Generating Context-Aware Natural Answers for Questions in 3D Scenes Supplementary Material

Mohammed Munzer Dwedari
munzer.dwedari@tum.de
Matthias Niessner
niessner@tum.de
Zhenyu Chen
zhenyu.chen@tum.de

Visual Computing Lab
Technical University of Munich
Munich, Germany

This supplementary material provides additional experiment results and evaluations, such as the performance of the SoftGroup [5] backbone trained on ScanRefer [2] classes (Section A.1.). We also include the question-answering scores on the different types of questions in comparison to ScanQA [1] (Section A.2.). Apart from that, we show additional ablation studies in Section B and further qualitative analysis results in Section C.

A Additional Quantitative Analysis Results

A.1. SoftGroup Trained on ScanRefer Classes

We show our evaluation results (Table 1) of SoftGroup [5] trained on ScanNet [3] scenes with different input features with ScanRefer [2] object classes. We see that having RGB and normals features yields the best overall scores.

A.2. Question Types

We compare our results with the ScanQA [1] baseline on the different types of questions in the validation set (Table 2). Since the question types split is not publicly available, we split the validation set based on the beginning words of every question, as mentioned in the ScanQA [1] paper. With that, we get the same number of questions as ScanQA [1] for each type. Overall, our model outperforms the baseline in all question types on all image

Point cloud features	AP	AP 50%	AP 25%	Bbox AP 50%	Bbox AP 25%	AR	RC 50%	RC 25%
xyz	40.4	60.6	72.1	54.3	66.8	49.8	72.1	83.6
xyz + rgb	40.6	60.9	74.2	53.5	68.1	49.7	71.6	**84.3**
xyz + rgb + normals	**42.0**	**62.2**	**74.5**	**57.1**	**69.3**	**51.3**	**73.6**	83.8

Table 1: Evaluation scores of SoftGroup [5] trained with ScanRefer [2] object classes. We report our scores on the ScanNet [3] validation set.

Model	BLEU-1	BLEU-4	ROUGE	METEOR	CIDEr
Object					
ScanQA [1]	23.94	0.00	**50.05**	10.62	26.01
Gen3DQA	**27.27**	0.00	27.23	**11.97**	**55.13**
Color					
ScanQA [1]	43.92	0.00	**84.42**	22.61	47.68
Gen3DQA	**45.76**	0.00	48.77	**22.92**	**83.22**
Object Nature					
ScanQA [1]	**41.65**	0.00	**73.26**	16.54	41.61
Gen3DQA	41.63	0.00	39.51	**17.61**	**73.72**
Place					
ScanQA [1]	28.78	9.55	**57.00**	11.49	28.19
Gen3DQA	**43.11**	**12.32**	38.32	**14.81**	**72.74**
Number					
ScanQA [1]	44.29	0.00	**72.15**	19.16	46.05
Gen3DQA	**51.97**	**0.04**	50.18	**20.99**	**74.93**
Other					
ScanQA [1]	22.26	0.00	**45.39**	9.96	26.30
Gen3DQA	**37.52**	**16.77**	30.40	**14.78**	**64.11**
Total					
ScanQA [1]	29.47	9.55	32.37	12.60	61.66
Gen3DQA	**39.53**	**12.70**	**35.97**	**15.11**	**71.97**

Table 2: Image captioning metrics scores for different types of questions in the ScanQA [1] validation set.

captioning metrics except ROUGE [4]. The biggest difference in scores can be observed in the "other" category, where our model has a BLEU-4 score of 16.77 compared to 0.00 of the baseline.

B Additional Ablation Studies

Model	BLEU-1	BLEU-4	ROUGE	METEOR	CIDEr
Gen3DQA (w/o target embeddings)	**35.4**	10.52	**33.39**	**13.62**	**64.91**
Gen3DQA (w/ target embeddings)	34.65	**11.07**	33.31	13.57	64.71

Table 3: Image captioning metrics scores of our model trained on XE loss once with and once without target embeddings. Evaluation is done on the validation set.

Do target embeddings help? Our aim in this experiment is to pass a signal from our object localization branch to the decoder by adding information about the target object proposal. Therefore, we train 0 & 1 embeddings and add the 1 embedding vector to the encoded object proposal with the highest confidence score and the 0 embedding vector to the rest. Our results in Table 3 show that there is no significant improvement when using the target embeddings. We assume the reason is the low object localization accuracy of our model (23.79 on Acc@0.5), because of which it does not get an accurate signal most of the time.

Model	BLEU-1	BLEU-4	ROUGE	METEOR	CIDEr
Gen3DQA (SCST switched)	38.25	13.01	35.36	14.82	70.96
Gen3DQA (w/o VQG)	**39.12**	**13.2**	**35.48**	**14.89**	**71.39**

Table 4: Experiment results on the validation set. Models are trained without VQG reward.

Does using beam search as a basesline for SCST help? In the SCST paper the authors use the greedy decoding output for the baseline reward. In our case, the sampled sentences are almost always worse than the ones generated by greedy decoding. As our model tries to make the reward gap positive, it becomes much worse after 5 epochs, where the CIDEr score goes below 22. Therefore, we experiment with switching the sampled sentence and the greedily generated one and report our findings in Table 4 (Gen3DQA (SCST switched)). As can be seen, using beam search for the baseline reward performs better.

C Additional Qualitative Analysis Results

In Figures 1 and 2 we show additional examples of our model compared to ScanQA [1]. We see that while our model localizes meaningful targets, it generates longer and/or better answers than ScanQA [1].

