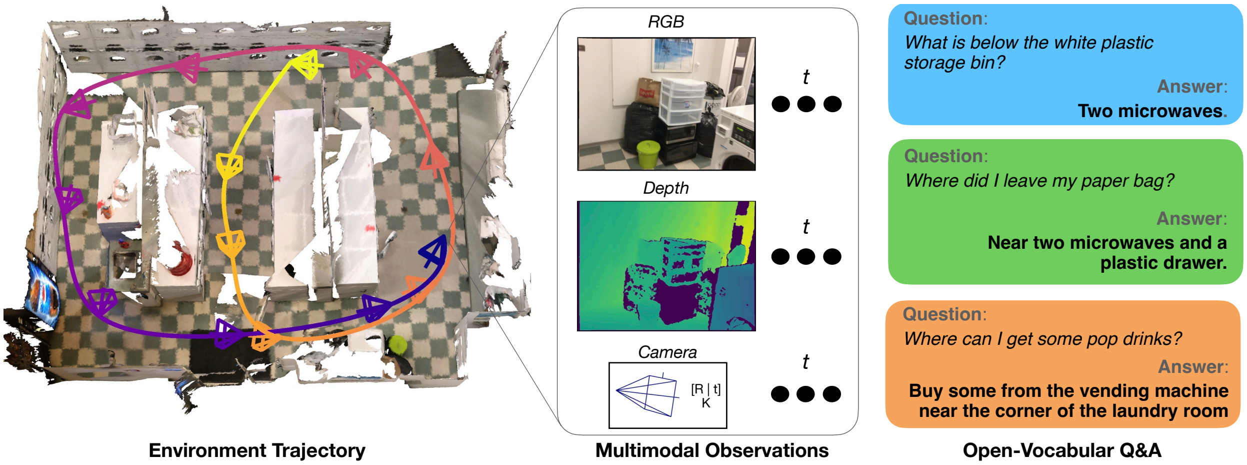
OpenEQA: Embodied Question Answering in the Era of Foundation Models

https://open-eqa.github.io

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] Alexander Sax[4] Aravind Rajeswaran[4]

[*] Equal contribution. 1. Georgia Tech 2. MIT 3. Binghamton University 4. Meta AI 5. UC Berkeley 6. CMU

Work done at Fundamental AI Research (FAIR), Meta.



Question: What is below the white plastic storage bin?
Answer: Two microwaves.

Question: Where did I leave my paper bag?
Answer: Near two microwaves and a plastic drawer.

Question: Where can I get some pop drinks?
Answer: Buy some from the vending machine near the corner of the laundry room

Figure 1. **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.

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 supporting both episodic memory and active exploration use cases. OpenEQA contains over 1600 *high-quality human generated questions drawn from over* 180 *real-world environments. In addition to the dataset, we also 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 including GPT-4V, 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.*

1. Introduction

AI agents are starting to transcend their digital origins and enter the physical world through devices like smartphones, smart glasses, and robots. These technologies are typically used by individuals who are not AI experts. To effectively assist them, Embodied AI (EAI) agents must possess a natural language interface and a type of "common

Table 1. **OpenEQA vs existing benchmarks.** OpenEQA has multiple modalities, real scenes, active agents, and automated scoring.

	Modalities			Open Vocab	Real Scenes	EM (video)	A(ctive)	LLM Scoring
	RGB	Depth	Camera					
EQA-v1 [7]	✔	✔	✔	✖	✖	✖	✔	✖
MP3D-EQA [41]	✔	✔	✔	✖	✔	✖	✔	✖
MT-EQA [50]	✔	✔	✔	✖	✖	✖	✔	✖
IQA [13]	✔	✔	✔	✖	✖	✖	✔	✖
SQA3D [31]	✔	✔	✔	✖	✔	✖	✖	✖
ScanQA [3]	✔	✔	✔	✖	✔	✔	✖	✖
RoboVQA [38]	✔	✖	✖	✔	✔	✔	✖	✖
SEED-Bench [23]	✔	✖	✖	✖	✔	✔	✖	✖
MMBench [29]	✔	✖	✖	✔	✔	✖	✖	✔
OpenEQA (Ours)	✔	✔	✔	✔	✔	✔	✔	✔

sense" rooted in human-like perception and understanding of the world. Recently, "foundation models" [4] trained on massive datasets have emerged as a promising approach to develop these capabilities. Against this backdrop, we propose that Embodied Question Answering (EQA) is both a useful end-application as well as a means to evaluate an agent's understanding of the world. Simply put, EQA is the task of understanding an environment well enough to answer questions about it in natural language as illustrated in Fig. 1. In this work, we present OpenEQA – the first open-vocabulary benchmark for EQA, and use it to study performance of various state of the art foundation models [15, 19, 26, 34, 35, 40, 47].

