SQA3D: SITUATED QUESTION ANSWERING IN 3D SCENES

Anonymous authors
Paper under double-blind review

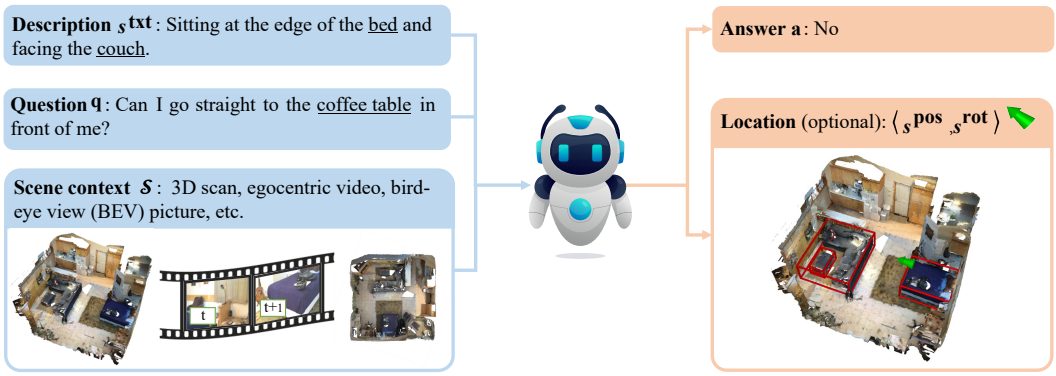


Figure 1: Task illustration of Situated Question Answering in 3D Scenes (SQA3D). Given scene context $\mathcal{S}$ (*e.g.*, 3D scan, egocentric video, bird-eye view picture), SQA3D requires an agent to first comprehend and localize its **situation** (position, orientation, *etc.*) in the 3D scene from a textual description s^{txt}, then answer a question q under that situation. **Note that understanding the situation and imagining the corresponding egocentric view correctly is necessary to accomplish our task.** We provide more example questions in Figure 2.

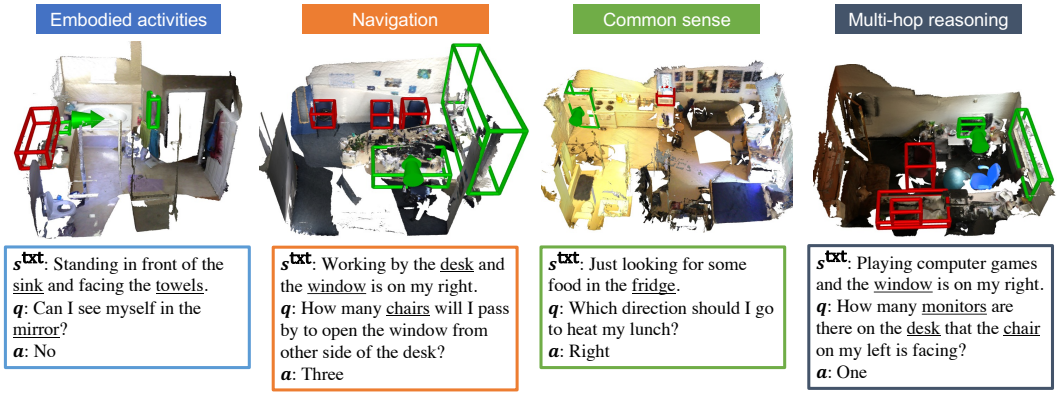
ABSTRACT

We propose a new task to benchmark scene understanding of embodied agents: Situated Question Answering in 3D Scenes (SQA3D). Given a scene context (*e.g.*, 3D scan), SQA3D requires the tested agent to first understand its **situation** (position, orientation, etc.) in the 3D scene as described by text, then reason about its surrounding environment and answer a question under that situation. Based upon 650 scenes from ScanNet, we provide a dataset centered around 6.8k unique situations, along with 20.4k descriptions and 33.4k diverse reasoning questions for these situations. These questions examine a wide spectrum of reasoning capabilities for an intelligent agent, ranging from spatial relation comprehension to commonsense understanding, navigation, and multi-hop reasoning. SQA3D imposes a significant challenge to current multi-modal especially 3D reasoning models. We evaluate various state-of-the-art approaches and find that the best one only achieves an overall score of 47.20%, while amateur human participants can reach 90.06%. We believe SQA3D could facilitate future embodied AI research with stronger situation understanding and reasoning capability.

1 INTRODUCTION

In recent years, the endeavor of building intelligent embodied agents has delivered fruitful achievements. Robots now can navigate (Anderson et al., 2018) and manipulate objects (Liang et al., 2019; Savva et al., 2019; Shridhar et al., 2022; Ahn et al., 2022) following natural language commands or dialogues. Albeit these promising advances, their actual performances in real-world embodied environments could still fall short of human expectations, especially on generalization to different situations (scenes and locations) and tasks that require substantial, knowledge-intensive reasoning. To diagnose the fundamental capability of realistic embodied agents, we investigate the problem of **embodied scene understanding**, where the agent needs to understand its situation and the surround-

Figure 2: **Examples from SQA3D.** We provide some example questions and the corresponding situations (s^{txt} and) and 3D scenes. The categories listed here do not mean to be exhaustive and a question could fall into multiple categories. The green boxes indicate relevant objects in situation description s^{txt} while red boxes are for the questions q.



ings in the environment from a *dynamic* egocentric view, then perceive, reason, and act accordingly, to accomplish complex tasks.

What is at the core of embodied scene understanding? Drawing inspirations from situated cognition (Greeno, 1998; Anderson et al., 2000), a seminal theory of embodiment, we anticipate it to be two-fold:

- **Situation understanding.** The ability to imagine what the agent will see from arbitrary situations (position, orientations, *etc*.) in a 3D scene and understand the surroundings anchored to the situation, therefore generalize to novel positions or scenes;
- **Situated reasoning.** The ability to acquire knowledge about the environment based on the agents' current situation and reason with the knowledge, therefore further facilitates accomplishing complex action planning tasks.

To step towards embodied scene understanding, we introduce **SQA3D**, a new task that reconciles the best of both parties, situation understanding, and situated reasoning, into embodied 3D scene understanding. Figure 1 sketches our task: given a 3D scene context (*e.g*., 3D scan, ego-centric video, or bird-eye view (BEV) picture), the agent in the 3D scene needs to first comprehend and localize its situation (position, orientation, *etc*.) from a textual description, then answer a question that requires substantial situated reasoning from that perspective. We crowd-sourced the situation descriptions from Amazon MTurk (AMT), where participants are instructed to select diverse locations and orientations in 3D scenes. To systematically examine the agent's ability in situated reasoning, we collect questions that cover a wide spectrum of knowledge, ranging from spatial relations to navigation, common sense reasoning, and multi-hop reasoning. In total, SQA3D comprises 20.4k descriptions of 6.8k unique situations collected from 650 ScanNet scenes and 33.4k questions about these situations. Examples of SQA3D can be found Figure 2.

Our task closely connects to the recent efforts on 3D language grounding (Dai et al., 2017; Chen et al., 2020; 2021; Achlioptas et al., 2020; Azuma et al., 2022). However, most of these avenues assume observations of a 3D scene are made from some third-person perspectives rather than an embodied, egocentric view, and they primarily inspect *spatial understanding*, while SQA3D examines scene understanding with a wide range of knowledge, and the problems have to be solved using an (imagined) first-person view. Embodied QA (Das et al., 2018; Wijmans et al., 2019a) draws very similar motivation as SQA3D, but our task adopts a simplified protocol (QA only) while still preserving the function of benchmarking embodied scene understanding, therefore allowing more complex, knowledge-intensive questions and a much larger scale of collection. Comparisons with relevant tasks and benchmarks are listed in Table 1.

