Supplementary Material

In this supplementary material, we first provide additional experimental details in Section A. Next, we provide additional experimental quantitative (Section B) and qualitative results (Section C). Finally, we have attached a snapshot of the annotation website (Section D).

A. Additional Experimental Details

We conducted additional experiments to demonstrate the effectiveness of the proposed method compared with various other methods. In this section, we introduce other baselines and an additional ScanQA model that considers multiple objects related to a question.

A.1. Additional Baseline Models

Additional 2D-QA Image. We prepared RandomImage (or OracleImage) + 2D-QA as a 2D-QA baseline method that uses three images captured in the environment per question. Fig. 6 shows the images randomly sampled from the video to build the ScanNet dataset. In addition to using such real images from the environments, we also used images of the mesh data of ScanNet captured at positions and angles similar to real images as in Fig. 7. We refer to models using real images "real" and ones using mesh images "mesh." We also used mesh images captured from a top-down view (referred to as TopDownImage) to view the entire room with a single image (Fig. 8).

Additional 2D-QA Model. In addition to MCAN [51], we evaluated 2D-QA using the BERT-based model Oscar [30] trained with many image-text pairs and has demonstrated high performance in various tasks, such as VQA, image retrieval, image captioning, and natural language visual reasoning. Although MCAN and Oscar use effective pretrained models, unlike our method, we can emphasize the effectiveness of the ScanQA model by comparing it to these models.

A.2. Additional ScanQA Model

The ScanQA model introduced in Section 4 predicts the object confidence (box confidences) and object classification for a single object. However, a given question is occasionally associated with more than one object. Thus, we extended the ScanQA model to perform object localization and labeling of multiple objects. Hence, we computed the final scores for all object proposals (or object labels) normalized by a sigmoid function and used BCE loss to train both the object localization and object classification modules. The proposed method used in this study was ScanQA (single), and the model that considers multiple objects was called ScanQA (multiple). Hereafter, unless otherwise specified, ScanQA is referred to as ScanQA (single).

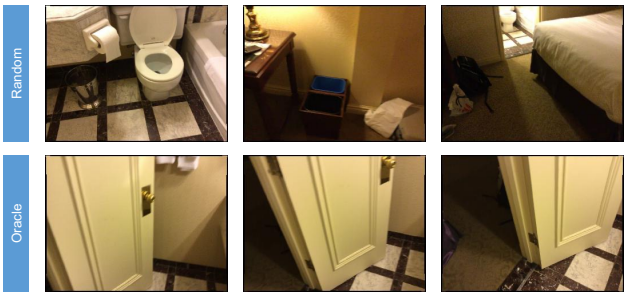

Figure 6. Example of real images about the question "What color is the bathroom door?" The upper panel is RandomImage, and the lower panel is OracleImage.

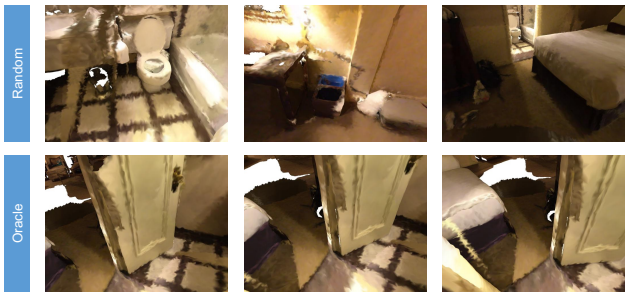

Figure 7. Example of mesh images about a question "What color is the bathroom door?" The upper panel is RandomImage, and the lower panel is OracleImage.

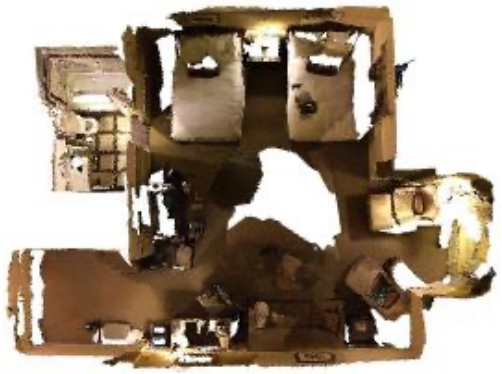

Figure 8. Example of a TopDownImage about the question "What color is the bathroom door?"

B. Additional Quantitative Experiments

B.1. Object Localization Results

Before introducing the additional QA results, we demonstrated the object localization performance using ScanRefer+MCAN (pipeline), ScanRefer+MCAN (end-to-end), and ScanQA. We used the Acc@K, where the positive predictions had a higher IoU with the ground truths than the threshold K (set to 0.25 and 0.5) [10]. As shown in Table 5, we refer to each accuracy as Acc@0.25 and Acc@0.5 in our

Model	Acc@0.25	Acc@0.5
Valid		
ScanRefer+MCAN (pipeline)	12.88	9.13
ScanRefer+MCAN (end-to-end)	23.53	11.76
ScanQA	**24.96**	**15.42**
Test w/ objects		
ScanRefer+MCAN (pipeline)	12.94	8.02
ScanRefer+MCAN (end-to-end)	21.97	10.41
ScanQA	**25.44**	**15.03**

Table 6. Object localization performance on the ScanQA dataset

experiments. The results in Table 6 show that the shared and end-to-end learning of QA, object localization, and object classification modules effectively predicted the target object for a given question.

