather information to answer questions. Natural language answers are scored using our proposed LLM-Match metric, which showed excellent agreement with human scoring.

APPENDIX III
LLM-BASED EVALUATION DETAILS

OpenEQA questions often require open-ended answers, we use an LLM to evaluate correctness of answer produced by EQA agents. We prompt an LLM to compare human annotated answer A_i^* and model output A_i given a question Q_i and output a score σ_i on a scale of 1 to 5. On this scale, 1 indicates an incorrect response, 5 is a correct response and intermediate values represent different levels of similarity. Since questions can often have multiple correct answers, we also provide extra answers to the LLM prompt during scoring. We show the LLM prompt in Figure 3. Given the scores σ_i, we calculate an aggregate LLM-based **correctness** metric (LLM-Match) using Equation eq:em-eqa-metric.

APPENDIX IV
EQA AGENT FUNCTION SIGNATURES

In this section, we describe the function signature that is expected from an agent by OpenEQA benchmark.

Box 1 describes the function signature for the EM-EQA task. An agent is expected to produce a text answer to a question based on an episode history. The episode history generally consists of RGB, depth, camera pose, and camera intrinsic information. The benchmark does not prescribe any specific way to use the history. A variety of different approaches and representations of the history can be pursued by researchers, such as point clouds, NeRFs, or instance maps. Since all methods have the same set of episode history information at their disposal, it allows for a fair comparison of methods. The final natural language answer is evaluated using LLM-Match metric described in section II-D and appendix III.

Similarly, Box 1 also describes the expected function signature for A-EQA task. Here, an agent does not receive an episode history and must generate its own experience through exploration. To allow standardization, we provide access to the simulation environment and start state as part of the benchmark. The state allows for instantiating an environment and fixing the starting location of the agent and various objects. We do not prescribe a specific navigation API for the benchmark, researchers are free to pursue different abstractions such as atomic navigation actions or navigation skills, as long as it doesn't use any privileged simulation information. The final answer is evaluated for correctness using LLM-Match, and the efficiency (see section II-E) is computed using the number of atomic actions taken by the agent (to allow for standardization).

APPENDIX V
BASELINE AGENT DETAILS

This section provides additional details and LLM prompts for the blind LLM baseline (appendix A), Socratic LLM w/ Frame Captions example (appendix B), and GPT-4V (appendix C).

A. Blind LLM Prompt and Details

The prompt used for both our LLaMA-2 and GPT-4 blind LLM baselines is illustrated in fig. 6. We use the 70B parameter version of LLaMA-2 that is fine-tuned for chat, and the `gpt-4-0613` version of GPT-4.

B. Socratic LLM w/ Frame Captions Example

Figure 7 shows how Socratic LLM w/ Frame Captions baseline produces an answer to a question given K frames sampled from episodic memory H. We use LLaVa-1.5 to generate image captions. We use the 70B parameter version of LLaMA-2 that is fine-tuned for chat, and the `gpt-4-0613` version of GPT-4 for large language model.

Fig. 3: **Prompt used for LLM-Match scoring.** The placeholders {question}, {answer}, {extra_answers}, and {prediction} are replaced by the question Q, ground truth answer A^*, additional answer, and the agent's predicted answer A, respectively. Note that the extra answers are only available for *object localization* questions. When not available, corresponding sections of the prompt are omitted.

```
You are an AI assistant who will help me to evaluate the response given the question, the correct answer, and
extra answers that are also correct. To mark a response, you should output a single integer between 1 and 5
(including 1, 5). 5 means that the response perfectly matches the answer or any of the extra answers. 1 means that
the response is completely different from the answer and all of the extra answers.

Example 1:
Question: Is it overcast?
Answer: no
Extra Answers: ['doesn't look like it', 'no',' it's sunny']
Response: yes
Your mark: 1

Example 2:
Question: Who is standing at the table?
Answer: woman
Extra Answers: ['a woman', 'a lady', 'woman']
Response: Jessica
Your mark: 3

Example 3:
Question: Are there drapes to the right of the bed?
Answer: yes
Extra Answers: ['yes, there are drapes', 'yeah', 'the drapes are to the right of the king bed']
Response: yes
Your mark: 5

Your Turn:
Question: {question}
Answer: {answer}
Extra Answers: {extra_answers}
Response: {prediction}
```

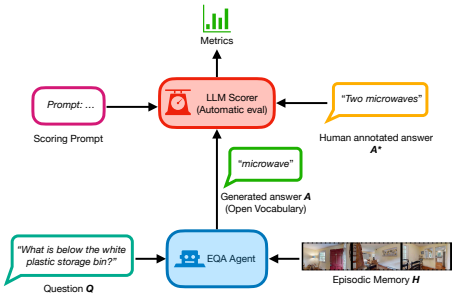


Fig. 4: Illustration of LLM-Match evaluation and workflow.

