DSPNet: Dual-vision Scene Perception for Robust 3D Question Answering

Supplementary Material

1. More Compared Methods

We further compare DSPNet with additional state-of-the-art methods that incorporate 3D-language alignment pre-training, external datasets, or large language models (LLMs). As summarized in Tab. 1, despite not leveraging any of these auxiliary enhancements, DSPNet achieves highly competitive performance on both ScanQA and SQA3D datasets. This demonstrates the effectiveness of our approach, highlighting its capability to perform well without relying on extensive pre-training or external resources.

2. Results on "3DQA" dataset

We have previously evaluated our method on ScanQA and SQA3D, two widely recognized 3D question answering (3D QA) benchmarks that encompass diverse reasoning tasks, including spatial attribute recognition, embodied activities, navigation, common sense reasoning, and multi-hop reasoning. To further assess the generalizability of our approach, we conduct additional experiments on another 3D QA benchmark introduced by Ye et al. [15], named "3DQA", which is a human-annotated free-form dataset. For fair comparison with our method, we fine-tune 3D-VisTA [16] from scratch on the "3DQA" dataset. As shown in Tab. 2, Our method achieves EM@1 scores of 52.0%, outperforming 3D-VisTA (49.3%), demonstrating its effectiveness across different 3D QA benchmarks.

3. More Ablation Studies

Here we provide more ablation studies on our model.

3D Encoder. To evaluate the impact of different pre-trained 3D encoders on our model's performance, we experimented with VoteNet [13] and PointNet++ [12]. PointNet++ extracts local geometric features by hierarchically partitioning point clouds into nested regions and recursively processing them into dense point-level visual features, without the utilization of an explicit object detection module. VoteNet, on the other hand, builds upon PointNet++ by introducing a voting mechanism and a detection head to perform 3D object detection within the point cloud. It generates object proposals by aggregating votes from dense point-level visual features and refines them to localize and classify objects. In our experiments, PointNet++ is initialized from the pre-trained VoteNet, which has been pre-trained on a 3D object detection task in ScanNet [3] dataset. In the VoteNet configuration, we input object-level visual features from object proposals into our Multimodal Context-guided Reasoning module, rather than using sparse candidate point-level visual features that are sampled from dense point-level visual features. As shown in Tab. 3, PointNet++ outperforms VoteNet, achieving higher accuracy on both the ScanQA [1] and SQA3D [10] datasets. This suggests that using a 3D encoder without an object detection head enhances the model's generalization ability in 3D QA tasks. The absence of an object detector allows the encoder to learn more generalized and holistic scene features, rather than focusing on specific object categories.

Image Encoder. To investigate the impact of different pre-trained image encoders on our model's performance, we conducted experiments with Vision Transformer (ViT) [5], BEiT [2] and Swin Transformer [9]. ViT directly applies a pure transformer structure by splitting images into fixed-size patches and processing them sequentially. BEiT employs a masked image modeling strategy for self-supervised pre-training, learning visual representations through predicting masked image patches. Swin Transformer introduces a hierarchical architecture with shifted windows for computing self-attention, which efficiently handles various image resolutions. For fair comparison, all experiments are conducted using the base size of these models. As shown in Tab. 4, Swin Transformer consistently outperforms other architectures, achieving the best performance. These results suggest that advanced image encoders can bolster a model's capabilities in 3D QA tasks, primarily due to their enhanced extraction of multi-view image features that deepen the perception of local texture details within 3D scenes.

Text Encoder. To evaluate the effectiveness of different pre-trained text encoders, we experimented with BERT [4], RoBERTa [8] and Sentence-BERT (SBERT) [14] architectures. BERT utilizes bidirectional training and masked language modeling to learn contextual representations. RoBERTa builds upon BERT by implementing optimized training strategies, such as extended training duration, increased batch sizes, removal of the next sentence prediction task, and dynamic masking. SBERT leverages siamese network structure to generate semantically meaningful sentence embeddings. In our experiments, we utilize the base size of each model for fair comparison. According to the results presented in Tab. 5, SBERT delivers the most notable performance enhancements. This improvement highlights the benefit of adopting a powerful text encoder, which helps to gain a deeper understanding of situation descriptions and questions through its strong semantic understanding at the sentence level, significantly improving the model's performance in 3D QA tasks.

