

SNAP: Towards Segmenting Anything in Any Point Cloud

Aniket Gupta^{1,*} Hanhui Wang^{1,*} Charles Saunders² Aruni RoyChowdhury²
Hanumant Singh¹ Huaizu Jiang¹

¹Northeastern University ²The Mathworks, Inc.

* Equal contribution

Abstract

Interactive 3D point cloud segmentation enables efficient annotation of complex 3D scenes through user-guided prompts. However, current approaches are typically restricted in scope to a single domain (indoor or outdoor), and to a single form of user interaction (either spatial clicks or textual prompts). Moreover, training on multiple datasets often leads to negative transfer, resulting in domain-specific tools that lack generalizability. To address these limitations, we present **SNAP** (Segment aNything in Any Point cloud), a unified model for interactive 3D segmentation that supports both point-based and text-based prompts across diverse domains. Our approach achieves cross-domain generalizability by training on 7 datasets spanning indoor, outdoor, and aerial environments, while employing domain-adaptive normalization to prevent negative transfer. For text-prompted segmentation, we automatically generate mask proposals without human intervention and match them against CLIP embeddings of textual queries, enabling both panoptic and open-vocabulary segmentation. Extensive experiments demonstrate that SNAP consistently delivers high-quality segmentation results. We achieve state-of-the-art performance on 8 out of 9 zero-shot benchmarks for spatial-prompted segmentation and demonstrate competitive results on all 5 text-prompted benchmarks. These results show that a unified model can match or exceed specialized domain-specific approaches, providing a practical tool for scalable 3D annotation. Project page is at <https://neu-vi.github.io/SNAP/>

1. Introduction

Inspired by the success of SAM [1] for 2D images, we study interactive segmentation for 3D point clouds in this paper, allowing users to create high-quality annotations at scale with minimal effort. While supervised deep learning has fueled significant progress in visual learning, its reliance on vast labeled datasets presents a major bottleneck in the

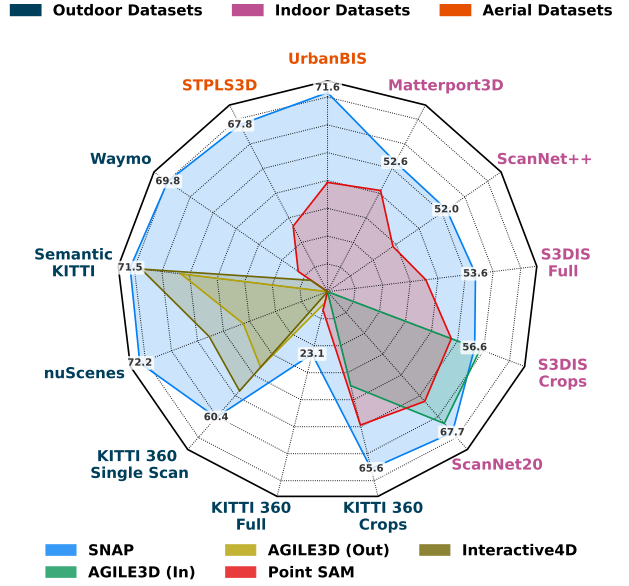


Figure 1. **Comparison of models on IoU@1 Click across multiple domains.** SNAP is a unified interactive point cloud segmentation model trained on multiple datasets spanning multiple domains. It generalizes robustly across a wide array of benchmarks.

3D domain, where manual annotation is notoriously difficult and time consuming.

Current interactive point cloud segmentation methods suffer from critical limitations that hinder their adoption as general-purpose annotation tools. Most existing approaches lack *generalizability*, being designed for specific domains such as indoor scenes [2–4] or outdoor environments [5]. This domain-specific design stems from the significant statistical differences between point cloud types: indoor scenes feature dense, structured environments with clear object boundaries, while outdoor scenes contain sparse, large scale structures with varying point densities. Training models across these diverse domains introduces negative transfer effects that degrades performance without careful architectural considerations [6]. Additionally, current methods lack *flexibility* in prompt modalities, typically supporting ei-

ther spatial inputs (*e.g.*, clicks) [2–5] or text descriptions[7], but rarely both. This limitation restricts users to specific annotation workflows, preventing adaptation to different labeling needs and use cases.