Benchmarking existing baselines: In our experiments, we examine state-of-the-art multi-modal reasoning models, including ScanQA from Azuma et al. (2022) that leverages 3D scan data, Clip-BERT (Lei et al., 2021) and MCAN (Yu et al., 2019) that exploits egocentric videos and BEV pictures. However, the results unveil that both models still largely fall behind human performances by a large margin (47.2% of the best model vs. 90.06% of amateur human testers). To understand

Table 1: **An overview of the different benchmark datasets covering grounded 3D scene understanding**. In general, we consider semantic grounding, language-driven navigation, and question answering in photo-realistic 3D scenes. In the first row, *situated* indicates whether the benchmark task is supposed to be completed by a "situated" agent with its egocentric perspective. *navigation*, *common sense* and *multi-hop reasoning* show whether the task requires certain capability or knowledge level of 3D understanding. *Rather than observing a complete 3D scan of the scene, the learner needs to navigate in a simulator to perceive the 3D scene incrementally.

dataset	task	situated?	3D type	text collection	navi-gation?	common sense?	multi-hop reasoning?	#scenes	#tasks
ScanNet (Dai et al., 2017)	seg.	✗	scan	n/a	✗	✗	✗	800 rooms	1.5k
ScanRefer (Chen et al., 2020)	det.	✗	scan	human	✗	✗	✗	800 rooms	52k
ReferIt3D (Achlioptas et al., 2020)	det.	✗	scan	human	✗	✗	✗	707 rooms	41k
ScanQA (Azuma et al., 2022)	q.a.	✗	scan	template	✗	✗	✗	800 rooms	41k
3D-QA (Ye et al., 2021)	q.a.	✗	scan	human	✗	✗	✗	806 rooms	5.8k
CLEVR3D (Yan et al., 2021)	q.a.	✗	scan	template	✗	✗	✓	478 rooms	60k
MP3D-R2R (Anderson et al., 2018)	nav.	✓	*nav.	human	✓	✗	✗	190 floors	22k
MP3D-EQA (Wijmans et al., 2019a)	q.a.	✓	*nav.	template	✓	✗	✗	146 floors	1.1k
SQA3D (Ours)	q.a.	✓	scan	human	✓	✓	✓	650 rooms	33.4k

the failure modes, we conduct experiments on settings that could alleviate the challenges brought by situation understanding. The improvement of these models confirms that the current models are indeed struggling with situation understanding, which is pivotal for embodied scene understanding. Finally, we explore whether powerful Large Language Models (LLMs) like GPT-3 (Brown et al., 2020) and Unified QA (Khashabi et al., 2020) could tackle our tasks by converting the multi-modal SQA3D problems into single-modal surrogates using scene captioning. However, our results read that these models can still be bottlenecked by the lack of spatial understanding and accurate captions.

Our contributions can be summarized as follow:

- We introduce SQA3D, a new benchmark for embodied scene understanding, aiming at reconciling the challenging capabilities of situation understanding and situated reasoning and facilitating the development of intelligent embodied agents.
- We meticulously curate the SQA3D to include diverse situations and interesting questions. These questions probe a wide spectrum of knowledge and reasoning abilities of embodied agents, ranging from spatial relation comprehension to navigation, common sense reasoning, and multi-hop reasoning.
- We perform extensive analysis on the state-of-the-art multi-modal reasoning models. However, experimental results indicate that these avenues are still struggling on SQA3D. Our hypothesis suggests the crucial role of proper 3D representations and the demand for better situation understanding in embodied scene understanding.

2 THE SQA3D DATASET

A problem instance in SQA3D can be formulated as a triplet $\langle \mathcal{S}, s, q \rangle$, where $\mathcal{S}$ denotes the scene context, *e.g.*, 3D scan, egocentric video, bird-eye view (BEV) picture, *etc.*; $s = \langle s^{\text{txt}}, s^{\text{pos}}, s^{\text{rot}} \rangle$ denotes a situation, where the textual situation description s^{txt} (*e.g.*, "*Sitting at the edge of the bed and facing the couch*" in Figure 1) depicts the position s^{pos} and orientation s^{rot} of an agent in the scene; q denotes a question. The task is to retrieve the correct answer from the answer set $a = \{a_1, \ldots, a_N\}$, while optionally predict the ground truth location $\langle s^{\text{pos}}, s^{\text{rot}} \rangle$ from the text. The additional prediction of location could help alleviate the challenges brought by situation understanding. The following subsections will detail how to collect and curate the data and then build the benchmark.

2.1 DATA FORMATION

The 3D indoor scenes are selected from the ScanNet (Dai et al., 2017) dataset. We notice that some scenes could be too crowded/sparse, or overall tiny, making situations and questions collection infeasible. Therefore, we first manually categorize these scenes based on the richness of objects/layouts and the space volume. We end up retaining 650 scenes after dropping those that failed to meet the requirement. We then develop an interactive web-based user interface (UI) to collect the data. Details of UI design can be found in *appendix*. All the participants are recruited on AMT.

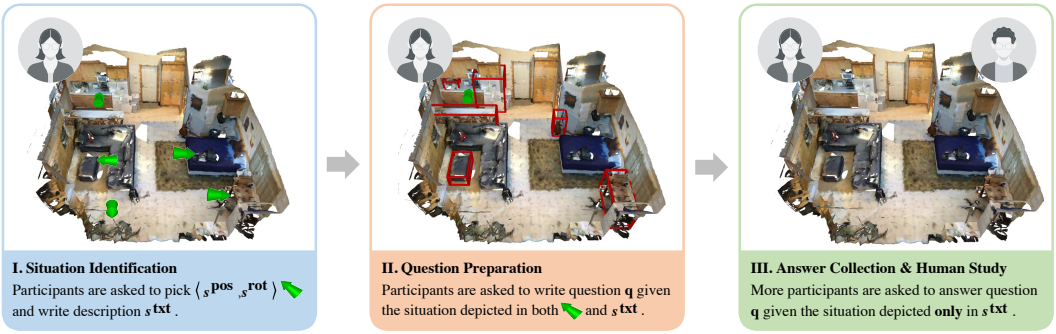


Figure 3: **Data collection pipeline of SQA3D.** Since our dataset comprises multiple types of annotations (situations and their descriptions, questions, answers, *etc.*), we found it more manageable to break down a single annotation task into three sub-tasks: i) Situation Identification; ii) Question Preparation; iii) Answer Collection & Human Study, where the participants recruited on AMT only need to focus on a relatively simple sub-task at a time.

Multi-staged collection. Compared to counterparts, the annotation load of a single SQA3D problem instance could be significantly heavier as participants need to explore the scene, pick a situation, make descriptions, and ask a few questions. All these steps also require dense interaction with the 3D scene. To ensure good quality, we introduce a multi-staged collection pipeline, which breaks down the load into more manageable sub-tasks. Figure 3 delineates this process. **I. Situation Identification.** We ask the workers to pick 5 situations by changing the location $\langle s^{\text{pos}}, s^{\text{rot}} \rangle$ of a virtual avatar in a ScanNet scene $\mathcal{S}$. The workers are then instructed to write descriptions s^{txt} that can uniquely depict these situations in the scene. We also use examples and bonuses to encourage more natural sentences and the use of human activities (*e.g.*, "*I'm waiting for my lunch to be heated in front of the microwave*"). All the collected situations are later manually curated to ensure diversity and the least ambiguity. If necessary, we would augment the data with more situations to cover different areas of the scene. **II. Question Preparation.** We collect a set of questions w.r.t. each pair of the 3D scene $\mathcal{S}$, and the situation description s^{txt} (the virtual avatar is also rendered at $\langle s^{\text{pos}}, s^{\text{rot}} \rangle$). To help prepare questions that require substantial situated reasoning, we tutor the workers before granting them access to our tasks. They are instructed to follow the rules and learn from good examples. We also remove & penalize the responses that do not depend on the current situation, *e.g.* "*How many chairs are there in the room?*" **III. Answer Collection & Human Study.** In addition to the answers collected alongside the questions, we send out the questions to more workers and record their responses. These workers are provided with the same interface as in stage **II** except showing in the scene to ensure consistency between question and answer collection. There is also mandatory scene familiarization in all three steps before the main job starts and we find it extremely helpful especially for more crowded scenes. More details can be found in *appendix*.

