ScanQA: 3D Question Answering for Spatial Scene Understanding

Daichi Azuma* Kyoto University
Taiki Miyanishi* ATR, RIKEN AIP
Shuhei Kurita* RIKEN AIP, JST PRESTO
Motoaki Kawanabe ATR, RIKEN AIP

Abstract

We propose a new 3D spatial understanding task for 3D question answering (3D-QA). In the 3D-QA task, models receive visual information from the entire 3D scene of a rich RGB-D indoor scan and answer given textual questions about the 3D scene. Unlike the 2D-question answering of visual question answering, the conventional 2D-QA models suffer from problems with spatial understanding of object alignment and directions and fail in object localization from the textual questions in 3D-QA. We propose a baseline model for 3D-QA, called the ScanQA[1], which learns a fused descriptor from 3D object proposals and encoded sentence embeddings. This learned descriptor correlates language expressions with the underlying geometric features of the 3D scan and facilitates the regression of 3D bounding boxes to determine the described objects in textual questions. We collected human-edited question–answer pairs with free-form answers grounded in 3D objects in each 3D scene. Our new ScanQA dataset contains over 41k question–answer pairs from 800 indoor scenes obtained from the ScanNet dataset. To the best of our knowledge, ScanQA is the first large-scale effort to perform object-grounded question answering in 3D environments.

1. Introduction

In recent years, significant advances have been achieved in vision-and-language tasks and datasets, and several new datasets have been created to develop models that understand textual expressions, such as captions or questions, which are grounded in two-dimensional (2D) images, such as image captioning [12, 42], understanding referring expressions [25, 50], image region and phrase correspondence [37], and visual question answering (VQA) [6, 20, 23]. VQA is successful in grasping object features visualized in 2D frames. However, when we develop models that understand the spatial information of 3D scenes, such as "What is between the table and TV set?" or "Where is the suitcase located?", the existing models based on 2D images have several challenges in accurately understanding the 3D world. For example, 2D images lack an accurate sense of the relative directions and distances in the 3D scenes, i.e., the stereoscopic attribute-perception problem. Some objects are hidden by other objects when they overlap, i.e., the occlusion problem. When multiple images are used in 2D-image-based question answering models, such models often encounter difficulties in tracking and recognizing whether some objects are the same object between images, i.e., the object localization and identification problem.

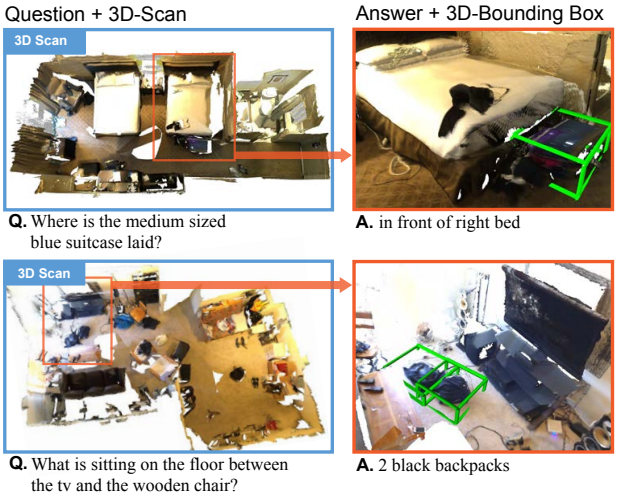

Figure 1. We introduce the new task of question answering for 3D modeling. Given inputs of an entire 3D modeling and a linguistic question, models predict an answer phrase and the corresponding 3D-bounding boxes.

Currently, 3D spatial-understanding models can be developed for the 3D object localization task of ScanRefer [10], the dialog-based localization of ReferIt3D [1], and the 3D object captioning task of Scan2Cap [13]. Embodied question answering [17, 46, 49] is an important task for navigating agents in a 3D scene. We consider that 3D spatial understanding datasets contribute to developing models that comprehend the embodied 3D scene and ask and answer questions about the 3D environment as humans do. However, unlike their 2D image counterparts, question answering datasets on 3D environmental annotations

* denotes equally contributed.
[1] https://github.com/ATR-DBI/ScanQA

are still limited in terms of dataset size and question variety because existing datasets often rely on template-based question–answer collections.

In this paper, we propose a 3D question answering (3D-QA) task that uses 3D spatial information instead of 2D images to comprehend real-world information through the question answering form. In the 3D-QA task, models answer a question for a 3D scene as well as the object localization described in the question. We present the overview of the task in Fig. 1. This 3D-QA task setting is reasonable when external sensors or mobile robots collect sufficient visual information to construct a 3D scene before the QA task. We assume that this is plausible when the model can use the preliminarily captured visual information from the 3D scene because of prior navigation in the scene, such as vision-and-language navigation [5]. This task is also applicable to real-world services that use preliminarily extracted 3D scenes, such as interactive virtual room-viewing services or searching in indoor scenes.