Specifically, we study two variants of EQA under a common umbrella: episodic memory (EM-EQA) and active exploration (A-EQA), depending on the agent platform. EM-EQA is applicable to devices like smart glasses that can leverage episodic memory generated by human wearers to answer questions. This has the potential to enhance the memory, perceptual capabilities, and general knowledge of the user. On the other hand, A-EQA is relevant to mobile robots that can autonomously explore environments to gather information to answer questions. For example, to answer the question: *'Q: Do I have Cayenne pepper left at home?'*, a robot can search the home before responding, *'A: I found a bottle of Cayenne pepper in the pantry.'*

The intersection of perception and language has long been a fertile ground for research in AI. While the broad problem of EQA [7, 50] and VQA [3, 5, 31] 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 Tab. 1. In particular, OpenEQA is the first open-vocabulary benchmark for EQA, and supports both the episodic-memory and active settings. The key technologies enabling this are: (1) videos and scans of real-world environments like ScanNet [6], Gibson [43], and HM3D [36], as well as simulators capable of rending these scenes [11, 21, 24, 32, 39]; and crucially (2) LLMs capable of scoring open-ended answers. This combination allows us to source questions from human annotators by watching episodes, and then automatically score responses of models against these annotated answers, enabling us to study a wide range of questions and methods (see Sec. 3). The combination of episodes from real-world environments and open-ended questions makes OpenEQA distinct from previous EQA [7, 13], 3DQA [3, 31], and VQA [1, 22, 33, 37] benchmarks that are either closed-vocabulary (i.e. a closed set of answer), require only a single frame, use simple procedurally-generated scenes, or non-interactive in nature.

1.1. Our Contributions

1. **Benchmark:** Our primary contribution is a modern reformulation of the EQA problem statement along with a concrete evaluation benchmark (OpenEQA) that contains over **1600 questions** sourced from over **180** real-world environments and scans [6, 36, 45]. The questions were meticulously crowd-sourced to be representative of real-world use cases. Each question was annotated by at least three individuals, ensuring the validity of the questions-answer pairs. **EM-EQA** requires answering questions by leveraging a provided episodic memory. For **A-EQA**, we focus on the subset of questions in simulation of photo-realistic scanned environments. The EAI agent is spawned at an initial location and must take any required exploratory actions to answer the question. The agent is rated on both the correctness of the answer as well as efficiency of its actions, to reward agents that perform targeted exploration specific to the question.
2. **Evaluation:** The open-vocabulary nature of our benchmark increases the complexity of evaluating answers generated by various models. While human evaluations have been the gold-standard in benchmarking LLMs, they can often be prohibitively slow and/or expensive. We thus utilize an LLM [34, 40] to score answers based on similarity to ground truth answers generated by humans. Through a double blind study, we find that there is a strong correlation between our LLM-Match metric and human preferences.
3. **Baselines:** Additionally, we provide a set of baseline results and implementations. These include the recently released GPT-4V [47] and Socratic use of LLMs [34, 40] that leverage captioning models [27] or generated scene-graph representations [15]. Through our evaluation, we find that GPT-4V is the strongest baseline achieving a score of 49.6%. While impressive, this significantly lags behind human-level score of 86.8% on our benchmark, underscoring the difficulty and relevance of the benchmark for our community. In particular, all the current generation of foundation models especially struggle at questions that require spatial understanding of objects and scenes, often performing no better than "blind" LLMs, highlighting a major deficiency.

