Vocabulary-Free 3D Instance Segmentation with Vision and Language Assistant

Supplementary Material

We first provide more details of the superpoint merging procedure via spectral clustering. Then we provide additional comparative analysis on the S3DIS dataset, and we perform an ablation on the hyperparameters of PoVo. Lastly, we present more qualitative results on ScanNet200 and Replica datasets and implementation details for VoF3DIS setting.

A. Spectral clustering

To cluster the superpoints more efficiently, we propose a hierarchical spectral clustering algorithm to generate masks. First, a Hilbert curve is applied to serialize the superpoints based on the coordinates of the center point $\mathbf{q}_i = \frac{1}{N_i} \sum_{v_j \in Q_i} \mathbf{p}_j$ for each superpoint Q_i, where $\mathbf{p}_j$ is a point in Q_i and N_i is the number of points in Q_i. Next, a sliding window is used to divide the serialized superpoints into K_s groups, with each group containing N_s superpoints. The window shifts by a stride of N_s. Spectral clustering is then applied to each group to generate coarse masks. This is an iterative spectral clustering process, where masks are merged with those from neighboring groups based on the overlap scores between two coarse masks, M_s and M_t. The overlap score between M_s and M_t is determined by selecting the maximum similarity (from the affinity matrix A) between superpoints, where one superpoint belongs to M_s and the other to M_t. Clustering combines region sets from merged coarse masks until $M_s * M_t$ is less than the threshod $\tau_{iou} * \tau_{sim}$. In our implementation, we first divide the superpoints into two groups for each 3D scene, followed by an iterative spectral clustering process.

B. Additional experiments

B.1. S3DIS dataset

S3DIS [1] consists of 271 scenes that cover 13 classes in 6 areas. We adopt the Open3DIS categorization strategy [4] for S3DIS, which splits the dataset into base and novel sets. The novel set includes two parts **N4** and **N6**. PLA [2], Lowis3D [3], and Open3DIS [4] train on the base classes of S3DIS data, whereas the novel classes are not seen during training. Note that our method is zero-shot and does not require any training (neither base nor novel classes).

Tab. 1 compares the performance of our method with other approaches on S3DIS, focusing on Average Precision (AP) at 50% (AP_{50}) IoU for novel classes. The table compares methods across two settings: open-vocabulary and vocabulary-free (VoF3DIS) setting. In the open-vocabulary setting, PoVo achieves the highest scores despite using 2D masks, with N4 AP_{50} of 29.1 and N6 AP_{50} of 33.4, outperforming methods like Open3DIS, which records 26.3 and 29.0 for N4 AP_{50} and N6 AP_{50}, respectively. Also in the vocabulary-free setting PoVo outperforms the other methods, achieving N4 AP_{50} of 28.4 and N6 AP_{50} of 29.7. Although this performance is slightly lower than in the open-vocabulary setting, it outperforms Open3DIS, which achieves 24.6 and 26.3 for N4 AP_{50} and N6 AP_{50}, respectively. These results demonstrate that PoVo provides superior performance in both open-vocabulary and vocabulary-free settings, highlighting its robustness and effectiveness in various 3D instance segmentation scenarios.

Table 1. OV-3DIS results on S3DIS in terms of AP_{50}.

Method	Mask type	**N4** AP_{50}	**N6** AP_{50}
Open-vocab. semantic			
PLA [2]	3D mask	8.6	9.8
Lowis3D [3]	3D mask	13.8	15.8
Open3DIS [4]	3D mask	26.3	29.0
PoVo [4]	2D mask	**29.1**	**33.4**
2D mask + Vocab.-free semantic			
Open3DIS [4]	3D mask	24.6	26.3
PoVo [4]	2D mask	**28.4**	**29.7**

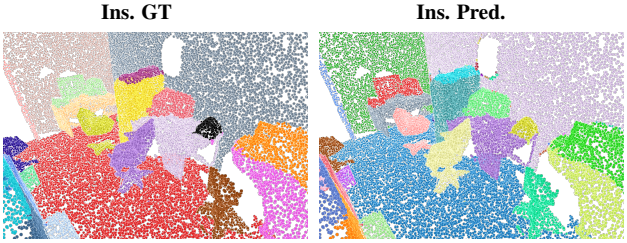

Figure 1. Qualitative results obtained by PoVo in the VoF3DIS setting. Left to right: ground truth instance, predicted instance.

Fig. 1 further presents two examples of instance segmentation results on S3DIS. Compared to the ground truth, our PoVo produces nearly identical results. This shows that our approach effectively leverages LLaVA-guided 2D mask prediction in conjunction with a superpoints-based spectral clustering strategy, resulting in high-quality instance segmentation results.