C. GPT-4V Details

Given an episodic memory H, we draw K frames and pass it to GPT-4V (through the API) in addition to question Q and prompt ω. We use chain-of-thought prompting in ω. We choose $K = 50$ for EM-EQA and $K = 75$ for A-EQA. Figure 8 shows the prompt ω and the input format passed to GPT-4V.

APPENDIX VI
SPARSE VOXEL MAPS

For building SVM, we use K uniformly-sampled frames from the episode history H. K varies across difference scenes but the principle is to find the minimum number of K (for reducing the run-time memory consumption) to cover the whole environment. We process each sampled frame with the following two steps:

Step 1. Detecting object views in the frame using Detic. Each object view v is a tuple of $\langle c, b \rangle$, where c is the 2D image crop of the object and b is the 3D bounding box in the world coordinate system. We first extract object masks from the frame by setting the vocabulary for Detic to more than 500 household object categories. Then we get the image crop c around each detected mask with an additional margin. We then use depth information to get a 3D point cloud where we run DBSCAN [41] to further filter out background points, and compute the bounding box b. Note that we only consider depths that are in the range of $[0.1m, 4m]$.

Step 2. Associating each object view v with a global object instance o. Most objects will be detected in more than one frame, and a main goal of SVM is to de-duplicate object views to create global object instances. Each global object instance o is a tuple of $\langle C, b^* \rangle$, where C is a list a image crops (i.e., c) from multiple viewpoints (i.e., v), and b^* is a re-computed 3D bounding box from a concatenated point cloud of different views. For matching v to o, SVM considers 3D bounding box overlapping and CLIP [42] embedding similarity.

After SVM is constructed, we then select the best crop from C per global instance o, where the object mask takes up the largest number of pixels. Each selected crop is passed to LLAVA-1.5 [21] to get the textual description, and all the descriptions with the instances' 3D coordinates (center of the bounding box b^*) are wrapped in a prompt for an LLM to answer the question Q. Limited by the LLM's capacity, we only consider topN ($N = 75$) instances ranked by the CLIP similarity between their visual feature and Q from all the instances we detect in SVM.