B.2. Question Answering Results

Table 8 shows the performance of additional baselines and proposed models on the ScanQA dataset. The best results in each column are indicated in bold. The results show that our ScanQA (single or multiple) models outperformed the baselines RandomImage + MCAN (mesh), RandomImage + Oscar (mesh), TopDownImage + MCAN, and TopDownImage + Oscar across all evaluation metrics on all splits. While RandomImage + Oscar (real) outperformed ours on EM@1 on the test without objects split owing to its effective pretrained model, the ScanQA models outperformed RandomImage + Oscar (real) on other evaluation metrics. Regarding the difference using real and mesh images, the performance tended to be better when using real images. For example, RandomImage + MCAN (real) outperformed RandomImage + MCAN (mesh) and RandomImage + Oscar (real) outperformed RandomImage + Oscar (mesh) in almost all evaluation measures on all splits. We observed no consistency in the advantage regarding the performance difference between ScanQA (single) and ScanQA (multiple).

B.3. Ablation Studies on ScanQA (multiple)

In this section, we describe ablation studies conducted on the ScanQA (multiple) model that predicted confidences and labels for multiple objects when performing QA. The effect of each major component of the proposed method is shown in Table 9, and the effect of different input data is shown in Table 7. The results in Table 9 suggest that using the object localization module (LOC) or the object classification module (OBJ) was effective for improving QA performance. Table 7 shows the object localization performance Top10-Acc@0.25 and QA performance EM@10 on ScanQA (multiple) using different input features, where Top10-Acc@0.25 is the accuracy at which the positive predictions had higher IoU with the ground truth than 0.25 (we compare the top 10 object boxes with the highest object localization scores with the ground true boxes and consider positive predictions for the box with the highest IoU.) We observed that RGB values were effective for QA, and the multiview image features were effective for both object localization and QA. We assumed that by predicting multiple objects, ScanQA (multiple) could utilize those various features.

Model	Top10-Acc@0.25	EM@10
Valid		
ScanQA (xyz)	67.83	49.58
ScanQA (xyz+rgb)	66.72	50.22
ScanQA (xyz+rgb+normal)	68.13	49.45
ScanQA (xyz+multiview)	67.85	49.86
ScanQA (xyz+multiview+normal)	**70.82**	**51.23**
Test w/ objects		
ScanQA (xyz)	68.23	55.18
ScanQA (xyz+rgb)	68.87	**56.23**
ScanQA (xyz+rgb+normal)	69.65	55.25
ScanQA (xyz+multiview)	70.76	55.65
ScanQA (xyz+multiview+normal)	**71.58**	55.99

Table 7. Feature ablation results on ScanQA (multiple)

B.4. Performance by Parameter

We also evaluated the performance of ScanQA with different parameters, the number of layers L, and the hidden size d in Tables 10 and 11, respectively. The results suggest that the number of layers $L = 1$ and hidden size $d = 128$ were suitable for the test splits.

B.5. Accuracy for Question Types

We classify the questions into six types: *object*, *color*, *object nature*, *place*, *number* and *other* by the beginning of the question sentences following Table 12. We evaluate the detailed question answering performance for each class for 2D-QA baselines and the ScanQA model. Table 13, Table 14 and Table 15 presents the detailed accuracy on valid, test w/ object and test w/o object respectively.

We notice that the performance scores for *place* questions is lower than other questions. We assume there are two reasons for this. First, there are several ways to answer the questions of *place*. Model predicts longer answers and therefor image captioning-based metrics are suitable rather than exact matching for *place* questions. For *color* questions, possible answers are limited and simple 2D-QA model can answer the questions from the sights of objects from several images. However, such answers are not grounded to objects in the questions.

C. Additional Qualitative Analysis

We conducted a qualitative evaluation by comparing our ScanQA model to the ScanRefer + MCAN (end-to-end) in addition to the evaluation on ScanRefer + MCAN (pipeline)

