MV3DIS: Multi-View Mask Matching via 3D Guides for Zero-Shot 3D Instance Segmentation

Supplementary Material

A. Implementation Details

Region-level affinity. Given the superpoint graph, let $\mathcal{N}(i)$ denote the neighbors of $\mathbf{S}_i$ in the superpoint graph, and let $\mathcal{N}_k(i) = \mathcal{N}(i) \cap \mathbf{U}_k$ be the subset of neighbors that belong to region $\mathbf{U}_k$. In region growing, the weighted affinity $A(\mathbf{U}_k, \mathbf{S}_i)$ between a candidate superpoint $\mathbf{S}_i$ and the region $\mathbf{U}_k$ is defined as the weighted average of affinity scores $A_{i,j}$ over $\mathbf{S}_j \in \mathcal{N}_k(i)$:

$$A(\mathbf{U}_k, \mathbf{S}_i) = \frac{\sum_{\mathbf{S}_j \in \mathcal{N}_k(i)} w_{ij}\, A_{i,j}}{\sum_{\mathbf{S}_j \in \mathcal{N}_k(i)} w_{ij}}, \tag{1}$$

where $w_{ij} = \frac{n_j}{d_{ij}}$, with d_{ij} denoting the Euclidean distance between $\mathbf{S}_i$ and $\mathbf{S}_j$, and n_j the number of points contained in superpoint $\mathbf{S}_j$. The weight function w_{ij} prioritizes neighbors that are closer and contain more points.

Algorithm 1 Superpoint Merging on the Superpoint Graph

Require: Superpoints $\{\mathbf{S}_i\}_{i=1}^{N_s}$, adjacency $\mathcal{N}(i)$, affinity matrix $\mathbf{A}$ with entries $A_{i,j}$, point counts $n_i = |\mathbf{S}_i|$, merging threshold τ_{merge}
Ensure: 3D segments $\{\mathbf{U}_k\}_{k=1}^{N_u}$

$Assigned[i] \leftarrow$ **false** for $i = 1, \ldots, N_s$
$k \leftarrow 0$
while there exists i with $Assigned[i]$ = **false do**
  $s \leftarrow$ any i with $Assigned[i] =$ **false**
  $k \leftarrow k + 1$
  $\mathbf{U}_k \leftarrow \{s\}$
  Initialize an empty queue Q
  Q.push(s)
  $Assigned[s] \leftarrow$ **true**
  while Q is not empty **do**
    $p \leftarrow Q$.pop()
    for each $i \in \mathcal{N}(p)$ with $Assigned[i]$ = **false do**
      $\mathcal{N}_k(i) \leftarrow \mathcal{N}(i) \cap \mathbf{U}_k$
      if $\mathcal{N}_k(i) \neq \emptyset$ **then**
        $a \leftarrow A(\mathbf{U}_k, \mathbf{S}_i)$ {by (1)}
        if $a \geq \tau_{\text{merge}}$ **then**
          $\mathbf{U}_k \leftarrow \mathbf{U}_k \cup \{i\}$
          $Assigned[i] \leftarrow$ **true**
          Q.push(i)
        end if
      end if
    end for
  end while
end while

Region refinement. Region growing Algorithm 1 greedily assigns each superpoint to the first region whose affinity exceeds the merging threshold. This may lead to suboptimal assignments, especially for boundary superpoints that are adjacent to multiple regions with comparable affinities. To further improve the segmentation quality, we recompute the superpoint affinity matrix using the refined 2D segmentation maps $\{\mathbf{L}_t^{\text{R}}\}_{t=1}^{T}$. We then refine the initial segments by reassigning superpoints to neighboring regions with the highest region-level affinity. To balance performance and efficiency, we limit the number of region changes per superpoint by $I_{\max}$, which is set to 2 by default and is not extensively tuned in our experiments. The refinement procedure is summarized in Algorithm 2.

Parameter settings. All experiments are conducted on a single NVIDIA RTX 3090 GPU. We set the merging threshold τ_{merge} to 0.7 for the ScanNet++ and Replica, and 0.5 for the ScanNetV2, ScanNet200, and Matterport3D.