2.2 Curation, Data Statistics, and Metrics

Curation. Our multi-staged collection ends up with around 21k descriptions of 6.8k unique situations and 35k questions. Although the aforementioned prompt did yield many high-quality annotations, some of them are still subject to curation. We first apply a basic grammar check to clean up the language glitches. Then we follow the practices in VQAv2 (Goyal et al., 2017) and OK-VQA (Marino et al., 2019) to further eliminate low-effort descriptions and questions. Specifically, we eliminate & rewrite template-alike descriptions (*e.g.*, repeating the same sentence patterns) and questions that are too simple or do not require looking at the scene. We also notice the similar answer bias reported in Marino et al. (2019) where some types of questions might bias toward certain answers. Therefore, we remove questions to ensure a more uniform answer distribution. A comparison of answer distribution before and after the balancing can be found in *appendix*. As a result, our final dataset comprises 20.4k descriptions and 33.4k diverse and challenging questions. Figure 2 demonstrates some example questions in SQA3D.

Statistics. Compared to most counterparts with template-based text generation, SQA3D is crowd-sourced on AMT and therefore enjoys more naturalness and better diversity. To the best of our knowledge, SQA3D is the **largest** dataset of grounded 3D scene understanding with the human-annotated question-answering pairs (a comparison to the counterparts can be found in Table 1).

Figure 4: Word cloud of s^{txt} in SQA3D.

Statistic	Value
Total s^{txt} (train/val/test)	16,229/1,997/2,143
Total q (train/val/test)	26,623/3,261/3,519
Unique q (train/val/test)	20,183/2,872/3,036
Total scenes (train/val/test)	518/65/67
Total objects (train/val/test)	11,723/1,550/1,652
Average s^{txt} length	17.49
Average q length	10.49

Table 2: SQA3D dataset statistics.

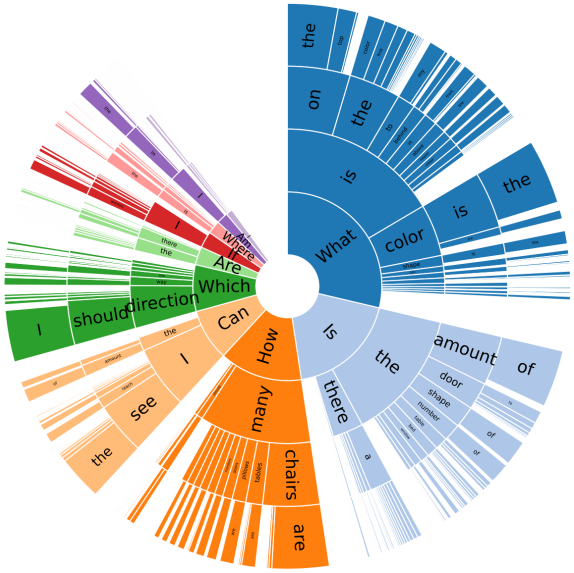


Figure 5: Question distribution in SQA3D

Table 2, Figure 4, and Figure 5 illustrate the basic statistics of our dataset, including the word cloud of situation descriptions and question distribution based on their prefixes. It can be seen that descriptions overall meet our expectations as human activities like "sitting" and "facing" are among the most common words. Our questions are also more diverse and balanced than our counterparts, where those starting with "What" make up more than half of the questions and result in biased questions (Azuma et al., 2022). More statistics like distributions over answer and length of the text can be found in *appendix*.

Dataset splits and evaluation metric. We follow the practice of ScanNet and split SQA3D into *train*, *val* and *test* sets. Since we cannot access the semantic annotations in ScanNet *test* set, we instead divide the ScanNet validation scenes into two subsets, and use them as our *val* and *test* sets, respectively. The statistics of these splits can be found in Table 2. Following the protocol in VQAv2 (Goyal et al., 2017), we provide a set of 706 "top-K" answer candidates by excluding answers that only appear very few times. Subsequently, we adopt the "exact match" as our evaluation metric, *i.e.*, the accuracy of answer classification in the *test* set. No further metric is not included as we find it sufficient enough to measure the differences between baseline models with "exact match".

3 Models for SQA3D

Generally speaking, SQA3D can be characterized as a multi-modal reasoning problem. Inspired by the recent advances in transformer-based (Vaswani et al., 2017) vision-language models (Lu et al., 2019; Li et al., 2020; Alayrac et al., 2022), we investigate how could these methods approach our task. Specifically, we study a recent transformer-based question answering system: ScanQA (Azuma et al., 2022), which maps 3D scans and questions into answers. We make a few adaptations to ensure its compatibility with the protocol in SQA3D. To further improve this model, we consider including some auxiliary tasks during training (Ma et al., 2022). For other types of 3D scene context, *e.g.* egocentric video clips and BEV pictures, we employ the corresponding state-of-the-art models. Finally, we explore the potential of recently-introduced LLMs like GPT-3 (Brown et al., 2020) and Unified QA (Khashabi et al., 2020) on solving SQA3D in a zero-shot fashion. An overview of these models can be found in Figure 6.

3D model. We use the term *3D model* to refer a modified version of the ScanQA model (Azuma et al., 2022), depicted in the blue box of Figure 6. It includes a VoteNet (Qi et al., 2019)-based 3D perception module that extracts object-centric features, LSTM-based language encoders for processing both questions q and situation description s^{txt}, and some cross-attention transformer blocks (Vaswani et al., 2017). The object-centric feature tokens attend to the language tokens of s^{txt} and q successively. Finally, these features will be fused and mapped to predict the answer. Optionally, we can add one head to predict the location $\langle s^{\text{pos}}, s^{\text{rot}} \rangle$ of the agent. Since the VoteNet

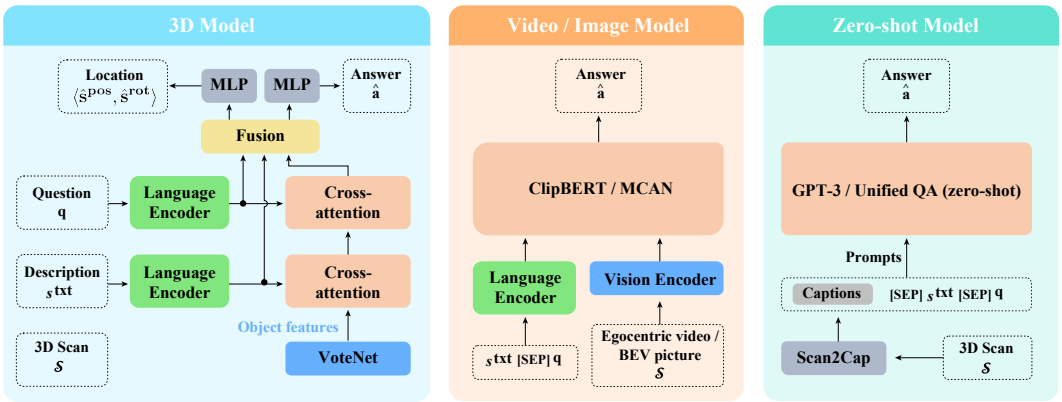


Figure 6: **Potential models for SQA3D.** We split the considered models into three groups: 3D model, video / image model, and zero-shot model. The 3D model is modified from the ScanQA model (Azuma et al., 2022) and maps 3D scan input to the answer. While the video / image models are effectively borrowed from canonical video QA and VQA tasks but we augment them with the additional situation input. The zero-shot model explores the potential of large pre-trained LLMs on our tasks. But they have to work with an additional 3D caption model that converts the 3D scene into text.

module is trained from scratch, we also employ an object detection objective (not shown in the figure).