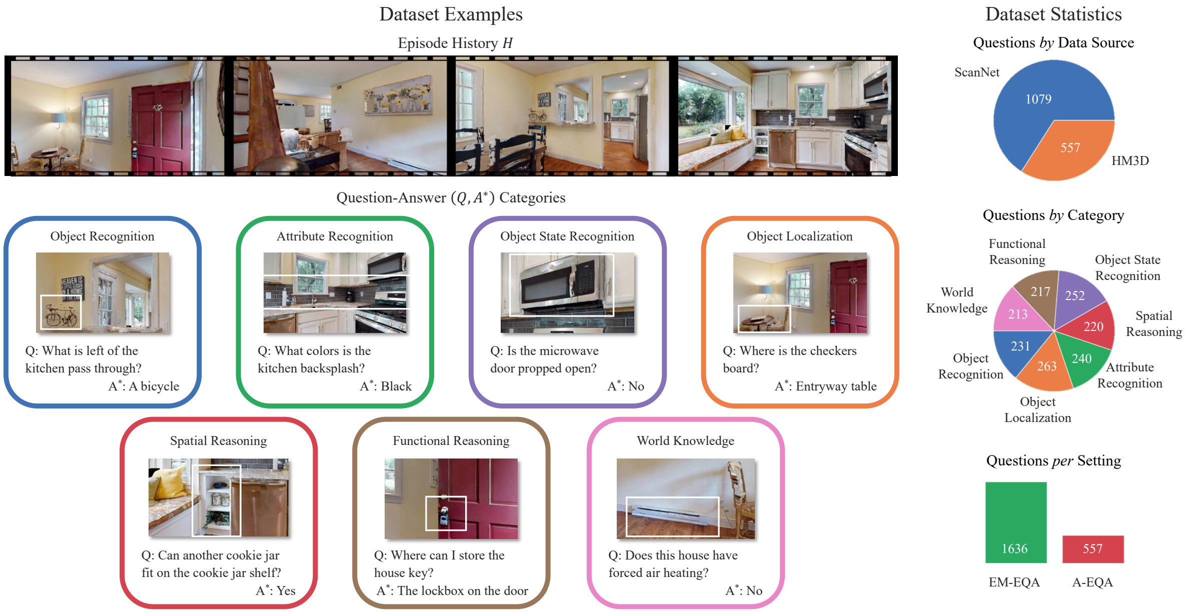


Figure 2. **Example questions and dataset statistics of OpenEQA.** The episode history H provides a human-like tour of a home. EQA agents must answer diverse, human-generated questions Q from 7 EQA categories, aiming match the ground answers A^*. Tours are collected from diverse environments including home and office locations (not shown above). Additional dataset examples are in Appendix O. 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.

2. 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.

2.1. Episodic-Memory Question Answering

The episodic memory EQA (EM-EQA) task is concerned with the setting where an agent must develop an understanding of the environment from its episodic memory to answer questions. This is particularly relevant for EAI agents embedded in devices such as smart glasses, which cannot autonomously explore the world and must rely on the history of observations to assist users (e.g. *'Q: I can't find my keys, where did I leave them? A: On the kitchen island.'*) An instance of EM-EQA is defined by the 3-tuple: (Q, H, A^*) where Q refers to an open-vocabulary question, H is a history of observations (i.e. episodic memory), and A^* is a ground truth answer (e.g. as annotated by a human). The agent's task is to generate an answer using the episodic memory, i.e. `A = EMEQA_Agent(Q, H)`, that is *"similar"* to the ground truth answer A^*. A concrete function signature that is expected for the agent is described in Algorithm 1 in Appendix D.

2.2. Active Embodied Question Answering

The *Active* EQA (A-EQA) problem studies the setting where an autonomous agent can answer questions by taking exploratory, information gathering actions when necessary (e.g. *'Q: Do we have canned tomatoes at home? A: Yes, I found canned tomatoes in the pantry.'*). For simplicity, our benchmark considers questions that require only navigation actions. In principle, this can be extended to mobile manipulators to allow for both navigation and manipulation actions (e.g. opening doors and cabinets) [48]. More formally, an instance of A-EQA is specified by the 3-tuple (Q, S, A^*). Similar to Section 2.1, Q and A^* denote the question and human annotated answer, respectively. S refers to the simulator initialized at the appropriate state state [39], and encompasses all details and assets needed to recreate the environment. Once the agent is spawned at S, it must take any necessary exploratory actions before producing an answer A. Please see Algorithm 1 in Appendix D for a concrete function signature of an A-EQA agent. Once the agent generates answer A, it is evaluated both for the correctness of the answer as well as the efficiency of actions.

2.3. 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 [6] and 3D scans of real-world environments [32, 36, 39, 45]. Then, we meticulously validate each question-answer pair to provide a high-quality benchmark for EM-EQA and A-EQA. The dataset is designed to reflect the types of questions that users might ask an AI assistant embedded in smart glasses or a mobile robot assistant. We present examples and dataset statistics in Fig. 2 and compare it to existing benchmarks in Tab. 1.

Data Sources. We collect episode histories H from two sources: ScanNet [6] and HM3D [36, 45]. For ScanNet, we utilize RGB-D data captured from human exploration in various indoor settings, such as bedrooms and offices, and translate these videos into episode history H. We selected 90 validation scenes and 10 test scenes from ScanNet. For the scans in HM3D rendered through Habitat, we define a heuristic exploration policy to mimic human behavior and manually verify that exploration trajectories adequately explore the space, ultimately resulting in episode histories for 87 validation scenes, as detailed in Appendix B.