Model	EM@1	EM@10	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr	SPICE
Valid										
RandomImage+MCAN (real)	19.19	48.15	23.71	15.41	11.81	0.00	28.90	10.92	53.83	9.35
RandomImage+MCAN (mesh)	18.59	47.81	22.12	14.49	11.72	7.69	27.61	10.32	50.93	8.37
RandomImage+Oscar (real)	19.38	46.37	22.91	14.52	11.20	0.56	28.70	10.66	52.86	8.91
RandomImage+Oscar (mesh)	17.97	43.55	20.67	12.01	11.51	0.57	26.51	9.82	47.82	7.75
TopDownImage+MCAN	12.71	41.50	14.82	8.21	16.54	0.74	19.33	7.57	33.14	5.72
TopDownImage+Oscar	17.20	43.81	19.75	11.21	15.31	0.71	25.39	9.43	45.21	7.32
VoteNet+MCAN	17.33	45.54	28.09	16.72	10.75	6.24	29.84	11.41	54.68	10.65
ScanRefer+MCAN (pipeline, real)	14.37	44.12	17.02	10.17	15.77	0.72	22.02	8.45	38.73	6.66
ScanRefer+MCAN (pipeline, mesh)	14.57	43.27	16.71	9.71	13.62	0.64	21.82	8.32	38.35	6.49
ScanRefer+MCAN (e2e)	18.59	46.76	26.93	16.59	11.59	7.87	30.03	11.52	55.41	11.28
ScanQA (single)	20.28	50.01	29.47	19.84	14.65	9.55	32.37	12.60	61.66	11.86
ScanQA (multiple)	**21.05**	**51.23**	**30.24**	**20.40**	**15.11**	**10.08**	**33.33**	**13.14**	**64.86**	**13.43**
OracleImage+MCAN (real)	22.59	49.43	26.58	18.32	15.37	8.50	33.23	12.45	63.44	12.56
OracleImage+MCAN (mesh)	20.66	48.04	24.35	17.00	14.23	0.00	30.34	11.23	57.01	10.15
OracleImage+Oscar (real)	21.39	45.05	24.29	14.19	14.21	0.67	31.24	11.33	58.23	10.51
OracleImage+Oscar (mesh)	22.27	46.59	23.01	13.96	14.23	0.00	31.37	11.53	57.98	11.11
Test w/ objects										
RandomImage+MCAN (real)	22.31	53.11	26.66	18.49	16.16	14.26	31.27	12.13	60.37	9.05
RandomImage+MCAN (mesh)	21.74	52.41	24.86	17.49	15.33	15.19	30.00	11.55	57.55	8.73
RandomImage+Oscar (real)	22.65	52.35	24.74	14.42	9.85	0.00	30.81	11.59	57.72	8.52
RandomImage+Oscar (mesh)	20.92	49.22	23.18	12.26	9.06	0.48	28.95	10.86	53.11	7.12
TopDownImage+MCAN	15.76	47.15	16.52	8.59	0.00	0.00	21.49	8.37	38.55	5.34
TopDownImage+Oscar	20.76	49.26	22.19	11.02	9.30	0.49	28.25	10.53	51.82	6.83
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, real)	17.52	49.92	19.17	10.66	0.00	0.00	24.40	9.38	44.25	6.24
ScanRefer+MCAN (pipeline, mesh)	17.44	48.83	18.45	9.49	0.00	0.00	23.90	9.11	42.97	5.93
ScanRefer+MCAN (e2e)	20.56	52.35	27.85	17.27	11.88	7.46	30.68	11.97	57.36	10.58
ScanQ (single)	**23.45**	**56.51**	**31.56**	**21.39**	**15.87**	**12.04**	**34.34**	13.55	**67.29**	11.99
ScanQA (multiple)	23.05	55.99	31.40	21.18	15.82	11.70	34.05	**13.60**	66.76	**12.30**
OracleImage+MCAN (real)	25.34	55.93	28.70	20.11	16.78	12.89	34.59	13.42	67.24	11.93
OracleImage+MCAN (mesh)	23.35	53.05	25.90	17.15	13.36	10.94	31.66	12.08	60.64	9.01
OracleImage+Oscar (real)	25.30	50.12	26.38	14.10	8.70	0.47	33.72	12.47	63.31	9.48
OracleImage+Oscar (mesh)	25.52	52.39	25.17	14.13	10.94	0.00	33.51	12.68	63.13	10.52
Test w/o objects										
RandomImage+MCAN (real)	20.82	51.23	26.29	17.90	14.27	9.66	29.23	11.54	55.64	8.87
RandomImage+MCAN (mesh)	20.34	51.20	24.72	16.93	14.04	7.55	28.10	10.95	53.41	8.61
RandomImage+Oscar (real)	**21.58**	49.85	24.86	16.16	13.29	0.00	28.99	10.99	54.62	8.62
RandomImage+Oscar (mesh)	19.91	48.74	23.02	13.24	12.91	0.64	26.94	10.17	49.83	7.45
TopDownImage+MCAN	14.39	45.94	15.70	8.86	12.56	0.62	19.26	7.71	34.59	5.22
TopDownImage+Oscar	19.13	48.06	21.93	11.66	14.64	0.70	25.98	9.67	47.37	6.93
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, real)	16.47	49.05	18.71	10.97	16.53	0.76	22.45	8.76	40.81	6.41
ScanRefer+MCAN (pipeline, mesh)	16.39	47.76	18.28	10.42	15.03	0.71	22.07	8.62	40.19	6.03
ScanRefer+MCAN (e2e)	19.04	49.70	26.98	16.17	11.28	7.82	28.61	11.38	53.41	10.63
ScanQA (single)	20.90	**54.11**	30.68	**21.20**	**15.81**	10.75	31.09	12.59	60.24	11.29
ScanQA (multiple)	21.30	53.05	**31.14**	**21.20**	**15.81**	**11.18**	**31.62**	**12.82**	**60.95**	**11.68**

Table 8. Performance comparison for question answering with image captioning metrics