For the 3D-QA task, we developed a novel ScanQA dataset based on RGB-D scans of an indoor scene and annotations derived from the ScanNet dataset [15]. We automatically generated questions from the object captions of ScanRefer [10] using question generation models. However, these auto-generated questions included many invalid questions; therefore, we filtered the invalid questions and refined them if necessary. We collected free-form answers and object annotations from humans using a newly developed interactive 3D scene viewer. In total, we gathered 41k question–answer pairs with 32k unique questions. We propose a 3D-QA model with textual and 3D scene encoding and several baseline models, including 2D image models (2D-QA), a combination of 3D object localization models [10, 38], and a question answering model [51]. We confirmed that the ScanQA model outperformed the baseline models in most evaluations, including exact matching and image captioning metrics in the proposed ScanQA dataset.

2. Related work

The 3D-QA task is similar to existing visual question answering and 3D embodied question answering. We place our task as a spatial comprehension of the entire 3D scene given linguistic questions.

2.1. Visual Question Answering

Visual question answering (VQA) is a task in which models are given a 2D image and a question about its content. They are expected to provide an appropriate answer. Question answering in 2D images was proposed by Malinowski *et al.* [34], and various inference methods [4, 6, 51] have since been proposed. One of the best VQA methods is Oscar [30], which uses Mask R-CNN to infer a solution by considering the relationship between individual objects in the image. In addition, Jang *et al.* [24] proposed a question answering method that considers more detailed vision and motion information using video. ClipBERT [29] improved its accuracy by dividing a video into clips and reasoning from them individually. VQA 360° [14] is a task for answering questions about a 360° image. Although VQA 360° contributes to understanding the 3D scene, the available information is limited compared with the ScanQA dataset. Our dataset also includes the object identification task, which is different from the existing VQA 360° dataset.

2.2. 3D Object Localization with Language

ScanRefer [10] localizes an object referred to in a free-form text description. The ScanRefer model identifies a 3D-bounding box for an object given the input description. This dataset is based on 800 scenes derived from the ScanNet dataset [15]. In addition to answering questions on 3D scenes, the 3D-QA task also includes an object localization task for objects that appear in the question answering. Unlike in the ScanRefer task, the objects in ScanQA object localization can be multiple because multiple objects can appear in a single question.

2.3. Question Answering in 3D scenes

Unlike 2D-QA, for which many datasets have been proposed, 3D question answering datasets are still limited. We notice that the existing question answering tasks on 3D scenes have interactive forms. The interactive QA dataset (IQUAD) [19] on AI2THOR [27] enables model agents to interact with objects in a scene to determine the answer to a question. Embodied question answering (EQA) [17, 46, 49] is a combination of visual question answering and navigation such as vision-and-language navigation tasks [5, 11, 43, 47] and models [18, 28]. In the original EQA dataset [17], the embodied model agent receives a question such as "What color is the car?" and navigate to the object described in the question in House 3D [47]. MP3D-EQA [46] is a photorealistic embodied QA for Matterport 3D scans [9]. MT-EQA [49] is a multitarget variation of EQA. We summarize the relation of these datasets with ScanQA in Table 1. Unlike these datasets, the ScanQA dataset is not created from fixed templates and hence includes more natural and a significantly larger number of unique questions, as discussed in Sec. 3.3.

3. ScanQA Dataset

We hereby define the 3D-QA task and describe the collection of the corresponding dataset.

3.1. 3D-QA Task

As illustrated in Fig. 1, a 3D-QA task requires models to answer a question when given all the information of a 3D

3D-QA Datasets	Type	Question Collection	Answer Collection	Environment	Photorealistic	# 3D Scenes
IQUAD	Interactive	Template-based	Template-based	AI2THOR	No	30 rooms
EQA	Navigation	Template-based	Template-based	House3D	No	588 scenes
MP3D-EQA	Navigation	Template-based	Template-based	Matterport 3D	Yes	144 floors
MT-EQA	Navigation	Template-based	Template-based	House3D	No	588 scenes
ScanQA dataset	3D Scan	AutoGen+HumanEdit	Human	ScanNet	Yes	800 rooms

Table 1. Comparison of 3D question-answering datasets.

scene. Here, models use the 3D spatial information, such as RGB-D scans or point cloud data. We also require models to specify the 3D-bounding boxes of objects that are related to this question answering. This prevents models from answering questions by relying on the textual priors of the trained questions without examining the scene. However, unlike the ScanRefer dataset, we do not require models to target one described object for each question. This is because multiple objects can be used to answer certain questions. For example, the question "What color is the chairs around the table?" is related to multiple objects. This question is also answerable as long as the chairs around the unique table in the scene have the same color. In such scenarios, we require models to answer the question addressing multiple 3D-bounding boxes.

