

3D Question Answering with Scene Graph Reasoning

Anonymous Author(s)

A ADDITIONAL QUANTITATIVE EXPERIMENTS

A.1 Localization Results on SQA3D

Situation Localization. We evaluate the situational localization results of our method on SQA3D dataset [3], which refers to predicting the specific location of the situation in the scene based on the input of situational description and the 3D scene. Specifically, we aim to predict the position of the current situation (s^{pos}) in 3D coordinate format ($\langle x, y, z \rangle$) and its orientation (s^{rot}) represented by an angle. We follow the same quantitative metrics defined by SQA3D for our predictions, i.e., for the target position (s^{pos}), we employ the metrics: Acc@0.5m and Acc@1m to assess whether the predicted position falls within a range of 0.5 meters and 1 meter from the ground truth, respectively. Regarding the rotation angle (s^{rot}), we utilized Acc@15° and Acc@30° to evaluate whether the predict direction is within $\pm 15^\circ$ and $\pm 30^\circ$ of the ground truth, respectively. We eventually compare the performance of our 3DGraphQA with that of SQA3D using these metrics, as shown in Table 1, where our method achieves state-of-the-art results across all metrics. Figure 1 also illustrates the the situation localization results of our method on workroom and bedroom scenes of the SQA3D dataset.

Table 1: The accuracy results of situation localization on SQA3D.

Method	Acc@0.5m	Acc@1m	Acc@15°	Acc@30°
SQA3D	14.60	34.21	22.39	42.28
3DGraphQA	16.12	37.73	22.74	44.42

Object Localization. We proceed to validate the object localization results on the SQA3D dataset, which requires predicting the location of the object based on the questions. Since the SQA3D method does not contain object localization function for questions, there are no specific quantitative metrics available. In this study, we refer to the object localization metrics used in ScanRefer, specifically, we use Acc@0.25 and Acc@0.5 as quantitative metrics, which represents the proportion of overlap between the predicted bounding boxes and the ground truth bounding boxes (IoU) within a range of 0.25 and 0.5, respectively. The specific results are presented in Table 2, from which our method also achieves state-of-the-art results across the metrics. Figure 1 illustrates the the object localization results of our method on workroom and bedroom scenes of SQA3D dataset.

Table 2: The accuracy results of object localization on SQA3D.

Method	Acc@0.25	Acc@0.5
3DGraphQA	28.82	17.11

A.2 Localization Results on ScanQA

We conduct the object localization experiments on the ScanQA dataset [1]. Since ScanQA also uses the overlap ratio between predicted bounding boxes and ground truth bounding boxes (IoU) as quantitative metrics, we similarly employ Acc@0.25 and Acc@0.5 as evaluation metrics. The comparative experimental results are presented in the table below:

Table 3: The object localization results on ScanQA.

Method	Acc@0.25	Acc@0.5
Scanrefer + MCAN	23.53	11.76
ScanQA	24.96	15.42
3DGraphQA	28.74	19.63

B ADDITIONAL EXPERIMENTAL DETAILS

B.1 Experimental Setting.

Several hyperparameters need to be set during experiments. First is the setting of the number of object proposals in VoteNet, which determines the size of the graph, i.e., how many graph nodes are included. Therefore, we experimented with values of 32, 64, 128, and 256 for the number of object proposals. The results are shown in Table 4, indicating that the best performance is achieved when the number of object proposals is set to 64. Additionally, the hidden size parameter affects the feature dimensions of various modalities and the overall model parameter size, thus influencing the experimental results. Following the settings in ScanQA, we experiment with hidden sizes of 128, 256, and 512. The results are presented in Table 5, indicating that the best performance was achieved when the hidden size is set to 256.

Table 4: Ablation on different settings of object proposals.

Method	Acc@1	Acc@10
3DGraphQA(num_proposal=32)	47.81	86.79
3DGraphQA(num_proposal=64)	49.04	88.75
3DGraphQA(num_proposal=128)	48.04	87.28
3DGraphQA(num_proposal=256)	48.74	88.24

Table 5: Ablation on different settings of hidden size.

Method	Acc@1	Acc@10
3DGraphQA(hidden_size=128)	48.52	88.22
3DGraphQA(hidden_size=256)	49.04	88.75
3DGraphQA(hidden_size=512)	47.93	86.81

C ADDITIONAL ABLATIONS

Ablation Study on Graph Edge. We also experiment with various settings for the edges in the graph. In addition to the neighborhood range settings mentioned in Section 3.2, we compare different weight settings for the edges. Initially, we set the neighborhood range to 8, and compare the effects of setting the weights to 0.5, random values between 0 and 1, and 1. The results show that the best performance is achieved when the weights are set to 1s. Furthermore, as mentioned in our method, we randomly add some edges to the graph structure to enhance its robustness, ensuring that objects are not connected to themselves. The results are shown in Table 6.