Inference Speed Analysis. We conducted an inference speed analysis by measuring the average processing time

Method	Pre-trained	LLMs-based	Extra dataset	ScanQA	SQA3D
LM4Vision [11]	×	√	×	- / -	48.1
PQ3D [17]	×	×	√	26.1 / 20.0	47.1
GPS [7]	√	×	√	25.0 / 23.5	49.9
LEO [6]	√	√	√	- / -	50.0
DSPNet (Ours)	×	×	×	**26.5 / 23.8**	**50.4**

Table 1. The QA accuracy (EM@1) on the "test w/ object" / "test w/o object" split of ScanQA and the test split of SQA3D.

Method	EM@1	EM@10
3D-VisTA [16]	49.3	88.6
DSPNet (Ours)	**52.0**	**90.5**

Table 2. The question answering accuracy on the validation split of "3DQA" dataset.

per sample for different settings of the number of image views. The results indicate that processing time per sample scales with the number of image views, increasing from **117 ms for 10 views to 171 ms for 15 views and 217 ms for 20 views**. These results demonstrate the significant impact of the number of image views on inference time, highlighting the importance of optimizing scene understanding with fewer multi-view images, which is a promising direction for future research.

Encoder	ScanQA	SQA3D
VoteNet [13]	22.65	49.84
PointNet++ [12]	**23.47**	**50.36**

Table 3. Ablation study of different 3D encoders. Conducted on the validation split of the ScanQA dataset and the test split of the SQA3D dataset, using EM@1 as the metric.

Encoder	ScanQA	SQA3D
ViT [5]	22.46	49.39
BEiT [2]	22.63	49.87
Swin Transformer [9]	**23.47**	**50.36**

Table 4. Ablation study of different image encoders. Conducted on the validation split of the ScanQA dataset and the test split of the SQA3D dataset, using EM@1 as the metric.

4. More Qualitative Results.

Qualitative Results of TGMF module. We visualized the intermediate results of the TGMF module in Fig. 1 to provide a clearer understanding of its functionality. Specifically, we showed the image that exhibits the highest context-specific importance weights. From the results, we can see that our TGMF module performs its intended function well.

Encoder	ScanQA	SQA3D
BERT [4]	22.57	48.68
RoBERTa [8]	23.22	49.47
SBERT [14]	**23.47**	**50.36**

Table 5. Ablation study of different text encoders. Conducted on the validation split of the ScanQA dataset and the test split of the SQA3D dataset, using EM@1 as the metric.

Qualitative Results of our model. Additional qualitative results demonstrating the performance of our model are provided in Fig. 2 and Fig. 3. These results illustrate our model's ability to handle a diverse range of tasks, including querying the locations of objects, identifying characteristics and states of specific objects, counting the number of objects within a scene, and responding to yes/no questions that require commonsense reasoning. From these results, we observe that our method remains robust in complex scenes, despite the varied shapes of the objects involved in the reasoning process, including objects with flat shapes (*e.g.*, whiteboard, TV, clock) and even objects with flexible shapes (*e.g.*, curtain, towel, jacket).

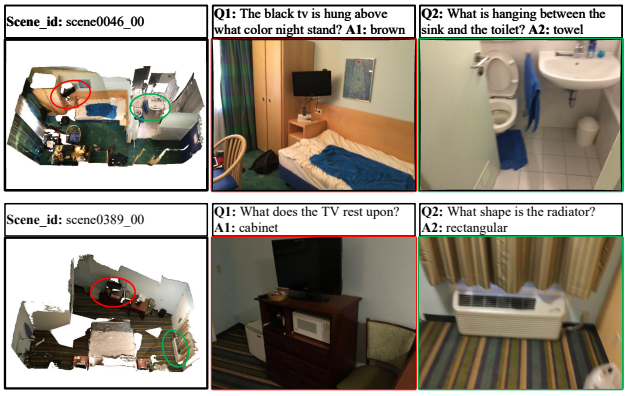

Figure 1. The TGMF module dynamically prioritizes different views according to question context within the same scene.

5. Future Work

Adapting to Dynamic Environments. In future developments of DSPNet, we plan to extend the model's functionality in dynamic environments where changes occur in real-