Evaluation settings. Following prior works [5, 10, 13, 14], we ignore instances labeled as "wall" or "floor" in the ScanNetV2 and ScanNet200 datasets. For ScanNet++, we employ all ground-truth labels provided for the instance segmentation task. For Matterport3D, we use the 160 most common categories from [6], excluding "wall", "floor", and "ceiling". For Replica, we follow the evaluation protocol of Open3DIS [5], and evaluate on the "office0", "office1", "office2", "office3", "office4", "room0", "room1", and "room2" scenes, with ground-truths provided by [12].

B. Additional Quantitative Results

Class-Agnostic 3D Instance Segmentation. To further validate the robustness and generalization capability of our method, we perform experiments on the Replica [11] and Matterport3D [1] datasets, as reported in Tabs. 1 and 2. For a fair comparison on Replica, we adopt SAM2 [7] as the 2D segmenter for all methods. Our method achieves the best performance, surpassing the baseline Open3DIS [5] by 3.6 mAP and 3.8 AP_{50}. Furthermore, on the Matterport3D dataset, under the same setting using SAM2 as the 2D segmenter, our method achieves a 2.0 mAP improvement over SAI3D [14]. These gains highlight the effectiveness and generalization of our method.

Analysis of small and long-tail instances. We further analyze the class-agnostic instance segmentation performance of MV3DIS on ScanNet200, broken down by category frequency (head/common/tail) and instance size (small/medium/large, split at the 33%/66% percentiles of

Algorithm 2 Region Refinement on the Superpoint Graph

Require: Adjacency $\mathcal{N}(i)$, affinity matrix $\mathbf{A}$ with entries $A_{i,j}$, initial segments $\{\mathbf{U}_k\}_{k=1}^{N_u}$ from region growing, maximum region changes per superpoint $I_{\max}$
Ensure: Updated 3D segments $\{\mathbf{U}_k\}_{k=1}^{N_u}$

Initialize $\text{Region}[i]$ such that $i \in \mathbf{U}_{\text{Region}[i]}$ for all i
Initialize $\text{Count}[i] \leftarrow 0$ for all i
changed $\leftarrow$ **true**
while changed $=$ **true do**
  changed $\leftarrow$ **false**
  for each superpoint index $i = 1, \ldots, N_s$ **do**
    if $\text{Count}[i] \geq I_{\max}$ **then**
      continue
    end if
    $k_{\text{cur}} \leftarrow \text{Region}[i]$
    $\mathcal{R}_i \leftarrow \{\text{Region}[j] \mid j \in \mathcal{N}(i), \text{Region}[j] \neq k_{\text{cur}}\}$
    if $\mathcal{R}_i = \emptyset$ **then**
      continue
    end if
    $\mathcal{C}_i \leftarrow \mathcal{R}_i \cup \{k_{\text{cur}}\}$
    $k^* \leftarrow k_{\text{cur}}, a_{\max} \leftarrow -\infty$
    for each region index $k \in \mathcal{C}_i$ **do**
      $\mathcal{N}_k(i) \leftarrow \mathcal{N}(i) \cap \mathbf{U}_k$
      if $\mathcal{N}_k(i) = \emptyset$ **then**
        continue
      end if
      $a_k \leftarrow A(\mathbf{U}_k, \mathbf{S}_i)$ {by (1)}
      if $a_k > a_{\max}$ **then**
        $a_{\max} \leftarrow a_k, k^* \leftarrow k$
      end if
    end for
    if $k^* \neq k_{\text{cur}}$ **then**
      $\mathbf{U}_{k_{\text{cur}}} \leftarrow \mathbf{U}_{k_{\text{cur}}} \setminus \{i\}$
      $\mathbf{U}_{k^*} \leftarrow \mathbf{U}_{k^*} \cup \{i\}$
      $\text{Region}[i] \leftarrow k^*$
      $\text{Count}[i] \leftarrow \text{Count}[i] + 1$
      changed $\leftarrow$ **true**
    end if
  end for
end while

GT instance point counts). Across all splits, our mAP consistently improves over Open3DIS across all splits: 34.4 / 40.6 / 43.9 vs. 29.1 / 34.2 / 37.8 for frequency, and 32.6 / 40.6 / 36.2 vs. 29.8 / 37.2 / 27.4 for size. These results indicate stronger robustness on long-tail categories and small instances.