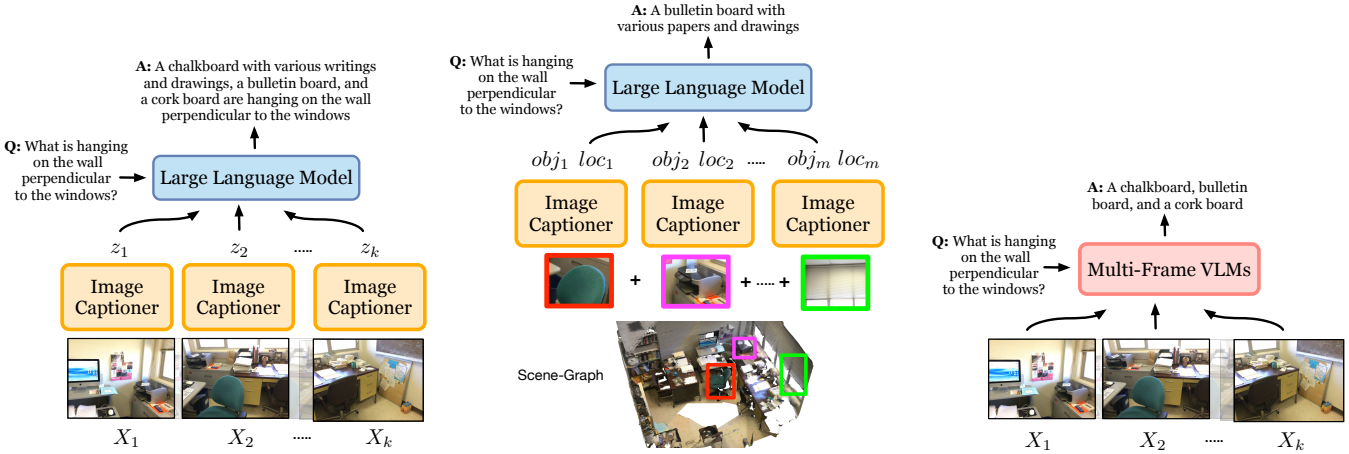


Fig. 5: **EQA Agents** (Left) Socratic LLMs w/ Image Captions generates captions for frames from episodic memory and provides it as context to an LLM to generate answer. (Middle) Socratic LLMs w/ Scene-Graph Captions leverage an object-centric scene-graph representation of episodic memory, which includes captions of object-centric crops and their 3D locations. (Right) Multi-Frame VLM directly processes visual frames from episodic memory to answer the question.

Algorithm 1 EQA Agent Signatures

```
def EMEQA_Agent(Q: str, H: dict) -> str:
    """ Function signature for EM-EQA Agents

    Args:
    - Q: EQA question
    - H: episodic memory (history)
        - keys -> rgb: image,
                  depth: image,
                  c_pose: camera pose,
                  c_in: camera intrinsics
        - H["rgb"] = np.array(T, H, W, 3)
        - H["depth"] = np.array(T, H, W, 1)
        - H["c_pose"] = np.array(T, 6)
        - H["c_in"] = np.array(T, 6)

    Returns:
    - answer: natural language
    """
    ...

    return answer

def AEQA_Agent(Q: str,
        S: dict) -> Tuple[str, int]:
    """ Function signature for A-EQA Agents

    Args:
    - Q: EQA question
    - S: initial state of simulator
        - keys -> metadata
        - S["metadata"] = Dict[str, Any]

    Returns:
    - answer: natural language
    - T: episode lifetime. Timesteps
         taken to answer the question
    """

    env = make_env(S["metadata"])
    env.set_state(S)
    ...
    return answer, T
```

Fig. 6: **Prompt used for Blind LLM baselines.** The placeholder {question} is replaced by the question Q. The same prompt is used for LLaMA-2 and GPT-4.

```
You are an intelligent question answering agent. I will
ask you questions about an indoor space and you must
provide an answer.

If the question does not provide enough information to
properly answer, provide an appropriate guess.

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: a bed
Explanation: bedrooms almost always contain a bed.

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

Q: {question}
```

APPENDIX VII
FORCE-A-GUESS DETAILS

As discussed in section III, we force baseline agents to guess an answer if they initial abstain – i.e. respond with an explanation for why the question is unanswerable. Specifically, we first ask an LLM if the initial answer is an abstaining response, and if so we replace the answer with a guess from a blind LLM. For step 1, use the prompt shown in fig. 9. We provide a comparison baseline performance with and without this procedure in appendix VIII.

APPENDIX VIII
FORCE-A-GUESS RESULTS

In table III, we present results illustrating the performance drop for baseline methods when they are allowed to abstain, rather than being forced to guess an answer. As expected,

