Supplementary Material for Four Ways to Improve Verbo-visual Fusion for Dense 3D Visual Grounding

Ozan Unal[1,2], Christos Sakaridis[1], Suman Saha[1,3], and Luc Van Gool[1,4,5]

[1]ETH Zurich, [2]Huawei Technologies, [3]PSI, [4]KU Leuven, [5]INSAIT
{ozan.unal, csakarid, suman.saha, vangool}@vision.ee.ethz.ch

1 Implementation Details

The visual inputs are formed via the concatenation of 3D coordinates and RGB color channels. For the 3D Unet backbone we use a voxel size of $2cm$ following set convention [7, 11, 13]. For the natural language encoding, we use the MPNet tokenizer and pre-trained model [9, 12]. Following Zhao *et al.* [15], we randomly mask the referred object nouns with a probability of 0.5 before the extraction of word embeddings in order to reduce overfitting and entice learning context to aid localization. We remap the output of MPNet onto a $d = 128$ dimensional vector using a single linear layer. BAF is built using a vanilla transformer encoder (2-layer) for the language encoding and a decoder (6-layer) for the attentive fusion [10]. After every other decoder layer, we increase the radius of the masking sphere r_l from $[1.0m, 2.5m, \infty]$, with $2.5m$ giving the approximate average inter-instance distance and ∞ providing global attention in the final two layers. We set $\gamma = 25$ following DKNet [13], and $\tau = 0.3$. We empirically choose $K = 5$ and $\tau = 0.9$ for MVE. We use a batch size of 4, with each sample consisting of a single scene and up to 32 utterances. We train for 400 epochs using the AdamW optimizer [8] with a learning rate of $3 \cdot 10^{-4}$ using a single Nvidia RTX 3090.

2 Segmentation vs. Detection for Grounding

In Tab. S1 we provide an extended analysis of the visual backbone change for 3D grounding (main paper Tab. 3).

3 NR3D

While ScanRefer tackles referral-based object localization, Nr3D [1] is built on referral-based identification. This means that in Nr3D, the ground-truth bounding boxes/instance masks are assumed to be given as input. We instead tackle the more challenging task of end-to-end referral-based object localization. Even though our method is not directly applicable to the Nr3D benchmark, the dataset can still be used to evaluate our method for our task of interest.

Table S1: Extension of Tab. 3: We replace the 3D object detector (outputs bounding box) of established 3D visual grounding models with our 3D instance segmentation backbone (outputs mask) to showcase the performance implications.

Method	Output	Unique		Multiple		Overall	
		Acc@25	Acc@50	Acc@25	Acc@50	Acc@25	Acc@50
ScanRefer	Box	67.64	46.19	**32.06**	**21.26**	**38.97**	26.10
	Mask	**70.49**	**63.26**	24.64	20.77	33.18	**28.69**
3DVG-T	Box	77.16	58.47	**38.38**	**28.70**	**45.90**	34.37
	Mask	**82.34**	**75.73**	33.12	27.64	42.29	**36.60**

Table S2: Comparison of SOTA on Nr3D evaluated on referral-based 3D object localization a la ScanRefer.

Method	SRefer [2]	3DVG-T [15]	D3Net [3]	HAM [4]	Ours
Acc@50	12.17	14.22	25.23	27.11	**33.66**

Table S3: Evaluating 3D dense visual grounding on the ScanRefer *val*-set, where the IoU is determined based on not the bounding boxes but the instance masks. Shown are the overall Acc@25/50.

Baseline	+BAF	+$\mathcal{L}_{con}$	+GCT	+MVE
44.2 / 39.1	46.9 / 43.3	49.2 / 44.9	50.7 / 46.8	51.4 / 48.6

As seen in Tab. S2, ConcreteNet significantly outperforms existing methods on Nr3D by +6.55% overall accuracy at 50% threshold, despite not being able to utilize the global camera token due to a lack of available camera information.

4 Evaluating 3D *Dense* Visual Grounding

In Tab. S3 we repeat the component-wise ablation study (see Tab. 2 of the main paper), but with the IoU computed on the instance masks rather than the axis-aligned bounding boxes. Here we again observe that each component significantly improves the *dense* 3D grounding performance as well.

Specifically, we see an improvement of +4.2%, +1.6%, +1.9% and +1.8% in accuracy at the 50% threshold when one-by-one introducing the bottom-up attentive fusion module (BAF), the contrastive loss ($\mathcal{L}_{con}$), the global camera token (GCT) and the multi-view ensembling (MVE) respectively.

5 Camera Rotation in GCT

We have only include camera *position* in GCT as it is sufficient for disambiguating view-dependent prompts (e.g. left, right relations). Tab. S4 shows a further comparison that extends GCT to include yaw and pitch. We supervise the unit look-at direction vector via a cosine similarity loss to the ground truth. This addition of rotational information slightly hurts the overall performance, especially

Table S4: Ablation study on GCT reporting Acc@50.

Method	Unique	Multiple	Overall
GCT (proposed)	75.62	**36.56**	**43.84**
GCT (with pose)	**76.35**	35.21	42.88

Fig. 1: Additional qualitative result from the ScanRefer *val*-set showing the benefits of a learned global camera token.

for *multiple* cases. We hypothesize that adding these uninformative (as justified above) rotational dimensions to the GCT target merely increases problem complexity and does not help disambiguation.

6 Additional Qualitative Results

In Fig. 1 further qualitative results from the ScanRefer *val*-set can be seen that demonstrate the benefits of the global camera token (GCT).

Additionally, in Fig. 2 we show a common failure case where, while the model predictions do not match the ground-truth object, the natural language description still fits the output.