Runtime analysis. As shown in Tab. 3, we measure the runtime of each component on a ScanNetV2 scene with about 147K points. Our method takes about 19 s, while SAI3D [14], which performs multiple rounds of region growing, requires about 15 s. The additional cost mainly arises from the region refinement and mask matching, which are critical for improving the 3D instance segmentation quality. To reduce computational overhead, we precompute the 2D projections and visibility of all 3D points, enabling efficient 3D-to-2D correspondence checks during superpoint graph construction. For each frame, we merge all masks into a single segmentation map, so that affinity matrix computation depends on the number of frames rather than the number of masks. Despite the increased runtime, our method achieves a clear improvement in segmentation performance.

Ablation study on visibility thresholds. We analyze the effect of the frame visibility threshold τ_f and mask visibility

Table 1. **Class-agnostic 3D instance segmentation on Replica.** Best and second-best results are bold and underlined, respectively. VFM represents the 2D vision foundation model.

Method	VFM	mAP	AP_{50}	AP_{25}
SAM3D [13]	SAM2	19.4	29.5	38.5
Open3DIS [5]	SAM2	22.2	<u>34.8</u>	42.3
SAI3D [14]	SAM2	<u>22.7</u>	34.5	<u>51.1</u>
Ours	SAM2	**25.8**	**38.6**	**52.2**

Table 2. **Class-agnostic 3D instance segmentation on Matterport3D.** Best and second-best results are bold and underlined, respectively. VFM represents the 2D vision foundation model.

Method	VFM	mAP	AP_{50}	AP_{25}
OVIR-3D [4]	SAM	6.6	15.6	28.3
SAM3D [13]	SAM	10.1	19.4	36.1
SAI3D [14]	SAM2	<u>19.2</u>	<u>36.7</u>	<u>57.8</u>
Ours	SAM2	**21.2**	**38.9**	**59.7**

Table 3. **Runtime of each component in MV3DIS.** SM indicates superpoint merging, 3DG-MM indicates 3D-guided mask matching, NMS indicates non-maximum suppression, and RR indicates region refinement.

	SM	3DG-MM	NMS	RR	Others	All
Time	3.1s	4.7s	0.3s	6.6s	4.6s	19.3s

Table 4. **Ablation study on τ_f and τ_m.**

τ_f	τ_m	mAP	AP_{50}	AP_{25}
0.1	0.9	38.1	59.3	75.6
0.95	0.9	36.8	57.2	74.3
0.3	0.7	38.3	**60.4**	75.8
0.3	0.05	37.5	58.1	74.9
0.3	0.9	**38.5**	60.2	**76.2**

Table 5. **Performance of MV3DIS with different 2D visual foundation models (VFMs)**. Best and second-best results are bold and underlined, respectively.

	ScanNet200			ScanNet++		
VFM	mAP	AP_{50}	AP_{25}	mAP	AP_{50}	AP_{25}
SAM2 [7]	34.3	52.5	68.3	**22.0**	**36.7**	**51.7**
YoloW-SAM [2]	<u>34.6</u>	<u>53.1</u>	<u>69.4</u>	21.2	35.5	50.9
GD-SAM [8]	**35.5**	**54.7**	**69.7**	<u>21.5</u>	<u>36.0</u>	<u>51.3</u>

threshold τ_m in Tab. 4. An overly high τ_f (0.95) excessively discards visible frames, leading to performance degradation. Furthermore, a low τ_m (0.05) introduces masks with low instance correlation, which compromises the mask consistency computation and hinders overall performance.

Ablation of different VFMs. We further evaluate the effectiveness of our method with different visual foundation models (VFMs), including SAM2 [7], YoloWorld-SAM (YoloW-SAM) [2], and Grounded-SAM (GD-SAM) [8]. As shown in Tab. 5, our method consistently achieves strong performance with all three VFMs on both ScanNet200 and ScanNet++, demonstrating robustness to the choice of VFM. Specifically, SAM2 tends to produce finer-grained segmentations, making it suitable for instance-dense, high-fidelity datasets like ScanNet++, whereas YoloWorld-SAM and Grounded-SAM are better suited for standard-resolution datasets such as ScanNet200.