3.2. Question-Answer Collection

The ScanQA dataset was created using multiple phrases, including automatic QA generation [2, 8, 32, 48], question filtering, question editing, and answer collection. First, we automatically generated question–answer pairs from the referring expressions to identify objects in 3D scenes obtained from the ScanRefer dataset [10]. We applied the question-and-answer generation model based on a T5-base model [40] trained on a text-based question answering dataset [41] [2] for ScanRefer captions and obtained the seed questions. However, these autogenerated question–answer pairs included many inadequate questions, such as those that are underspecified or not grounded in the scene, as presented in Fig. 2. Auto-generated questions also include easy questions that can be answered with common sense. Therefore, we decided to remove such questions as much as possible. We also did not include auto-generated answers in the final dataset because it was not clear that they were grounded in scenes. Then, we applied filtering and editing to the seed questions with basic rules and human editing in addition to the answer collection via Amazon Mechanical Turk (MTurk). For human editing and answer collection, we developed an interactive visualization website for each 3D scene that enables workers to interact with the 3D scene and check the object names and IDs if they are available. Following the ScanRefer dataset, we attached the object names and IDs for the objects in the 3D scene. We embedded this site into the MTurk task page (Fig. 2).

[2] We used the weights available at https://huggingface.co/valhalla/t5-base-qa-qg-hl.

Split	# Question	# Unique Question	# 3D Scenes
Train	25,563	20,546	562
Val	4,675	4,306	71
Test w/ objects	4,976	4,552	70
Test w/o objects	6,149	5,484	97
Total	41,363	32,337	800

Table 2. ScanQA dataset statistics.

The filtering and editing of the seed questions were conducted as follows. First, we filtered the inadequate questions from the auto-generated seed questions using basic rules. Subsequently, we asked the workers to classify the remaining questions into four classes: *valid*, *too easy*, *unanswerable*, and *unclear* questions. Each question was evaluated by at least three workers. We selected questions in which two or more workers were marked as valid for the next phrase of the editing and answer collection process. In the editing and answer collection, we first presented the filtered questions to workers and requested them to rewrite the questions themselves if they were inadequate for the scene before writing a free-form answers. Multiple answers are collected when necessary. We also collected the object IDs that were used in the question to identify the object in the scene in this phrase. See SM D for details.

3.3. Dataset Statistics

We collected 41,363 questions and 58,191 answers, including 32,337 unique questions and 16,999 unique answers. Table 2 presents the statistics of the ScanQA dataset. This dataset is an order of magnitude larger than existing embodied question-answering datasets in terms of both question size and variation. For example, the EQA dataset [17] contains 4,246 questions, consisting of 147 unique questions in its training set. The EQA-MP3D dataset [46] contains 767 questions consisting of 174 unique questions in its training set. Considering that our dataset contains not only question–answer pairs but also 3D object localization annotations, we assume that this is the largest dataset to specify the nature of objects in 3D scenes with the question answering form. The distribution of the questions based on their first word is shown in Fig. 3. We collected various types of questions through question auto-generation and editing by humans.

We followed the training, validation, and test set splits

Underspecified questions

Q: What is in the corner?
- Several objects at corners!

Q: What color is the chair?
- Three chairs at the scene!

Valid questions

Q: What is over the chair beneath the blackboard?
- Answer: jacket

Q: What color is the office chair next to the desk with a monitor?
- Answer: green

Figure 2. Underspecified and valid questions for an office room scene. We presented scenes with object IDs and names to MTurk workers for the dataset collection.

used in ScanRefer. However, as the object IDs for the test set of ScanRefer are not publicly available, we further split the validation set of ScanRefer into two-holds as the validation set and test set with object annotations in the ScanQA dataset. Therefore, the ScanQA dataset includes two test sets with and without object annotations. We collected at least two answers for each question in the validation and two test sets to evaluate the free-form answers. As our dataset includes the question type of "Where is," writing expressions for answers can vary. Therefore, we adopted evaluation metrics for image captioning in addition to an exact match to the annotated answers in the evaluation.

4. ScanQA Model

We introduce the baseline model of ScanQA for the 3D-QA task. The 3D-QA is formalized as follows: given inputs of the point cloud $p \in \mathcal{P}$ and question $q \in \mathcal{Q}$ about the 3D scene, the 3D-QA model aims to output $\hat{a}$ that semantically matches true answer a^*.

3D feature representation. We primarily use the input point cloud p consisting of point coordinates $c \in \mathbb{R}^3$ in the 3D space for 3D representation. Following previous 3D and language research [10, 13], we use additional point features such as the height of the point, colors, normals, and multiview image features [16] that project 2D appearance features to the point cloud. We use these combined point features as 3D features $r \in \mathbb{R}^{135}$.

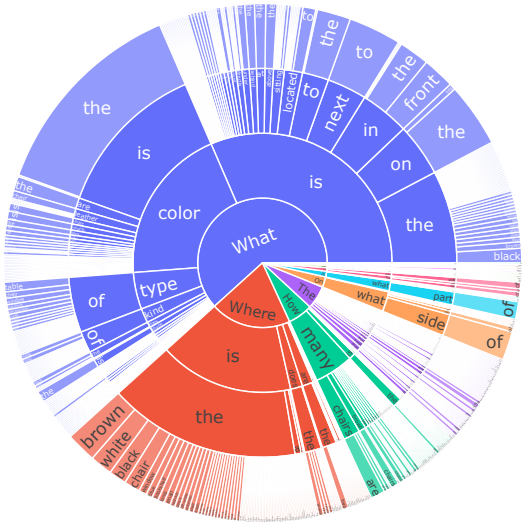

Figure 3. The distribution of the question types by the beginning of the question writing.