Question Generation. In a preliminary experiment, we showed human annotators the history H and asked them to generate question-answer pairs (Q, A^*) while playing the role of end users. This exercise led to the identification of seven EQA question categories that broadly encompass the range of questions asked of AI assistants. They test an agent's ability to (1) *recognize objects* (e.g. what is on the coffee table?), (2) *recognize object attributes* (e.g. colors or shapes), (3) *recognize object states* (e.g. open or closed), (4) *localize objects* (e.g. where are my keys?), (5) perform *spatial reasoning* (e.g. I'm sitting on the couch watching TV, in which direction should I turn to find the kitchen?), (6) perform *functional reasoning* (e.g. how can I cool down this room?), and (7) utilize *outside world knowledge* (e.g. who/what is depicted in a painting?). The final OpenEQA dataset focuses on these seven categories. Annotators were asked to generate two questions and answers per category after viewing H. Illustrations of the question categories are provided in Fig. 2, and additional details on the dataset collection and interface are in Appendix B.

Dataset Validation. Each question created by humans underwent further examination by two independent annotators. Validators watched the episode history and assessed whether the question was unanswerable, ambiguous, or if the answer was incorrect. Any question-answer pair with identified issues was discarded. The interface for validation is provided in Appendix B. The final dataset includes 1636 questions following the statistics in Fig. 2.

Dataset Splits. The validated (Q, A^*) pairs are used for EM-EQA, and reused for A-EQA since we recorded S in addition to H for simulated scenes. Specifically, A-EQA agents are initialized at the same start state S that was used to generate the episodic memory H for EM-EQA. The existence of a feasible trajectory H provides proof that A-EQA questions are answerable. However, they can potentially be answered more efficiently through targeted exploration.

Additional Object Localization Answers. Among the 7 EQA categories, *object localization* questions pose a unique challenge for evaluation, because they often have multiple correct answers with differences that go beyond rephrasing. For example, the question *'Q: Where is the toaster?'* may have multiple correct answers such as *'A1: to the right of the stove'* or *'A2: to the left of the fridge'*. Thus we collect 4 additional answers from 2 additional annotators resulting in 5 answers per object localization question that reflect a natural distribution of answers that humans would expect from each localization question.

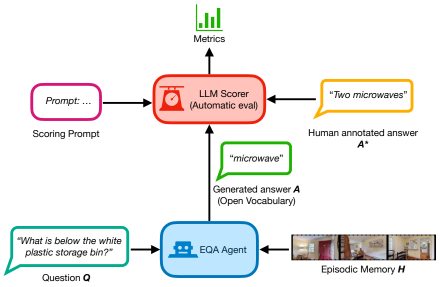

Metrics
Prompt: ...
Scoring Prompt
LLM Scorer (Automatic eval)
"Two microwaves"
Human annotated answer A*
"microwave"
Generated answer A (Open Vocabulary)
"What is below the white plastic storage bin?"
Question Q
EQA Agent
Episodic Memory H

Figure 3. Illustration of LLM-Match evaluation and workflow.

2.4. 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, especially for benchmarks. As an alternative, we use an LLM to evaluate the correctness of open-vocabulary answers produced by EQA agents. Specifically, we adapt the evaluation protocol introduced in MMBench [29] 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, and intermediate values represent levels of similarity. 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 illustrated in Fig. 3, detailed in app. C, and validated against human judgement in Sec. 5.

2.5. Evaluating Efficiency for A-EQA

In A-EQA, we evaluate an agent based on two criteria: (a) **correctness** of the answer based on similarity with human annotation A^* as described in Eq. (1); and (b) **efficiency**, which measures how quickly the agent answered the question and favors agents that perform targeted exploration necessary for the question.

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:

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

which can be interpreted as modified version of SPL [2] (a metric commonly used to measure the efficiency of navigation agents). We note that the ground-truth paths are generated by a scripted exploratory agent. This path was used to construct the (Q, A) pairs, so it is guaranteed to contain sufficient information to answer. However, we note that these paths are not necessarily optimal, and thus $E > 100\%$ is theoretically possible.

3. EQA Agents

This section describes the different EQA agents we study and evaluate in this work. Our guiding principle is to explore different ways in which foundation models (specifically LLMs and VLMs) can be used for EQA without any additional fine-tuning. Towards this goal, the family of agents studied are: (1) blind LLMs [34, 40], (2) Socractic LLMs w/ frame captions [26], (3) Socractic LLMs w/ scene-graph representations [15], and (4) VLMs that can directly process multiple frames (e.g. GPT-4V [47]). For simplicity, we first describe the agents in the EM-EQA setting, and subsequently discuss extensions to A-EQA. All agents have the general signature of $A = \text{Agent}(Q, H)$ and contain a language model component that generates the answer. The agents primarily differ in their perception capabilities and how they process H. In addition to these agents, we also study how humans perform in our benchmark.

Blind LLMs. The text-only or 'blind' LLM agent simply produces an answer based on the question Q without considering any visual information about the scene, i.e. $A = \text{LLM}([\omega, Q])$, where ω is a generic prompt that we prepend to the question. See Appendix E for additional details. This agent provides a reference for how far we can get purely using prior world knowledge and/or random guessing (e.g. yes/no questions). For the LLM choice, we present results with both GPT-4 [34] and LLaMA-2-70B [40].