Robustness to noisy depth. We evaluate the sensitivity of our method to noisy depth inputs on ScanNet200 by perturbing the input depth map (in meters) with additive Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$, where $\sigma \in \{0.01, 0.05, 0.10\}$ m. As shown in Tab. 6, the performance drops only slightly as the noise level increases, demonstrating robustness to depth noise.

Robustness of the merging threshold. In the main paper, we analyzed the merging threshold τ_{merge} on the ScanNetV2 dataset, showing that high values impose a strict merging criterion and cause over-segmentation, whereas smaller values allow low-confidence merges and result in under-segmentation. To further evaluate the robustness of MV3DIS, we tested $\tau_{\text{merge}} \in \{0.4, 0.5, 0.6, 0.7\}$ on ScanNet200 [9] and KITTI-360 [3], observing stable performance with at most 1.6 mAP variation on ScanNet200 (35.3–36.9) and 1.2 on KITTI-360 (22.3–23.5). These results demonstrate that MV3DIS is robust to moderate changes in the merging threshold across diverse scene densities.

C. Additional Qualitative Results

We first visualize the results from both the coarse and refined 3D segmentation stages. As illustrated in Fig. 1, the coarse 3D segments often suffer from over-segmentation due to insufficient multi-view consistency. In contrast, by consolidating multi-view consistency, the refined segmentation yields more complete instance masks.

Table 6. Robustness to noisy depth on ScanNet200.

Noise std. σ (m)	0	0.01	0.05	0.10
mAP	35.5	35.4	35.2	34.8

Figures 2 and 3 further present visual comparisons between our method and SAM3D [13] and SAI3D [14] on the ScanNetV2 and ScanNet++ datasets. SAM3D and SAI3D process multiple views in a frame-by-frame manner, leading to inconsistent masks across views. In contrast, our method explicitly consolidates multi-view consistency, resulting in more robust and coherent 3D instance segmentation.

In Fig. 4, we present qualitative results for semantic instance segmentation. We employ OpenMask3D [12] to extract open-vocabulary CLIP features for each segmented instance. Given text prompts such as "bag", "book", and "mouse", our method accurately segments and identifies the corresponding instances in complex 3D scenes. Furthermore, our method demonstrates robustness in handling less frequent classes, successfully segmenting objects like "door hanger" and "projector". This highlights the effectiveness and flexibility of our approach for open-vocabulary 3D semantic instance segmentation.

D. Limitations

Although our method achieves strong performance, it still has several limitations. First, it relies heavily on the quality of the initial 3D over-segmentation. When multiple objects fall into the same superpoint, the resulting instance predictions may exhibit blurry boundaries or under-segmentation. In addition, the computational and memory costs grow linearly with the number of superpoints and views, which may limit its applicability to large-scale scenes. Future work should develop finer-grained superpoint over-segmentation methods, along with more efficient superpoint graph construction and aggregation strategies, to further enhance the scalability of the proposed approach.