Overview of network architecture. To solve the 3D-QA task, we developed a ScanQA model consisting of a 3D & language encoder, 3D & language fusion, and object localization & QA layers. An overview of the proposed ScanQA network is presented in Fig. 4. The 3D & language encoder layer transforms the question into contextualized word representations and point clouds into object proposals. The 3D & language fusion layer combines multiple 3D object features guided by language information using transformer-based encoder and decoder layers [44, 51]. The object localization & QA layer estimates the target object box and object labels and predicts answers associated with questions and scene content.

3D & language encoder layers. This layer encodes the question words $\{w_i\}_{i=1}^{n_q}$ using GloVe [36], and we obtain word representation $Q \in \mathbb{R}^{n_q \times 300}$, where n_q is the number of words in question, and feeds them into a one-layer bidirectional long short-term memory (biLSTM) [22] for word sequence modeling. We project a series of output states from the LSTM using a nonlinear layer with GELUs [21] activation to obtain the contextualized word representation $Q' \in \mathbb{R}^{n_q \times d}$, where d is the hidden size of the biLSTM (set to 256). In addition, this layer detects objects in a scene based on point cloud features $r \in \mathbb{R}^{135}$ using VoteNet [38], which uses PointNet++ [39] as a backbone network. We obtain the object proposals (object boxes) from VoteNet and project them using a nonlinear layer with GELUs activation to obtain the object proposal representation $V \in \mathbb{R}^{n_v \times d}$, where n_v is the number of object proposals (set to 256.)

3D & language fusion layer. Inspired by the architecture of deep modular co-attention networks of MCAN [51], often used for VQA, we use transformer blocks [44] to represent the relationships between object proposals and between question words. After feeding contextual question represen-

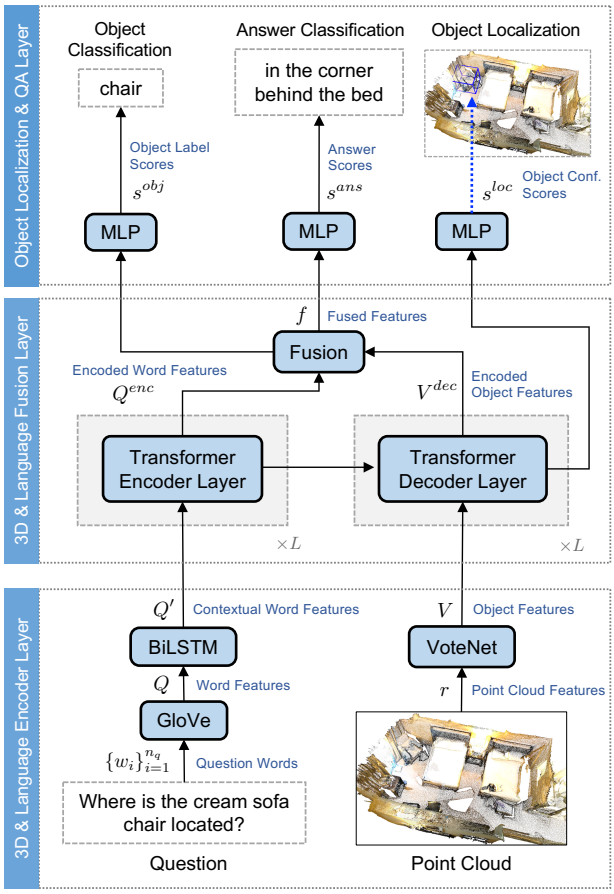

Figure 4. ScanQA model for answering 3D environments. Given a point cloud and RGB frame sequence that capture indoor scenes, the QA model outputs a corresponding answer by fusing 3D and language information through three layers: the 3D scene and language encoder layer, fusion layer, and classification layers.

tation Q' into a stack of L (set to two) transformer encoder layers, we obtain a deeply contextualized question representation $Q^{enc} \in \mathbb{R}^{n_q \times d}$. In addition, we use transformer decoder layers to represent the features of object proposals related to the question words by using the final output of the transformer encoder as the decoder's keys and values. We obtain a question-aware-object proposal representation $V^{dec} \in \mathbb{R}^{n_v \times d}$ after feeding the question and object proposal pairs into a stack of L transformer decoder layers. Subsequently, the final outputs of the transformer layers Q^{enc} and V^{dec} are fused by a fusion layer that uses two-layer multi-layer perceptron (MLP) with an attention mechanism [33] (for details, see [51]). We obtain the fused feature $f \in \mathbb{R}^d$, which simultaneously represents a 3D scene and linguistic question information.

Object localization & QA layer. This layer consists of object localization, object classification, and answer classification modules. Each module is described as follows.

Object localization module aims to predict which of the proposed object boxes corresponds to the question. The question-aware-object proposal representation, $V^{dec} \in \mathbb{R}^{n_v \times d}$, is fed into a two-layer MLP to determine the likelihood of each object box being related to the question. Following [10], we compute the localization confidence $s^{loc} \in \mathbb{R}^{n_v}$ for the proposed n_v object boxes with the cross-entropy (CE) loss to train this module.

