EG-3DVG: Expression and Geometry Aware Grounding Decoder for 3D Visual Grounding

Supplementary Material

S-1. Implementation of two-stage model

Following prior two-stage approaches [4, 6, 8, 9], we adopt GroupFree [5] as our 3D object detector and encode its detection outputs as box tokens $\mathrm{B} \in \mathbb{R}^{\mathrm{N_{box}} \times \mathrm{C}}$. These box tokens provide object-level proposals that complement the point-level visual tokens. To effectively exploit these proposals, we inject bounding-box cues into the output features of the PECA module via an additional cross-attention layer. The resulting refined representation is then processed by two parallel branches, a cross-attention module that aggregates contextual information from the refined superpoint tokens and the GMA module.

S-2. Ablation Study on Source Tokens for Candidate Token Refinement

To understand the contribution of each source token type and attention mechanism to candidate token refinement, we perform a series of ablation experiments. Specifically, we remove text tokens, visual tokens, or superpoint tokens from the expression and geometry aware grounding decoder, and additionally replace PECA or GMA with standard cross-attention to examine their functional roles. Table S-1 shows that removing text tokens yields a substantial performance drop, indicating that linguistic cues are essential for guiding candidate tokens toward the correct expression-aligned semantics. While replacing PECA with standard cross-attention recovers part of this loss, its performance still lags behind the full EG-3DVG, demonstrating the importance of injecting spatial cues into text features.

Table S-2 further indicates that excluding visual tokens or substituting GMA with vanilla cross-attention reduces geometric consistency. This degradation occurs because visual tokens provide crucial spatial structure, and GMA selectively aggregates geometry-consistent cues, enabling more reliable grounding. Finally, Table S-3 shows that removing superpoint tokens noticeably decreases 3DRES accuracy. This highlights the role of superpoint tokens in supplying higher-level object-region cues, which in turn support more accurate and stable mask prediction.

S-3. GT-based Candidate Filtering in SR3D/NR3D

A notable but often overlooked characteristic of the SR3D/NR3D benchmarks [1] is that their standard evaluation protocol is not fully automatic. Each scene provides ground-truth (GT) bounding boxes for several objects, and prior 3DVG methods [3, 4, 6–9] rely on these GT boxes during inference. Specifically, after generating candidate tokens and their predicted bounding boxes, existing approaches compute their IoU against all GT boxes, and discard candidates whose maximum IoU is below 0.25. This GT-based filtering effectively removes localization errors and reduces the task to selecting the correct object among pre-filtered candidates. Although widely used, this protocol introduces an additional source of supervision at test time and therefore diverges from fully automatic grounding settings, such as ScanRefer [2], where no GT information is available during inference.

For consistency with prior work, we follow the same GT-based filtering protocol when reporting SR3D/NR3D results in the main paper. However, to more faithfully assess the localization capability of the model, we additionally evaluate performance without using any GT information during inference. As shown in Table S-4, prior methods experience significant performance drops when GT-based filter-

Method	3DREC		3DRES
	Acc@0.25	Acc@0.5	mIoU
w/o text token	56.64	49.35	45.81
w/o PECA	57.86	51.64	46.59
EG-3DVG	**58.54**	**52.36**	**47.28**

Table S-1. Ablation study for text tokens and PECA in the decoder on ScanRefer.

Method	3DREC		3DRES
	Acc@0.25	Acc@0.5	mIoU
w/o visual token	56.57	48.99	45.91
w/o GMA	58.18	50.95	46.34
EG-3DVG	**58.54**	**52.36**	**47.28**

Table S-2. Ablation study for visual tokens and GMA in the decoder on ScanRefer.

Method	3DREC		3DRES
	Acc@0.25	Acc@0.5	mIoU
w/o superpoint token	57.91	51.20	45.13
EG-3DVG	**58.54**	**52.36**	**47.28**

Table S-3. Ablation study for superpoint tokens in the decoder on ScanRefer.