ANS	OBJ	LOC	EM@1	EM@10	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr	SPICE
Valid												
✓			7.53	27.70	8.30	6.03	0.15	0.02	10.14	4.47	21.04	2.21
✓	✓		19.87	49.43	29.26	18.75	13.03	7.30	31.98	12.28	60.63	11.98
✓		✓	20.47	50.05	28.34	19.03	14.16	9.95	31.91	12.25	61.02	11.69
✓	✓	✓	**21.05**	**51.23**	**30.24**	**20.40**	**15.11**	**10.08**	**33.33**	**13.14**	**64.86**	**13.43**
Test w/ objects												
✓			8.20	32.68	8.10	5.57	0.14	0.02	10.09	4.29	21.02	1.89
✓	✓		22.07	55.10	30.44	19.85	14.33	10.18	32.89	13.06	63.93	11.51
✓		✓	**23.77**	55.31	30.65	**21.42**	**16.61**	**13.37**	**34.15**	13.40	66.69	10.88
✓	✓	✓	23.05	**55.99**	**31.40**	21.18	15.82	11.70	34.05	**13.60**	**66.76**	**12.30**
Test w/o objects												
✓			8.41	31.42	8.40	5.48	0.14	0.02	10.32	4.35	20.64	1.68
✓	✓		20.80	53.70	**31.30**	20.51	14.58	10.48	31.51	12.75	60.09	**11.75**
✓		✓	20.90	**53.78**	30.01	**22.39**	**17.98**	**13.79**	30.69	12.51	60.37	11.34
✓	✓	✓	**21.30**	53.05	31.14	21.20	15.81	11.18	**31.62**	**12.82**	**60.95**	11.68

Table 9. Performance comparison between the different experimental conditions of the ScanQA (multiple) model

Model	EM@1	EM@10	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr	SPICE
Valid										
ScanQA ($L = 1$)	19.96	49.78	**29.49**	19.16	13.23	8.44	32.35	12.59	61.14	**12.61**
ScanQA ($L = 2$)	**20.28**	**50.01**	29.47	**19.84**	**14.65**	**9.55**	**32.37**	**12.60**	**61.66**	11.86
ScanQA ($L = 3$)	11.94	39.14	13.93	8.82	7.23	0.00	18.64	6.98	33.30	6.55
Test w/ objects										
ScanQA ($L = 1$)	**23.83**	55.63	**32.64**	**21.80**	15.63	11.67	**35.20**	**14.15**	**69.70**	**12.65**
ScanQA ($L = 2$)	23.45	**56.51**	31.56	21.39	**15.87**	**12.04**	34.34	13.55	67.29	11.99
ScanQA ($L = 3$)	13.83	44.71	15.05	7.90	5.84	0.00	19.62	7.34	36.38	5.62
Test w/o objects										
ScanQA ($L = 1$)	**21.01**	52.50	**31.23**	**21.37**	**15.97**	**11.20**	**31.55**	**12.84**	**61.11**	**11.82**
ScanQA ($L = 2$)	20.90	**54.11**	30.68	21.20	15.81	10.75	31.09	12.59	60.24	11.29
ScanQA ($L = 3$)	11.63	40.30	13.42	7.75	5.96	0.00	16.75	6.43	29.86	5.13

Table 10. Performance comparison for the ScanQA model with difference number of layers L

Model	EM@1	EM@10	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr	SPICE
Valid										
ScanQA ($d = 128$)	**20.88**	**50.70**	**30.08**	**20.62**	**15.72**	**11.18**	**33.25**	**12.97**	**64.09**	**12.77**
ScanQA ($d = 256$)	20.28	50.01	29.47	19.84	14.65	9.55	32.37	12.60	61.66	11.86
ScanQA ($d = 512$)	13.99	41.54	17.01	11.02	8.26	0.00	21.97	8.11	38.78	6.89
Test w/ objects										
ScanQA ($d = 128$)	**24.38**	**56.71**	**32.30**	**22.47**	**17.98**	**14.96**	**35.24**	**14.07**	**69.53**	**12.61**
ScanQA ($d = 256$)	23.45	56.51	31.56	21.39	15.87	12.04	34.34	13.55	67.29	11.99
ScanQA ($d = 512$)	14.75	46.02	16.89	9.18	7.06	6.77	21.17	7.76	38.56	5.47
Test w/o objects										
ScanQA ($d = 128$)	**21.17**	**54.20**	**31.79**	**22.23**	**16.65**	**11.32**	**31.83**	**13.01**	**61.92**	**12.13**
ScanQA ($d = 256$)	20.90	54.11	30.68	21.20	15.81	10.75	31.09	12.59	60.24	11.29
ScanQA ($d = 512$)	12.83	43.28	16.67	9.56	7.01	4.70	18.97	7.31	33.63	5.08

Table 11. Performance comparison for the ScanQA model with difference hidden size d

(Fig. 5). Fig. 9 shows the results of the object localization and QA correctness with the visualization of the scene and bounding boxes using ScanRefer + MCAN (end-to-end), ScanQA, and ground truth. The results also indicated that QA correctness and localization were closely related. The leftmost case shows that models were to answer the color of the bathrobe. However, the baseline model, ScanRefer + MCAN (end-to-end), localized a different object "table" and answered its color. The baseline model used the word "wall" to determine an object near the wall and localized it. However, our model could localize the correct object, "bathrobe," on the wall. In the second case from the left, the question was about the object placed on the chair, but the baseline model provided the wrong answer "table" because the word "seat" is frequently associated with the table. In contrast, our model understood the meaning "in the seat" and correctly selected a backpack in the seat. In the second case from the right, the baseline model incorrectly localized a TV close to the wall and answered "chair" close to the TV. Our model correctly recognized the meaning "corner," localized it, and indicated the correct object "trash can" at the corner. In the final case, there were multiple ottomans in the scene, and the models were to correctly understand the positional relationship between the ottoman, chair, and pillow. The baseline model localized the lamp and incorrectly answered its color "white." In contrast, our model correctly understood the positional relationship between the ottoman, chair, and pillow, localizing the ottoman in front of the chair with a pillow.