Object classification module predicts what objects are associated with a question. Note that many questions do not contain target object names related to the answer, in contrast to a 3D localization task [10, 52, 53]. We use the 3D and question-aware fused feature f and feed it into a two-layer MLP to predict 18 ScanNet benchmark classes. We compute the object classification scores $s^{obj} \in \mathbb{R}^{18}$ with a softmax function and use the CE loss to train this module.

Answer classification module predicts an answer corresponding the question and scene. We project the fused feature f into a vector $s^{ans} \in \mathbb{R}^{n_a}$ for the n_a answer candidates in the training set. To consider multiple answers, we compute final scores with the binary cross-entropy (BCE) loss function to train the module.

Loss function. We use a loss function similar to ScanRefer [10], such as the localization loss $\mathcal{L}_{loc}$ of the object localization module, object detection loss $\mathcal{L}_{det}$ of VoteNet [38], and object classification loss $\mathcal{L}_{obj}$ of the object classification module. To answer the 3D scene content, we additionally use the answer loss $\mathcal{L}_{ans}$ of the answer classification module. We set the final loss as a simple linear combination of these losses, computed as $\mathcal{L} = \mathcal{L}_{ans} + \mathcal{L}_{obj} + \mathcal{L}_{loc} + \mathcal{L}_{det}$.

5. Experiments

5.1. Experimental Setup

In this experimental setup, we referred to the experimental setup of existing studies on scene understanding of ScanNet [15] through languages such as ScanRefer and Scan2Cap [10, 13].

Data augmentation. We applied data augmentation to our training data and applied rotation about all three axes using a random angle in $[-5°, 5°]$ and randomly translated the point cloud within 0.5 m in all directions. Because the ground alignment in ScanNet was incomplete, we rotated it on all axes (not just the top).

Training. To train the ScanQA model, we used Adam [26], a batch size of 16, and an initial learning rate of $5e{-}4$. We trained the model for 30 epochs until it converged and decreased the learning rate by 0.2 times after 15 epochs. To mitigate the fitting of the model against its training data, we set the weight decay factor to $1e{-}5$.

Evaluation. To evaluate the QA performance, we used exact matches EM@1 and EM@10 as the evaluation metric, where EM@K is the percentage of predictions in which the top K predicted answers exactly match any one of the

Model	EM@1	EM@10	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr	SPICE
Test w/ objects										
RandomImage+MCAN	22.31	53.11	26.66	18.49	16.16	14.26	31.27	12.13	60.37	9.05
VoteNet+MCAN	19.71	50.76	29.46	17.23	10.33	6.08	30.97	12.07	58.23	10.44
ScanRefer+MCAN (pipeline)	17.52	49.92	19.17	10.66	0.00	0.00	24.40	9.38	44.25	6.24
ScanRefer+MCAN (e2e)	20.56	52.35	27.85	17.27	11.88	7.46	30.68	11.97	57.36	10.58
ScanQA	**23.45**	**56.51**	**31.56**	**21.39**	**15.87**	**12.04**	**34.34**	**13.55**	**67.29**	**11.99**
OracleImage+MCAN	25.34	55.93	28.70	20.11	16.78	12.89	34.59	13.42	67.24	11.93
Test w/o objects										
RandomImage+MCAN	20.82	51.23	26.29	17.90	14.27	9.66	29.23	11.54	55.64	8.87
VoteNet+MCAN	18.15	48.56	29.63	17.80	11.57	7.10	29.12	11.68	53.34	10.36
ScanRefer+MCAN (pipeline)	16.47	49.05	18.71	10.98	16.53	0.76	22.45	8.76	40.81	6.41
ScanRefer+MCAN (e2e)	19.04	49.70	26.98	16.17	11.28	7.82	28.61	11.38	53.41	10.63
ScanQA	**20.90**	**54.11**	**30.68**	**21.20**	**15.81**	**10.75**	**31.09**	**12.59**	**60.24**	**11.29**

Table 3. Performance comparison of question answering with image captioning metrics. **e2e** represents an end-to-end model.

ANS	OBJ	LOC	EM@1	EM@10	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr	SPICE
Test w/ objects												
✓			12.16	42.77	12.86	5.45	0.12	0.02	17.55	6.71	29.17	4.05
✓	✓		18.31	49.18	22.32	14.53	11.15	7.92	26.37	9.94	49.10	6.36
✓		✓	20.46	51.67	25.06	16.91	14.06	11.37	29.22	11.13	55.17	8.21
✓	✓	✓	**23.45**	**56.51**	**31.56**	**21.39**	**15.87**	**12.04**	**34.34**	**13.55**	**67.29**	**11.99**
Test w/o objects												
✓			10.78	39.44	11.94	5.02	0.12	0.02	15.34	5.91	25.51	3.51
✓	✓		16.23	46.30	21.37	13.49	10.71	7.64	23.63	9.10	43.21	6.13
✓		✓	18.12	49.60	25.12	17.58	14.68	10.23	26.50	10.43	49.93	8.16
✓	✓	✓	**20.90**	**54.11**	**30.68**	**21.20**	**15.81**	**10.75**	**31.09**	**12.59**	**60.24**	**11.29**

Table 4. Performance comparison of different experimental conditions of the ScanQA model.

