PV-Ground: Text-Guided Point-Voxel Interaction for 3D Visual Grounding

Supplementary Material

This supplementary material provides further details and experimental results to complement the main manuscript. The content is organized as follows:

- Section 6 presents more detailed ablation studies, specifically focusing on the ReferIt3D dataset.
- Section 7 validates the generalizability of our proposed framework to existing point-based methods, demonstrating universal performance improvements across various models.
- Section 8 showcases additional visualization results, including comparisons with other methods across diverse and complex scenarios.
- Section 9 discusses the broader impact and potential limitations of this work.

6. Ablation Study on ReferIt3D

To further validate the effectiveness of each proposed module, we conducted additional experiments on the ReferIt3D dataset to supplement the ablation study in the main manuscript. Similar to ScanRefer, queries in ReferIt3D can be divided into several subsets, including "Easy" samples (with one distractor) and "Hard" samples (with multiple distractors). Additionally, samples are classified as "view-dependent" or "view-independent" depending on whether the localization requires viewer-centric perspective information. In this section, we provide these detailed results on the Nr3D and Sr3D subsets to precisely analyze the impact of PV-Ground's components across these varied scenarios.

The results are presented in Table 5. It is evident that both the proposed PVI framework and the TGS module yield effective performance gains over the baseline. Specifically, the PVI framework contributes the primary performance boost, validating the superiority of the high-fidelity scene feature representation inherent to the point-voxel interaction paradigm. Additionally, the TGS module can achieve further improvements by adaptively re-guiding the keypoint distribution to focus on text-relevant regions.

Table 5. Ablation on the proposed point-voxel interaction 3D VG framework and the TGS module. Evaluated on Nr3D and Sr3D.

Dataset	PVI	TGS	Easy	Hard	Dep.	Indep.	Overall
Nr3D			51.4	40.9	43.6	47.3	45.7
		✓	53.9	42.2	46.2	48.9	48.0
	✓		**55.7**	45.5	47.4	52.1	50.5
	✓	✓	55.2	**47.5**	**48.3**	**52.9**	**51.3**
Sr3D			55.9	47.8	48.5	53.7	53.4
		✓	56.4	48.7	46.3	54.5	54.1
	✓		58.2	50.1	48.1	56.1	55.8
	✓	✓	**59.0**	**50.6**	**51.3**	**56.7**	**56.5**

7. Adaptation to Existing Baselines

As discussed in the main text, the robust keypoint features generated by PV-Ground can be directly fed into the decoders and prediction heads of most existing point-based methods, yielding universal performance gains. To verify this general adaptability, we integrated PV-Ground's keypoint features into several representative point-based models—including BUTD-DETR [15], EDA [54], and MCLN [32]—replacing their original backbone features while retaining their specific multi-modal interaction decoders and prediction heads. We conducted ablation studies on the ScanRefer dataset for these baselines, with results summarized in Table 6.

The results clearly demonstrate that our framework consistently improves performance across different point-based baselines. This gain is particularly significant in single-stage pipelines. Notably, integrating PV-Ground features into BUTD-DETR [15] resulted in a remarkable 8% performance increase, achieving 57.72% on Acc@0.25 and 44.28% on Acc@0.50, respectively. The accuracy significantly surpass even the original baselines of more recent methods like EDA and MCLN. These results empirically validate that the high-quality keypoint features produced by the PV-Ground framework can deliver substantial and universal performance improvements to the broader field of point-based 3D visual grounding.

8. Visualization and Qualitative Comparisons

Figure 5 presents extensive visualization results and detailed qualitative comparisons, specifically benchmarking our grounding results against existing state-of-the-art methods [12, 32]. We showcase performance across a variety of challenging scenarios, including "small targets" (g, j), targets with "incomplete or occluded point clouds" (d, h, i), and scenes containing "multiple distracting objects" (a, c). These complex, real-world scenarios pose significant challenges to the 3D VG task.

It can be observed that while existing methods are highly prone to failure in these contexts, PV-Ground consistently achieves accurate target localization and segmentation. This robustness is attributable to our high-fidelity scene representation combined with the superior multi-modal interactive reasoning capabilities of target-aware keypoints, enabling precise grounding even in the most demanding environments.