D. MTurk Annotation Details

We developed a visualizer website for 3D modeling. MTurk workers can interactively rotate and zoom in 3D modeling. Fig. 10 presents the snapshot of the MTurk website for the editing and answer collection phrase of the QA annotation.

We filtered the auto-generated questions as follows. We first applied rule-based filtering for removing potential underspecified or noisy expressions such as "this" and "image" and direction words, namely "north," "west," "south," and "east." We also filtered odd question types, such as "What is the name of..." Furthermore, we filtered meaningless questions in 3D scenes using MTurk. We asked three workers to evaluate each question and filtered those questions that were evaluated as "valid" in the scene by at least two of the three workers. Finally, we asked workers to rewrite those questions that, according to them, were underspecified before they filled out the answers.

We consider that our dataset covers a broad range of

Question type	Question beginning	# Instances in the valid set
Object	*What is* (except questions classified as color)	1476
Color	*What color* *What is the color*	838
Object nature	*What type* *What shape* *What kind*	358
Place	*Where is*	963
Number	*How many*	224
Other	(remaining questions)	814

Table 12. Question type and the beginning of the question sentence

Model	EM@1	EM@10	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr	SPICE
Object										
RandomImage+MCAN	15.02	45.06	20.33	16.22	**17.13**	**0.79**	43.03	9.18	22.18	14.03
VoteNet+MCAN	12.31	41.07	20.81	13.51	6.47	0.00	39.81	8.88	21.37	14.09
ScanRefer+MCAN (pipeline)	11.84	38.09	15.91	12.73	0.23	0.03	32.64	7.37	17.49	12.40
ScanRefer+MCAN (e2e)	14.82	42.76	20.43	15.57	8.19	0.00	41.74	9.00	22.12	14.87
ScanQA (single)	17.52	45.47	23.94	18.19	0.00	0.00	50.05	10.62	26.01	17.57
ScanQA (multiple)	**18.27**	**47.70**	**26.15**	**19.19**	14.46	0.00	**53.84**	**11.53**	**27.52**	**18.46**
Color										
RandomImage+MCAN	44.03	**86.04**	45.92	31.38	0.44	**0.05**	86.65	23.57	49.36	2.19
VoteNet+MCAN	41.65	81.62	**47.01**	28.19	**20.58**	0.01	86.54	23.32	49.01	**3.09**
ScanRefer+MCAN (pipeline)	30.79	83.41	32.75	0.05	0.01	0.00	59.01	16.14	34.87	0.00
ScanRefer+MCAN (e2e)	**44.99**	83.77	46.72	**35.91**	0.00	0.00	**87.75**	**24.16**	**50.13**	1.85
ScanQA (single)	42.60	83.53	43.92	29.48	0.00	0.00	84.42	22.61	47.68	1.39
ScanQA (multiple)	42.60	85.44	45.23	17.78	0.00	0.00	85.20	22.76	48.34	2.41
Object nature										
RandomImage+MCAN	18.16	47.49	31.06	26.40	**26.83**	**1.07**	57.15	13.18	33.69	7.11
VoteNet+MCAN	17.04	47.77	35.70	23.15	16.71	0.00	62.83	14.66	35.21	16.42
ScanRefer+MCAN (pipeline)	15.36	43.30	20.14	16.72	18.38	0.77	38.60	9.25	27.09	2.27
ScanRefer+MCAN (e2e)	13.41	48.32	36.50	20.03	16.41	0.00	58.03	14.50	34.75	16.08
ScanQA (single)	18.44	51.40	**41.65**	25.65	18.81	0.00	73.26	16.54	**41.61**	17.16
ScanQA (multiple)	**20.11**	**54.75**	41.34	**29.72**	25.80	0.01	**78.31**	**17.31**	40.54	**18.67**
Place										
RandomImage+MCAN	3.95	17.24	17.87	11.59	8.29	0.00	33.58	8.05	19.00	11.55
VoteNet+MCAN	4.88	18.48	28.31	17.84	12.07	7.47	45.08	9.97	26.41	14.68
ScanRefer+MCAN (pipeline)	2.08	11.53	8.26	5.08	0.11	0.02	14.34	5.29	9.65	5.74
ScanRefer+MCAN (e2e)	4.98	17.86	25.31	15.64	11.20	8.01	44.72	9.87	25.06	15.98
ScanQA (single)	6.85	23.16	**28.78**	**19.48**	**14.41**	9.55	57.00	**11.49**	28.19	16.66
ScanQA (multiple)	**7.79**	**23.99**	28.21	19.13	14.27	**9.92**	**59.34**	11.46	**28.30**	**18.13**
Number										
RandomImage+MCAN	39.29	**86.16**	45.01	0.06	0.01	0.00	72.26	19.55	46.70	0.00
VoteNet+MCAN	36.16	**86.16**	43.77	20.59	0.33	0.04	67.56	18.13	44.02	0.40
ScanRefer+MCAN (pipeline)	38.39	85.71	44.26	0.06	0.01	0.00	67.76	19.37	45.69	0.00
ScanRefer+MCAN (e2e)	34.82	84.82	39.01	0.06	0.01	0.00	61.06	17.01	40.78	0.00
ScanQA (single)	39.29	85.71	44.29	0.00	0.00	0.00	72.15	19.16	46.05	0.00
ScanQA (multiple)	**40.62**	85.71	**46.00**	**44.09**	**0.55**	**0.06**	**75.68**	**19.80**	**47.72**	**0.45**
Other										
RandomImage+MCAN	14.13	41.15	18.93	13.19	**11.63**	**0.54**	42.28	8.83	24.75	9.17
VoteNet+MCAN	11.06	36.36	20.87	7.91	0.00	0.00	36.71	8.61	23.29	7.68
ScanRefer+MCAN (pipeline)	10.69	37.22	17.28	10.45	0.18	0.02	33.89	8.12	21.74	6.98
ScanRefer+MCAN (e2e)	12.16	38.94	21.10	12.82	8.29	0.00	42.69	9.32	24.54	9.91
ScanQA (single)	14.13	43.37	22.26	13.72	8.02	0.00	45.39	9.96	26.30	10.78
ScanQA (multiple)	**14.62**	**43.61**	**23.52**	**15.64**	9.61	0.00	**47.89**	**10.39**	**27.26**	**11.34**

