OpenTrack3D: Towards Accurate and Generalizable Open-Vocabulary 3D Instance Segmentation

Supplementary Material

1. More Implementation Details

In this section, we provide additional implementation details of our MLLM components to improve understanding and reproducibility. We deploy Qwen3-4B using vLLM [5] on a single A100 GPU and use the default generation settings unless otherwise noted.

For the SceneFun3D [3] dataset, the user-provided task descriptions cannot be directly processed by open-vocabulary detectors. To bridge this gap, we use the same MLLM to convert each textual query into a list of directly operable affordance nouns using the following prompt: `Extract directly operable affordance elements from the given task description (e.g., handle, knob, switch, latch, lever). Do not include whole objects. Output a comma-separated list of nouns. Task: {}.` These extracted nouns are then fed into the open-vocabulary detector.

For datasets [2, 10, 12] with predefined category sets, we adopt a unified prompt template: `Identify the object shown in the red rectangle. Return the class name from {}. If no object is shown, return 'none'` In contrast, for SceneFun3D, we employ a specialized prompt that maps all task descriptions in a scene to their corresponding task indices in a single forward pass: `The interactive element is shown in the red rectangle. Determine whether the element in the image can accomplish any task in the task list. If so, return the task index; otherwise return 'no match'. Tasks: {}.` This batch-style prompting significantly accelerates inference compared with querying each proposal independently.

2. Hyperparameter Sensitivity Analysis

τ_{match}	AP	AP_{50}	AP_{25}
0.6	26.4	38.3	46.4
0.5	28.2	39.6	47.9
0.4	**28.6**	**40.5**	**48.8**
0.3	27.3	38.2	45.9

Table 1. Sensitivity analysis of the τ_{match} parameter on subsets of the ScanNet200 validation set.

τ_{exp}	AP	τ_{vis}	AP	γ	AP	τ_{merge}	AP
0.00	28.0	0.1	28.6	0.5	27.6	0.8	27.6
0.03	28.2	0.2	**29.1**	0.4	28.6	0.7	28.4
0.05	**28.6**	0.3	27.2	0.3	**28.6**	0.6	**28.6**
0.07	28.0	0.4	24.9	0.2	27.0	0.5	27.0

Table 2. Sensitivity analysis of hyperparameters in the Proposal Refinement module on subsets of the ScanNet200 validation set.

	ScanNet200			SceneFun3D	
K	AP	AP_{50}	AP_{25}	AP_{50}	AP_{25}
1	28.6	40.5	48.8	9.8	20.6
2	28.8	40.8	49.0	10.2	20.8
3	**28.9**	**41.8**	**49.5**	**10.5**	**21.2**

Table 3. Ablation study on the MLLM top-K parameter on subsets of the ScanNet200 and SceneFun3D validation sets.

We analyze the sensitivity of key hyperparameters used in our framework. As shown in Tab. 1 and Tab. 2, the results remain stable across a wide range of values in both the Proposal Generation and Proposal Refinement modules.

We also evaluate the effect of varying the MLLM top-K parameter, as summarized in Tab. 3. Larger K generally improves performance but increases inference latency; in practice, users can choose an appropriate trade-off based on computational constraints.

3. Runtime Analysis

We analyze the runtime of our framework and compare it with the previous state-of-the-art, DetailMatters [4]. Most of our runtime is spent on model inference. The cost of the 2D open-vocabulary segmenter and DINO scales with the number of video frames, while the proposal-classification cost (via either MLLM or CLIP [9]) mainly depends on the number of proposals.

DetailMatters [4] follows the Open3DIS [7] pipeline and employs Grounding-DINO [6] as its 2D open-vocabulary detector. Because our method additionally incorporates a DINO [8] model, we choose to pair it with a more lightweight 2D open-vocabulary detector, YOLO-World [1], to balance the runtime overhead. As shown in Tab. 4, despite using one additional model, our overall runtime is approximately half that of DetailMatters.

The gap becomes even more pronounced in the proposal-classification stage. DetailMatters extracts the top 20 images for each proposal and applies multi-scale cropping, re-

	2D OV-SEG	DINO	Frames	Total
DetailMatters [4]	1.4	-	346	484
Ours	0.52	0.25	346	266

Table 4. Runtime of the 2D open-vocabulary segmenter and DINO. We report the per-frame runtime (in seconds), the average number of frames per scene, and the average runtime per scene on the ScanNet200 validation set.

	Classification	Proposals	Total
DetailMatters [4]	5.9	56	330.4
Ours (K=3)	1.6	56	89.6
Ours (K=1)	1.1	56	61.6

Table 5. Runtime of the proposal-classification stage. We report the per-proposal runtime (in seconds), the average number of proposals per scene, and the average runtime per scene on the ScanNet200 validation set.

sulting in roughly 60 Alpha-CLIP [11] forward passes per proposal, in addition to substantial preprocessing overhead. In contrast, our pipeline directly overlays a red bounding box on the original image to highlight the target and only feeds the top 3 images into the MLLM. As shown in Tab. 5, our approach achieves substantially lower inference latency compared to DetailMatters.

For fairness, all reported DetailMatters timings are obtained from our custom re-implementation, and are generally lower than the values reported in the original paper due to differences in implementation and hardware.

References

[1] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024. 1

[2] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 1

[3] Alexandros Delitzas, Ayca Takmaz, Federico Tombari, Robert Sumner, Marc Pollefeys, and Francis Engelmann. Scenefun3d: Fine-grained functionality and affordance understanding in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14531–14542, 2024. 1

[4] Sanghun Jung, Jingjing Zheng, Ke Zhang, Nan Qiao, Albert YC Chen, Lu Xia, Chi Liu, Yuyin Sun, Xiao Zeng, Hsiang-Wei Huang, et al. Details matter for indoor open-vocabulary 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9627–9637, 2025. 1, 2

[5] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. 1

[6] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024. 1

[7] Phuc Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4018–4028, 2024. 1

[8] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[9] 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

[10] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 1

[11] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13019–13029, 2024. 2

[12] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes, 2023. 1