Model	Acc@0.25	EM@10
Test w/ objects		
ScanQA (xyz)	23.67	55.67
ScanQA (xyz+rgb)	23.45	55.43
ScanQA (xyz+rgb+normal)	23.35	54.50
ScanQA (xyz+multiview)	25.40	55.51
ScanQA (xyz+multiview+normal)	25.44	56.51

Table 5. Feature ablation results

ground-truth answers. We also included sentence evaluation metrics frequently used for image captioning models because some of the questions had multiple possible answer expressions, as discussed in Sec. 3.3. We added the BLEU [35], ROUGE-L [31], METEOR [7], CIDEr [45], and SPICE [3] metrics to evaluate robust answer matching.

Baselines. To validate our 3D-QA model (ScanQA), we prepared several baselines. Empirical experiments were conducted using the following methods.

RandomImage+2D-QA First, we prepared several 2D-QA models as baselines to demonstrate how our 3D-QA models outperformed 2D-based VQA models for the 3D-QA task. We used a pretrained MCAN model [51] as the 2D-QA model. MCAN is a transformer network [44] that uses a cross-attention mechanism to represent the relationship between question words and objects in an image. The proposed method uses some of the modules used in MCAN, such as transformer encoder and decoder layers, to create 3D and language features. By comparing these two, we can confirm the importance of creating a model specialized for 3D-QA. Because 2D-QA models cannot be directly applied to a 3D environment, we randomly sampled three images from the video captured to build the ScanNet dataset. We used a bottom-up top-down attention model [4] to extract the appearance features of the objects. We applied pretrained 2D-QA models to these images and computed the answer scores for each image. Finally, we selected the most probable answer according to the averaged answer scores of these images. We experimented with 2D-QA using three images captured in the environment per question.

OracleImage+2D-QA To investigate the upper bound on the performance of 2D-QA for questions in 3D space, we used images around a target object associated with a question–answer pair. We set the camera's position based on the coordinates of the bounding box of the correct object and captured images from the direction and distance at which the bounding box was most visible. Because the object may not be visible depending on the camera's position, we used three images per question. We applied 2D-QA models to these images similar to RandomImage+2D-QA. Note that it is difficult to obtain such images in actual QA scenarios. By examining the performance of this method, we can determine the difficulty of solving the 3D-QA task using 2D-QA models.

VoteNet+MCAN VoteNet [38] is a 3D object detection method that locates and recognizes objects in a 3D scene.

This method detects objects in a 3D space, extracts their features, and uses them in a standard VQA model (MCAN). Unlike our method, this method does not consider the target object or its location in the 3D space.

ScanRefer+MCAN (pipeline) ScanRefer [10] is a 3D object localization method for localizing a given linguistic description to a corresponding target object in a 3D space. ScanRefer internally uses VoteNet to detect objects in a room and estimates the object corresponding to the linguistic description from among the candidate objects. Note that ScanRefer cannot be used directly for QA. Thus, we used a pretrained ScanRefer model to identify the object corresponding to the question and then applied 2D-QA (MCAN) to the image surrounding the object localized by ScanRefer. Note that ScanRefer and MCAN were run separately, and end-to-end learning was not possible.

ScanRefer+MCAN (end-to-end) This method is more sophisticated and closer to the proposed method. Although ScanRefer+MCAN (pipeline) conducts object localization and QA separately, this method simultaneously learns localization and QA modules. Specifically, the input to the method is the object proposal feature of ScanRefer for VQA models; subsequently, the model predicts answers based on object box features and question content. Unlike the ScanQA model, this uses the output of VoteNet separately for object localization and QA modules (although the information for both tasks is mutually useful) and does not learn both modules in common.

5.2. Quantitative Analysis

The performance of 3D-QA on the ScanQA dataset and image caption metrics are presented in Table 3. The best results in each column are shown in bold. We compared our ScanQA model with competitive baselines VoteNet+MCAN, ScanRefer+MCAN (pipeline), and ScanRefer+MCAN (end-to-end). These baselines share some of the components of the proposed method and aided us in understanding useful components for 3D-QA. The results indicated that our ScanQA method significantly outperformed all baselines across all data splits over all evaluation metrics. In particular, ScanQA significantly outperformed ScanRefer+MCAN (end-to-end), which learns QA and object localization separately across all data splits over all evaluation metrics, indicating that the ScanQA model succeeded in synergistic learning by sharing object localization and QA modules. ScanQA outperformed VoteNet+MCAN by a large margin. This result suggested that using object localization in a 3D space and predicting object categories related to questions are important for 3D-QA. We will clarify this point in the section on the ablation study. Interestingly, VoteNet+MCAN, ScanRefer+MCAN (end-to-end), and ScanQA significantly outperformed ScanRefer+MCAN (pipeline), which detects target objects related to a question using a pretrained ScanRefer and then applies 2D-QA to the surrounding images of a target object. The results indicated that end-to-end training with 3D and language information is suitable for solving 3D-QA model problems. In addition, we observed that our 3D-QA model, ScanQA, is superior to a 2D-QA model, RandomImage+MCAN, which uses an effective pretrained model. We also observed that the 2D-QA baseline with oracle object identification of OracleImage+MCAN performed better or more competitively than the ScanQA model. Although this suggests that accurate object identification for questions indeed boosts 3D-QA results, this is an oracle setting for real-world applications. We finally evaluated the human performance on the sampled questions in the test set with objects using MTurk. The exact matching score (EM@1) is 51.6 for the best-performing MTurk worker.