Ablation Study on Graph Reasoning. We note that graph reasoning can be implemented by various neural network architectures, such as GNN, GCN[2], and GAT[4]. When initially setting up the network architecture, we also compared GNN and GCN. In addition to the GAT mentioned in our ablation experiments, we also experimented with different numbers of layers in the GAT. Specifically, we set up 1, 2, and 3 layers for comparison of effectiveness, the results are shown in Table 7.

Table 6: Ablation study on graph edge.

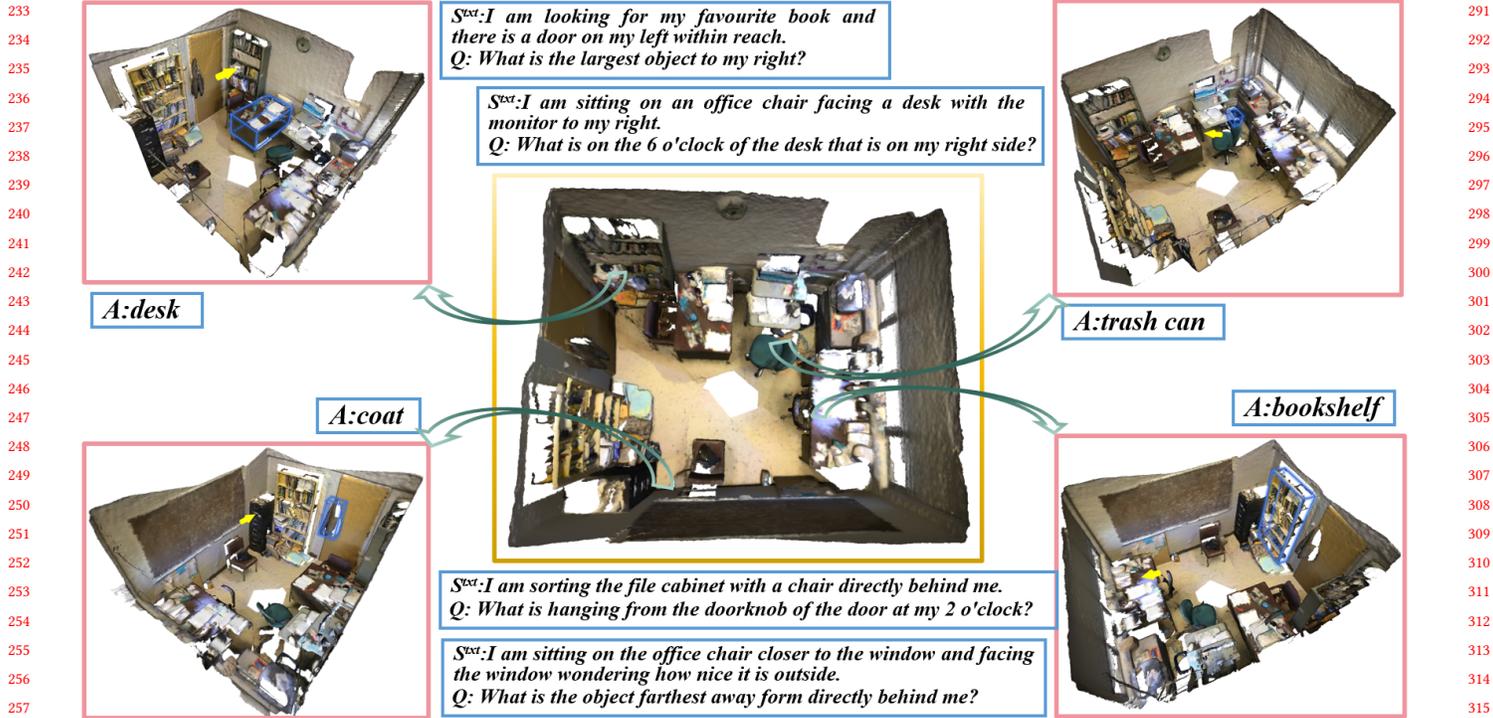
Method	Acc@1	Acc@10
3DGraphQA(edge_weight=0.5)	47.66	87.84
3DGraphQA(edge_weight=random)	48.31	87.98
3DGraphQA(edge_weight=1)	49.04	88.75
3DGraphQA(no random edges)	48.91	88.70
3DGraphQA(connect self)	48.10	88.09

Table 7: Ablation study on graph reasoning.