Table 6. Adaptation on different point-based baselines. Evaluated on ScanRefer. The proposed PV-Ground framework significantly boosts existing baselines, achieving general performance improvement. Especially, we observed a remarkable 8% increase for BUTD-DETR.

baseline	Venue	Pipeline	PV-Ground	Unique (∼19%)		Multiple (∼81%)		Overall	
				Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	**Acc@0.25**	**Acc@0.5**
BUTD-DETR [15]	ECCV'22	two-stage	×	82.88	64.98	44.73	33.97	50.42	38.60
			✓	**87.53**	**70.40**	**53.15**	**40.88**	**58.28** (+7.9%)	**45.29** (+6.7%)
EDA [54]	CVPR'23	two-stage	×	85.76	68.57	49.13	37.64	54.59	42.26
			✓	**87.32**	67.65	**51.63**	**38.52**	**56.95** (+2.4%)	**42.87** (+0.6%)
MCLN [32]	ECCV'24	two-stage	×	86.89	72.73	51.96	40.76	57.17	45.53
			✓	**87.67**	**72.94**	**54.99**	**43.11**	**59.87** (+2.7%)	**47.56** (+2.0%)
BUTD-DETR [15]	ECCV'22	single-stage	×	81.47	61.24	44.20	32.81	49.76	37.05
			✓	**87.32**	**69.63**	**52.53**	**39.83**	**57.72** (+8.0%)	**44.28** (+7.2%)
EDA [54]	CVPR'23	single-stage	×	86.40	69.42	48.11	36.82	53.83	41.70
			✓	85.55	68.08	**51.17**	**39.29**	**56.30** (+2.5%)	**43.58** (+1.9%)
MCLN [32]	ECCV'24	single-stage	×	84.43	68.36	49.72	38.41	54.30	42.64
			✓	**86.11**	**72.45**	**54.61**	**43.44**	**59.31** (+5.0%)	**47.77** (+5.1%)

9. Discussion

9.1. Limitations and Future Works

Despite the significant advancements, our framework presents certain limitations that outline directions for future research:

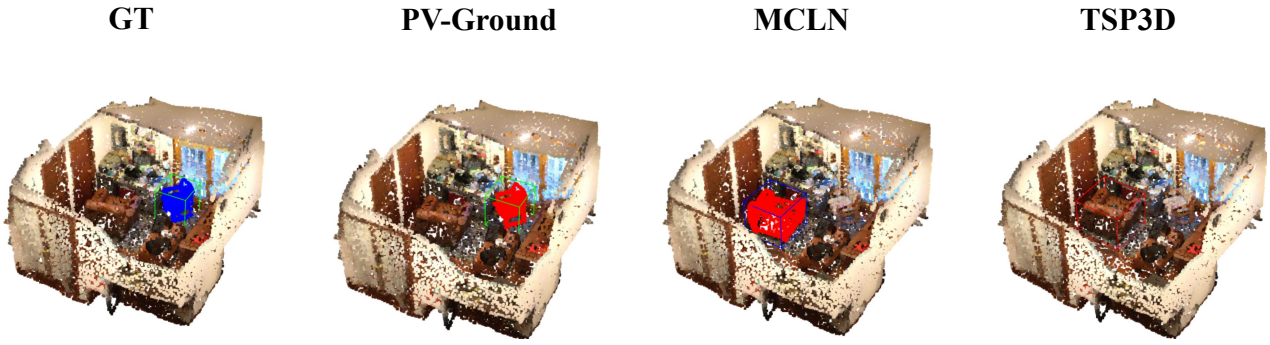
Integration with LLMs. PV-Ground primarily focuses on optimizing the visual feature representation problem in 3D VG. Currently, there is a surge in methods leveraging Large Language Models (LLMs) for deep semantic and spatial reasoning [22, 49]. Our framework is complementary to these approaches; future work could explore synergizing PV-Ground's high-quality visual features with the reasoning capabilities of LLMs to achieve even stronger performance.

Inference Efficiency. While the sparse voxel convolution backbone is inherently efficient, our framework employs existing point-based decoder to handle complex multi-modal interactions and regression. The computational overhead introduced by the keypoint aggregation and the deep attention-based interaction layers results in an inference speed that is comparable to previous point-based baselines, rather than matching the impressive efficiency of pure voxel pruning-based methods like TSP3D [12].