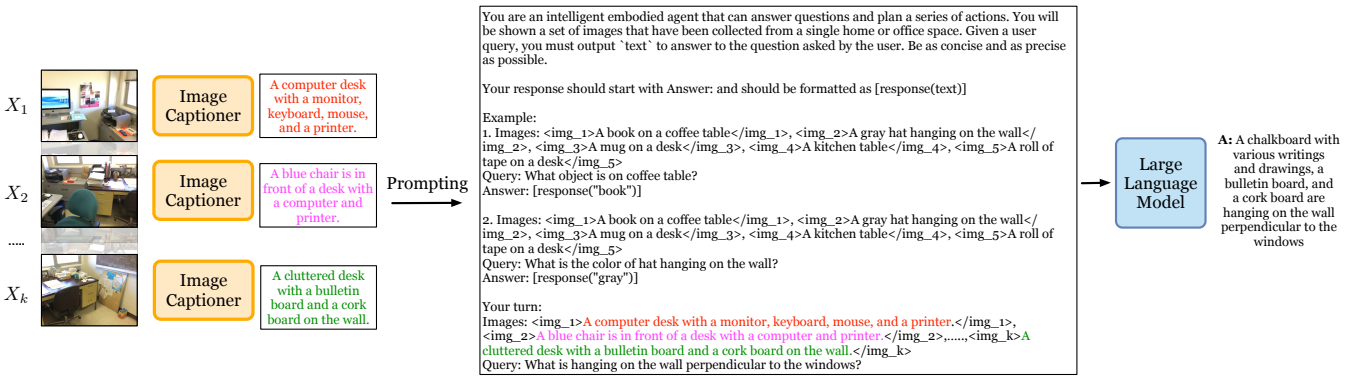


Fig. 7: **Input example for Socratic LLMs w/ Frame Captions baseline.** We first caption each of the K frames with an image captioner and then prompt the LLM with those captions along with the question. The large langauge model produces an answer.

Fig. 8: GPT4V input prompt.

```
You are an intelligent embodied agent that can answer
questions. You will be shown a set of images that have
been collected from a single location. Given a user
query, you must output 'text' to answer to the question
asked by the user.

User Query: {question}
Think step by step.
```

Fig. 9: **Prompt used for Force-A-Guess.** The placeholders `{question}` and `{old_answer}` are replaced by the question Q and initial answer A, respectively. The same prompt is used for LLaMA-2 and GPT-4.

```
You are an intelligent question answering agent. I need
you to fix the answers to these question.

If the proposed answer says the question is
unanswerable you should output the action ''guess''.
Otherwise, output the action ''keep''.

Question: What machine is on top of the stove?
Proposed Answer: the microwave
Action: keep

Question: What piece of furniture is in the middle of
the bedroom?
Proposed Answer: The question is unanswerable from the
provided images.
Action: guess

Question: {question}
Proposed Answer: {old_answer}
```

performance drops for most methods. We find that GPT-4-based methods (rows 3, 5, and 7) show the largest drop in performance, which corresponds with GPT-4's propensity to abstain. Specifically, for EM-EQA, GPT-4 abstains 36% to 55% of the time (as measured by GPT-4). LLaMA-2-based methods abstain 3% to 12% of the time (as measured by LLaMA-2). Thus, we observe minimal changes in LLaMA-2-based method scores. Finally, GPT-4V abstains 12% of the time (as measured by GPT-4), corresponding with a small drop in LLM-Match scores. Similar trends are observe in the A-EQA setting for all methods.

TABLE III: **LLM-Match scores without forcing agents to guess.** *GPT-4V results are calculated on a subset of 500 examples.

#	method	EM-EQA	EM-EQA (w/o guess)	A-EQA	A-EQA (w/o guess)
	Blind LLMs				
1	GPT-4	33.5	-	35.5	-
2	LLaMA-2	27.7	-	28.8	-
	Socratic LLMs w/ Frame Captions				
3	GPT-4 w/ LLaVA-1.5	43.6	29.3 (-14.3)	38.1	23.7 (-14.3)
4	LLaMA-2 w/ LLaVA-1.5	36.7	36.2 (-0.6)	30.9	31.2 (+0.4)
	Socratic LLMs w/ Scene-Graph Captions				
5	GPT-4 w/ ConceptGraphs	36.5	18.5 (-18.0)	34.4	12.4 (-21.9)
6	LLaMA-2 w/ ConceptGraphs	28.7	26.6 (-2.0)	23.8	18.9 (-4.8)
7	GPT-4 w/ Sparse Voxel Maps	38.9	27.3 (-11.5)	34.2	21.2 (-13.0)
8	LLaMA-2 w/ Sparse Voxel Maps	34.3	34.6 (+0.3)	29.9	29.3 (-0.6)
	Multi-Frame VLMs				
9	GPT-4V*	49.5	46.7 (-2.8)	41.8	40.6 (-1.2)
	Human	86.8	-	85.1	-

APPENDIX IX
FULL RESULTS

Table IV and Table V breaks down performance of different EQA agents, as described in Section III, on EM-EQA and A-EQA respectively by the seven question categories described in Section II-C. Due to API limitations, we only evaluate GPT4V on a subset of 500 OpenEQA questions in EM-EQA and 184 OpenEQA questions in A-EQA. We find that EQA agents with visual information excel at localizing and recognizing objects and attributes, and make better use of this information to answer questions that require world knowledge. However, on other categories, their performance is closer to the blind LLM baseline (GPT-4), indicating substantial room for improvement on OpenEQA.

