Appendix of Zoo3D

Andrey Lemeshko[1] Bulat Gabdullin[1] Nikita Drozdov[2] Anton Konushin[2]
Danila Rukhovich[3] Maksim Kolodiazhnyi[2†]
[1]Higher School of Economics [2]Lomonosov Moscow State University
[3]M:3L Lab, Institute of Mechanics, Armenia

In Appendix, we provide additional quantitative scores, including evaluation results on ScanNet++ [13] and ARKitScenes [1], in Sec. A, report results of ablation experiments in Sec. B, and show more visualizations in Sec. C.

A. Quantitative Results

ScanNet++ As can be seen from Tab. 1, Zoo3D$_0$ sets state-of-the-art in ScanNet++ [13] even compared with fully-supervised methods. Despite not being exposed to any annotations or even training scenes, it outperforms models that utilize both scans and labeled 3D bounding boxes during the training.

ScanNet60 results are reported in Tab. 2. For objects of *base* categories, Zoo3D falls beyond training-based competitors, which could be expected, since neither of our methods has access to ground truth 3D bounding boxes. For *novel* objects not seen during training, Zoo3D scores first in the leaderboard.

3D segmentation baselines Open-vocabulary 3D object detection metrics on ScanNet are listed in Tab. 3. To establish baselines, we adapt state-of-the-art 3D instance segmentation approaches by simply enclosing each predicted mask with a 3D bounding box. Obviously, while Zoo3D$_0$ is built upon MaskClustering, our open-vocabulary assignment strategy is way more effective compared to the one used in the original MaskClustering.

Method	mAP$_{25}$	mAP$_{50}$
TR3D† [7]	26.2	14.5
UniDet3D† [4]	26.4	17.2
Zoo3D$_0$	**26.5**	**18.3**

Table 1. 3D object detection results from points clouds on ScanNet++ dataset. † is for fully-supervised method utilized labeled 3D bounding boxes during training.

Method	Zero-shot	Novel	Base	All
Det-PointCLIPv2† [14]	✗	0.1	1.0	0.2
3D-CLIP† [6]	✗	2.5	11.2	3.9
CoDA† [2]	✗	6.5	21.6	9.0
INHA† [3]	✗	7.8	25.1	10.7
OV-Uni3DETR† [11]	✗	13.7	**48.1**	19.4
Zoo3D$_0$	✓	29.3	16.2	27.1
Zoo3D$_1$	✗	**33.6**	24.4	**32.0**

Table 2. Results of open-vocabulary 3D object detection of *base*, *novel*, and *all* object categories on ScanNet60. Methods marked with † use ground truth 3D bounding boxes for objects of *base* classes.

Method	Venue	mAP$_{25}$	mAP$_{50}$
MaskClustering [12]	CVPR'24	13.4	8.5
OnlineAnySeg [8]	CVPR'25	19.0	12.3
Zoo3D$_0$	-	**21.1**	**14.1**

Table 3. Comparison with zero-shot 3D instance segmentation methods on ScanNet200.

ARKitScenes was used to train DUSt3R, so we cannot use it in zero-shot experiments for fair comparison. In this series of experiments, DUSt3R is replaced with VGGT to preserve methodological purity. Since VGGT does not support camera poses as inputs natively, we only report quality in point cloud-based and unposed images-based tracks in Tab. 5.

Per-category scores on ScanNet20 are given in Tab. 4. Evidently, both our methods outperform the competitors by a large margin, with the most significant gains achieved for *chair* (+45.7 mAP$_{25}$ for Zoo3D$_1$ over previous state-of-the-art), *sofa* (+20.9), *bookshelf* (+19.7), *refridgerator* (+43.1), *nightstand* (+44.0), *lamp* (+37.7). The average mAP$_{25}$ of Zoo3D$_0$ is 9.4 higher than of any existing approach, while Zoo3D$_1$ further expands the gap to 11.9 mAP$_{25}$.