Method	SR3D		NR3D	
	Overall	Hard	Overall	Hard
EDA [9]	68.1	62.9	52.1	46.1
EDA w/o GT-based filtering	65.4 (2.7↓)	60.3 (2.6↓)	50.0 (2.1↓)	44.1 (2.0↓)
MCLN [6]	68.4	-	59.8	-
MCLN † [6]	67.2	58.8	54.9	48.7
MCLN † w/o GT-based filtering	64.1 (3.1↓)	55.9 (2.9↓)	52.6 (2.3↓)	45.9 (2.8↓)
EG-3DVG (Ours)	75.2	66.6	59.9	52.4
EG-3DVG w/o GT-based filtering	73.3 (1.9↓)	64.9 (1.7↓)	58.6 (1.3↓)	51.2 (1.2↓)

Table S-4. 3DREC results with and without GT-based filtering on SR3D/NR3D. We assess performance using Acc@0.25. † denotes result obtained by training the model from scratch, as the pretrained weights are not provided.

PECA	GMA	ECL	3DREC						3DRES
			Unique (∼19%)		Multiple (∼81%)		Overall		mIoU
			Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	
			89.43	81.40	51.86	44.16	57.04	49.62	44.19
✓			90.63	82.17	52.18	44.68	57.73	50.92	45.69
	✓		**90.84**	83.94	51.01	44.32	57.01	50.35	46.19
		✓	90.53	84.43	52.26	45.01	57.56	50.85	45.84
✓	✓		90.67	84.34	52.39	45.94	58.06	51.84	46.57
✓		✓	90.12	83.65	52.53	45.21	58.18	50.95	46.34
	✓	✓	90.19	84.27	52.03	45.87	57.86	51.64	46.59
✓	✓	✓	**90.84**	**84.57**	**52.87**	**46.71**	**58.54**	**52.36**	**47.28**

Table S-5. Ablation study for combinations of position-guided expression cross-attention (PECA), geometry-aware masked attention (GMA), and expression-aware contrastive learning (ECL) on ScanRefer.

ing is removed. EDA decreases by 2.7 on SR3D and 2.1 on NR3D, and MCLN drops even more sharply, 3.1 on SR3D and 2.3 on NR3D, highlighting their reliance on the GT-assisted evaluation protocol. In contrast, EG-3DVG maintains considerably more stable performance, with only 1.9 degradation on SR3D and 1.3 on NR3D, demonstrating that our method can robustly localize objects even in a fully automatic setting. For MCLN, pretrained weights are not publicly available, so we report results reproduced by training the model from scratch.

S-4. Detailed Ablation Study

Table S-5 presents a detailed ablation study of the three proposed components. First, PECA shows consistent improvements across both the Unique and Multiple settings, showing that injecting 3D positional cues into text features effectively strengthens cross-modal alignment. Second, GMA yields gains particularly at Acc@0.5, showing that incorporating geometric consistency enables the model to produce more accurate and spatially precise box predictions. Finally, ECL provides improvements in the Multiple setting, where fine-grained discrimination among many semantically similar objects is required, showing its effectiveness in reducing intra-class confusion.

Method	Inside rate	Distance
w/o PECA	57.63	0.127
EG-3DVG	**58.37**	**0.118**

Table S-6. Comparison of candidate token selection accuracy with and without PECA on ScanRefer.

S-5. Analysis on PECA

Effect of PECA on Candidate Token Selection: We assess whether PECA improves the model's ability to select the correct candidate token representing the target object. Since the correct candidate should lie within the GT bounding box and ideally be positioned close to its geometric center, we evaluate two criteria: whether the selected candidate token falls inside the ground-truth box, and its Euclidean distance to the box center for those that fall inside. As reported in Table S-6, EG-3DVG equipped with PECA more frequently selects candidates that locate within the correct bounding box and consistently chooses tokens located closer to the object center. These results indicate that PECA provides stronger spatial guidance, leading to more accurate and stable target-object selection.

Effect of PECA on Attention Weight: To further under-

stand how PECA shapes the interaction between textual expressions and candidate tokens, we visualize the cross-attention patterns.

