DEGround: An Effective Baseline for Ego-centric 3D Visual Grounding with a Homogeneous Framework

Supplementary Material

Yani Zhang[1*†], Dongming Wu[2*], Hao Shi[3], Yingfei Liu[4‡], Tiancai Wang[4], Xingping Dong[1✉]
[1] School of Computer Science, Wuhan University
[2] MMLab, The Chinese University of Hong Kong,
[3] Tsinghua University, [4] Dexmal
{zebrazyn, xinpngdong}@whu.edu.cn, wudongming97@gmail.com

A. Implementation Details

Model. For feature extraction, we use ResNet50 [3] for 2D semantic features, Minkowski ResNet34 [2] for 3D geometric features, and RoBERTa [5] for textual features as the respective backbones. During the construction of 3D feature representations, all feature maps are projected into a consistent dimension of 256. The decoder is composed of 6 Transformer decoder layers, while the box, classification, and grounding heads, built on top of the decoder, employ a linear layer. The early fusion module is implemented as a single-layer transformer. Moreover, the category number C^{det} and the grounding number C^{grd} are set to 284 and 1, respectively. The number of queries is 1024 for detection and 512 for grounding.

Training. We adopt the AdamW optimizer [6] for network training. The learning schedule incorporates a warmup strategy, followed by cosine decay. Data augmentation includes random flipping and random rotation in 3D space. All RGB-D frames are resized to 480×480. For 3D detection, the model is trained for 36 epochs on 8 Nvidia A100 GPUs with a batch size of 8. The learning rate is set to 1e-4 for the backbone and 1e-3 for the remaining parameters. Both loss weights λ_{cls} and λ_{box} are set to 1. For 3D grounding, the model is initialized from a checkpoint pretrained on 3D detection. Training proceeds for 3 epochs with a learning rate of 8e-5. The loss weights are set as: $\lambda_{ground} = 1$, $\lambda_{box} = 1$, and $\lambda_{spatial} = 0.01$.

Inference. Since there is no gap between training and inference in our method, we directly output the corresponding bounding boxes using the well-trained model without post-processing.

*Equal contribution. ✉Corresponding author: *Xingping Dong*. †This work was done during the internship at Dexmal. ‡Project lead. This work was supported in part by the New Generation Artificial Intelligence-National Science and Technology Major Project (No. 2025ZD0123501), the National Natural Science Foundation of China under Grant 62471342, and WHU-Kingsoft Joint Lab.

Table 1. Performance of PQ3D with RAG and QIM on ScanRefer

RAG	QIM	Overall IOU_{50}	Unique IOU_{50}	Multiple IOU_{50}
-	-	47.4	76.6	42.0
✔	-	48.4	**77.2**	43.1
-	✔	48.0	76.1	42.8
✔	✔	**48.6**	77.0	**43.3**

B. Generalization Study on ScanRefer

To further validate the effectiveness of the proposed method, we conduct experiments on ScanRefer [1]. While the query-sharing mechanism is specifically designed for ego-centric 3D visual grounding, we integrate RAG and QIM into the strong baseline PQ3D [8] and evaluate on the ScanRefer validation set. As shown in Tab. 1, both modules consistently enhance the overall performance over the PQ3D baseline, indicating that these designs remain effective beyond the ego-centric setting.

C. Computational Cost Analysis

To further evaluate the efficiency of our framework, we compare the model parameters, training time and inference speed with different DETR-based methods.

Detection. As shown in Tab. 2, our model shows stronger performance with lower computational cost. Compared to BIP3D [4], it reduces training time by about half (32h36m vs. 64h48m) and achieves over 4× faster inference (138.56 ms vs. 652.14 ms), while using 40% fewer parameters (102.8 M vs. 175.0 M). Despite the lighter architecture,