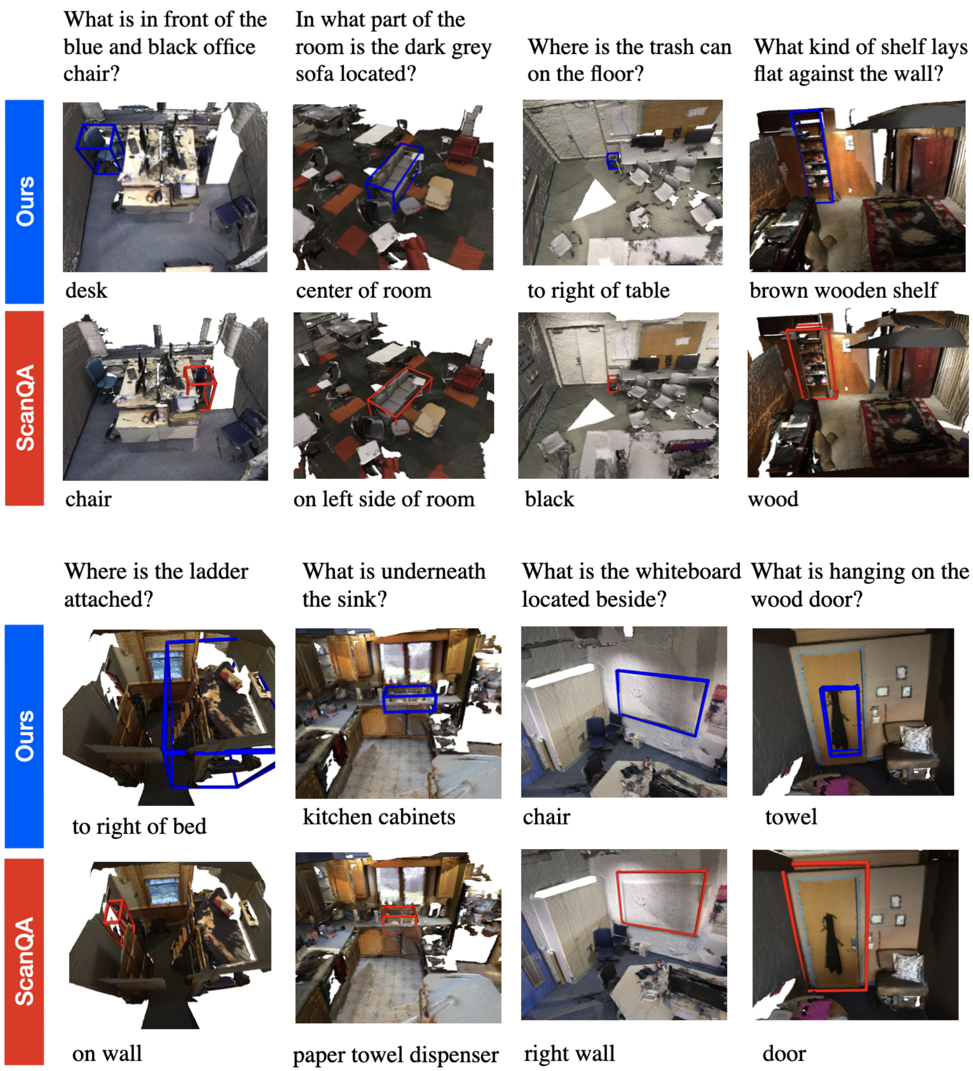

Figure 1: Example questions and answers from the test set without object IDs. We compare the results of our model (blue) to ScanQA [1] (red). Below every image is the predicted or generated answer. Since we do not axis-align our scenes, the bounding boxes in our model look tilted. Best viewed in color.

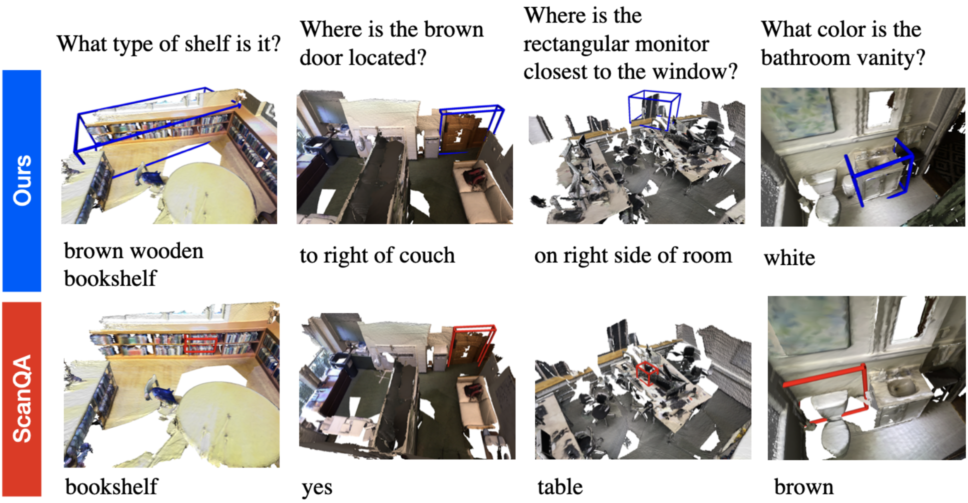

Figure 2: Example questions and answers from the test set with object IDs. We compare the results of our model (blue) to ScanQA [1] (red). Below every image is the predicted or generated answer. Since we do not axis-align our scenes, the bounding boxes in our model look tilted. Best viewed in color.

References

[1] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19129–19139, 2022.

[2] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX*, pages 202–221. Springer, 2020.

[3] 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 conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[4] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

[5] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, and Chang D Yoo. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2708–2717, 2022.