Supplementary Materials: Spatial Matters: Position-Guided 3D Referring Expression Segmentation

Yabing Wang[1] Zhuotao Tian[2] Le Wang[1*] Zheng Qin[1] Sanping Zhou[1]
[1]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
National Engineering Research Center for Visual Information and Applications,
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
[2]Harbin Institute of Technology, Shenzhen

1. Impact of the Number of Point Proxies N_{key}.

We study the influence of the number of point proxies used in the spatial-aware query generation module. As shown in Table 1, increasing N_{key} gradually improves the performance by enabling richer aggregation of local context and spatial cues. However, excessively large N_{key} introduces redundant points and hampers optimization, leading to a performance drop. The best performance is achieved when $N_{key} = 256$, which provides more appropriate spatial coverage and enables the model to aggregate sufficiently rich local context.

2. Efficiency Comparison.

As shown in Table 2, our method achieves the best runtime efficiency while maintaining a comparable model size. This improvement stems from avoiding full global attention, which substantially reduces computational overhead. As a result, inference is significantly accelerated, particularly in long-sequence point cloud scenarios. These results demonstrate that our method achieves a favorable efficiency–performance trade-off.

Table 1. Ablation study on the number of nearest neighbors in the local context aggregation layer.

N_{key}	Acc@0.25	Acc@0.5	mIoU
64	59.9	52.6	48.5
128	59.9	55.2	50.2
256	**61.5**	**56.1**	**51.0**
512	59.3	54.3	49.5

Table 2. Efficiency comparison between our method and existing methods.

Method	Params(M)	FPS
MDIN [2]	150.87	3.71
3D-STMN [3]	135.94	6.28
IPDN [1]	156.53	9.74
Ours	150.39	12.68

3. Visualization of the Multiple Referred Objects

In Figure 1, we present qualitative results on Multi3DRefer, in which scenes contain multiple referred objects. We observe that our method successfully distinguishes and segments each target object by leveraging both semantic cues and spatial relationships. For instance, in (e), our approach accurately segments the chairs despite the presence of multiple distractor objects, demonstrating its strong capability in resolving complex spatial dependencies and differentiating objects with similar appearance. These results further validate the effectiveness of explicit spatial relation modeling in improving segmentation reliability under challenging multi-object scenarios, especially when semantic cues alone are insufficient.

4. Visualization and Failure Cases

In Figure 2, we present the results of our Position3D on ScanRefer. We observe that our method accurately segments the referred objects, benefiting from the proposed explicit spatial relation modeling. Moreover, we find that our predictions are sometimes even more precise than the provided ground-truth annotations. This improvement is largely attributed to the point proxy mechanism, which effectively aggregates local contextual features around key regions and enables more discriminative segmentation boundaries, demonstrating the effectiveness of our design.

We also provide several failure cases to further analyze the limitations of our method. As illustrated in (g) and (h), we observe that some referring expressions in the dataset are inherently ambiguous and do not provide sufficiently precise descriptions to clearly identify the referred target

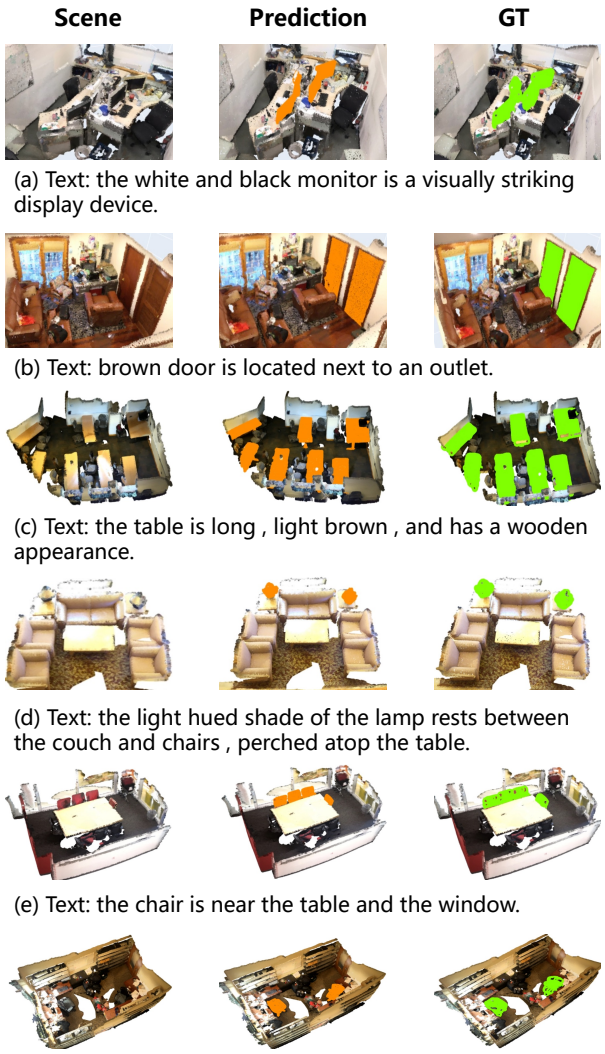


Figure 1. Visualization of our Position3D on the Multi3DRefer dataset that contains multiple referred objects.

Table 3. Ablation study on the accuracy of each layer.

Layer	Acc@0.25	Acc@0.5	mIoU
Layer1	56.9	50.8	46.3
Layer2	60.4	54.2	49.3
Layer3	60.6	55.4	50.3
Layer4	**61.5**	**56.1**	**51.0**

object. In future work, we believe that there is still room for improvement in the formulation of textual expressions, such as complexity and specificity.