Table 13. Valid set performance comparison for question answering with image captioning metrics. **e2e** represents an end-to-end model.

questions with distinct meanings as they are constructed through human (re)annotation; they reflect various questions that humans may ask. The number of unique answers is also high corresponding to the number of unique questions. We noticed that the question length in terms of tokens had a fat tail distribution. Interestingly, there are both very short and long questions in the dataset such as "How many chairs?" and "What color is the chair that is located to the right of another brown chair with a red bag on it?"

Model	EM@1	EM@10	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr	SPICE
Object										
RandomImage+MCAN	16.47	46.59	22.68	18.15	0.00	0.00	47.16	9.69	23.49	15.65
VoteNet+MCAN	13.23	43.84	21.32	13.72	7.47	5.36	40.98	8.80	21.05	14.25
ScanRefer+MCAN (pipeline)	12.74	41.24	15.87	12.36	0.23	0.03	33.82	7.10	17.32	13.24
ScanRefer+MCAN (e2e)	15.97	46.24	21.12	15.76	6.96	0.00	43.01	9.13	22.28	15.87
ScanQA (single)	18.86	49.96	24.75	18.78	13.73	10.62	51.81	10.68	26.00	17.09
ScanQA (multiple)	**19.14**	**52.15**	**26.78**	**20.97**	**17.86**	**17.64**	**55.24**	**11.25**	**26.98**	**18.47**
Color										
RandomImage+MCAN	45.56	**89.56**	49.39	**50.81**	0.62	**0.07**	91.84	25.87	50.08	2.08
VoteNet+MCAN	40.57	84.66	48.05	23.43	**11.85**	0.00	85.40	23.56	47.41	**2.28**
ScanRefer+MCAN (pipeline)	31.24	86.69	36.03	0.06	0.01	0.00	62.00	17.61	35.67	0.00
ScanRefer+MCAN (e2e)	41.96	88.08	47.76	27.33	0.00	0.00	84.52	24.04	47.74	2.22
ScanQA (single)	**45.75**	88.45	**50.27**	22.71	0.00	0.00	**92.54**	**26.06**	**50.96**	2.10
ScanQA (multiple)	43.25	88.54	48.71	12.31	0.00	0.00	88.00	24.53	48.83	2.14
Object nature										
RandomImage+MCAN	19.91	50.98	35.93	25.12	**23.14**	**0.98**	56.79	13.73	35.93	6.54
VoteNet+MCAN	19.91	49.45	38.95	22.65	0.00	0.00	61.17	14.90	36.42	**18.18**
ScanRefer+MCAN (pipeline)	19.69	46.17	27.85	23.18	0.00	0.00	47.31	11.42	33.08	2.33
ScanRefer+MCAN (e2e)	19.04	50.33	39.23	23.34	17.01	0.00	61.06	14.71	35.99	15.24
ScanQA (single)	**22.98**	**56.67**	45.09	28.98	21.27	0.96	**72.81**	16.85	41.52	14.79
ScanQA (multiple)	22.76	55.14	**45.76**	**29.72**	18.14	0.00	72.16	**17.15**	**41.77**	16.95
Place										
RandomImage+MCAN	3.97	16.10	19.50	14.32	13.07	11.67	39.74	8.68	19.87	10.97
VoteNet+MCAN	4.43	17.62	28.82	18.10	11.84	7.05	46.59	10.31	26.70	15.02
ScanRefer+MCAN (pipeline)	1.40	9.68	7.33	3.65	0.09	0.01	11.45	5.11	8.52	4.45
ScanRefer+MCAN (e2e)	4.08	18.20	24.31	15.62	11.18	7.35	43.72	9.64	24.54	14.32
ScanQA (single)	**7.12**	**22.17**	**29.59**	**20.93**	**16.47**	**12.59**	**61.89**	**12.06**	**29.35**	**18.17**
ScanQA (multiple)	5.95	21.94	26.89	18.87	14.53	10.60	55.13	11.11	27.10	16.74
Number										
RandomImage+MCAN	**43.14**	90.20	44.97	0.06	0.01	0.00	**83.02**	23.81	46.94	0.00
VoteNet+MCAN	39.61	**90.98**	44.01	16.07	0.29	0.04	75.94	21.71	44.29	**0.72**
ScanRefer+MCAN (pipeline)	43.53	89.41	**45.32**	0.06	0.01	0.00	81.52	**23.89**	**47.48**	0.00
ScanRefer+MCAN (e2e)	34.12	87.84	36.29	0.00	0.00	0.00	63.70	18.28	37.43	0.00
ScanQA (single)	40.78	90.59	44.94	0.00	0.00	0.00	77.33	22.44	45.80	0.00
ScanQA (multiple)	37.25	90.59	40.25	**22.86**	**0.36**	**0.05**	72.15	20.07	41.37	0.39
Other										
RandomImage+MCAN	16.37	45.46	19.02	12.93	0.00	0.00	42.83	9.01	24.99	9.02
VoteNet+MCAN	13.72	41.81	20.85	9.92	3.53	0.00	41.08	8.81	24.41	8.72
ScanRefer+MCAN (pipeline)	15.04	42.48	18.00	10.61	0.19	0.03	37.78	8.51	23.46	6.42
ScanRefer+MCAN (e2e)	14.71	42.59	21.09	11.01	7.21	0.00	41.46	9.23	24.70	9.36
ScanQA (single)	**17.92**	**46.79**	**24.50**	**15.76**	**9.92**	**7.06**	**50.88**	**10.66**	**28.78**	**11.69**
ScanQA (multiple)	17.37	46.02	23.94	14.30	8.00	0.00	50.63	10.55	28.06	11.58