5.3. Ablation Studies

We conducted ablation studies concerning the design of the ScanQA model. Table 4 lists the effects of each major component of the proposed method. The effect of different input data is shown in Table 5.

Does object classification help? We demonstrated the effectiveness of object classification by conducting an experiment using ScanQA combined with and without the object classification module. We compared our method trained with the answer and object classification modules (ANS+OBJ) with a model trained with only the answer module (ANS), and we also compared a model trained with full modules (ANS+OBJ+LOC) with one trained with the answer and object localization modules (ANS+LOC). The results in Table 4 show that the models with the OBJ outperformed the other models. This suggests that predicting the category of a target object is effective for 3D-QA.

Does object localization help? We demonstrated the effectiveness of object localization by conducting an experiment with ScanQA and without an object localization module. We compared our method trained with the answer and object localization modules (ANS+LOC) with a model trained with only the answer module (ANS), and we also compared a model trained with full modules (ANS+OBJ+LOC) with one trained with the answer and object classification modules (ANS+OBJ). The results in Table 4 show that the model with LOC consistently outperformed the other models. This suggests that localizing target objects is also important for improving 3D-QA performance.

Do colors help for 3D-QA? According to the ScanRefer study, models that use color information have better object localization performance than models that use only geometry [10]. Motivated by this finding, we evaluated several models using different features. We compared our method trained with geometry (xyz) and multiview image features (xyz+multiview) with a model trained with only geome-

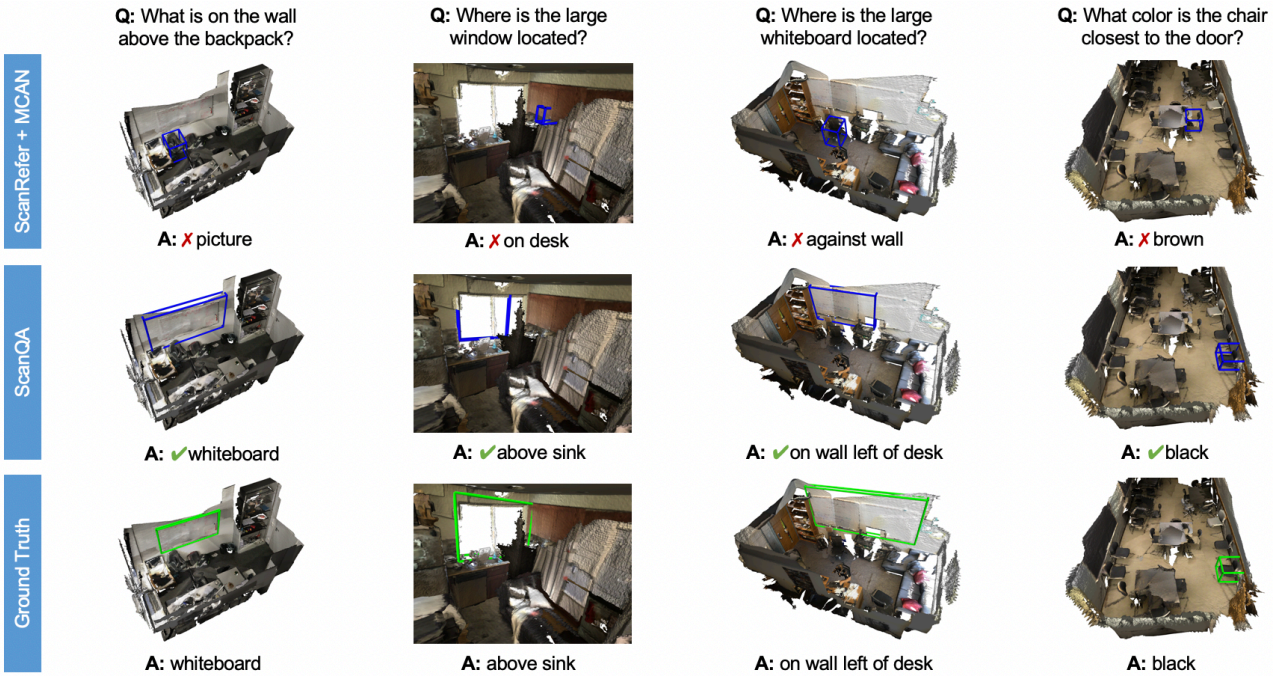

Figure 5. **Qualitative results.** Predicted answers are described below each figure. Predicted boxes are marked blue and the ground truth is marked green. We show examples in which ScanQA produced good object localization when predicting correct answers, whereas ScanRefer+MCAN (pipeline) could not.

try (xyz) and one trained with RGB values (xyz+rgb). To validate the object localization performance, we used accuracy (Acc@0.25), in which the positive predictions have a higher intersection over union with the ground truths than the threshold of 0.25 used in [10]. As Table 5 shows, RGB values were not effective for both object localization and QA. The multiview image features were slightly effective for object localization but not on QA. This is because it is more difficult for the ScanQA dataset to associate language and object information than the ScanRefer dataset because multiple objects may apply to a single question. Fortunately, ScanQA trained with geometry, preprocessed multiview image features, and normals which is demonstrated in Table 3 outperformed the other models, but the effect was limited. This result suggested that subsequent studies should consider a more balanced selection of features in terms of computational cost and performance.