Auxiliary task. As we mentioned before, situation understanding plays a crucial role in accomplishing SQA3D tasks. To encourage a better understanding of the specified situation, we introduce two auxiliary tasks: the model is required to make predictions about the s^{pos} and s^{rot} of the situation. We use mean-square-error (MSE) loss for these tasks. The overall loss for our problem therefore becomes $\mathcal{L} = \mathcal{L}_{\text{ans}} + \alpha\mathcal{L}_{\text{pos}} + \beta\mathcal{L}_{\text{rot}}$, where $\mathcal{L}_{\text{ans}}$, $\mathcal{L}_{\text{pos}}$, and $\mathcal{L}_{\text{rot}}$ depicts the losses of the main and auxiliary tasks, α and β are balancing weights.

Video and Image-based model. The orange box in the middle of Figure 6 demonstrates the models for video and image-based input. SQA3D largely resembles a video question answering or visual question answering problem when choosing to represent the 3D scene context $\mathcal{S}$ as egocentric video clips or BEV pictures. However, SQA3D also requires the model to take both question q and the newly added situation description s^{txt} as input. We therefore follow the practice in the task of context-based QA (Rajpurkar et al., 2018) and prepend s^{txt} to the question as a *context*. For the model, we use the state-of-the-art video QA system ClipBERT (Lei et al., 2021) and VQA system MCAN (Yu et al., 2019). We adopt most of their default hyper-parameters and the details can be found in *appendix*.

Zero-shot model. We explore to which extent the powerful LLMs like GPT-3 (Brown et al., 2020) and Unified QA (Khashabi et al., 2020) could tackle our tasks. Following prior practices that apply GPT-3 to VQA (Changpinyo et al., 2022; Gao et al., 2022), we propose to convert the 3D scene into text using an emerging technique called 3D captioning (Chen et al., 2021). We provide the caption, s^{txt}, and q as part of the prompt and ask these models to complete the answer. For GPT-3, we further found providing few-shot examples in the prompt helpful with much better results. Minor post-processing is also needed to ensure answer quality. We provide more details on prompt engineering in the *appendix*.

4 Experiments

4.1 Setup

We benchmark the models introduced in Section 3 to evaluate their performances on SQA3D. As mentioned before, we examine three types of scene context $\mathcal{S}$: 3D scan (point cloud), egocentric video, and BEV picture. Both the 3D scan and egocentric video for each scene are provided by ScanNet (Dai et al., 2017). However, we down-sample the video to allow more efficient computation per the requirement of the ClipBERT model (Lei et al., 2021). The BEV pictures are rendered by placing a top-down camera on top of the scan of each 3D scene. We also conduct additional experiments that investigate factors that could contribute to the results, *e.g.*, situation and auxil-

	$\mathcal{S}$	Format	test set						Avg.
			What	Is	How	Can	Which	Others	
Blind test	-	SQ→A	26.75	63.34	43.44	69.53	37.89	43.41	43.65
ScanQA (w/o s^{txt})	3D scan	VQ→A	28.58	65.03	47.31	66.27	43.87	42.88	45.27
ScanQA	3D scan	VSQ→A	31.64	63.80	46.02	69.53	43.87	45.34	46.58
ScanQA + aux. task	3D scan	VSQ→AL	33.48	66.10	42.37	69.53	43.02	46.40	47.20
MCAN	BEV	VSQ→A	28.86	59.66	44.09	68.34	40.74	40.46	43.42
ClipBERT	Ego. video	VSQ→A	30.24	60.12	38.71	63.31	42.45	42.71	43.31
Unified QA_{Large}	ScanRefer	VSQ→A	33.01	50.43	31.91	56.51	45.17	41.11	41.00
Unified QA_{Large}	ReferIt3D	VSQ→A	27.58	47.99	34.05	59.47	40.91	39.77	38.71
GPT-3 (10% data)	ScanRefer	VSQ→A	31.36	46.97	44.64	51.28	40.62	41.67	40.67
GPT-3 (10% data)	ReferIt3D	VSQ→A	28.00	56.06	27.78	36.11	20.51	37.50	34.82
Human (amateur)	3D scan	VSQ→A	88.53	93.84	88.44	95.27	87.22	88.57	90.06

Table 3: **Quantitative results on the SQA3D benchmark**. Results are presented in accuracy (%) on different types of questions. In the "Format" column: V = 3D visual input $\mathcal{S}$; S = situation description s^{txt}; Q = question q; A = answer a; L = location $\langle s^{\text{pos}}, s^{\text{rot}} \rangle$. In ScanQA, *aux. task* indicates the use of both $\mathcal{L}_{\text{pos}}$ and $\mathcal{L}_{\text{rot}}$ as additional losses. We use the *Large* variant as Unified QA (Khashabi et al., 2020) as it works better. Note that we only evaluate GPT-3 on 10% of our testing data due to the limited API access resources.

iary tasks. In our early experiments, we found that the 3D model overall performs better than the video or image-based models. Therefore we only conduct these additional experiments with the variants of our 3D model due to the limit of computational resources. We use the official implementation of ScanQA, ClipBERT, and MCAN and include our modifications for SQA3D. For the zero-shot models, we extract 3D scene captions from two sources: ScanRefer (Chen et al., 2020) and ReferIt3D (Achlioptas et al., 2020). Considering the limit on the length of the input prompt, these 3D captions are also down-sampled. The Unified QA model weights are obtained from its Huggingface official repo. All the models are tuned using the validation set and we only report results on the test set. More details on model implementation can be found in *appendix*.

4.2 Quantitative Results

We provide the quantitative results of the considered models (detailed in Section 3) on our SQA3D benchmark in Table 3. The findings are summarized below:

Question types. In Table 3, we demonstrates accuracy on six types of questions based on their prefixes. Most models tend to perform better on the "Is" and "Can" questions while delivering worse results on "What" questions, likely due to a smaller number of answer candidates – most questions with binary answers start with "Is" and "Can", offering a better chance for the random guess. Moreover, we observe the hugest gap between the blind test (model w/o 3D scene context input) and our best model on the "What" and "Which" categories, suggesting the need for more visual information for these two types of questions. This also partially echoes the finding reported in Lei et al. (2018).

Situation understanding and reasoning. At the heart of SQA3D benchmark is the requirement of situation understanding and reasoning. As we mentioned in Section 2.1, the model will be more vulnerable to wrong answer predictions if ignoring the situation that the question depends on (*e.g.* "*What is in front of me*" could have completely different answers under different situations). In Table 3, removing situation description s^{txt} from the input leads to worse results, while adding the auxiliary situation prediction tasks boosts the overall performance, especially on the challenging "What" questions. The only exception is "How" questions, where a majority of them are about counting. Our hypothesis is that most objects in each ScanNet scene only have a relatively small number of instances, and the number could also correlate to the object category. Therefore, guessing/memorization based on the question only could offer better results than models with the situation as input if the situation understanding & reasoning are still not perfect yet. Additionally, we also provide an inspection of the relation between situation understanding and QA using attention visualization in Section 4.3.