5. Analysis of each layer

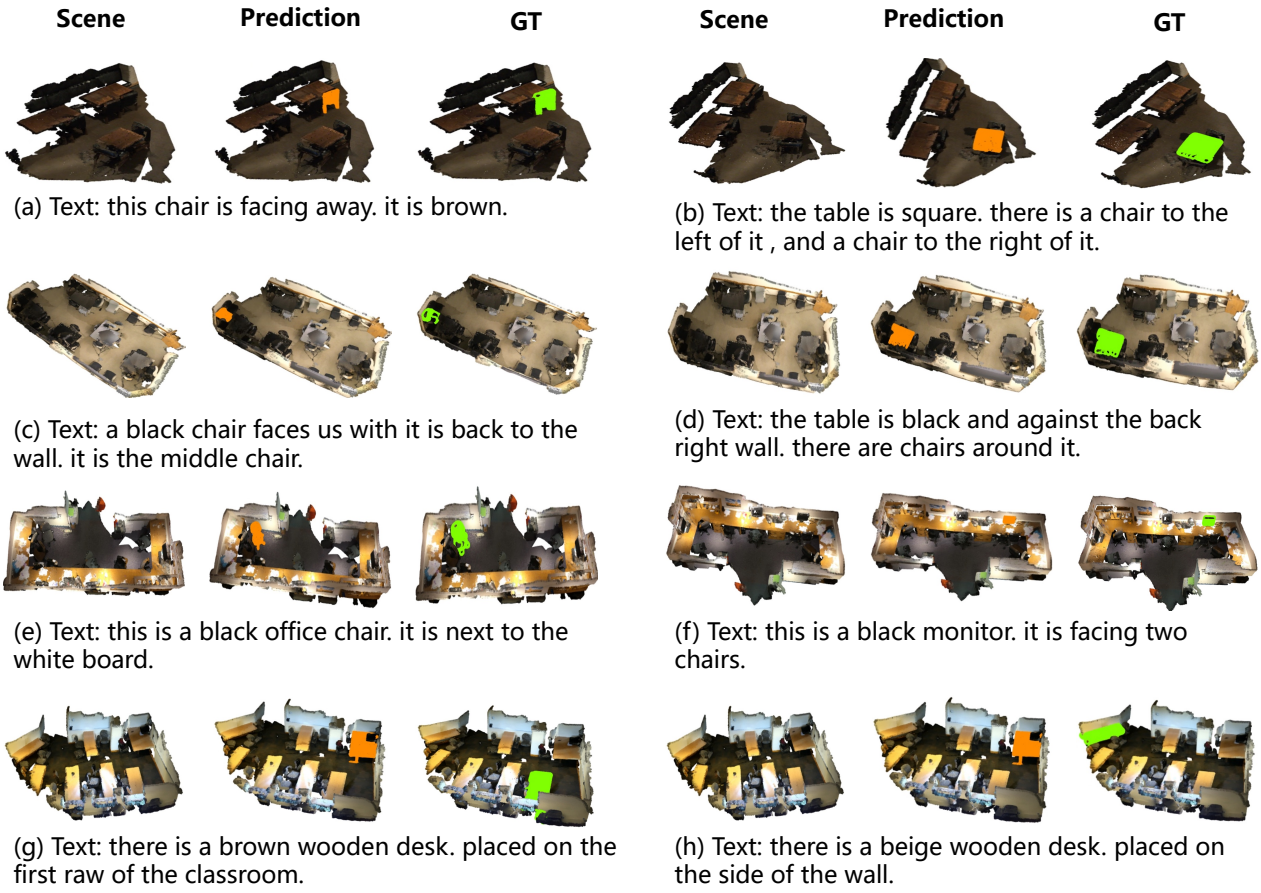
As shown in Table 3, the accuracy of each layer shows consistent improvements at each layer, confirming that the later decoder layers provide meaningful refinement.

References

[1] Qi Chen, Changli Wu, Jiayi Ji, Yiwei Ma, Danni Yang, and Xiaoshuai Sun. Ipdn: Image-enhanced prompt decoding network for 3d referring expression segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 2132–2140, 2025. 1

[2] Changli Wu, Yihang Liu, Jiayi Ji, Yiwei Ma, Haowei Wang, Gen Luo, Henghui Ding, Xiaoshuai Sun, and Rongrong Ji. 3d-gres: Generalized 3d referring expression segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7852–7861, 2024. 1

[3] Changli Wu, Yiwei Ma, Qi Chen, Haowei Wang, Gen Luo, Jiayi Ji, and Xiaoshuai Sun. 3d-stmn: Dependency-driven superpoint-text matching network for end-to-end 3d referring expression segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5940–5948, 2024. 1



(a) Text: this chair is facing away. it is brown.

(b) Text: the table is square. there is a chair to the left of it , and a chair to the right of it.

(c) Text: a black chair faces us with it is back to the wall. it is the middle chair.

(d) Text: the table is black and against the back right wall. there are chairs around it.

(e) Text: this is a black office chair. it is next to the white board.

(f) Text: this is a black monitor. it is facing two chairs.

(g) Text: there is a brown wooden desk. placed on the first raw of the classroom.

(h) Text: there is a beige wooden desk. placed on the side of the wall.

Figure 2. Visualization and failure cases of our Position3D on the ScanRefer dataset.