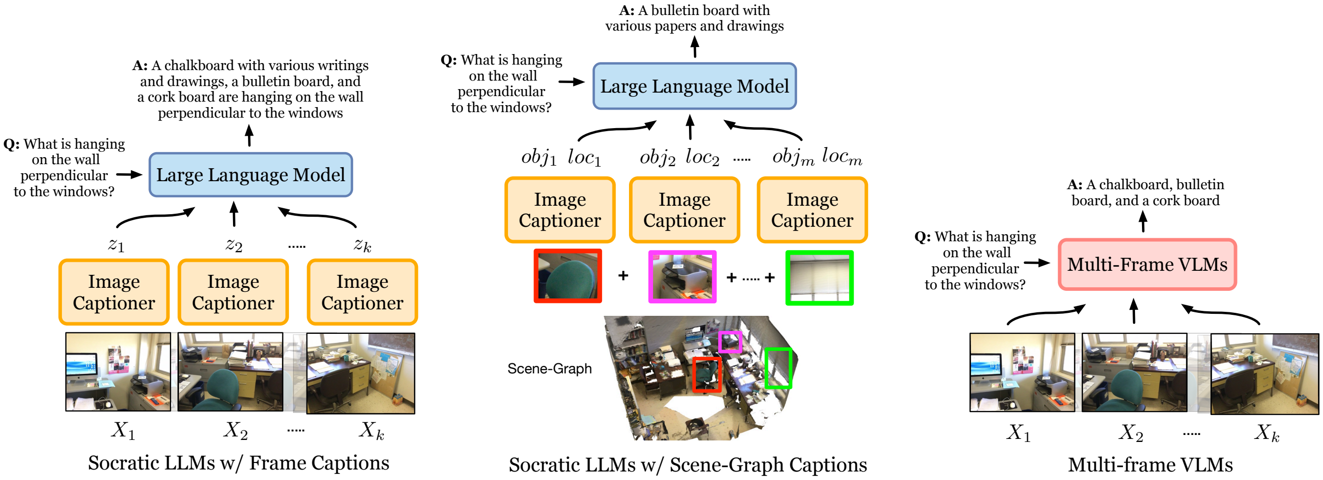
Socratic LLMs w/ Frame Captions. This is the simplest agent we study that leverages the perceptual information from the episodic memory H. Let $\{X_1, X_2, \ldots X_K\}$ be K frames drawn from the episodic memory H. We first leverage an image captioning model (e.g. LLaVA [26, 27]) to generate $z_i = \text{Captioner}(X_i),\ i = 1, \ldots K$. These captions provide a language description of the episodic memory to the LLM, which could allow it to answer better than a blind agent. The final answer is computed by the agent using a generic prompt, the aforementioned frame captions, and the question, i.e. $A = \text{LLM}([\omega, z_1, z_2, \ldots z_K, Q])$. See Appendix E for an example of the input. In practice, we sample K frames uniformly over time from H, with $K = 50$ for EM-EQA and $K = 75$ for A-EQA. For the captioning model, we use LLaVA-v1.5 [26], and for the LLM we study both GPT4 [34] and LLaMA-2-70B [40].

Socratic LLMs w/ Scene-Graph Captions. We next study agents that leverage an object-centric scene-graph representation of H. The motivation for such agents is that an object-centric representation might allow for a more fine-grained perceptual understanding of objects, and provide a textual representation that might be easier for LLMs to reason over. Object-centric 3D world representations involve constructing a scene graph $G = \text{SceneGraph}(H)$ that contains a description of the objects in the scene, their semantic attributes such as color and 3D locations, and their relationships. We study two methods of constructing such a scene-graph: (1) ConceptGraph [15]; and (2) Sparse Voxel Map (SVM). ConceptGraph (CG) generates a textual scene-graph representation by first detecting various objects in the scene, extracting the 3D location of objects using camera pose and depth information, and sematic descriptions of objects by using an image captioning model on crops of the object extracted from the video. We use the publicly released implementation of CG, which uses Grounded-SAM [18, 28] with RAM [52] for object detection and LLaVA-v1 [27] for image captioning. SVMs are constructed similarly to CGs, but differ in the post-processing of object detections and in the image captioning model used. See Appendix F for details. Once a textual scene graph G is generated, we use it for EQA as $A = \text{LLM}([\omega, G, Q])$.