Methods	mAP_{25}	toilet	bed	chair	sofa	dresser	table	cabinet	bookshelf	pillow	sink
OV-3DET [5]	18.0	57.3	42.3	27.1	31.5	8.2	14.2	3.0	5.6	23.0	31.6
CoDA [2]	19.3	68.1	44.1	28.7	44.6	3.4	20.2	5.3	0.1	28.0	45.3
OV-Uni3DETR [11]	25.3	86.1	50.5	28.1	31.5	18.2	**24.0**	6.6	12.2	29.6	**54.6**
$Zoo3D_0$	34.7	**91.1**	51.2	53.4	60.5	31.9	20.2	12.9	**41.9**	32.2	25.7
$Zoo3D_1$	**37.2**	78.4	**54.4**	**74.4**	**65.5**	**33.6**	19.1	**14.1**	32.3	**46.1**	27.3
		bathtub	refrigerator	desk	nightstand	counter	door	curtain	box	lamp	bag
OV-3DET [5]		56.3	11.0	19.7	0.8	0.3	9.6	10.5	3.8	2.1	2.7
CoDA [2]		50.5	6.6	12.4	15.2	0.7	8.0	0.0	2.9	0.5	2.0
OV-Uni3DETR [11]		63.7	14.4	**30.5**	2.9	**1.0**	1.0	**19.9**	12.7	5.6	**13.5**
$Zoo3D_0$		50.0	50.5	11.2	**59.2**	0.1	21.1	18.2	17.8	34.8	9.8
$Zoo3D_1$		**64.6**	**57.5**	10.7	58.8	0.2	**27.4**	8.0	**20.0**	**43.3**	9.1

Table 4. Per-class 3D object detection scores on the ScanNet20.

Method	Zero-shot shot	mAP_{25}	mAP_{50}
Point cloud + posed images			
$Zoo3D_0$	✓	24.4	11.0
$Zoo3D_1$	✗	**34.2**	**24.2**
Unposed images			
VGGT [9] → $Zoo3D_0$	✓	13.0	2.6
VGGT [9] → $Zoo3D_1$	✗	**16.1**	**3.5**

Table 5. Results of open-vocabulary 3D object detection from points clouds on ARKitScenes.

B. Ablation Experiments

VGGT vs DUSt3R for point cloud reconstruction is evaluated in Tab. 6. Evidently, VGGT surpasses DUSt3R by a large margin, which can be expected, since VGGT was trained on ScanNet. Unfortunately, this also means that we cannot use it in the zero-shot setting; so despite the superior performance of VGGT, we employ DUSt3R as our primary method in our ScanNet-based experiments.

Level assignment strategy In Tab. 7, we ablate assigner parameters. In the class-agnostic mode, object classes remain unknown during the training, so we cannot apply the category-aware assignment scheme used in the original TR3D. Namely, we try assigning all objects to the 16 cm-level or 32 cm-level, or split the objects based on their size 50/50. The results demonstrate that assigning all objects to the 16-cm level yields the best performance.

Alignment with g/t in image-based scenarios In Tab. 8, we report results obtained with different alignment strategies. Here, point clouds are reconstructed from images; still, ground truth annotations are needed for evaluation, since scans must be transformed into common coordinate space before computing the metrics. To estimate an affine transformation that aligns ground truth and predicted point clouds, we apply two strategies. In the first strategy, we use both ground truth and predicted depth and pose for the first frame. The predicted pose is aligned with the ground truth pose, giving the rotation and translation. The scale is estimated from a single ground-truth depth map as the median of per-pixel ratios of ground-truth to predicted depth. In the second strategy, we use ground truth and predicted camera poses of two first frames in a sequence. The first predicted pose is aligned with the first ground truth pose, giving the rotation and translation. The relative scale coefficient is derived as a ratio of distances between two camera poses in predicted and ground truth scans.

Method	Trained on ScanNet	mAP_{25}	mAP_{50}
DUSt3R [10] → $Zoo3D_1$	✗	19.0	3.9
VGGT [9] → $Zoo3D_1$	✓	**28.2**	**6.4**

Table 6. Results of class-agnostic $Zoo3D_1$ from unposed images on ScanNet200 with different pose estimation methods.

Object assignment	mAP_{25}	mAP_{50}
all objects to 32-cm level	31.2	9.8
50/50	34.5	12.3
all objects to 16-cm level	**36.1**	**13.9**

Table 7. Results of class-agnostic $Zoo3D_1$ from posed images on ScanNet200 with different assignment strategies.