APPENDIX X
LLM-MATCH HUMAN ALIGNMENT AND DETAILS

Evaluating open-vocabulary responses to questions is an open challenge in AI, and in particular for question-answering. While human evaluation remains the gold-standard, it is also expensive and time consuming. An automatic evaluation metric is preferable for benchmarking,

TABLE IV: **Category-level Performance on EM-EQA** Rows represent the different agents as described in Section III and columns represent the different category of questions in the dataset, as described in Section II-C. *GPT-4V scores are calculated on a subset of 500 OpenEQA question due to API limitations. Bold numbers indicate max in section.

# method	EQA Category							LLM-Match (C)
	object recognition	object localization	attribute recognition	spatial understanding	object state recognition	functional reasoning	world knowledge	
Blind LLMs								
1 GPT-4	**15.4**	**20.3**	**31.5**	**31.4**	51.0	**52.2**	**34.2**	**33.5±1.0**
2 LLaMA-2	10.7	15.3	22.3	25.0	**51.7**	44.1	29.7	28.3±1.0
Average	13.0	17.8	26.9	28.2	51.3	48.2	31.9	
Socratic LLMs w/ Frame Captions								
3 GPT-4 w/ LLaVA-1.5	**36.5**	**31.9**	**45.8**	**36.1**	**56.0**	**54.8**	**44.8**	**43.6±1.1**
4 LLaMA-2 w/ LLaVA-1.5	30.5	18.8	39.4	31.4	50.1	47.4	41.7	36.8±1.1
Average	33.5	25.4	42.6	33.8	53.0	51.1	43.3	
Socratic LLMs w/ Scene-Graph Captions								
5 GPT-4 w/ ConceptGraphs	26.4	17.0	40.6	29.1	**55.5**	**48.4**	39.9	36.5±1.0
6 LLaMA-2 w/ ConceptGraphs	17.1	13.9	24.4	27.2	43.5	38.1	39.0	28.7±1.0
7 GPT-4 w/ Sparse Voxel Maps	**30.0**	**20.0**	**49.6**	**31.7**	**55.5**	45.4	**40.8**	**38.9±1.0**
8 LLaMA-2 w/ Sparse Voxel Maps	23.4	11.7	38.9	30.8	52.8	45.4	39.1	34.3±1.1
Average	24.2	15.6	38.4	29.7	51.8	44.3	39.7	
Multi-Frame VLMs								
9 GPT-4V*	43.4	42.0	57.2	33.6	63.2	57.4	50.7	49.6±2.0
Average All Agents	29.6	22.2	42.3	31.4	53.8	48.1	42.3	
Human	87.9	77.3	87.9	86.7	98.7	81.8	87.2	86.8±0.6

TABLE V: **Category-level Performance on A-EQA.** Rows represent the different agents as described in Section III and columns represent the different category of questions in the dataset, as described in Section II-C. *GPT-4V scores are calculated on a subset of 184 OpenEQA question due to API limitations. Bold numbers indicate max in section.

# method	EQA Category							LLM-Score (C)
	object recognition	object localization	attribute recognition	spatial understanding	object state recognition	functional reasoning	world knowledge	
Blind LLMs								
1 GPT-4	**25.3**	**28.4**	**27.3**	**37.7**	47.2	**54.2**	**29.5**	**35.5±1.7**
2 LLaMA-2	13.7	22.1	16.2	29.7	**43.3**	50.4	28.8	29.0±1.6
Average	19.5	25.2	21.8	33.7	45.3	52.3	29.2	
Socratic LLMs w/ Frame Captions								
3 GPT-4 w/ LLaVA-1.5	**25.0**	**24.0**	**34.1**	**34.4**	**56.9**	**53.5**	**40.6**	**38.1±1.7**
4 LLaMA-2 w/ LLaVA-1.5	19.7	11.7	31.2	28.3	48.1	46.1	35.8	30.9±1.7
Average	22.3	17.8	32.6	31.3	52.5	49.8	38.2	
Socratic LLMs w/ Scene-Graph Captions								
5 GPT-4 w/ ConceptGraphs	25.3	16.5	29.2	37.0	52.2	**46.8**	37.8	34.4±1.8
6 LLaMA-2 w/ ConceptGraphs	13.3	11.9	18.8	27.9	31.7	31.7	36.5	23.9±1.6
7 GPT-4 w/ Sparse Voxel Maps	**29.0**	**17.2**	**31.5**	**31.5**	**54.2**	39.8	**38.9**	**34.2±1.8**
8 LLaMA-2 w/ Sparse Voxel Maps	16.7	9.7	33.4	29.0	47.2	40.5	37.5	29.9±1.7
Average	21.1	13.8	28.2	31.3	46.3	39.7	37.7	
Multi-Frame VLMs								
9 GPT-4V*	34.0	34.3	51.5	39.5	51.9	45.6	36.6	41.8±3.2
Average All Agents	23.3	17.9	32.8	32.5	48.9	43.4	37.7	
Human	89.7	72.8	85.4	84.8	97.8	78.9	88.5	85.1±1.1