B.2. Ablation study on different values of IoU and similarity threshold

We use mask IoU and text similarity to guide 3D instance generation (superpoint merging). This involves two hyperparameters: the IoU threshold τ_{iou} and the similarity threshold τ_{sim}, which determine how IoU and text similarity are con-

Table 2. Ablation study on IoU threshold (τ_{iou}) and its impact on Average Precision (AP) and AP at 50% IoU (AP_{50}).

τ_{iou}	0.5	0.7	0.8	0.9	0.95
Open-vocabulary					
AP	21.9	22.2	22.3	**22.4**	22.1
AP_{50}	27.4	27.6	**27.9**	**27.9**	27.5
Vocabulary-free					
AP	21.2	21.3	18.0	**21.6**	16.9
AP_{50}	26.2	26.4	26.6	**26.7**	26.4

Table 3. Ablation study on IoU threshold (τ_{sim}) and its impact on Average Precision (AP) and AP at 50% IoU (AP_{50}).

τ_{iou}	0.5	0.6	0.7	0.8	0.9	0.95
Vocabulary-free						
AP	20.5	20.9	21.2	21.4	**21.6**	21.4
AP_{50}	25.6	26.2	26.4	26.5	**26.7**	26.3

sidered when deciding whether two superpoints belong to the same 3D instance. To evaluate the impact of the IoU threshold τ_{iou} and the similarity threshold τ_{sim} on the performance of instance segmentation, we report the performance of 3D instance mask formation from 2D masks extracted by Grounded-SAM from multi-view RGB images, using different values for the IoU threshold τ_{iou} and similarity threshold τ_{sim} in Tab. 2 and Tab. 3. The tests were conducted on the ScanNet200 validation set.

For the open-vocabulary setting, Tab. 2 reports the results on τ_{iou} quantified using average precision (AP), and AP at 50% IoU (AP_{50}), by varying τ_{iou} values from 0.5 to 0.95. AP gradually increases from 21.9 at $\tau_{iou} = 0.5$ to a peak of 22.4 at $\tau_{iou} = 0.9$, before slightly decreasing at $\tau_{iou} = 0.95$. AP_{50} follows a similar trend, peaking at 27.9 for both $\tau_{iou} = 0.8$ and $\tau_{iou} = 0.9$. In the vocabulary-free setting, AP also shows an increase, reaching its highest value of 21.6 at $\tau_{iou} = 0.9$ before dropping at $\tau_{iou} = 0.95$. The AP_{50} increases from 26.2 at $\tau_{iou} = 0.5$ to a maximum of 26.7 at $\tau_{iou} = 0.9$, and then slightly decreases at $\tau_{iou} = 0.95$. These results indicate that using a higher τ_{iou} threshold up to 0.9 generally improves performance in both settings. However, pushing the threshold to 0.95 can lead to performance degradation, suggesting that $\tau_{iou} = 0.9$ provides the best results.

For the setting without vocabulary, Tab. 3 reports the results of τ_{sim} using AP, and AP_{50}. By varying τ_{sim} values from 0.5 to 0.9, both AP and AP_{50} improve, with AP peaking at 21.6 and AP_{50} reaching its highest value of 26.7 at $\tau_{\text{sim}} = 0.9$. However, further increasing the threshold to 0.95 results in a slight decrease in both metrics, with AP dropping to 21.4 and AP_{50} to 26.3. These results suggest that while a higher similarity threshold generally enhances performance, setting the threshold too high can lead to lower performance, indicating that $\tau_{\text{sim}} = 0.9$ offers the best results.

B.3. Visualizations on Replica

Fig. 2 shows the qualitative results of text-driven 3D instance segmentation. Our model can successfully segment instances based on various types of input text prompts, which can include object categories (such as pottery) not present in the predefined labels, or objects' functionality (throw away garbage).