To address these limitations, we present **SNAP** (Segment Anything in aNy Point cloud), a unified model that supports both spatial and text-based prompts. To achieve cross-domain generalizability, we train on seven diverse datasets [8–14] spanning indoor, outdoor, and aerial domains, employing domain-wise normalization to mitigate negative transfer effects caused by statistical shifts between datasets [15]. For text-prompted segmentation, we first introduce a simple iterative algorithm to automatically generate prompt points without human intervention. These prompt points are used via the spatial-prompted segmentation pipeline to generate mask proposals, which are then matched against CLIP [16] embeddings of textual prompts to run text-prompted segmentation. During inference, this approach supports both panoptic segmentation with pre-defined categories and open-vocabulary segmentation with novel classes.

Extensive experiments demonstrate that SNAP consistently predicts high-quality spatial masks and correct class labels across a diverse range of indoor [11, 17–19], outdoor [8–10, 20–22], and aerial [13, 14, 23, 24] point clouds. For point-prompted segmentation, SNAP sets a new state-of-the-art on 8 out of 9 zero-shot benchmarks. For text-prompted segmentation, SNAP shows competitive results on all 5 evaluated benchmarks. These results demonstrate that a single, unified model can match or exceed the performance of specialized, domain-specific approaches, shown in **Fig. 1**. In summary, our contributions are as follows:

- We introduce SNAP, a model for interactive point cloud segmentation that works across indoor, outdoor, and aerial domains.
- SNAP offers flexibility to use multi-modal prompts like points and text to segment objects of interest, and predicts classes as well as spatial masks.
- We introduce an automatic prompt points generation algorithm that bootstraps the panoptic labeling process without human intervention.
- SNAP achieves state-of-the-art performance across multiple datasets across various domains and is usable as an out-of-the-box semi-automated labeling tool.

2. Related Work

Interactive 3D segmentation with spatial prompts. Interactive segmentation has become well established in 2D since the introduction of SAM [1], but remains comparatively underexplored in 3D. Early works such as InterObject3D[2] and AGILE3D[3] focused on indoor point clouds, using positive/negative clicks for single or multi-object segmentation. Point-SAM[4] leveraged SAM-

generated pseudo-labels to support indoor and part-level segmentation. Interactive4D [5] adopted a 4D setup on outdoor LiDAR sequences. However, these works are generally restricted in their usability to either indoor or outdoor scenes and show limited generalizability to out of domain point clouds. SNAP, in contrast, is designed to perform robustly across different domains.

Text-based 3D segmentation. Another line of research introduces natural language as a flexible interface for 3D segmentation. Within this area, *open-vocabulary segmentation* methods [25–28] aim to recognize novel or user-specified categories beyond a fixed label set. OpenScene [25] achieves this directly in the 3D domain by predicting per-point CLIP embeddings without relying on images. In contrast, [26–28] utilize either original RGB images or rendered multiview images to extract CLIP image features, which are aligned with text embeddings to enable text-based segmentation. A complementary direction is *panoptic segmentation*, where the goal is to provide instance-level predictions across all categories. [7, 29] exemplify this approach by predicting per-instance CLIP tokens and aligning them with a predefined class vocabulary to support class-specific instance segmentation. SNAP is able to accommodate both *open-vocabulary* and *panoptic* settings within a single framework.

Lifting 2D foundation models for 3D segmentation. Due to the limited availability of annotated 3D data, recent works like SAM3D [30] and SAMPro3D [31] focus on lifting robust 2D foundation models into the 3D domain. However, they require paired image and point cloud data, which limits their usability in real-world use cases with LiDAR only datasets or legacy datasets. [4, 7, 32] employs SAM [1] to generate pseudo-labels provided RGB images and use them to train 3D models on indoor scenes, with PointSAM [4] further addressing part-level segmentation. While this process helps generate significant amounts of training data, this pseudo-labeling process invariably introduces label noise. SNAP avoids this problem altogether by pooling publicly available datasets for training and establishes state-of-the-art performance on several unseen datasets.