fast iteration, and model selection. We thus use an automatic LLM-Based evaluation metric in this work as described in Section II-D. We performed analysis experiments to test the quality of this metric along two axis: (1) How closely aligned is the LLM-Match metric with human evaluators? (2) How sensitive is the LLM-Match metric towards specific choice of prompts and the LLM?

Human Alignment. To answer the first question, we designed an experiment to measure the agreement between LLM-Match metric and human evaluators. For this analysis,

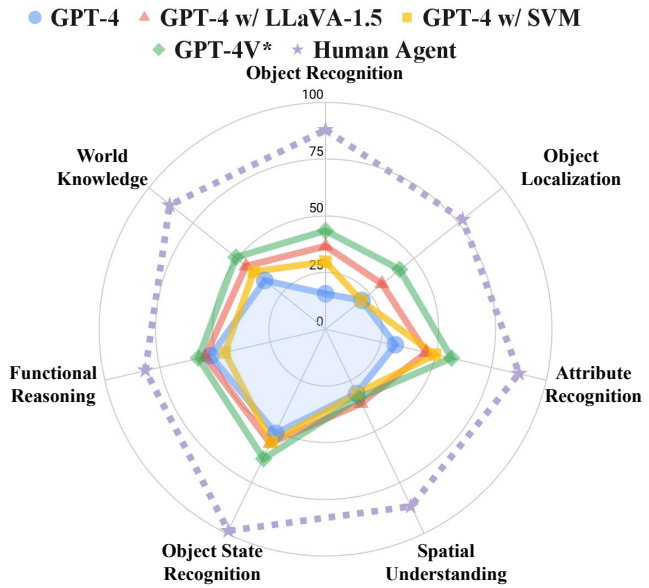


Fig. 10: **Category-level performance on EM-EQA.** We find that agents with access to visual information excel at localizing and recognizing objects and attributes, and make better use of this information to answer questions that require world knowledge. However, on other categories performance is closer to the blind LLM baseline (GPT-4), indicating substantial room for improvement on OpenEQA. See scores for all methods in appendix IX.

TABLE VI: **Varying LLM used for scoring.** On a subset of 100 questions with answers from GPT-4, GPT-4 scoring shows excellent agreement with human judgement, while using other LLMs shows lower correlation (Spearman correlation coefficient).

Scorer LLM	ChatGPT-4	ChatGPT3.5	LLaMA 2	Human
ChatGPT-4	1.00	0.66	0.68	0.88
ChatGPT3.5	-	1.00	0.66	0.61
LLaMA 2	-	-	1.00	0.63
Human	-	-	-	1.00

we uniformly sampled a subset of 300 questions from OpenEQA. To ensure coverage of the answer distributions (i.e. poor, fair, and good answers), we sampled 100 responses from a blind LLM (LLaMA-2), multi-frame VLM (GPT-4V), and human baseline answers. In a double blind study, we then asked 4 human evaluators to score the 300 responses using an evaluation prompt similar to the one used by LLM-Match. The evaluators were provided no information about the source of the response (except an MD5 hash of the question ID, response source, and annotator ID). We found a **Spearman's** $\rho = 0.909$ **between human and LLM evaluation** (bootstrap CI=(0.883,0.928), N=9999), indicating excellent agreement with human judgement. Table VII shows the Spearman's ρ (a measure of correlation) between (1) each annotator and other humans and (2) each annotator and GPT-4 scoring. Human evaluators correlated with other humans in $\rho \in [0.91, 0.93]$, and with LLMs in $\rho \in [0.90, 0.94]$.