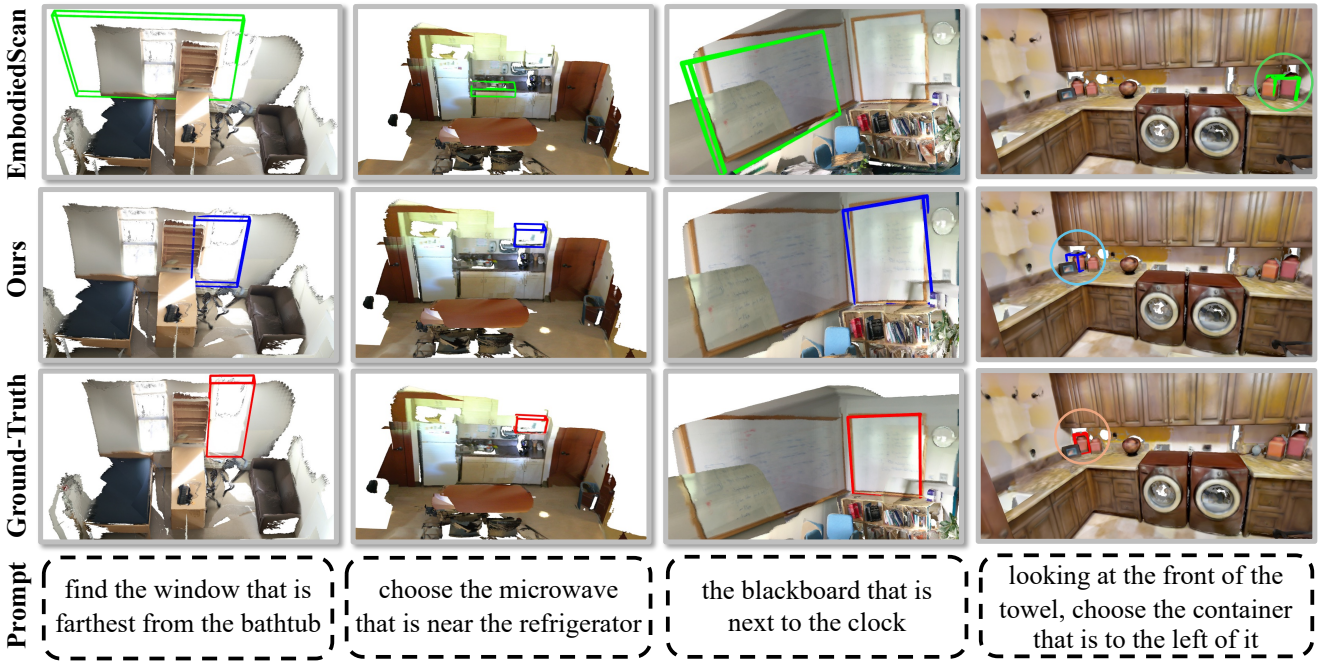

Figure 1. **Qualitative comparisons of our method and EmbodiedScan.** Best viewed in color and with zoom.

Table 2. Comparison of computational cost for detection models on the EmbodiedScan detection benchmark.

Method	Params	Epoch	Training	Inference	AP@25
BIP3D [4]	175.01M	24	64h48m	652.14ms	20.91
Ours	102.80M	36	32h36m	138.56ms	24.68

Table 3. Comparison of computational cost for grounding models on the EmbodiedScan mini validation set.

Method	Params	Epoch	Training	Inference	AP@25
EmbodiedScan [7]	229.58M	12	6h33m	152.64ms	35.84
BIP3D [4]	175.01M	2	8h46m	598.70ms	45.79
Ours	239.40M	3	2h13m	163.72ms	61.28

our model attains a higher AP@25 of 24.68, indicating a superior balance between efficiency and accuracy.

Grounding. For the grounding task, although there is a slight increase in parameter count and inference time compared to EmbodiedScan [7], this overhead is acceptable given the substantial performance gains. As shown in Tab. 3, BIP3D exhibits a much longer inference time than both EmbodiedScan and ours. We hypothesize that BIP3D's slower inference is attributable to its additional depth-prediction branch and a Swin-Tiny backbone operating at a higher input resolution, which together make the overall pipeline more time-consuming in practice.

D. More Qualitative Comparisons

We provide additional qualitative comparisons between DEGround and EmbodiedScan [7] in Fig. 1. Compared with EmbodiedScan, our method yields more accurate and fine-grained localization results that align better with the textual instructions. It more effectively distinguishes between same-category objects and focuses on regions consistent with the textual cues (*e.g.*, identifying the correct window or container as specified in the prompt). These results highlight the improved instruction alignment and discrimination ability of our approach.

References

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

[2] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[4] Xuewu Lin, Tianwei Lin, Lichao Huang, Hongyu Xie, and Zhizhong Su. Bip3d: Bridging 2d images and 3d perception for embodied intelligence. In *CVPR*, 2025. 1, 2

[5] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 1

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2019. 1

[7] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *CVPR*, 2024. 2

[8] 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 *ECCV*, 2024. 1