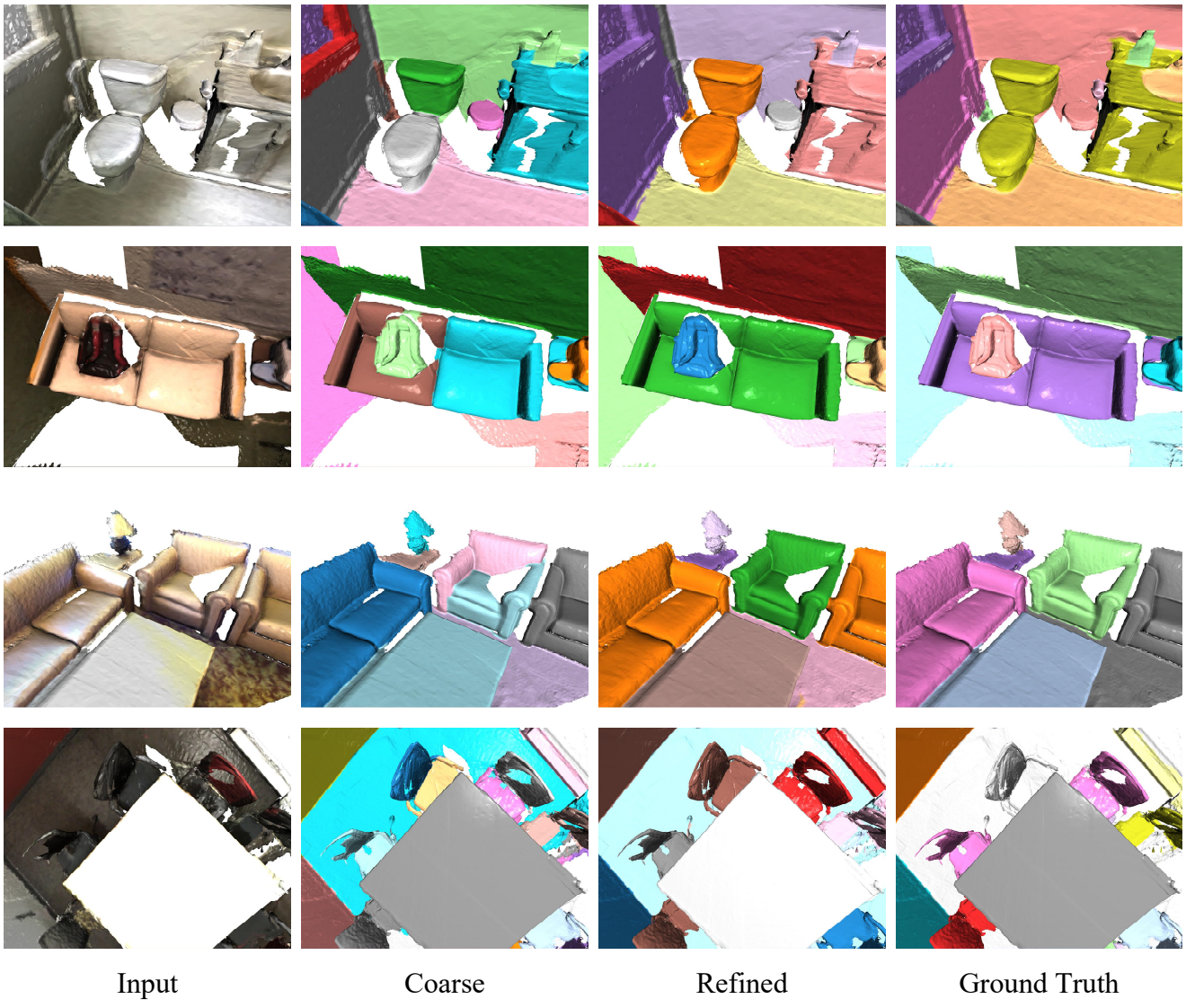

Figure 1. Visual comparison between the coarse 3D segments and the final refined results.

Figure 2. Additional visual results of our method compared to SAM3D [13] and SAI3D [14] on the ScanNetV2 dataset.

Figure 3. Additional visual results of our method compared to SAM3D [13] and SAI3D [14] on the ScanNet++ dataset.