Table 14. Test w/ object set performance comparison for question answering with image captioning metrics. **e2e** represents an end-to-end model.

Model	EM@1	EM@10	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr	SPICE
Object										
RandomImage+MCAN	15.85	44.84	22.37	17.97	15.07	**0.73**	41.56	9.29	22.39	14.41
VoteNet+MCAN	13.80	42.52	20.65	12.39	5.86	0.00	37.06	8.71	20.69	14.79
ScanRefer+MCAN (pipeline)	12.48	41.14	17.70	14.87	0.27	0.04	32.31	7.44	17.52	12.95
ScanRefer+MCAN (e2e)	14.30	44.56	20.26	15.31	9.77	0.00	37.90	8.67	20.56	15.61
ScanQA (single)	17.39	**49.75**	24.79	18.87	11.47	0.00	46.97	10.36	24.89	16.73
ScanQA (multiple)	**18.83**	48.48	**25.96**	**21.26**	**16.90**	0.00	**49.72**	**11.00**	**25.95**	**17.00**
Color										
RandomImage+MCAN	**43.31**	**89.38**	**45.74**	0.00	0.00	0.00	**87.49**	**24.55**	**46.66**	0.90
VoteNet+MCAN	38.62	83.00	44.74	**15.35**	**11.79**	**0.63**	79.21	22.16	44.35	**1.03**
ScanRefer+MCAN (pipeline)	33.46	87.77	36.88	0.06	0.01	0.00	67.21	18.88	36.90	0.00
ScanRefer+MCAN (e2e)	40.15	85.23	44.53	0.00	0.00	0.00	81.64	22.92	44.72	0.71
ScanQA (single)	40.00	87.85	43.58	0.00	0.00	0.00	80.77	22.60	43.89	0.54
ScanQA (multiple)	40.92	87.46	44.15	14.64	0.00	0.00	82.22	22.96	44.62	0.77
Object nature										
RandomImage+MCAN	15.17	41.67	29.06	25.84	23.07	0.98	47.21	11.46	27.54	7.27
VoteNet+MCAN	10.68	41.45	33.29	18.02	11.91	0.00	45.85	12.37	28.44	15.82
ScanRefer+MCAN (pipeline)	13.46	39.74	20.73	22.67	**25.65**	**1.03**	35.67	8.29	22.18	3.46
ScanRefer+MCAN (e2e)	13.25	39.32	34.51	19.23	13.25	0.00	51.52	12.96	30.13	18.28
ScanQA (single)	**15.60**	46.15	**41.96**	**28.75**	19.96	0.92	**65.48**	**15.62**	**35.60**	**20.89**
ScanQA (multiple)	13.68	**47.65**	41.40	25.58	20.39	0.01	59.70	14.75	34.43	19.06
Place										
RandomImage+MCAN	4.76	19.34	21.88	14.48	11.08	7.76	41.12	9.18	22.08	12.85
VoteNet+MCAN	5.85	20.20	32.15	20.62	13.86	8.80	50.54	11.16	29.06	16.43
ScanRefer+MCAN (pipeline)	1.64	11.93	8.87	4.58	0.11	0.02	12.82	5.34	9.14	5.42
ScanRefer+MCAN (e2e)	4.99	21.14	24.47	15.03	10.76	7.95	44.46	9.57	23.99	15.05
ScanQA (single)	**6.79**	**26.37**	**32.47**	**23.04**	**18.34**	**13.74**	**63.40**	**12.27**	**29.67**	**19.28**
ScanQA (multiple)	**6.79**	24.57	30.18	20.39	15.19	10.81	56.65	11.37	27.97	17.71
Number										
RandomImage+MCAN	**42.51**	**90.64**	45.13	0.07	0.01	0.00	**77.60**	**22.40**	**45.21**	0.00
VoteNet+MCAN	33.69	**90.64**	40.35	25.05	0.40	**0.05**	63.81	18.53	38.63	**0.27**
ScanRefer+MCAN (pipeline)	41.44	**90.64**	44.12	0.06	0.01	0.00	74.31	22.01	44.17	0.00
ScanRefer+MCAN (e2e)	37.97	89.30	41.12	0.00	0.00	0.00	70.08	19.69	40.83	0.00
ScanQA (single)	36.90	**90.64**	42.69	28.69	0.43	**0.05**	68.30	19.57	41.43	0.00
ScanQA (multiple)	40.91	**90.64**	**46.20**	**29.85**	**0.44**	**0.05**	77.26	21.84	45.16	0.00
Other										
RandomImage+MCAN	15.21	**43.11**	18.74	**14.85**	**15.93**	0.72	**40.50**	8.65	22.32	8.13
VoteNet+MCAN	12.36	37.75	20.15	9.26	3.22	0.00	33.82	8.25	20.73	7.64
ScanRefer+MCAN (pipeline)	11.82	40.81	16.74	11.26	0.00	0.00	32.49	7.23	19.04	5.50
ScanRefer+MCAN (e2e)	13.35	38.51	20.31	9.82	6.21	0.00	37.70	8.73	22.34	9.12
ScanQA (single)	**15.32**	41.79	**21.99**	12.74	6.92	0.00	40.30	**9.30**	**23.99**	**10.33**
ScanQA (multiple)	14.55	40.48	20.69	12.93	8.57	**6.81**	38.74	8.69	22.53	9.22