Multi-Frame VLMs. The most generic agent for EQA is one that can directly process the entire episodic memory to answer questions, i.e. $A = \text{MultiFrameVLM}([\omega, Q, H])$. The recently released GPT-4V model [47] is capable of processing up to 50 frames (through the API) in addition to textual queries. We thus extract 50 frames uniformly spaced from H and provide it to GPT-4V in addition to prompts for generating the answer. See Appendix E for details on implementations and prompts.

Human Agent. Finally, we also run a study with human participants to establish human-level performance metrics on our benchmark. We collect answers from a set of human annotators by providing each annotator with a video of the episode history H and asking them to answer all of the questions Q for that scene. We enrolled two independent participants for this benchmarking exercise and found strong agreement in responses.

Agents for A-EQA. So far, we have described agents that can answer questions Q given an episode history H. However, in the case of A-EQA, no explicit H is provided, and agent must generate its own observations through exploration. In this work, we provide the simplest baseline for A-EQA that explores environments in a **task or ques-**


Q: What is hanging on the wall perpendicular to the windows?
Large Language Model
A: A chalkboard with various writings and drawings, a bulletin board, and a cork board are hanging on the wall perpendicular to the windows
z_1
z_2
z_k
Image Captioner
Image Captioner
Image Captioner
X_1
X_2
X_k
Q: What is hanging on the wall perpendicular to the windows?
Large Language Model
A: A bulletin board with various papers and drawings
obj_1 loc_1
obj_2 loc_2
obj_m loc_m
Image Captioner
Image Captioner
Image Captioner
Scene-Graph
Q: What is hanging on the wall perpendicular to the windows?
Multi-Frame VLMs
A: A chalkboard, bulletin board, and a cork board
X_1
X_2
X_k

Socratic LLMs w/ Frame Captions

Socratic LLMs w/ Scene-Graph Captions

Multi-frame VLMs

Figure 4. **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.

tion agnostic manner. Specifically, we use frontier exploration [46] and use the agent's observations as the episodic memory H to answer questions about the scene. This effectively allows us to re-use all the aforementioned agents, just with a different and self-generated episodic memory. We note that the efficiency score of such an agent is expected to be poor, and we leave open the challenge of more efficient A-EQA agents to future work.

Force-A-Guess when Agents Abstain. To recall, all EQA agents we study involve an LLM component and differ primarily in how the episode history is used. In our experiments, we observed that such agents can often be overly conservative and **abstain** from answering, especially when a model thinks it lacks sufficient context. In our evaluation metric, we do not make a special provision for abstaining, and consider abstaining an incorrect answer. Thus, we force the agent to take a guess to give it at least an informed random chance, instead of immediately counting it as a failure.

Despite our best efforts, we were not able to force non-blind agents to guess through prompt engineering. However, blind LLMs are able to guess purely based on prior knowledge, and seldom abstain. Thus, for non-blind agents we first check if the agent abstained. If it did, we use the answer generated by the corresponding blind LLM. Full details of this protocol are in Appendix G and an analysis of the effects of this procedure are in Appendix H. All results in the main paper use this force-a-guess protocol.

Table 2. **LLM-Match and efficiency scores on OpenEQA.** EM-EQA results are broken down by data source (ScanNet, HM3D, and ALL). A-EQA results include both LLM-Match scores (Eq. 1) and agent efficiency (Eq. 2). *GPT-4V scores are calculated on a subset of 500 OpenEQA questions due to API limitations.

	EM-EQA			A-EQA	
# method	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	-

4. Experimental Results on OpenEQA

We present evaluation results of agents described in Sec. 3. Table 2 reports the overall LLM-Match scores (C) (see Eq. 1) of the baselines evaluated on the EM-EQA and A-EQA benchmarks, where EM-EQA results are separately reported on each of the data sources (i.e., ScanNet and HM3D). It also presents the efficiency score on A-EQA, as described in Eq. 2, along with bootstrapped standard errors. Based on the results, we first share some observations and remarks, and discuss specific observations in Sec. 5.

1. Humans achieve excellent performance on the benchmark (>85%), which confirms the validity of the benchmark and correctness of evaluation metrics.
2. Multi-frame VLMs (i.e., GPT-4V) outperform other

agents. This suggests that a tight integration of perception and language may significantly benefit EQA.