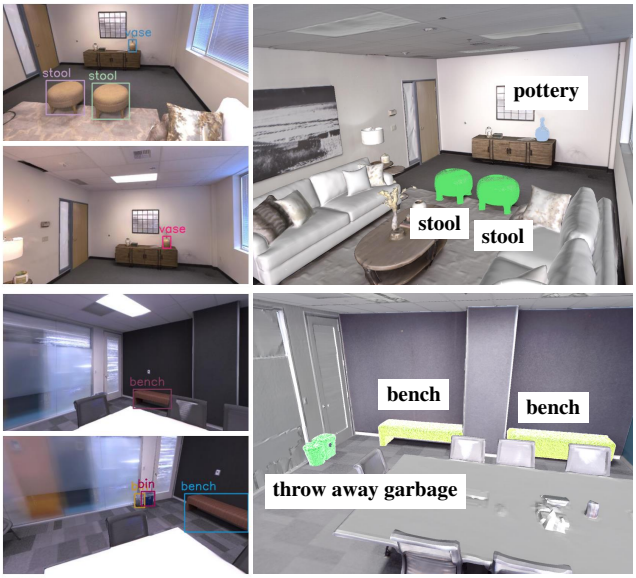

Figure 2. Qualitative results of two examples obtained by PoVo in the VoF3DIS setting on Replica dataset. The instance with the highest similarity score to the query's embedding is highlighted in the point clouds. In the images, each box outlines the regions of the objects detected by Grounded-SAM based on the queries.

Fig. 3 presents qualitative results from two examples of instance segmentation achieved using our method in the VoF3DIS setting. For each example, we provide 2D instance segmentation results from two different view RGB images of a 3D scene, 3D instance ground truth and our prediction. Our method demonstrates visually robust performance without relying on predefined categories.

B.4. Visualizations on ScanNet200.

To illustrate the quality of segmentation, we provide additional visualizations on ScanNet200. Fig. 4 presents four examples of instance segmentation results. Ideally, different instances should have distinct colors, while the same instance should maintain consistent coloring. It's not necessary for objects to match colors exactly between the ground truth and predictions, but semantic success is achieved when objects match the ground truth colors. As shown, PoVo accurately segments most of the scene for both instance and semantic segmentation using only the LLaVA-provided vocabulary.

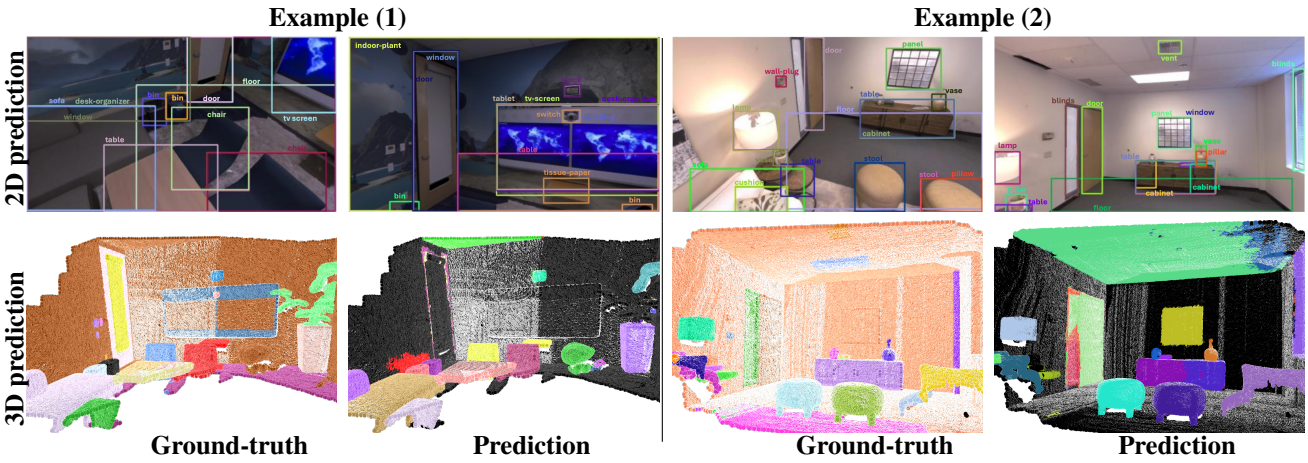

Figure 3. Qualitative results were obtained using PoVo in the VoF3DIS setting on Replica. The process begins by generating 2D instance masks for each image by querying a vocabulary established by LLaVA (as shown in the top row, where the words are displayed in each image). These 2D masks (highlighted by boxed regions in each image) are then integrated into 3D masks (bottom row: Prediction columns) using spectral clustering techniques.