For every candidate token, we first compute the average cross-attention weight in PECA over all expression-relevant words (*e.g.*, "main", "attribute", "pronoun", and "relationship"). To obtain an object-level visualization in the scene, we then aggregate these candidate-level scores by averaging the attention weights of all candidate tokens belonging to the same object. This yields a per-object attention intensity that reflects how strongly the textual expressions attend to each object. Figure S-1 compares these attention maps with and without PECA, showing that PECA substantially enhances the model's ability to focus on the correct object described in the input text. Without PECA, the model often places high attention scores on distractor objects or even on instances from different categories. In contrast, EG-3DVG with PECA consistently aligns its attention with the correct semantic target. For instance, in the first row of Figure S-1, the description refers to an "ottoman," but the model without PECA incorrectly assigns strong attention to a nearby "table." With PECA enabled, the attention becomes correctly concentrated on the ottoman, enabling accurate grounding of the intended object.

S-6. Analysis on GMA

To better understand how GMA selectively aggregates geometry-consistent visual cues, we visualize the activated visual tokens that remain after applying the attention mask. Figure S-2 compares three masking strategies: (1) an attention-threshold mask, (2) a geometry-only mask, and (3) the proposed GMA. The attention-threshold mask represents a baseline that does not use GMA at all. It is obtained by applying a small threshold to the raw cross-attention weights between candidate tokens and all visual tokens. Since cross-attention without geometric constraints tends to diffuse broadly, this mask produces wide and noisy activations, often including tokens far outside the target object. The geometry-only mask captures spatial structure more consistently by relying solely on geometric affinity. However, it still activates many irrelevant tokens when nearby distractors exist, as it lacks semantic and expression-driven cues to isolate the correct object.

In contrast, GMA produces a notably more concentrated and semantically aligned activation pattern. By combining geometric relationships with expression-guided priors, GMA compensates for the limitations of candidate tokens, whose initial representations may miss fine-grained or spatially distributed cues, and selectively retains visual tokens that truly correspond to the intended target. This enables GMA to suppress spatially close but irrelevant tokens while enriching each candidate token with the geometry-consistent information it lacks on its own.

S-7. Analysis on ECL

Figure S-3 presents qualitative comparisons between the proposed EG-3DVG with and without ECL, focusing particularly on intra-category misidentification. EG-3DVG without ECL often selects objects of the same category but with different attributes. For example, as shown in the first row of Figure S-3, the description refers to a "yellow chair" but the model without ECL selects a "red chair" as the target object. On the other hand, incorporating ECL enables the model to correctly identify the target object that matches the expression-specific attributes described in the text.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *ECCV*, pages 422–440, 2020. 1

[2] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, pages 202–221, 2020. 1

[3] Zoey Guo, Yiwen Tang, Ray Zhang, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. Viewrefer: Grasp the multi-view knowledge for 3d visual grounding. In *ICCV*, pages 15372–15383, 2023. 1

[4] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *ECCV*, pages 417–433, 2022. 1

[5] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *ICCV*, pages 2949–2958, 2021. 1

[6] Zhipeng Qian, Yiwei Ma, Zhekai Lin, Jiayi Ji, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Multi-branch collaborative learning network for 3d visual grounding. In *ECCV*, pages 381–398, 2024. 1, 2

[7] Xiangxi Shi, Zhonghua Wu, and Stefan Lee. Viewpoint-aware visual grounding in 3d scenes. In *CVPR*, pages 14056–14065, 2024.