3. We find that blind LLMs are surprisingly strong baselines, with GPT-4 and LLaMA-2 achieving a score of 33.5 and 28.3 respectively on EM-EQA. While this is substantially lower than GPT-4V or human-level performance, it suggests a large degree of regularity in the world and that answers to several questions can be "guessed" without explicit visual context of a specific environment. We note that early works in VQA [1] also found blind agents to be strong baselines.
4. Within each family of agents we consistently find that agents that use GPT-4 as the LLM outperform LLaMA-2. This suggests that larger LLMs can be a key enabling factor for good EQA performance.
5. In the EM-EQA benchmark, we find that all agents with access to perceptual information in the form of frame captions or scene-graphs outperform blind LLMs (under the force-a-guess protocol). This again underscores the importance of perception for EQA.
6. When comparing the performance of agents in EM-EQA and A-EQA, we generally observe lower scores in A-EQA. In part, this is due to longer trajectories due to the use of exhaustive exploration in our A-EQA agents, which forces a longer history representation often with irrelevant information for a specific question. In several situations, this makes the performance of various agents comparable to that of blind LLMs or even lower (e.g. GPT-4 w/ ConceptGraphs). This underscores the challenging nature of the A-EQA benchmark and the importance of efficient exploration in interactive settings.

Figure 5 breaks down performance on EM-EQA (human-like trajectories) by the seven question categories described in Sec. 2.3. Among all the categories, functional reasoning questions are the easiest for EQA agents to answer, reaching an average LLM-Match score of 45.6. Additionally, EQA agents also feel comfortable when answering object state recognition and world knowledge types of questions. These categories require the agent to have commonsense understanding of the world, which is what the current large models are good at. 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 understanding of space and geometry into large models. While most models achieve nontrivial performance on all categories, there remains a large gap between even the best method and human-level performance.

5. Analysis and Discussions

Human Alignment and Robustness of LLM-Match. Evaluating open-vocabulary answers is an open challenge in AI. While human evaluation remains the gold-standard, it is also expensive and time consuming. An automatic evaluation metric is preferable for benchmarking, fast iteration, and model selection. For this, we proposed the LLM-Match metric in Sec. 2.4. We now test 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?

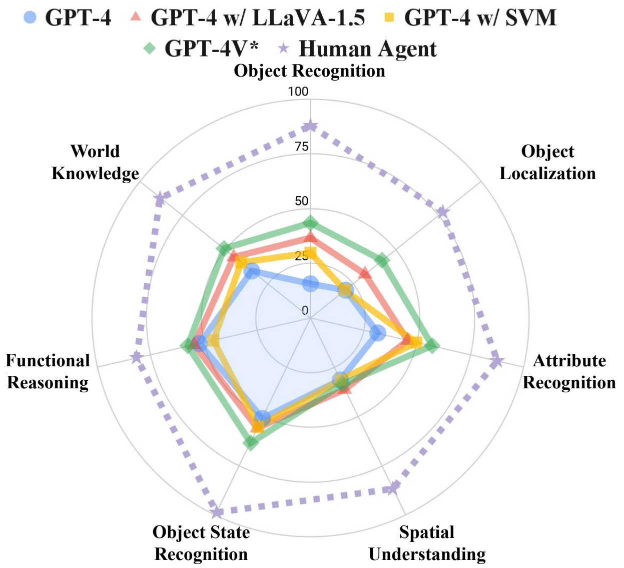


Figure 5. **Category-level performance on EM-EQA.** We find that agents with access to visual information excel at localizing and recognizing objects and attributes, and make better use of this information to answer questions that require world knowledge. However, on other categories performance is closer to the blind LLM baseline (GPT-4), indicating substantial room for improvement on OpenEQA. See scores for all methods in Appendix I.

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 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. 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 of the prompt (see Appendix M). Table 8 in Appendix M shows that changing the LLM's role from 'AI' to 'Score Master' or 'professional evaluator' does not significantly change results, the scores have a tight correlation with a Spearman's $\rho > 0.95$. Similarly, Table 9 in Appendix M 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. When studying Socractic LLMs augmented with perceptual information (image or scene-graphs captions), we found that agents often abstained from answering the question (e.g. *'Not enough information to answer the question.'*). As noted in Sec. 3, our LLM-Match metric does not give preferential scoring of abstaining vis-a-vis an incorrect answer. Thus, we defaulted to the answer from the blind LLM powering an agent when it abstained. In Appendix H, we provide statistics on how frequently each agent abstained, and study performance without defaulting to a blind LLM. In general, we find that GPT-4-based Socratic agents abstain frequently (up to 55% of the time), and thus, rely more 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% of the time), and thus the differences between the two variants is minimal.

6. Related Work

The intersection of perception and language [3, 9, 12, 16, 20, 23, 25, 30, 54] 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 [1], VQA-v2 [14], OK-VQA [33] and A-OKVQA [37], that focus on answering questions from a single image. Later works extended question answering tasks to videos [22, 51, 53] and 3D scenes [3, 5, 16, 31]. These include benchmark such as VideoQA [53], SQA3D [31] and ScanRefer [5]. 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 [5, 17, 22, 23, 31, 42, 44, 51]. Another closely related line of work is prior benchmarks on Embodied QA [7, 8, 41, 50] and is conceptually similar to our A-EQA setting. They focus on leveraging RGB-D to accomplish navigation tasks in simulation [41], in which the agent must seek out multiple target locations or objects sampled from a closed vocabulary set [50] Our work takes inspiration from such prior works [7] 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 [23] or during evaluation time [29, 49]. 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 [29]. 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.