Table 15. Test w/o object set performance comparison for question answering with image captioning metrics. **e2e** represents an end-to-end model.

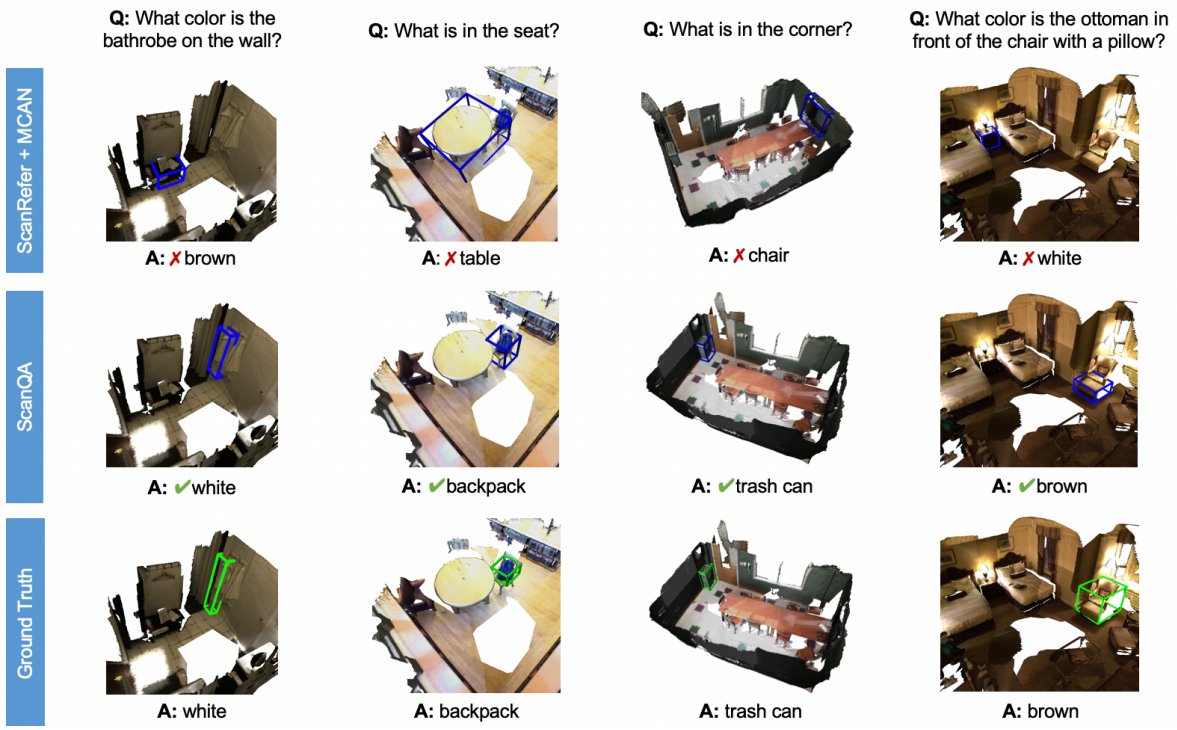

Figure 9. Qualitative analysis comparing ScanQA and ScanRefer + MCAN (end-to-end).

Figure 10. Example of 3D modeling viewer and QA form on the MTurk website.