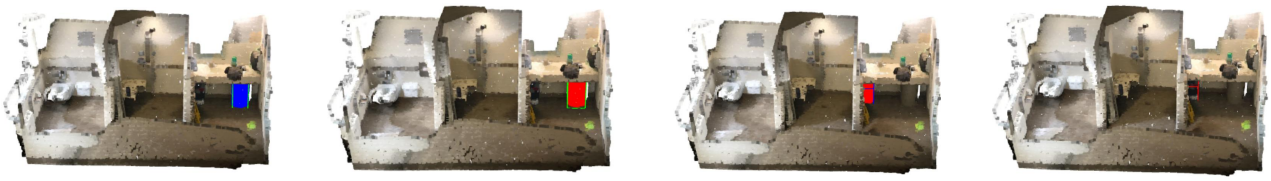
Dynamic Scene Modelling. Our method is currently designed for static 3D scans. Extending the text-guided point-voxel interaction mechanism to 4D spatio-temporal modeling for processing video point clouds or dynamic environments represents a significant and necessary direction for future development.

9.2. Broader Impact

The core contribution of this work extends beyond the immediate performance gains on standard 3D VG benchmarks. We have demonstrated that the point-voxel interaction framework holds significant potential as a high-quality and effective visual representation paradigm for generic 3D multi-modal tasks. By combining the fine-grained geometric details of voxels with the flexible, compact and token-like keypoint representations, our method offers a generalizable solution for downstream tasks such as 3D Dense Captioning, 3D Question Answering, and Embodied Intelligence. This task-specific hybrid representation facilitates high-fidelity and efficient multi-modal interaction without the prohibitive computational costs associated with dense voxel grids, potentially serving as a universal visual encoder for future 3D multi-modal foundation modeling.



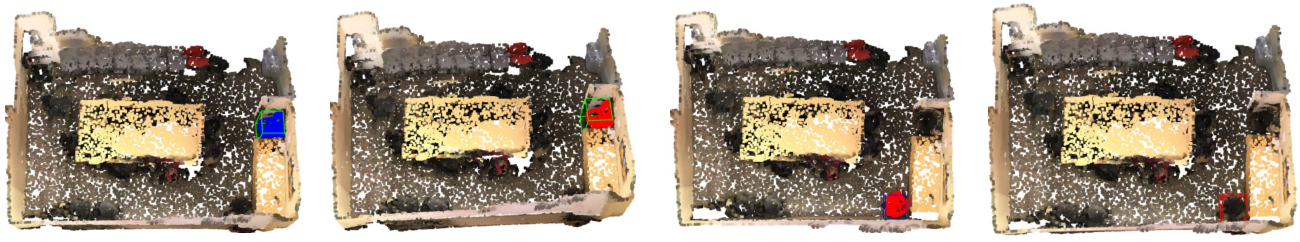
(a) a wooden **chair** with a cushioned seat and back sits in fronts of a wall with wood and glass doors. a desk is to its right. a brown leather sofa is to its left.



(b) this is a **round trash can**. it is under a bathroom counter .

(c) the **chair** is black and has arms. the chair is sideways and in front of two others.

(d) this is a stainless steel mini **fridge**. it is to the left of the desk.



(e) the **chair** is to the left of the cabinet below the picture. the chair has five legs and a curved backside .

(f) the coffee **table** is just to the left of center of the back wall of the room. the coffee table is directly in front of a chair with circular black workspace arm and to the right of a red seat.

(g) the **stool** placed on hardwood floor is dark blue. it has four legs .

(h) the **bookshelf** is to the left of the sofa chair. the bookshelf is brown and has four sections.

(i) there is a square beige **table**. it is next to a large rectangular table at the side of the room.

(j) the **trash can** is left of the curtains. the trash can is a gray cylinder.

Figure 5. Visualization and qualitative comparisons of PV-Ground and existing state-of-the-art methods. The target in the text is annotated in red and the key clues in the textual description are highlighted in gold. PV-Ground consistently excels in challenging scenarios, including "small targets" (g, j), targets with "incomplete or occluded point clouds" (d, h, i), and scenes with "multiple distracting objects" (a, c).