7. Conclusion

We introduce OpenEQA, the first realistic benchmark to study open-vocabulary EQA in both episodic memory and active settings. Specifically, 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-graph construction, and multi-frame VLMs such as GPT-4V. 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] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering, 2016. 2, 7, 8

[2] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018. 5

[3] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding, 2022. 2, 8

[4] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 2

[5] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. *16th European Conference on Computer Vision (ECCV)*, 2020. 2, 8

[6] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 2, 3, 4

[7] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 8

[8] Abhishek Das, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Neural Modular Control for Embodied Question Answering. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2018. 8

[9] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 8

[10] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, pages 226–231, 1996. 3

[11] Chuang Gan, Jeremy Schwartz, Seth Alter, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, Megumi Sano, Kuno Kim, Elias Wang, Damian Mrowca, Michael Lingelbach, Aidan Curtis, Kevin Feigelis, Daniel M. Bear, Dan Gutfreund, David Cox, James J. DiCarlo, Josh McDermott, Joshua B. Tenenbaum, and Daniel L. K. Yamins. Threedworld: A platform for interactive multi-modal physical simulation, 2020. 2

[12] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 8

[13] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4089–4098, 2018. 2

[14] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017. 8

[15] Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. *arXiv*, 2023. 2, 5

[16] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models, 2023. 8

[17] Zhijian Hou, Lei Ji, Difei Gao, Wanjun Zhong, Kun Yan, Chao Li, Wing-Kwong Chan, Chong-Wah Ngo, Nan Duan, and Mike Zheng Shou. Groundnlq@ ego4d natural language queries challenge 2023. *arXiv preprint arXiv:2306.15255*, 2023. 8

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 5

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2

[20] Dohwan Ko, Ji Soo Lee, Wooyoung Kang, Byungseok Roh, and Hyunwoo J Kim. Large language models are temporal and causal reasoners for video question answering. *arXiv preprint arXiv:2310.15747*, 2023. 8

[21] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv*, 2017. 2

[22] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 2, 8

[23] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023. 2, 8

[24] Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Elliott Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, Andrey Kurenkov, Karen Liu, Hyowon Gweon, Jiajun Wu, Li Fei-Fei, and Silvio Savarese. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. In *5th Annual Conference on Robot Learning*, 2021. 2

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 8

[26] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 2, 5, 4

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2, 5

[28] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 5

[29] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2023. 2, 4, 8

[30] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 8

[31] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations*, 2023. 2, 8

[32] Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2, 3

[33] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge, 2019. 2, 8

[34] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. 2, 5

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2, 3

[36] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv 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. 2, 3, 4

[37] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge, 2022. 2, 8

[38] Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. Robovqa: Multimodal long-horizon reasoning for robotics. *arXiv preprint arXiv:2311.00899*, 2023. 2

[39] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 3

[40] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2, 5

[41] Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. Embodied Question Answering in Photorealistic Environments with Point Cloud Perception. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 8

[42] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang 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. 8

[43] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: real-world perception for embodied agents. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018. 2

[44] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng 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*, pages 9777–9786, 2021. 8

[45] Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, et al. Habitat-matterport 3d semantics dataset. *arXiv preprint arXiv:2210.05633*, 2022. 2, 3, 4

[46] Brian Yamauchi. A frontier-based approach for autonomous exploration. In *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97.'Towards New Computational Principles for Robotics and Automation'*, pages 146–151. IEEE, 1997. 6

[47] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9, 2023. 2, 5

[48] Sriram Yenamandra, Arun Ramachandran, Karmesh Yadav, Austin Wang, Mukul Khanna, Theophile Gervet, Tsung-Yen Yang, Vidhi Jain, Alexander William Clegg, John Turner, Zsolt Kira, Manolis Savva, Angel Chang, Devendra Singh Chaplot, Dhruv Batra, Roozbeh Mottaghi, Yonatan Bisk, and Chris Paxton. Homerobot: Open vocabulary mobile manipulation. *arXiv preprint arXiv:2306.11565*, 2023. 3

[49] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang,

Zhiyong Wang, Jing Shao, and Wanli Ouyang. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark, 2023. 8

[50] Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L. Berg, and Dhruv Batra. Multi-target embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 8

[51] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134, 2019. 8

[52] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023. 5

[53] Yaoyao Zhong, Junbin Xiao, Wei Ji, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answering: Datasets, algorithms and challenges. *arXiv preprint arXiv:2203.01225*, 2022. 8

[54] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 8