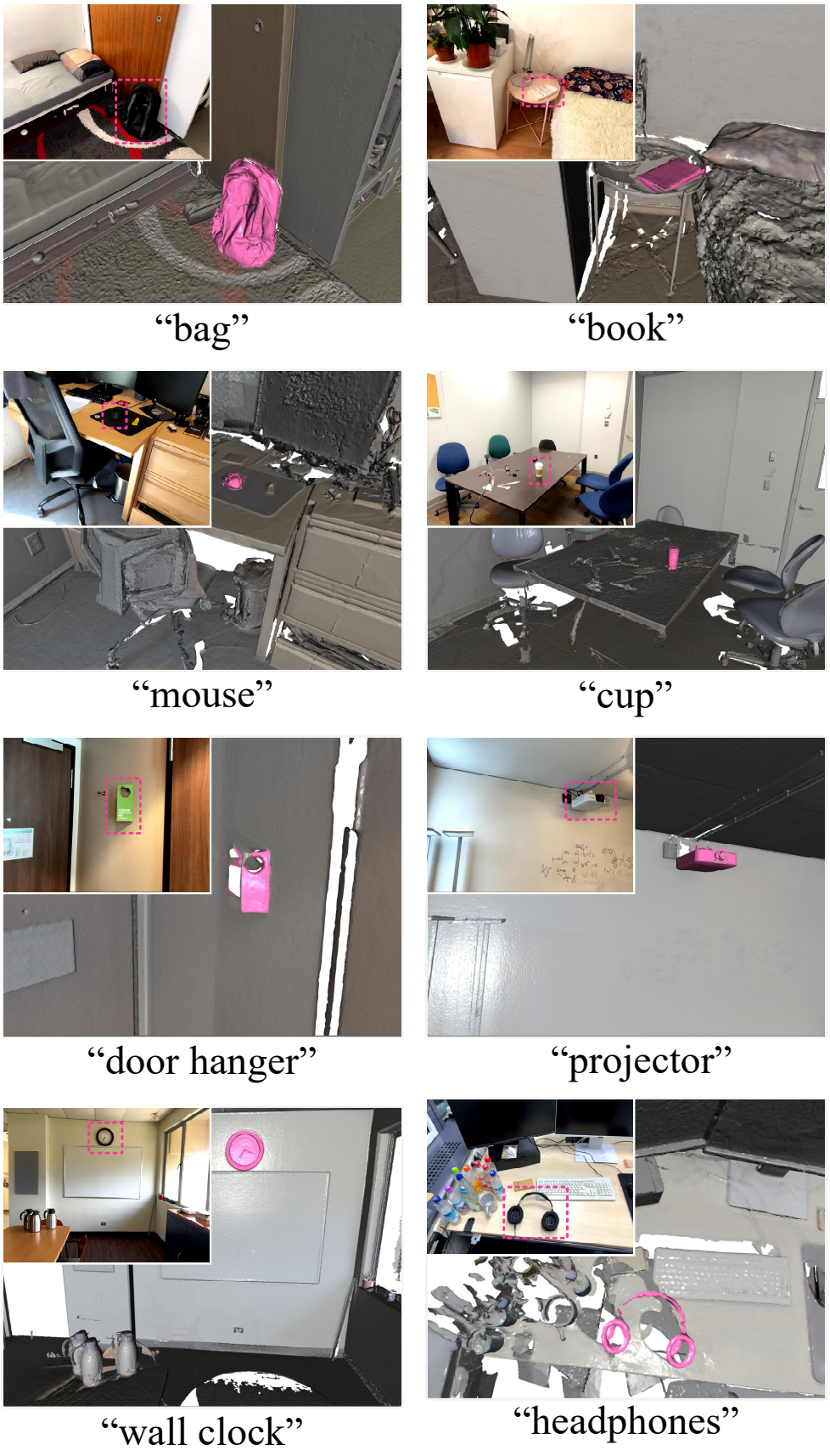

Figure 4. Visual results of open-vocabulary 3D semantic instance segmentation. Given a text prompt, we can segment and retrieve the target instances in the scene.

References

[1] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 1

[2] 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. 3

[3] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2022. 3

[4] Shiyang Lu, Haonan Chang, Eric Pu Jing, Abdeslam Boularias, and Kostas Bekris. Ovir-3d: Open-vocabulary 3d instance retrieval without training on 3d data. In *Conference on Robot Learning*, pages 1610–1620. PMLR, 2023. 2

[5] 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, 2

[6] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 815–824, 2023. 1

[7] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1, 3

[8] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. arxiv 2024. *arXiv preprint arXiv:2401.14159*. 3

[9] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *European conference on computer vision*, pages 125–141. Springer, 2022. 3

[10] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. *arXiv preprint arXiv:2210.03105*, 2022. 1

[11] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 1

[12] Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*, 2023. 1, 3

[13] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*, 2023. 1, 2, 3, 5, 6

[14] Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, and Baoquan Chen. Sai3d: Segment any instance in 3d scenes. In *CVPR*, pages 3292–3302, 2024. 1, 2, 3, 5, 6