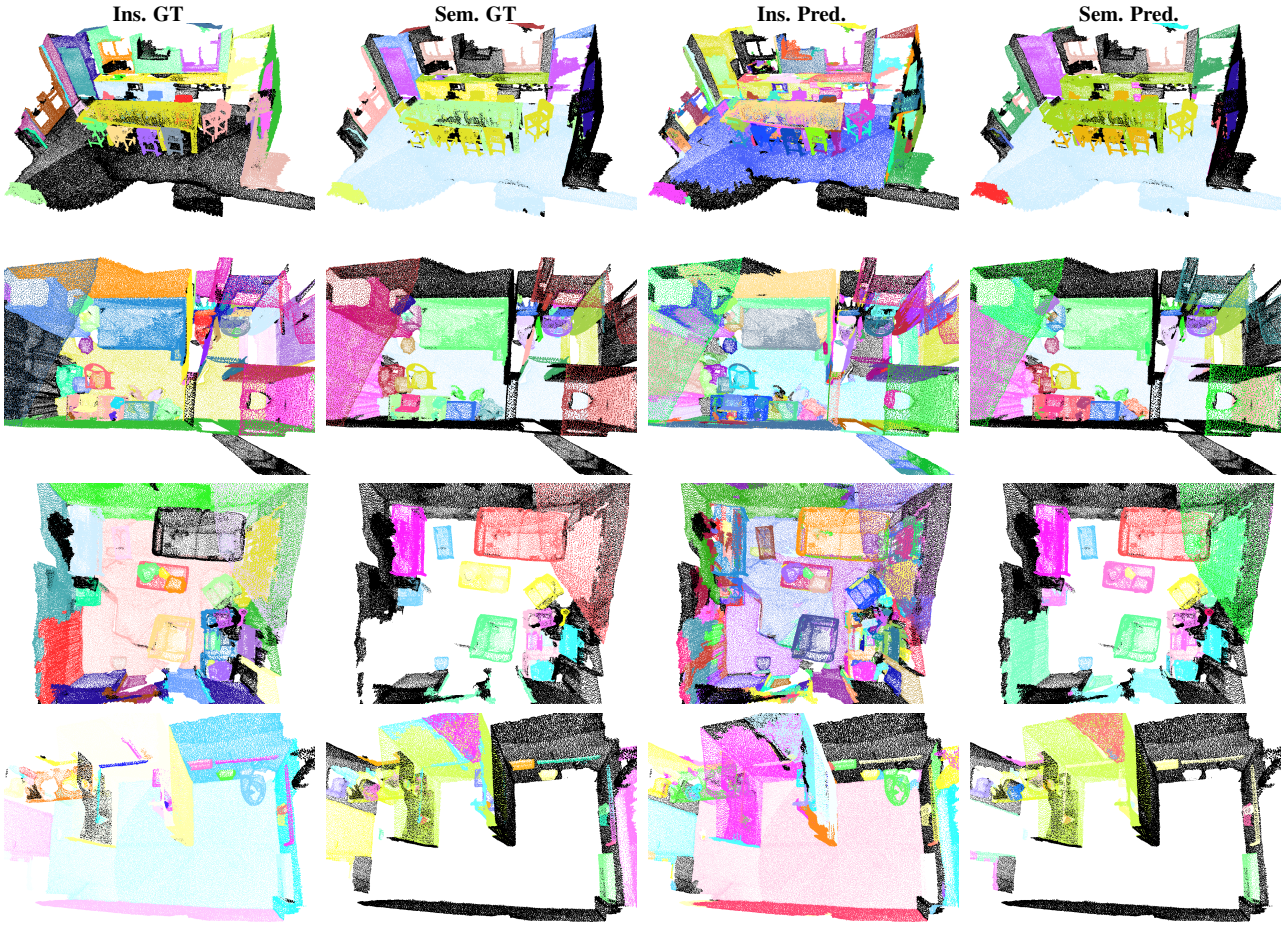

Figure 4. Qualitative results obtained by PoVo in the VoF3DIS setting on ScanNet200 are presented. From left to right: ground truth instance labels, ground truth semantic labels, predicted 3D instance labels, and predicted 3D semantic labels. In ideal instance segmentation within a 3D scene, different instances should be colored differently, while the same instance should have a consistent color. It is not necessary for the same object to have the same color in both the ground truth and predicted results. For semantic prediction, success is indicated when each object matches the ground truth color.

C. Implementation details for VoF3DIS setting

In our experiments, because the categories provided by the annotated datasets are fewer than the actual objects they contain, we evaluate our method in the VoF3DIS setting by restricting it to the categories detected by both LLaVA and those provided in the datasets.

References

[1] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *CVPR*, 2016. 1

[2] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. PLA: Language-Driven Open-Vocabulary 3D Scene Understanding. In *CVPR*, 2023. 1

[3] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Lowis3d: Language-driven open-world instance-level 3d scene understanding. *IEEE TPAMI*, 2024. 1

[4] 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 *CVPR*, pages 4018–4028, 2024. 1