[8] Yuan Wang, Yali Li, and Shengjin Wang. Gˆ 3-lq: Marrying hyperbolic alignment with explicit semantic-geometric modeling for 3d visual grounding. In *CVPR*, pages 13917–13926, 2024. 1

[9] Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In *CVPR*, pages 19231–19242, 2023. 1, 2

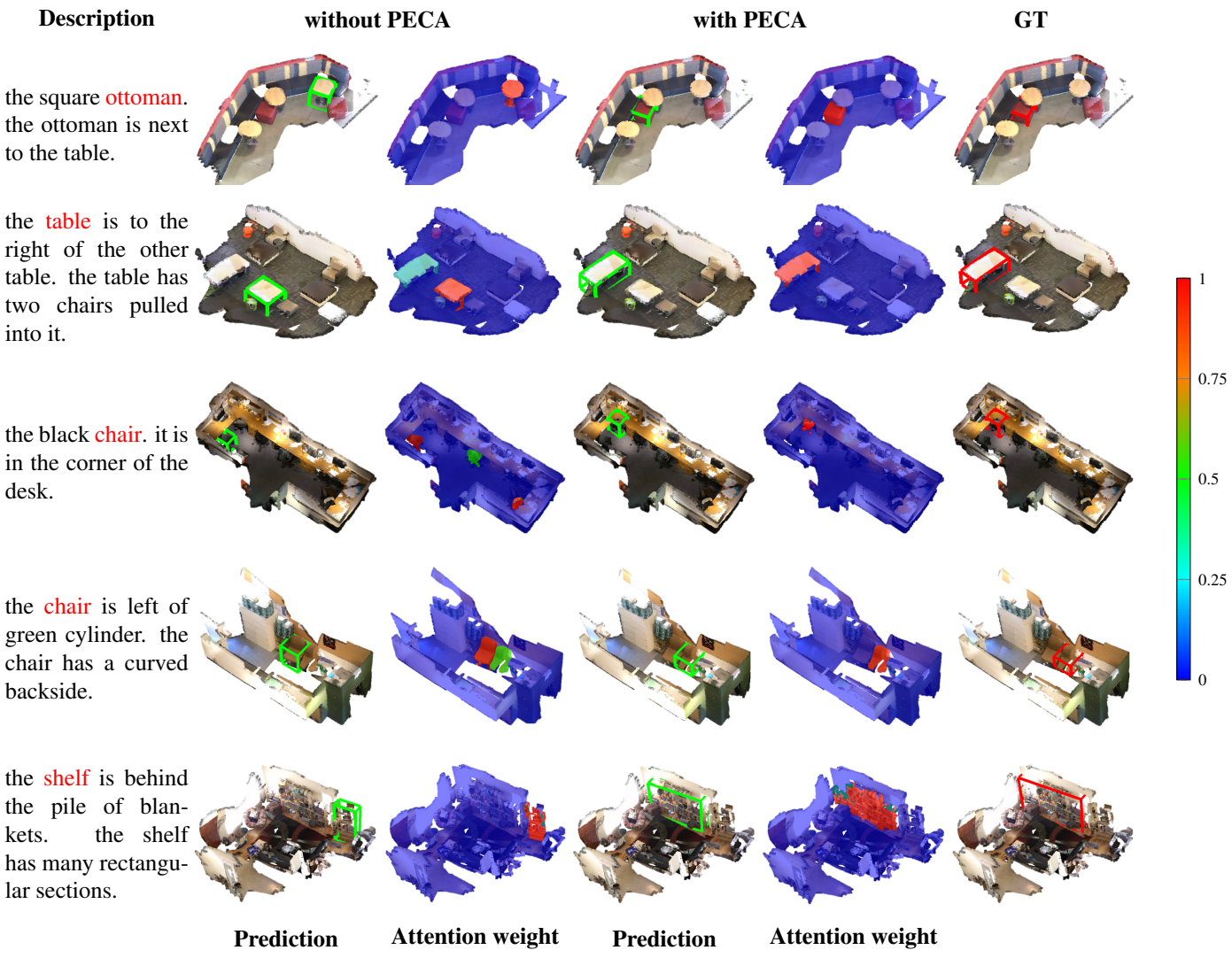

Figure S-1. Attention weight comparison between EG-3DVG with and without PECA on ScanRefer.