Representations of 3D scenes. Indeed, SQA3D does not limit the input to be 3D scan only, as we also offer options of egocentric videos and BEV pictures. Compared to models with the 3D scan as

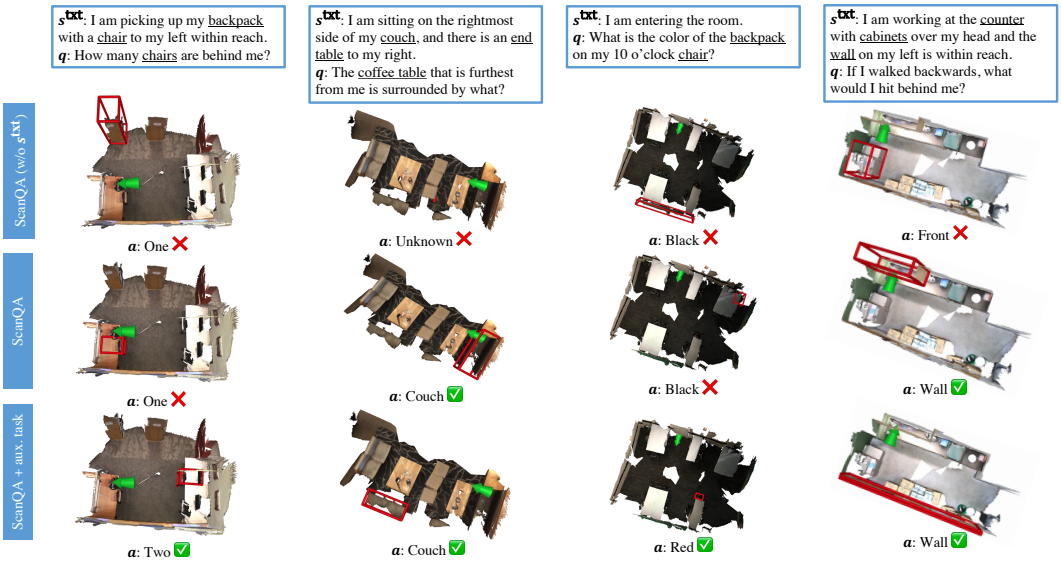


Figure 7: **Qualitative results.** We show the predicted answer and bbox with highest attention for the variants of ScanQA (Azuma et al., 2022) models. We anticipate the bbox to indicate the object that situation description s^{txt} or question q refers to. We observe that better situation understanding (via comprehension on s^{txt} or auxiliary tasks) could result in more reasonable attention over objects, which positively correlates to more robust answer prediction.

input, the tested models with other 3D representations (*i.e.*, MCAN and ClipBERT) deliver much worse results, implying that the 3D scan so far could still be a better representation for the 3D scene when the reasoning models are probed with questions that require a holistic understanding of the scene. On the other hand, MCAN and ClipBERT are general-purpose QA systems, while ScanQA is designed for 3D-language reasoning tasks. The generalist-specialty trade-off could also partially account for the gap.

Zero-shot vs. training from scratch. The success of pre-trained LLMs like GPT-3 on myriads of challenging reasoning tasks (Wei et al., 2022b;a) suggests that these models could possibly also understand embodied 3D scenes with language-only input (Landau & Jackendoff, 1993). However, SQA3D imposes a grand challenge to these models. The powerful Unified QA (*Large* variant) and GPT-3 both fail to deliver reasonable results on our tasks. Further, we hypothesize the bottleneck could also be on the 3D captions, as the results verify the consistent impact on model performances brought by a different source of captions (ScanRefer→ReferIt3D).

Human vs. machine. Finally, all the machine learning models largely fall behind amateur human participants (47.2% of ScanQA + aux. task vs. 90.06%). Notably, we only offer a limited number of examples for the testers before sending them the SQA3D problems. Our participants promptly master how to interact with the 3D scene, understand the situation from the textual description, and answer the challenging questions. The human performance also shows no significant bias for different question types.

4.3 Qualitative Results

Finally, we offer some qualitative results of the variants of our 3D model in Figure 7. We primarily focus on visualizing both the answer predictions and the transformer attention over the object-centric feature tokens (bounding boxes) generated by the VoteNet (Qi et al., 2019) backbone. We highlight the most-attended bounding box among all the predictions by the transformer-based model, in the hope of a better understanding of how these models perceive the 3D scene in order to comprehend the situations and answer the questions. In Figure 7, the correct predictions are always associated with attention over relevant objects in the situation description s^{txt} and questions. Moreover, in case there are multiple instances of the same object category, it is also crucial to identify the correct instance. For example, only ScanQA + aux. task makes the correct prediction for the first question and also attends to the right chair behind , while ScanQA focuses on a wrong instance. These

results confirm our findings in Section 4.2 about the critical role of situation understanding. We also provide some failure modes in *appendix*.

5 Related Work

Embodied AI. The study of embodied AI (Brooks, 1990) emerges from the hypothesis of "*ongoing physical interaction with the environment as the primary source of constraint on the design of intelligent systems*". To this end, researchers have proposed a myriad of AI tasks to investigate whether intelligence will emerge by acting in virtual or photo-realistic environments. Notable tasks including robotic navigation (Das et al., 2018; Anderson et al., 2018; Savva et al., 2019; Chen et al., 2019; Wijmans et al., 2019b; Qi et al., 2020; Deitke et al., 2022) and vision-based manipulation (Kolve et al., 2017; Puig et al., 2018; Xie et al., 2019; Shridhar et al., 2020a;b; 2022). These tasks are made more challenging as instructions or natural-dialogues are further employed as conditions. Sophisticated models have also been developed to tackle these challenges. Earlier endeavors usually comprise multi-modal fusion (Tenenbaum & Freeman, 1996; Perez et al., 2018) and are trained from scratch (Wang et al., 2018; Fried et al., 2018; Wang et al., 2019), while recent efforts would employ pre-trained models (Pashevich et al., 2021; Hong et al., 2021; Suglia et al., 2021). However, the agents still suffer from poor generalization to novel and more complex testing tasks (Shridhar et al., 2020a) compared to results on training tasks. More detailed inspection has still yet to be conducted and it also motivates our SQA3D dataset, which investigates one crucial capability that the current embodied agents might need to improve: **embodied scene understanding**.

Grounded 3D understanding. Visual grounding has been viewed as a key to connecting human knowledge, which is presumably encoded in our language, to the visual world, so as enable the intelligent agent to better understand and act in the real environment. It is natural to extend this ability to 3D data as it offers more immersive representations of the world. Earlier work has examined word-level grounding with detection and segmentation tasks on 3D data (Gupta et al., 2013; Song & Xiao, 2014; Dai et al., 2017; Chang et al., 2017). Recent research starts to cover sentence-level grounding with complex semantics (Chen et al., 2020; Achlioptas et al., 2020; Chen et al., 2021). More recently, new benchmarks introduce complex visual reasoning to 3D data (Azuma et al., 2022; Ye et al., 2021; Yan et al., 2021). However, these tasks mostly assume a passive, third-person's perspective, while our SQA3D requires problem-solving with an egocentric viewpoint. This introduces both challenges and chances for tasks that need a first-person's view, *e.g.* embodied AI.

Multi-modal question answering. Building generalist question answering (QA) systems has long been a goal for AI. Along with the progress in multi-modal machine learning, VQA (Antol et al., 2015; Zhu et al., 2016) pioneers the efforts of facilitating the development of more human-like, multi-modal QA systems. It has been extended with more types of knowledge, *e.g.* common sense (Zellers et al., 2019) and factual knowledge (Marino et al., 2019). Recent research has also introduced QA tasks on video (Lei et al., 2018), and 3D data (Ye et al., 2021; Azuma et al., 2022; Yan et al., 2021). We propose the SQA3D benchmark also in hope of facilitating multi-modal QA systems with the ability of embodied scene understanding. Notably, models for SQA3D could choose their input from a 3D scan, egocentric video, or BEV picture, which makes our dataset compatible with a wide spectrum of existing QA systems.