Alignment strategy	mAP_{25}	mAP_{50}
first 2 poses	4.1	0.8
first depth + first pose	**8.3**	**2.9**

Table 8. Results of $Zoo3D_0$ from unposed images on ScanNet200 with different alignment stratagies.

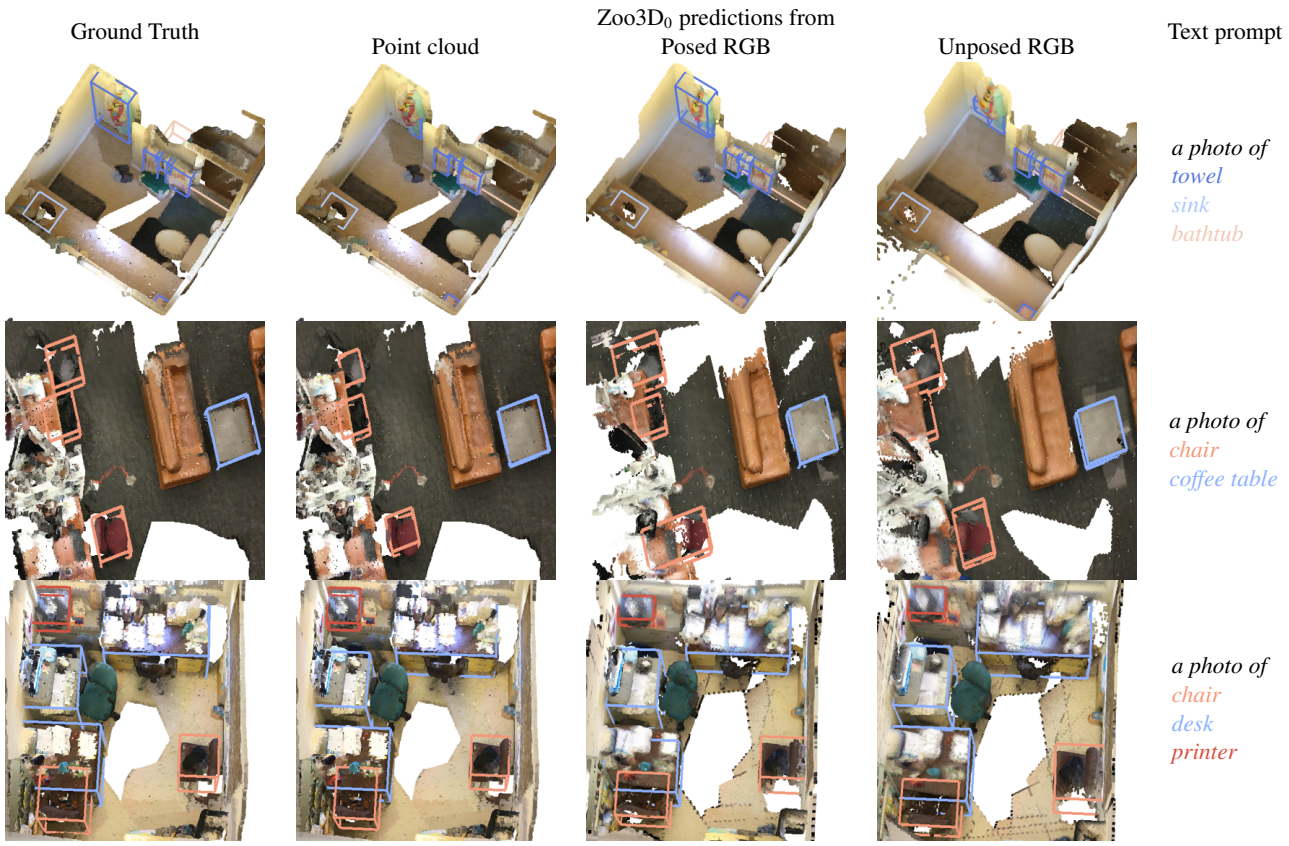

Figure 1. Qualitative results of Zoo3D$_0$ on ScanNet200.