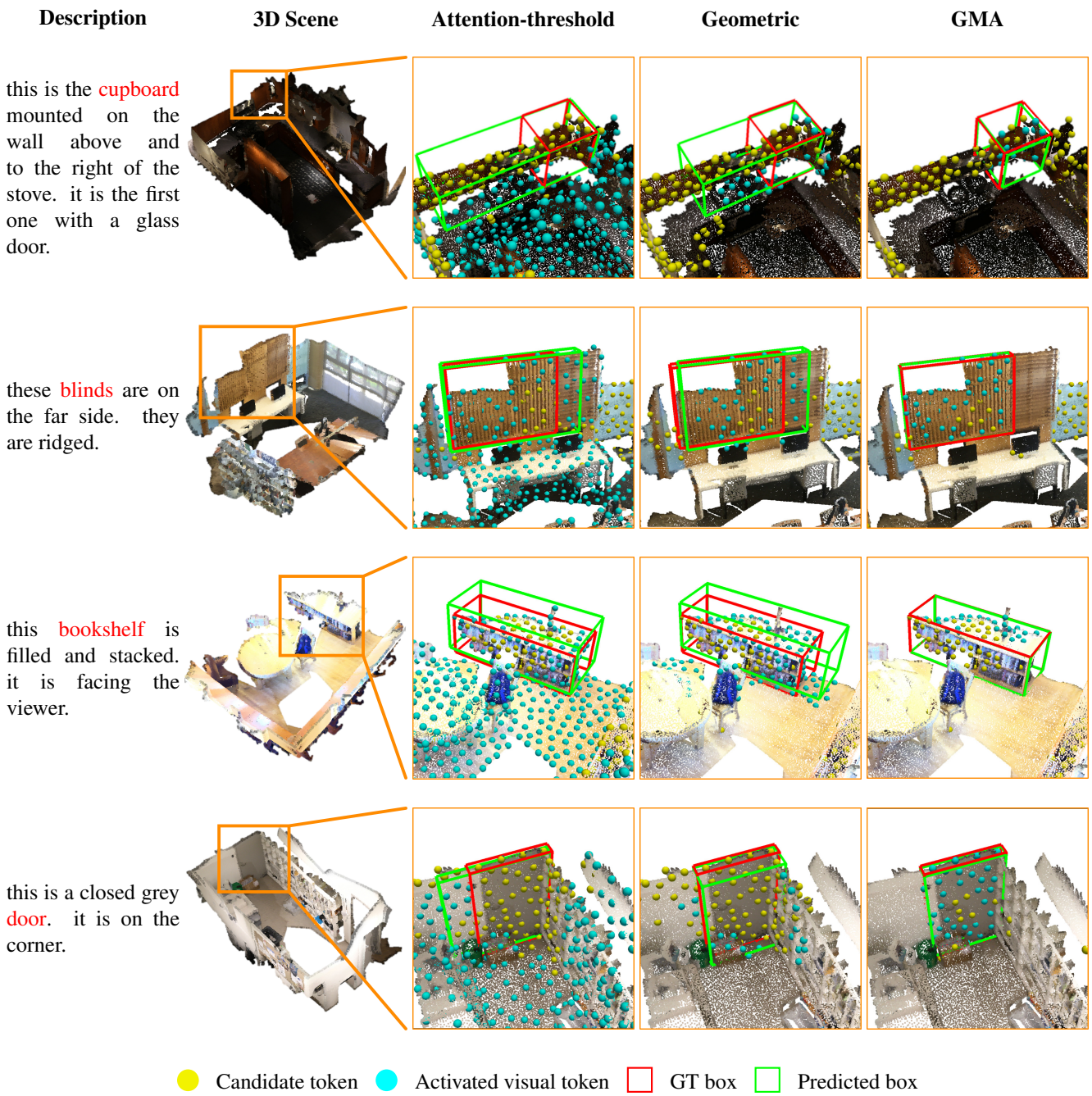

Figure S-2. Qualitative comparison of different attention masking strategies in GMA, including attention-threshold mask, geometric-only mask, and our geometry-aware masked attention (GMA) on ScanRefer.