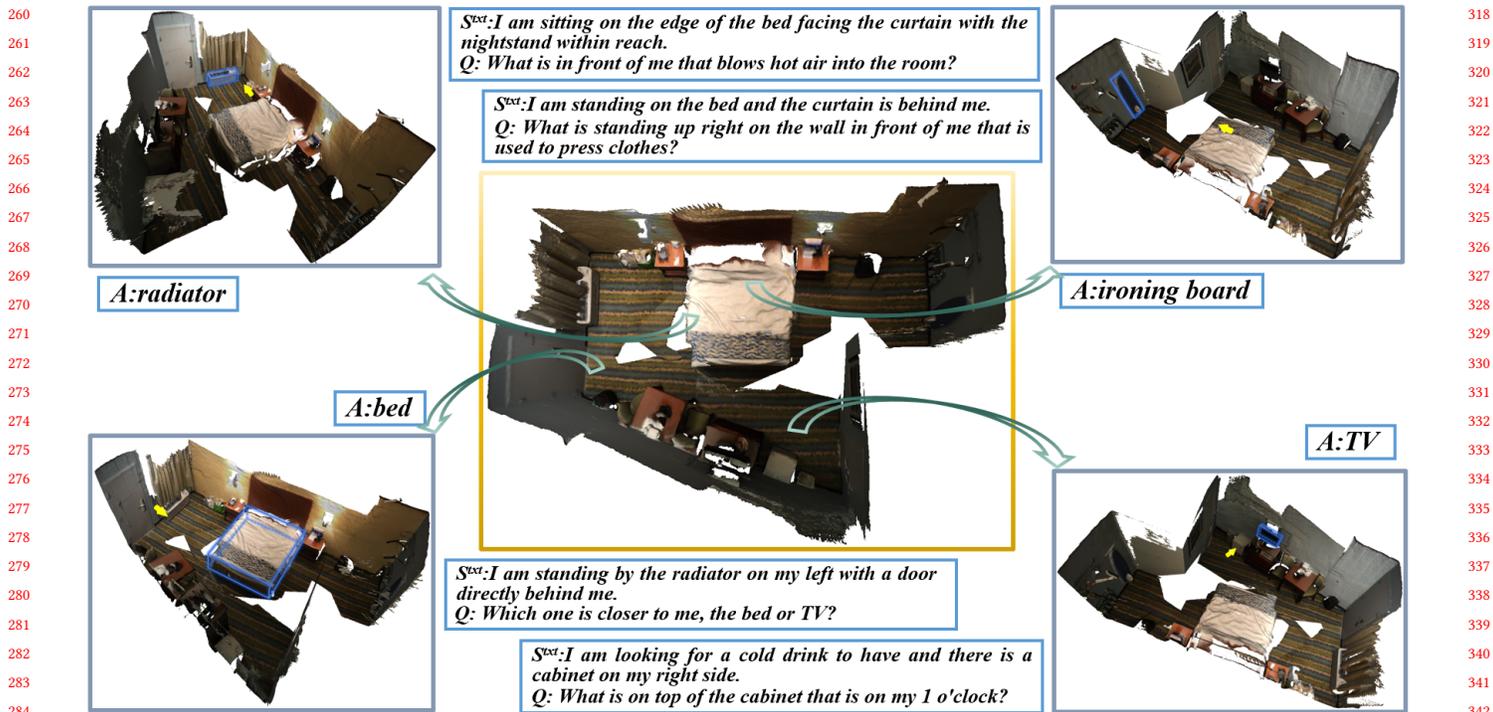
Method	Acc@1	Acc@10
3DGraphQA(GNN)	43.37	82.50
3DGraphQA(GCN)	44.96	83.83
3DGraphQA(GAT 1L)	46.06	84.32
3DGraphQA(GAT 2L)	47.88	84.97
3DGraphQA(GAT 3L)	47.80	85.05

REFERENCES

- [1] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. 2022. ScanQA: 3D Question Answering for Spatial Scene Understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE, 19107–19117.
- [2] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations, ICLR*. OpenReview.net.
- [3] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. 2023. SQA3D: Situated Question Answering in 3D Scenes. In *ICLR*. OpenReview.net.
- [4] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *International Conference on Learning Representations, ICLR*. OpenReview.net.



(a) One example of 3DQA in the workroom scene.



(b) One example of 3DQA in the bedroom scene.

Figure 1: Our visualization results on the SQA3D dataset. As depicted in the figure, our approach is capable of generating accurate answers based on the problem and various situation descriptions. Moreover, it accurately outputs the positions of situations and bounding boxes for objects.