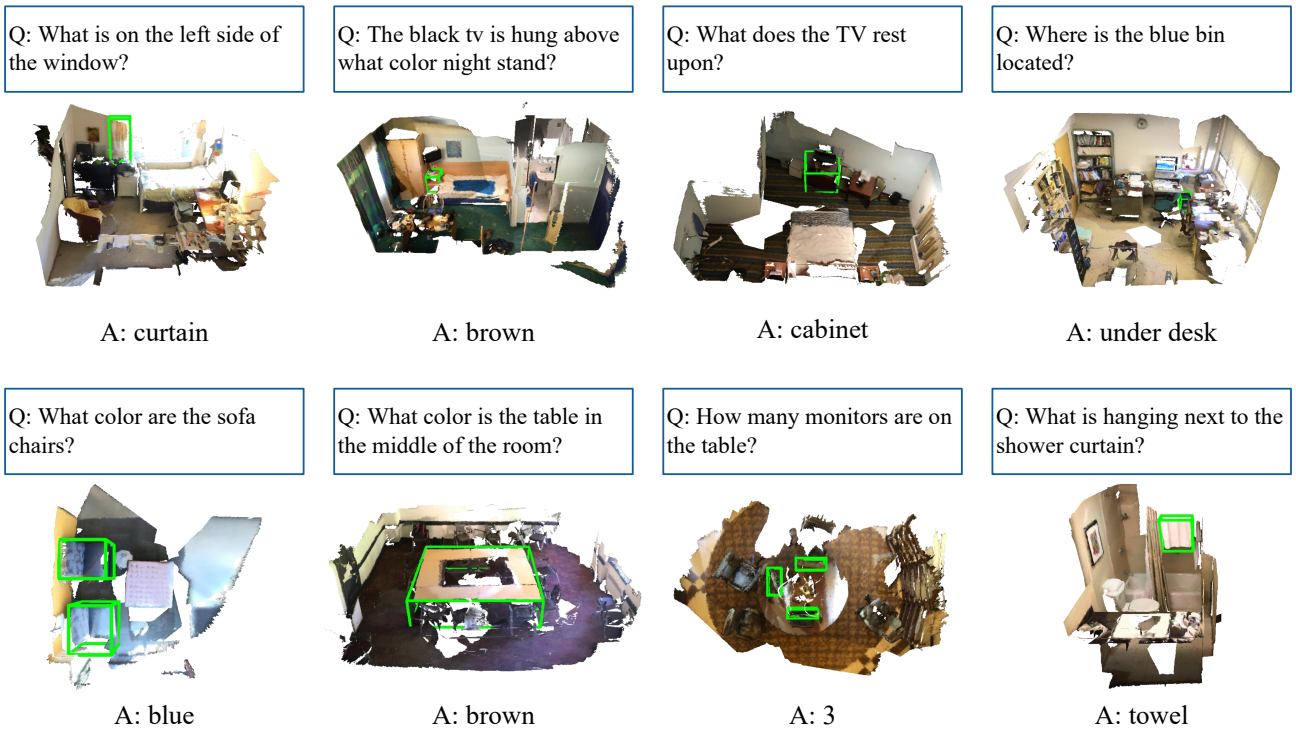

Figure 2. We present more qualitative results on ScanQA dataset.