7 Analysing the Semantic Class Accuracy of the Model Predictions

An analysis of the predicted instance semantics can be found in Tab. S5. Specifically, we extract the semantic label of the predicted referral-based instance mask from the ground-truth semantic labels. We then report the accuracy when comparing the predicted semantic class to the ground-truth counterpart. It can be seen that with $\sim$ 86% accuracy, ConcreteNet is able to correctly identify the semantic class of the referred object instance in most cases. Furthermore, from the marginal gap between the *unique* semantic class accuracy and the *unique* accuracy at 25% IoU (Learned GCT on Tab. 5 of the main paper), the effectiveness and robustness of the 3D visual backbone can be inferred.

Fig. 2: Failure case from the ScanRefer *val*-set. While the predictions from both ConcreteNet and 3DVG-Transformer do not match the ground truth, given the symmetric nature of the scene along with the vagueness of the description, it can be seen that the cue does match both predictions and the ground truth.

Table S5: Further analysis on the semantic class of the predicted instance without the inclusion of MVE. We extract the semantic class of the predicted mask and measure the accuracy compared to the semantic class of the target object instance.

Unique	Multiple	Overall
85.33	86.22	86.05

8 Limitations and Discussion

In this section, we dive into the limitations of our work and discuss possible remedies and counterpoints.

Global Camera Token: To employ a learned global camera token (GCT), ConcreteNet requires ground-truth camera information to be provided. As this can be trivially collected during a standard data annotation pipeline, we hope that the presented benefits of GCT aid in conveying the necessity of such additions in future 3D visual grounding datasets. Nevertheless, as seen in Tab. S2, our method continues to show state-of-the-art performance when trained and evaluated without the inclusion of GCT on datasets that lack such information.

Multi-view Ensembling: Exploiting multi-view representation for 3D visual grounding has been studied before. As mentioned in the main manuscript, MVT [6], ViewRefer [5] and Multi3DRefer [14] aggregate 3D, textual or 2D features from multiple views to reduce dependence on a specific viewpoint. Thus, each of the aforementioned works requires multiple forward passes of their respective encoder to extract multi-view features, resulting in a trade-off of performance versus run-time efficiency. Compared to the aforementioned works, our multi-view ensembling (MVE) directly operates on selected referred objects rather than each individual predicted object. This means that not only does MVE require multiple forward passes of the 3D encoder but also the decoder to extract mask information. While this is a notable limitation of the module, we

would like to provide further advantages and counterpoints that emerge from such a limitation:

1. The parallelization of multi-forward pass methods for real-world applications is accomplished by constructing multiple copies of a single model. Essentially this maps the trade-off of performance versus run-time efficiency to that of memory efficiency. Given that the majority of the model's complexity stems from the encoder, the additional decoder forward passes required by MVE compared to existing approaches only incur minimal additional memory costs.
2. Previous methods propose decoders specifically designed to handle multi-view features to tackle 3D visual grounding, while MVE only operates on the outputs of a single model. Thus MVE is a more flexible contribution as it can be utilized with any dense 3D visual grounding model, under the assumption that the model employs a grounding-by-selection approach.

References

1. Achlioptas, P., Abdelreheem, A., Xia, F., Elhoseiny, M., Guibas, L.: ReferIt3D: Neural listeners for fine-grained 3D object identification in real-world scenes. In: ECCV (2020) 1
2. Chen, D.Z., Chang, A.X., Nießner, M.: ScanRefer: 3D object localization in RGB-D scans using natural language. In: ECCV (2020) 2
3. Chen, D.Z., Wu, Q., Nießner, M., Chang, A.X.: D3Net: a speaker-listener architecture for semi-supervised dense captioning and visual grounding in RGB-D scans. In: ECCV (2022) 2
4. Chen, J., Luo, W., Wei, X., Ma, L., Zhang, W.: HAM: Hierarchical attention model with high performance for 3D visual grounding. arXiv preprint arXiv:2210.12513 (2022) 2
5. Guo, Z., Tang, Y., Zhang, R., Wang, D., Wang, Z., Zhao, B., Li, X.: Viewrefer: Grasp the multi-view knowledge for 3d visual grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15372–15383 (2023) 4
6. Huang, S., Chen, Y., Jia, J., Wang, L.: Multi-view transformer for 3D visual grounding. In: CVPR (2022) 4
7. Jiang, L., Zhao, H., Shi, S., Liu, S., Fu, C.W., Jia, J.: Pointgroup: Dual-set point grouping for 3d instance segmentation. In: CVPR (2020) 1
8. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 1
9. Song, K., Tan, X., Qin, T., Lu, J., Liu, T.Y.: Mpnet: Masked and permuted pre-training for language understanding. NIPS (2020) 1
10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NIPS (2017) 1
11. Vu, T., Kim, K., Luu, T.M., Nguyen, X.T., Yoo, C.D.: Softgroup for 3d instance segmentation on 3d point clouds. In: CVPR (2022) 1
12. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Transformers: State-of-the-art natural language processing. In: EMNLP (2020) 1
13. Wu, Y., Shi, M., Du, S., Lu, H., Cao, Z., Zhong, W.: 3D instances as 1D kernels. In: ECCV (2022) 1
14. Zhang, Y., Gong, Z., Chang, A.X.: Multi3drefer: Grounding text description to multiple 3d objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15225–15236 (2023) 4
15. Zhao, L., Cai, D., Sheng, L., Xu, D.: 3DVG-Transformer: Relation modeling for visual grounding on point clouds. In: ICCV (2021) 1, 2