6 Conclusion

We've introduced SQA3D, a benchmark that investigates the capability of embodied scene understanding by combining the best of situation understanding and situated reasoning. We carefully curate our dataset to include diverse situations and interesting questions while preserving the relatively large scale (20.4k situation descriptions and 33.4k questions). Our questions probe a wide spectrum of knowledge and reasoning abilities of embodied agents, notably navigation, common sense, and multi-hop reasoning. We examine many state-of-the-art multi-modal reasoning systems but the gap between the best ML model and human performances so far is still significant. Our findings suggest the crucial role of proper 3D representations and better situation understanding. With SQA3D, we hope of fostering the research efforts in developing better embodied scene understanding methods and ultimately facilitate the emergence of more intelligent embodied agents.

REFERENCES

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European Conference on Computer Vision (ECCV)*, pp. 422–440, 2020. 2, 3, 7, 9

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 1

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 5

John R Anderson, James G Greeno, Lynne M Reder, and Herbert A Simon. Perspectives on learning, thinking, and activity. *Educational Researcher*, 29(4):11–13, 2000. 2

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3674–3683, 2018. 1, 3, 9

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*, pp. 2425–2433, 2015. 9, 14

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19129–19139, 2022. 2, 3, 5, 6, 8, 9, 15

Rodney A Brooks. Elephants don't play chess. *Robotics and autonomous systems*, 6(1-2):3–15, 1990. 9

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901, 2020. 3, 5, 6

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 9

Soravit Changpinyo, Doron Kukliansky, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. All you may need for vqa are image captions. *arXiv preprint arXiv:2205.01883*, 2022. 6

Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European Conference on Computer Vision (ECCV)*, pp. 202–221, 2020. 2, 3, 7, 9, 14

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12538–12547, 2019. 9

Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3193–3203, 2021. 2, 6, 9

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5828–5839, 2017. 2, 3, 6, 9, 15

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–10, 2018. 2, 9

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, et al. Procthor: Large-scale embodied ai using procedural generation. *arXiv preprint arXiv:2206.06994*, 2022. 9

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018. 9

Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5067–5077, 2022. 6

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. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6904–6913, 2017. 4, 5

James G Greeno. The situativity of knowing, learning, and research. *American psychologist*, 53(1): 5, 1998. 2

Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 564–571, 2013. 9

Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. Vln bert: A recurrent vision-and-language bert for navigation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1643–1653, 2021. 9

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*, 2020. 3, 5, 6, 7

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 preprint arXiv:1712.05474*, 2017. 9

Barbara Landau and Ray Jackendoff. "what" and "where" in spatial language and spatial cognition. *Behavioral and brain sciences*, 16(2):217–238, 1993. 8

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 7, 9

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7331–7341, 2021. 2, 6, 15

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision (ECCV)*, pp. 121–137. Springer, 2020. 5

Hongzhuo Liang, Xiaojian Ma, Shuang Li, Michael Görner, Song Tang, Bin Fang, Fuchun Sun, and Jianwei Zhang. Pointnetgpd: Detecting grasp configurations from point sets. In *International Conference on Robotics and Automation (ICRA)*, pp. 3629–3635. IEEE, 2019. 1

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 (NeurIPS)*, 32, 2019. 5

Xiaojian Ma, Weili Nie, Zhiding Yu, Huaizu Jiang, Chaowei Xiao, Yuke Zhu, Song-Chun Zhu, and Anima Anandkumar. Relvit: Concept-guided vision transformer for visual relational reasoning. *arXiv preprint arXiv:2204.11167*, 2022. 5

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3195–3204, 2019. 4, 9, 14

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 163–170, Seattle, WA, USA â†' Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.bea-1.16. 14

Alexander Pashevich, Cordelia Schmid, and Chen Sun. Episodic transformer for vision-and-language navigation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15942–15952, 2021. 9

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 32, 2018. 9

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8494–8502, 2018. 9

Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9277–9286, 2019. 5, 8

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9982–9991, 2020. 9

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018. 6

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *International Conference on Computer Vision (ICCV)*, pp. 9339–9347, 2019. 1, 9

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10740–10749, 2020a. 9

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020b. 9

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning (CoRL)*, pp. 894–906. PMLR, 2022. 1, 9

Shuran Song and Jianxiong Xiao. Sliding shapes for 3d object detection in depth images. In *European conference on computer vision*, pp. 634–651. Springer, 2014. 9

Alessandro Suglia, Qiaozi Gao, Jesse Thomason, Govind Thattai, and Gaurav Sukhatme. Embodied bert: A transformer model for embodied, language-guided visual task completion. *arXiv preprint arXiv:2108.04927*, 2021. 9

Joshua Tenenbaum and William Freeman. Separating style and content. *Advances in Neural Information Processing Systems (NeurIPS)*, 9, 1996. 9

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 5

Xin Wang, Wenhan Xiong, Hongmin Wang, and William Yang Wang. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. In *European Conference on Computer Vision (ECCV)*, pp. 37–53, 2018. 9

Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6629–6638, 2019. 9

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022a. 8

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022b. 8

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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6659–6668, 2019a. 2, 3

Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. *arXiv preprint arXiv:1911.00357*, 2019b. 9

Xu Xie, Hangxin Liu, Zhenliang Zhang, Yuxing Qiu, Feng Gao, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Vrgym: A virtual testbed for physical and interactive ai. In *Proceedings of the ACM Turing Celebration Conference-China*, pp. 1–6, 2019. 9

Xu Yan, Zhihao Yuan, Yuhao Du, Yinghong Liao, Yao Guo, Zhen Li, and Shuguang Cui. Clevr3d: Compositional language and elementary visual reasoning for question answering in 3d real-world scenes. *arXiv preprint arXiv:2112.11691*, 2021. 3, 9

Shuquan Ye, Dongdong Chen, Songfang Han, and Jing Liao. 3d question answering. *arXiv preprint arXiv:2112.08359*, 2021. 3, 9

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6281–6290, 2019. 2, 6, 15

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6720–6731, 2019. 9

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4995–5004, 2016. 9

A DATA COLLECTION

A.1 DATA COLLECTION WEB UI

We present the Web UI of our data collection in Figure 8 (Stage I), Figure 9 (Stage II) and Figure 11 (Stage III) respectively. We developed our UI based on Chen et al. (2020). These UIs share some common components: a 3D scene viewer, where the user can drag, rotate, and zoom in/out the scene; clickable objects/tags, where users might click on either the object mesh directly or the tag on the sidebar to highlight it in the scene; and an instruction set that guide the user through the task. Users may also switch between a full scene or object-mesh only to focus on the tasks. The users are also required to submit multiple responses with the same scene.

Notably, we create detailed tutorials for each stage (not shown in the UI) with examples and animated demonstrations. We found tutorials and instruction sets with clear criteria on rejection and bonus(*e.g.* Figure 10) helpful with high quality data. Finally, all the testers need to pass a test before the qualification for our task is granted.

A.2 DATA POST-PROCESSING

There are two major data post-processing steps in SQA3D: **cleaning** and **balancing**. For cleaning, we primarily focus on the grammatical correction. We adopt both rule-based cleaning and an ML-based tool called GECToR (Omelianchuk et al., 2020) in our grammatical correction pipeline. We adjust the correction threshold based on human judgment over the corrected data samples.

In the balancing step, our goal is to reduce the question-answer bias in the dataset. Therefore we follow the practice in Antol et al. (2015); Marino et al. (2019) and re-sample the questions based on their prefixes and answer type, in hope of a more balanced answer distribution. We provide answer distribution before and after balancing in Section B.1.

A.3 MORE MTURK DETAILS

We provide the detailed MTurk job settings as below:

Region. We enable access to our tasks in the following countries/regions:

US, DE, GB, AU, CA, SG, NZ, NO, SE, FI, DK, IE

Approval rate & Number of approved jobs. The testers are required to have at least a 95% approval rate and have completed more than 1000 tasks. However, we relax this requirement to a 90% approval rate for Stage III as it is simpler than the other annotation tasks.