Choice of LLM. Table VI shows rho between human evaluators and different LLMs, on the subset of 100 questions from GPT4V. GPT-4 scoring shows good agreement with human scoring ($\rho = 0.88$), while GPT3.5 (ρ=0.66) and LLaMA 2 (ρ=0.68) show lower correlation. We believe that future LLMs will show higher agreement with human annotators, and in the meantime we recommend only using GPT-4 for scoring.

TABLE VII: **Per-annotator Spearman-ρ.** Human scoring has excellent agreement with both other humans and with LLM scoring.

Annotator	vs. Other Humans	vs. LLM
0	0.91	0.91
1	0.91	0.91
2	0.92	0.90
3	0.93	0.94

APPENDIX XI
LLM-MATCH ROBUSTNESS DETAILS

TABLE VIII: **LLM Role.** Correlation between scores when changing the 'role' of the LLM in the scoring prompt (Spearman correlation coefficient).

Role	AI	"Score Master"	Professional
AI	1.00	0.97	0.96
"Score Master"	-	1.00	0.97
Professional	-	-	1.00

TABLE IX: **Match criterion for a '5'.** Correlation between scores when changing the criterion in the scoring prompt (Spearman correlation coefficient).

Match Crit.	Perfect	Contains	Pro	Person
Perfect	1.00	0.96	0.95	0.96
Contains	-	1.00	0.97	0.97
Pro	-	-	1.00	0.98
Person	-	-	-	1.00

TABLE X: **Temperature of scoring LLM.** Changing the temperature of GPT-4 used in scoring (Spearman correlation coefficient).

Temp	0.01	0.1	0.2	0.3
0.01	1.00	0.98	0.98	0.98
0.1	-	1.00	0.97	0.98
0.2	-	-	1.00	0.97
0.3	-	-	-	1.00

Our LLM-Match uses the specific evaluation prompt described in fig. 3. The metric is stable under small permutations of the prompt and LLM-Match settings as illustrated in Table VIII, Table IX and Table X, which show the correlation in LLM-Match scores using different prompting strategies, assessed on 500 GPT-4V answers.

Role: Table VIII demonstrates that changing the LLM's role from 'AI' to 'Score Master' or 'professional evaluator' does not significantly change the results, and scores between any two treatments have a tight correlation with a Spearman's ρ all above 0.95.

Match criterion: Similarly, Table IX shows analogous results ($\rho > 0.95$) when changing the description of a '5' from 'perfect match' to 'contains correct answer', 'similar to a reasonable person', or 'reasonable professional'.

Temperature: The stochasticity in the evaluation function has negligible impact as well, as shown by varying the temperature and seed. Table X shows results when varying the temperature used in the GPT-4 scorer from 0.01-0.3, with results all >0.97.

APPENDIX XII
3D COORDINATE ABLATION

TABLE XI: **Ablating 3D location for scene-graph agents.** Removing bounding box locations and extent had no significant effect for agents using either LLM.

	LLM-Match	
method	w/ 3D BBox	Crop-Only
GPT-4 w/ Sparse Voxel Maps	38.9±1.0	**39.6±1.0**
LLaMA-2 w/ Sparse Voxel Maps	34.3±1.1	36.6±1.1

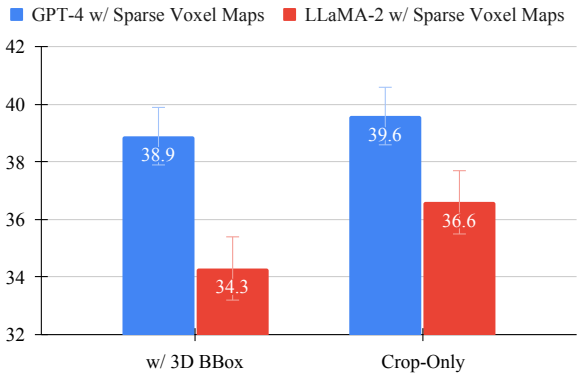


Fig. 11: **Ablating 3D location for scene-graph agents.** Removing bounding box locations and extent had no significant effect for agents using either LLM.