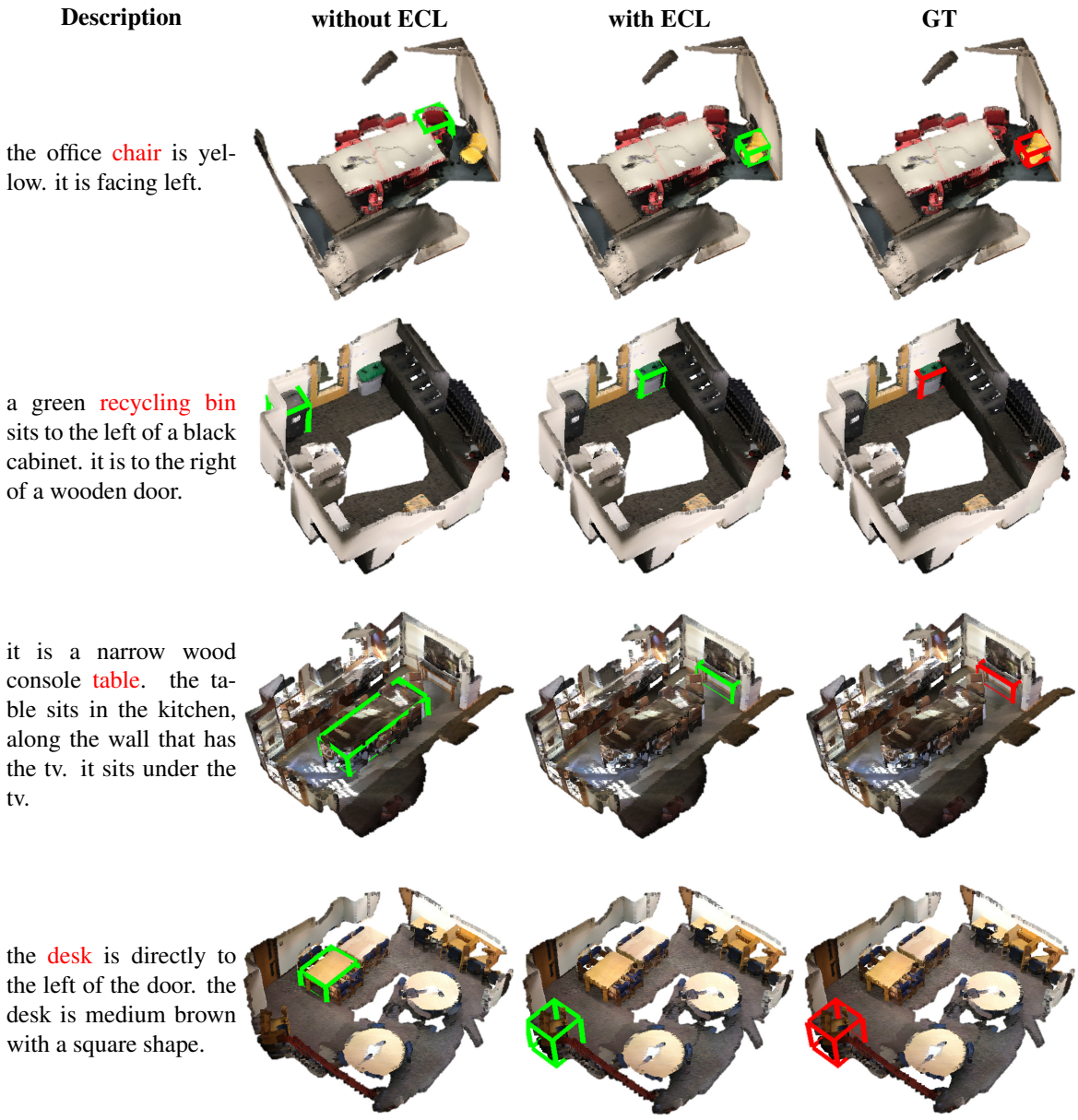

Figure S-3. Qualitative comparisons between EG-3DVG with and without ECL on ScanRefer.