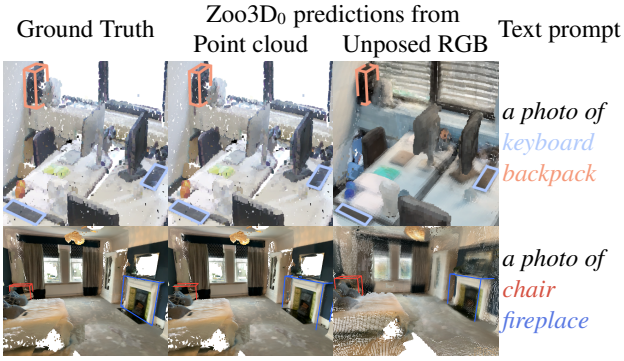

Figure 2. Qualitative results of Zoo3D$_0$ on ScanNet++ (top row) and ARKitScenes (bottom row).

C. Qualitative Results

Open-vocabulary 3D object detection results on ScanNet200 are shown in Fig. 1, while results on ARKitScenes and ScanNet++ are presented in Fig. 2. We visualize predictions from different input modalities to provide intuition of how the prediction accuracy depends on the amount of information passed to the model.

Failure cases are depicted in Fig. 3. Here we challenge our model in restrictive posed and unposed image-based scenarios. As metric values suggest, our model is prone to errors of various types: it can misclassify detected objects or even miss them in the scene entirely.

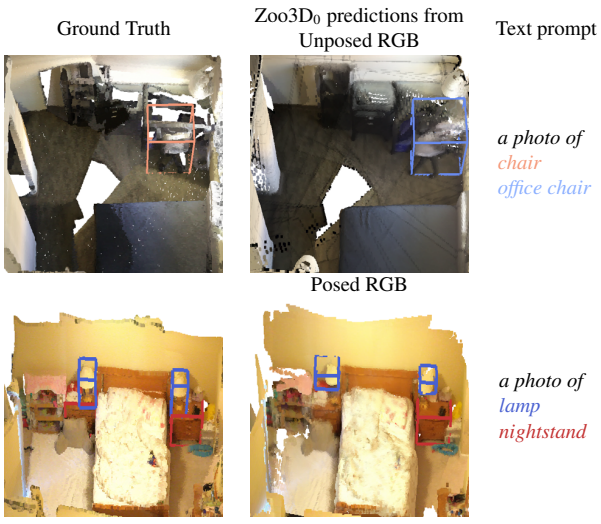

Figure 3. Failure cases of Zoo3D$_0$ on ScanNet200.

References

[1] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 1

[2] Yang Cao, Zeng Yihan, Hang Xu, and Dan Xu. Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection. *Advances in Neural Information Processing Systems*, 36:71862–71873, 2023. 1, 2

[3] Pengkun Jiao, Na Zhao, Jingjing Chen, and Yu-Gang Jiang. Unlocking textual and visual wisdom: Open-vocabulary 3d object detection enhanced by comprehensive guidance from text and image. In *European Conference on Computer Vision*, pages 376–392. Springer, 2024. 1

[4] Maksim Kolodiazhnyi, Anna Vorontsova, Matvey Skripkin, Danila Rukhovich, and Anton Konushin. Unidet3d: Multi-dataset indoor 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4365–4373, 2025. 1

[5] Yuheng Lu, Chenfeng Xu, Xiaobao Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. Open-vocabulary point-cloud object detection without 3d annotation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1190–1199, 2023. 2

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1

[7] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Tr3d: Towards real-time indoor 3d object detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 281–285. IEEE, 2023. 1

[8] Yijie Tang, Jiazhao Zhang, Yuqing Lan, Yulan Guo, Dezun Dong, Chenyang Zhu, and Kai Xu. Onlineanyseg: Online zero-shot 3d segmentation by visual foundation model guided 2d mask merging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3676–3685, 2025. 1

[9] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 2

[10] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 2

[11] Zhenyu Wang, Yali Li, Taichi Liu, Hengshuang Zhao, and Shengjin Wang. Ov-uni3detr: Towards unified open-vocabulary 3d object detection via cycle-modality propagation. In *European Conference on Computer Vision*, pages 73–89. Springer, 2024. 1, 2

[12] Mi Yan, Jiazhao Zhang, Yan Zhu, and He Wang. Maskclustering: View consensus based mask graph clustering for open-vocabulary 3d instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28274–28284, 2024. 1

[13] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 1

[14] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2639–2650, 2023. 1