5.4. Qualitative Analysis

Finally, we demonstrated the excellent performance of our model by visualizing qualitative examples of ScanQA, ScanRefer+MCAN (pipeline), and ground truth. Fig. 5 shows the representative QA results of a baseline method and ScanQA. The results suggested that answering questions requires object localization related to the answer according to the question content and point cloud matching. For example, the leftmost case shows that a whiteboard located above a backpack could not be answered by ScanRefer+MCAN (pipeline), which localized the backpack in error. This is because the QA and localization modules were separated in ScanRefer+MCAN (pipeline). In contrast, our model successfully localized target objects related to answers and predicted correct answers by simultaneously learning QA and object localization.

6. Conclusion

Spatial understanding using language expression is a core technology for models deployed in the real world and interacting with humans. We introduce a novel task of 3D question answering (3D-QA), in which models observe an entire 3D scan and answer a question about the 3D scene in addition to the object localization. Based on the ScanRefer dataset, we created a new ScanQA dataset which consists of 41,363 questions and 32,337 unique answers from 800 scenes derived from the ScanNet scenes. We propose a 3D-QA baseline model of ScanQA. We confirm that the ScanQA baseline performs better than the counterpart 2D-based VQA baselines in most of the evaluation measures, including the exact match and image captioning metrics.

Acknowledgements

This work was supported by NEDO JPNP20006, JSPS KAKENHI 21H03516 and 18KK0284, and JST PRESTO JPMJPR20C2.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. ReferIt3D: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, 2020. 1

[2] Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6168–6173, 2019. 3

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 6

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 6

[5] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015. 1, 2

[7] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Association for Computational Linguistics, 2005. 6

[8] Ying-Hong Chan and Yao-Chung Fan. A recurrent BERT-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, 2019. 3

[9] 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. *International Conference on 3D Vision (3DV)*, 2017. 2

[10] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020. 1, 2, 3, 4, 5, 7, 8

[11] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[12] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015. 1

[13] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 4, 5

[14] Shih-Han Chou, Wei-Lun Chao, Wei-Sheng Lai, Min Sun, and Ming-Hsuan Yang. Visual question answering on 360deg images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020. 2

[15] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 5

[16] Angela Dai and Matthias Nießner. 3DMV: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 4

[17] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3

[18] 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. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 2

[19] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018. 2

[20] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1

[21] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *CoRR*, abs/1606.08415, 2016. 4

[22] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4

[23] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[24] Yunseok Jang, Yale Song, Chris Dongjoo Kim, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Video Question Answering

with Spatio-Temporal Reasoning. *International Journal of Computer Vision (IJCV)*, 2019. 2

[25] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, 2014. Association for Computational Linguistics. 1

[26] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 5

[27] 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

[28] Shuhei Kurita and Kyunghyun Cho. Generative language-grounded policy in vision-and-language navigation with bayes' rule. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[29] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: ClipBERT for video-and-language learningvia sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[30] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 1

[31] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics, 2004. 6

[32] Luis Enrico Lopez, Diane Kathryn Cruz, Jan Christian Blaise Cruz, and Charibeth Cheng. Simplifying paragraph-level question generation via transformer language models. In Duc Nghia Pham, Thanaruk Theeramunkong, Guido Governatori, and Fenrong Liu, editors, *Proceedings of the Pacific Rim International Conference on Artificial Intelligence (PRICAI)*, 2021. 3

[33] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421, 2015. 5

[34] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2014. 2

[35] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. 6

[36] Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global vectors for word representation. In *Proceeding of Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 4

[37] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2641–2649, 2015. 1

[38] Charles R. Qi, Or Litany, Kaiming He, and Leonidas J. Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2, 4, 5, 6

[39] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2017. 4

[40] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 3

[41] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. 3

[42] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2018. 1

[43] 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 *Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2017. 4, 6

[45] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575. IEEE Computer Society, 2015. 6

[46] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3

[47] Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. Building generalizable agents with a realistic and rich 3d environment. *arXiv preprint arXiv:1801.02209*, 2018. 2

[48] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of*

the IEEE/CVF International Conference on Computer Vision (ICCV), 2021. 3

[49] Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L. Berg, and Dhruv Batra. Multi-target embodied question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2

[50] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 69–85, 2016. 1

[51] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 4, 5, 6, 1

[52] Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Sheng Wang, Zhen Li, and Shuguang Cui. Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1791–1800, 2021. 5

[53] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2928–2937, 2021. 5