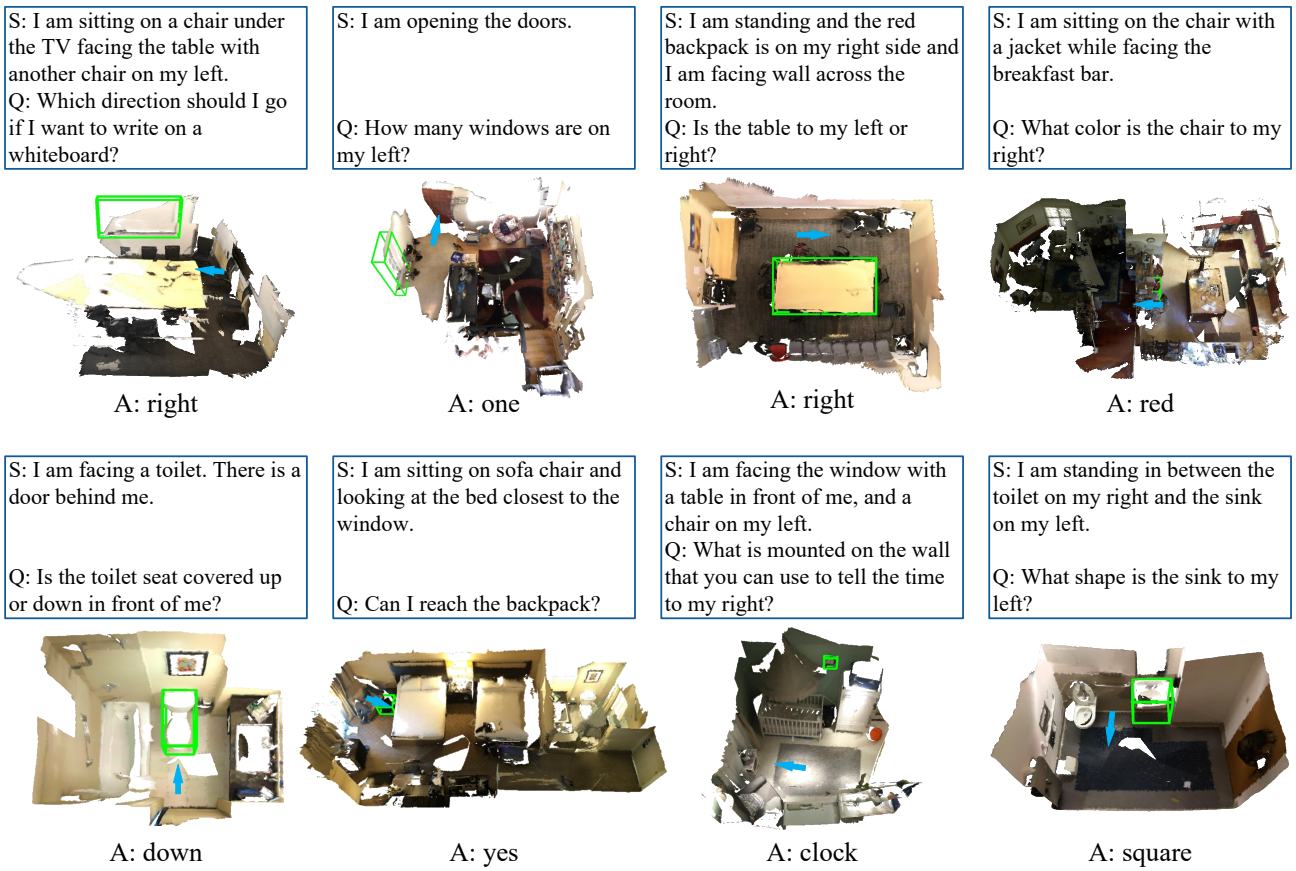

Figure 3. We present more qualitative results on SQA3D dataset.

time. This improvement requires evolving our framework to accommodate real-time data acquisition and processing, reducing the dependency on pre-scanned point clouds and pre-captured multi-view images. Such advancements will involve integrating adaptive streaming algorithms that can handle continuous input from moving cameras and sensors.

Multi-modal Alignment. Further, we intend to enhance DSPNet's ability to perceive and reason within 3D scenes comprehensively through multi-modal integration. We will investigate the alignment of pre-training across 3D scenes, multi-view images, and text related to scenes. This effort will focus on developing a comprehensive multi-modal pre-training approach that utilizes the inherent relationships among these modalities. By applying strategies like contrastive learning and cross-modal distillation, we aim to improve the semantic consistency and contextual understanding across visual and textual data.

Integration with Large Models. In this paper, we haven't adopt large models due to the limited size of available 3D QA datasets, which restricts the effective training and generalization capabilities of such models. Large models usually require large-scale datasets to avoid overfitting and fully utilize their capacity. In addition, the computational limitations of current devices make it challenging to deploy large models. However, with the improvement of computing power of modern hardware and the emergence of larger 3D QA datasets in the future, exploring large 3D QA models with dual-vision becomes a promising direction. Our future research will focus on developing scalable architectures to effectively utilize expanded datasets and enhance the model's ability to comprehensively perceive and reason in 3D scenes.

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. 1

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. 1, 2

[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. 1

[4] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1, 2

[6] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. 2

[7] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In *European Conference on Computer Vision*, pages 289–310. Springer, 2024. 2

[8] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364, 2019. 1, 2

[9] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1, 2

[10] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *The Eleventh International Conference on Learning Representations*, 2022. 1

[11] Ziqi Pang, Ziyang Xie, Yunze Man, and Yu-Xiong Wang. Frozen transformers in language models are effective visual encoder layers. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[12] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 1, 2

[13] 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*, pages 9277–9286, 2019. 1, 2

[14] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019. 1, 2

[15] Shuquan Ye, Dongdong Chen, Songfang Han, and Jing Liao. 3D Question Answering . *IEEE Transactions on Visualization & Computer Graphics*, 30(03):1772–1786, 2024. 1

[16] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2911–2921, 2023. 1, 2

[17] Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, Yixin Chen, Baoxiong Jia, Zhidong Deng, Siyuan Huang, and Qing Li. Unifying 3d vision-language understanding via promptable queries. In *European Conference on Computer Vision*, pages 188–206. Springer, 2024. 2