Figure 11 and table XI compares the EM-EQA performance of the Socratic baseline that uses Sparse Voxel Map captions with and without including 3D bounding box information in the text descriptions. Results show that explicit bounding box location and size information from the scene graph does not significantly change the performance of scene-graph based agents. This suggests that neither LLM, trained with only text information, is able to effectively use the 3D location information.

APPENDIX XIII
OPENEQA DATASET EXAMPLES

Additional examples from the ScanNet and HM3D splits of OpenEQA are provided in the Figures 12, 13, and 14.

Episode History H

Question-Answer Pairs (Q, A^*)

[Object Recognition]
Q: What is the red object on the chari?
A*: a backpack

[Attribute Recognition]
Q: Among all the chairs, what is the unique color of the chair?
A*: green

[Spatial Understanding]
Q: Can 10 people sit in this room?
A*: yes

[Object State Recognition]
Q: Is the plastic water bottle open?
A*: no

[Functional Reasoning]
Q: What can I use to write something on using my pencil?
A*: the piece of paper

[World Knowledge]
Q: Were students here lately?
A*: yes

[Object Localization]
Q: Where is my unfinished Starbucks drink?
A*: on the table near the front whiteboard
Extra Answers: ['On the second table from the front',
'In the center of the second table.',
'On the second table from the front',
'On the table near the windows']

Fig. 12: **OpenEQA dataset examples from a ScanNet scene.** Note that only a subset of frames from the episode history H are displayed. Thus, some questions may require additional visual information to answers.

Episode History H

Question-Answer Pairs (Q, A^*)

[Object Recognition]
Q: What object is on top of the work desk?
A*: shelf

[Attribute Recognition]
Q: what color is the task chair?
A*: black

[Spatial Understanding]
Q: which object is closer to the window, the bed or the trash can?
A*: bed

[Object State Recognition]
Q: Is the closet door fully closed?
A*: no

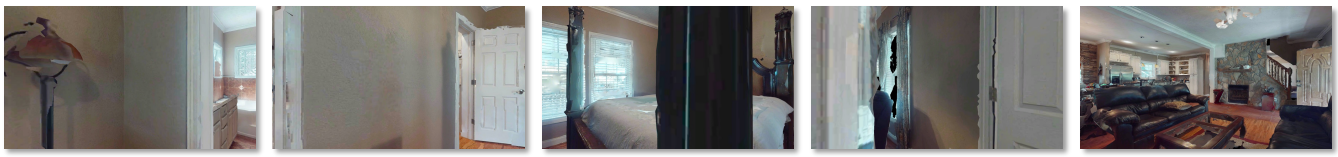
[Functional Reasoning]
Q: The closet is full. Where can I store a suitcase in this room?
A*: Under the bed

[World Knowledge]
Q: What are the two main functions or purposes of this room? Why did you arrive at those conclusions?
A*: Sleeping, since there is a bed. Working, since there is a task chair and desk.

[Object Localization]
Q: Where is the two-tier shelf?
A*: on top of the desk
Extra Answers: ['On top of the desk.',
'On the left when you first walk in to the room.',
'above the desk infront of the chair',
'on top of the desk']

Fig. 13: **OpenEQA dataset examples from a ScanNet scene.** Note that only a subset of frames from the episode history H are displayed. Thus, some questions may require additional visual information to answers.

Episode History H



Question-Answer Pairs (Q, A^*)

[Object Recognition]
Q: what is on the chair?
A*: a soft pillow

[Attribute Recognition]
Q: is the outside door open or closed?
A*: open

[Spatial Understanding]
Q: is the table in the living room clean?
A*: yes

[World Knowledge]
Q: what is special about the wall in the living room?
A*: it seems to be made of stone

[Object Localization]
Q: where is the standing lamp?
A*: next to the bed in the bedroom
Extra Answers: ['in the bedroom',
'to the left of the bed',
'the bedroom',
'The room with the bed and the bathroom']

Fig. 14: **OpenEQA dataset examples from an HM3D scene.** Note that only a subset of frames from the episode history H are displayed. Thus, some questions may require additional visual information to answers.