Reward. The participants will be rewarded $0.5 for each task in Stage I and II, and $0.2 for the QA tasks in Stage III, with a possibility of an extra bonus depending on the overall quality. We actively monitor the response quality and send bonuses/rejections on a daily basis. Note that we collected 5 responses for each task in all three stages.

Task lifetime. We set the lifetime as 10 days for tasks in Stage I and 20 days for those in Stage II and III. However, we found most of the tasks can be completed in less than 7 days.

B DATASET DETAILS

B.1 MORE STATISTICS

We provide the histogram of the answer distribution before & after balancing in Figure 12 and Figure 13, respectively. It can be seen that we manage to ensure there is no single answer that dominates any type of question (categorized by their prefixes). However, we do acknowledge that prefix-based balancing might still not be sufficient since models could also learn to use the n-grams pattern. A more effective avenue is collecting more questions with less-frequent answers, which we leave as future work.

Figure 8: Dataset collection Web UI for Stage I.

In Figure 14a and Figure 14b, we show the histogram of the length of situation description s^{txt} and question q. Overall most of the descriptions and questions are middle-length sentences (10-20 words).

B.2 Details on egocentric video and BEV image

For egocentric videos, we uniformly downsample the frames of the original ScanNet (Dai et al., 2017) video by using the first frame of every 20 frames. Afterwards, we resize all the frames to 224×224 to create the video used for training ClipBERT(Lei et al., 2021). Blender is used for rendering all BEV images. We compute the radius of the bounding sphere of the scene and put the camera to the top of the scene with a distance of 7 times the radius to the center of the bounding sphere. Images of size 1920×1080 are rendered for clarity while the input to the MCAN(Yu et al., 2019) model is the resized version of the images to 224×224.

C Model details

C.1 Input pipeline

We follow the input pipeline in ScanQA(Azuma et al., 2022) without further modification. As for MCAN, we only transform the images to fit the ImageNet-pretrained encoder. In ClipBERT, we randomly sample 8 clips with each clip consisting of 2 frames of the video to feed into the model as the scene representation. Note that each frame is resized to 1000×1000 following the practice of original ClipBERT(Lei et al., 2021).

Figure 9: Dataset collection Web UI for Stage II.

Assignment with questions that can be answered without considering the context will be rejected.

-- How many chairs are there in the room? ×

-- Is the amount of monitors on the desk I'm facing at odd or even? √

Assignment with questions that can be answered by merely reading the description (no need to look at the 3D scene) will be rejected.

You may ask at most 3 "simple" questions ((you need to ask 5 questions in total) as we encourage creative questions; otherwise your assignment will be rejected. Questions below are viewed as "simple":

- Questions about simple object category/property or counting, ex. Is there a chair on my 6 o'clock direction?; What is the color of the table on my right?; How many chairs are there behine me?

- Questions that repeat the same pattern

- Questions that can be answered with "Yes" or "No" (Note: some creative questions can also have answer "yes" or "no", but you still need to control the overall amount of questions with answer "yes" or "no" in your assignment)

Figure 10: Additional instruction set to the AMT participants in Stage II.

C.2 Hyper-parameters

We provide the hyper-parameters of the considered models in Table 4.

C.3 Additional details on zero-shot models

We uniformly sample 30 sentences from our 3D caption sources for both models. When testing with the Unified QA_{Large} model, we employ a simple greedy sampling method and the following prompt:

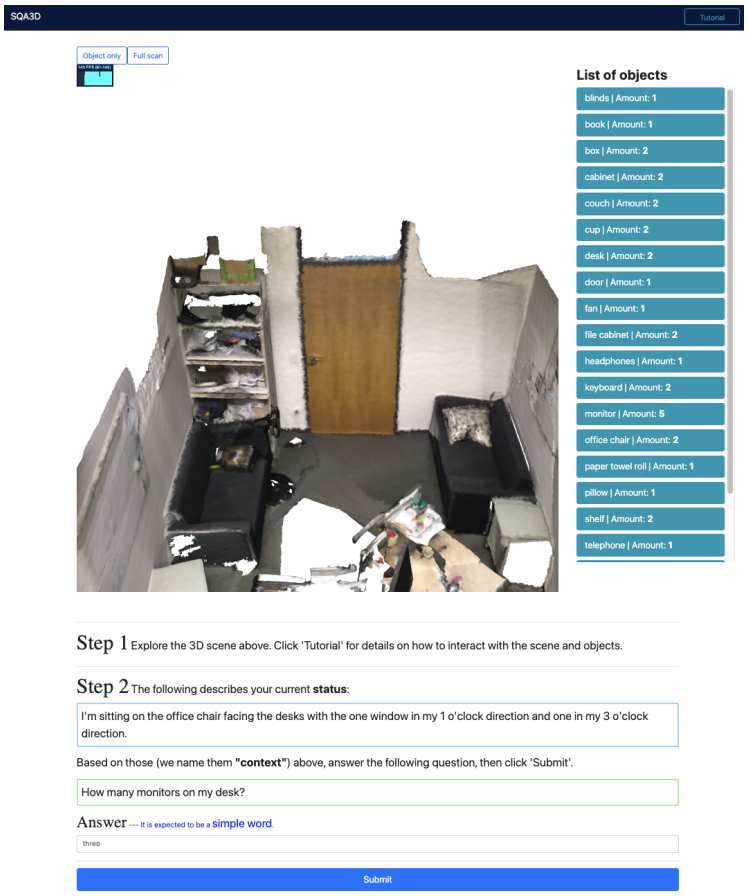


Figure 11: Dataset collection Web UI for Stage III.

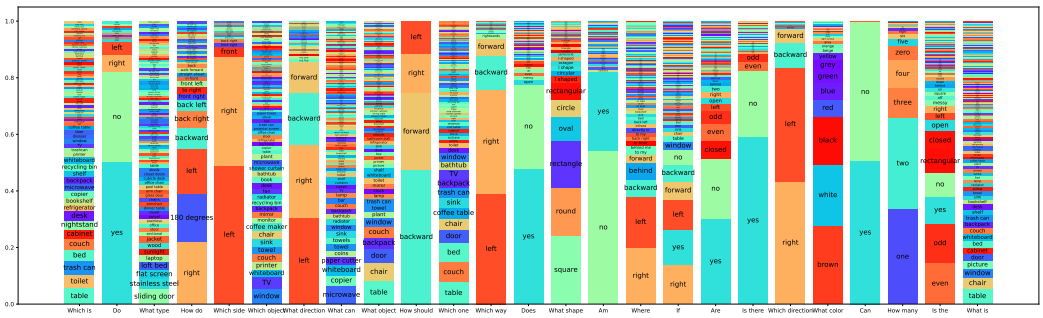


Figure 12: Answer distribution (organized by question prefixes) before balancing.

```
{s^txt}
Q: {q}
A:
```

, where $\{s^{\text{txt}}\}$ and $\{q\}$ are replaced by the situation description and question. For GPT-3, we use the `text-davinci-002` variant and the following prompt:

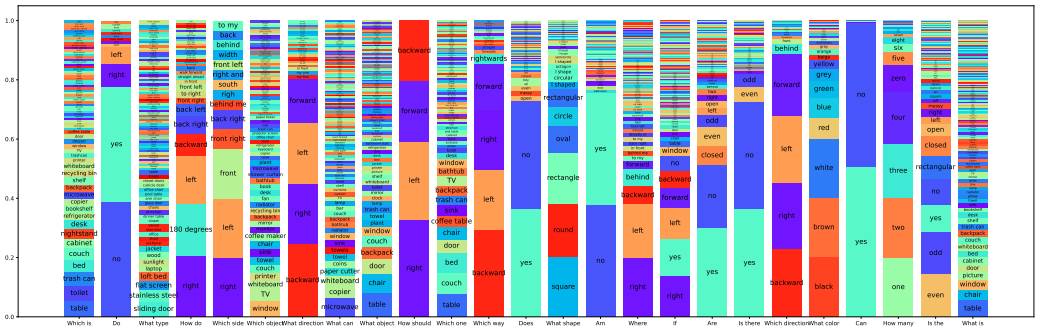
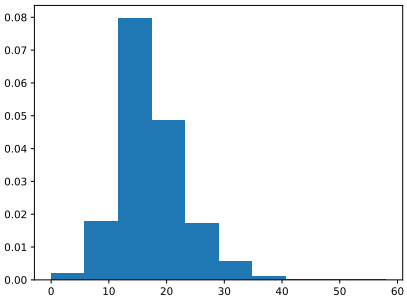
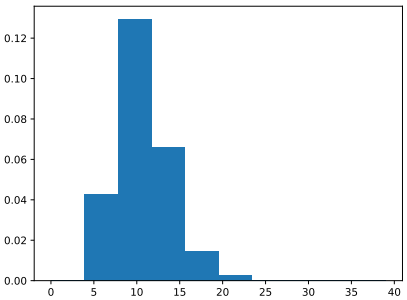


Figure 13: Answer distribution (organized by question prefixes) after balancing.



