Supplementary Material

Table 1. Ablation on Cluster Drop Radio β. A smaller β indicates that a greater number of submanifolds are used for subsequent transformations. The reported values are mAP for predictions greater than 25% IoU.

Drop Radio	Easy	Hard	Overall
0.5	40.55	34.17	40.03
0.6	**41.66**	**34.38**	**41.08**
0.7	40.67	32.70	40.02
0.8	39.65	31.02	38.95
0.9	36.40	27.02	35.64

Table 2. Ablation on the Max Value of Offsets. The reported values are mAP for predictions greater than 25% IoU.

Max	Easy	Hard	Overall
3	38.14	29.65	37.46
4	**41.66**	**34.38**	**41.08**
5	36.67	26.71	35.87

1. Additional Visualization

Additional visualization is provided in **??**, demonstrating the qualitative effectiveness of our method.

2. Additional Experiments

Ablation on Cluster Drop Radio. As shown in Tab. 1, an excessively high drop ratio reduces model performance, highlighting the effectiveness of our submanifold transformations. A lower drop ratio retains more submanifolds for subsequent Proxy Transformation, enhancing results. However, experiments show that an overly low drop ratio also degrades performance, as point cloud enhancement should complement rather than overly alter the original structure.

Ablation on s. As discussed when introducing deformable offsets, s represents the maximum value of the offsets. Before generating the grid prior, we scale down the cuboid defined by the maximum and minimum coordinates of the point cloud based on s, ensuring that reference points, when adjusted by deformable offsets, do not move outside the boundaries of the point cloud. As shown in Tab. 2, we selected an optimal maximum value for s. An excessively large s results in the reduced preset grid losing essential prior information, while an overly small s prevents reference points from shifting toward more critical target regions, thereby reducing the flexibility of the model.

3. Ego-Centric 3D Visual Grounding

In real-world applications, intelligent agents interact with their surroundings without prior knowledge of the entire scene. Instead of relying on pre-reconstructed 3D point clouds or other scene-level priors commonly used in previous studies [1, 3], they primarily depend on ego-centric observations, such as multi-view RGB-D images.

Following the definition in [2], we formalize the ego-centric 3D visual grounding task as follows: Given a natural language query $L \in \mathbb{R}^T$, along with V RGB-D image views $\{(I_v, D_v)\}_{v=1}^V$, where $I_v \in \mathbb{R}^{H\times W\times 3}$ denotes the RGB image and $D_v \in \mathbb{R}^{H\times W}$ represents the depth map for the v-th view, and their corresponding sensor intrinsics $\{(K_v^I, K_v^D)\}_{v=1}^V$ and extrinsics $\{(T_v^I, T_v^D)\}_{v=1}^V$, the goal is to predict a 9-degree-of-freedom (9DoF) bounding box $B = (x, y, z, l, w, h, \theta, \phi, \psi)$.

In this context, (x, y, z) specify the 3D center coordinates of the target object, (l, w, h) define its dimensions, and (θ, ϕ, ψ) represent its orientation angles. The task is to determine B such that it accurately localizes the object described by L within the scene captured by $\{(I_v, D_v)\}_{v=1}^V$.

4. Details about Proxy Bias

As mentioned in the methodology, to compensate for the lack of positional information and the diversity of features, we propose a novel **Proxy Bias**:

$$F = F_0 + B^{\mathrm{P}}, \tag{1}$$

where $F \in \mathbb{R}^{N\times C}$ is the input of Proxy Block and $F_0 \in \mathbb{R}^{N\times C}$ is our deformable cluster features. $B^{\mathrm{P}} \in \mathbb{R}^{N\times C}$ is our novel proxy bias.

Initially, we set three learnable parameters $B_d \in \mathbb{R}^{N\times D\times D}$, $B_c \in \mathbb{R}^{N\times 1\times S}$ and $B_r \in \mathbb{R}^{N\times S\times 1}$. Here, $C = S^2 = D^4$. Therefore, our parameters are way less than directly setting B^{P} as a learnable parameter, thus improving our parameter efficiency.

We first interpolate B_d into $B_1 \in \mathbb{R}^{N\times S\times S}$, mapping the low-dimensional subspace into a higher-dimensional feature space to enhance feature diversity. Subsequently, we add B_c and B_r to obtain the final $B_2 \in \mathbb{R}^{N\times S\times S}$, representing the linear union of two low-dimensional subspaces to form the final high-dimensional space, expressed as $V = V_1 \cup V_2$. Finally, we get $B^{\mathrm{P}} = (B_1 + B_2).\mathrm{reshape}(N, C)$, which can enrich the feature space with positional information and guide ProxyAttention to focus on diverse regions.

References

[1] Rui Huang, Songyou Peng, Ayca Takmaz, Federico Tombari, Marc Pollefeys, Shiji Song, Gao Huang, and Francis Engelmann. Segment3d: Learning fine-grained class-agnostic 3d segmentation without manual labels. *arXiv preprint arXiv:2312.17232*, 2023. 1

[2] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, Xihui Liu, Cewu Lu, Dahua Lin, and Jiangmiao Pang. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *CVPR*, 2024. 1

[3] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *CVPR*, 2024. 1