(a) Histogram of situation description s^{txt} length.



(b) Histogram of question q length.

```
Context:  There is a book on the desk.  A laptop with a green cover
is to the left of the book.
Q: I'm working by the desk.  What is on the desk beside the book?
A: laptop
Context:{s^txt}
Q: {q}
A:
```

, where we use a 1-shot example to demonstrate the format of our task. Interestingly, we found only GPT-3 would benefit from few-shot examples.

D Additional Empirical Results

We provide additional qualitative results and failure modes in Figure 15 and Figure 16.

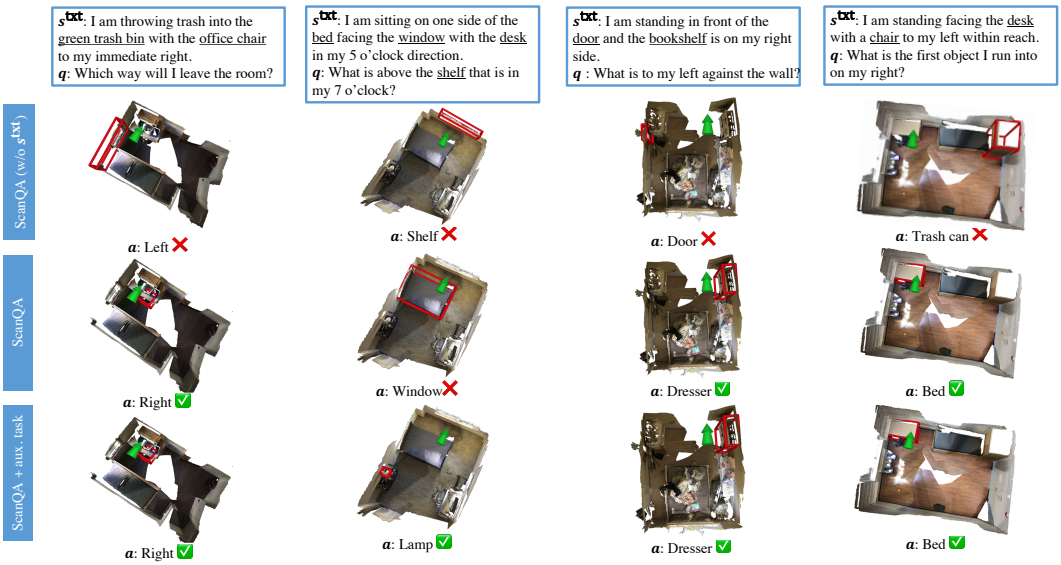


Figure 15: Additional qualitative results.

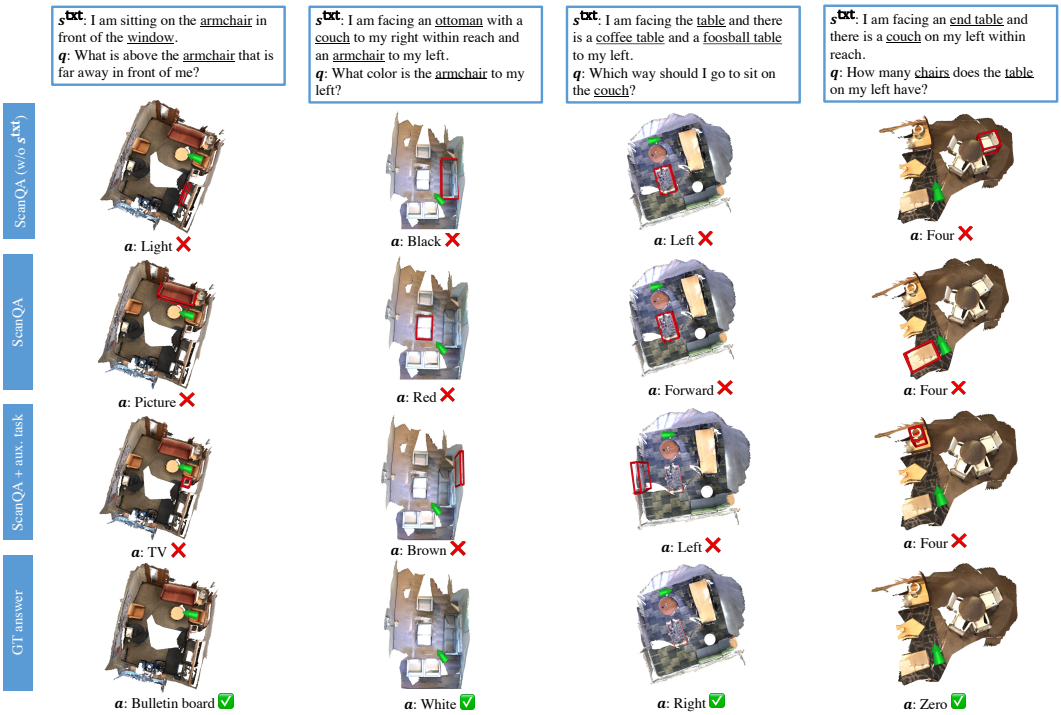


Figure 16: **Failure mode**. Models are likely to predict the wrong answers when they do not attend to relevant objects.

Table 4: Hyper-parameters for the considered models.

Parameter	Value
ScanQA(modified)	
Optimizer	Adam
Gradient clipping norm	1.0
Epsilon	$1e^{-8}$
Weight decay factor	$1e^{-5}$
Beta hyperparameters for Adam	$[0.9, 0.999]$
Learning rate	$5e^{-4}$
Learning rate schedule	No learning rate schedule
Batch size	16
Total training epochs	50
Number of layers for transformer	2
Number of heads for transformer	8
MLP hidden size in MCAN	256
MCAN flatten output size	512
Model hidden size	256
Number of VoteNet output proposals	256
Position regression loss weight α for auxiliary task	1.0
Rotation regression loss weight β for auxiliary task	1.0
MCAN	
Optimizer	AdamW
Gradient clipping norm	0.5
Epsilon	$1e^{-8}$
Weight decay factor	0
Beta hyperparameters for Adam	$[0.9, 0.999]$
Learning rate	$1e^{-4}$
Learning rate schedule	No learning rate schedule
Batch size	16
Total training epochs	12
Number of layers for transformer	6
Number of heads for transformer	8
MLP hidden size in MCAN	512
MCAN flatten output size	1024
Model hidden size	512
ClipBERT	
Optimizer	AdamW
Gradient clipping norm	5.0
Epsilon	$1e^{-6}$
Weight decay factor	$1e^{-3}$
Beta hyperparameters for Adam	$[0.9, 0.98]$
Learning rate	$5e^{-5}$
Learning rate schedule	No learning rate schedule
Batch size	16
Total training epochs	10