3. Method

In this section, we present the overall framework of SNAP, which is organized into 4 parts: (1) Point Cloud Encoding, (2) Spatial-prompted Segmentation, (3) Text-prompted Segmentation, and (4) Training. **Fig. 2** provides an overview of our approach.

3.1. Point Cloud Encoding

The input point cloud is represented by its XYZ coordinates, denoted as $\mathbf{P} = \{\mathbf{p}_i \in \mathbb{R}^3\}_{i=1}^N$, where N is the number of points. We use the Point Transformer V3 (PTv3 [33])

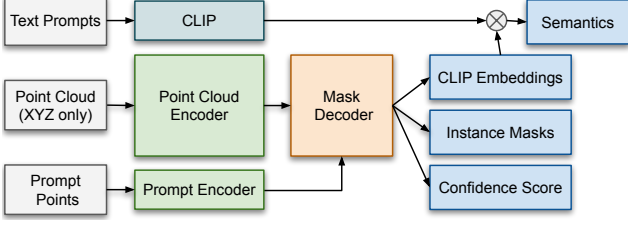


Figure 2. **Overview of SNAP.** SNAP encodes point clouds and prompts separately, then uses a Mask Decoder to generate segmentation masks. Text prompts are handled by matching CLIP embeddings with predicted mask embeddings for semantic classification.

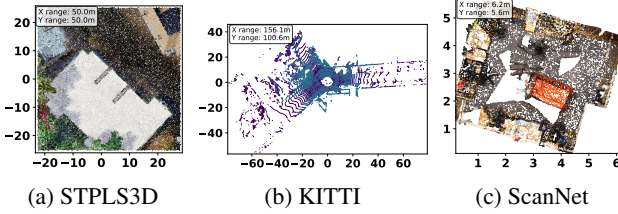


Figure 3. **Point clouds from different domains vary significantly in their properties.** (a) STPLS3D provides dense point clouds from an aerial view using RGB photogrammetry in the 50m range, (b) KITTI provides lidar data in the 150m range, (c) ScanNet provides point clouds in the 10m range.

to extract point-wise embeddings $\mathbf{F}_{pc} \in \mathbb{R}^{N \times D}$. To support cross-domain generalization, we replace the regular batch normalization in PTV3 with domain normalization.

Domain Normalization for Multi-Dataset Training. Training a *single* model on multiple point cloud datasets often leads to lower performance than *multiple per-dataset* models due to negative transfer caused by significant distributional differences between various datasets [6], as shown in Fig. 3. A straightforward approach to mitigate this is *dataset-specific* normalization [6], which learns unique normalization parameters for each dataset in the training set. While effective at addressing distributional differences among *known* datasets, this method presents practical challenges: users need to select appropriate normalization parameters at test time, which can be ambiguous when the source of a point cloud at test-time differs from the training datasets. Moreover, dataset-specific normalization may limit knowledge transfer between related datasets that could benefit from shared representations.

To address the limitation, we propose *domain-specific* normalization, which groups datasets into broader *domains* with similar statistical properties (e.g., indoor, outdoor, or aerial) and learns a separate set of normalization parameters for each domain. This strategy allows our model to effectively adapt to different data distributions while maintaining the flexibility to be applied to new datasets by identifying their general domain, a more intuitive decision than selecting specific dataset parameters as shown in Fig. 4.

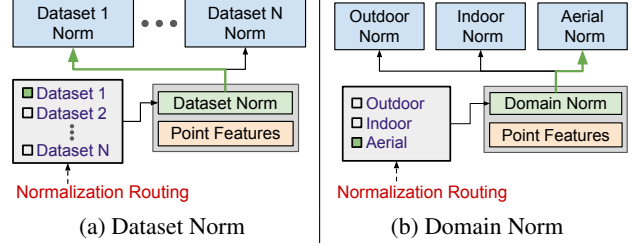


Figure 4. **Dataset Norm vs Domain Norm.** Domain-norm simplifies the overall architecture and improves zero-shot generalization.

Formally, we denote the set of domains as $\mathbb{D} = \{\mathcal{D}_j\}$. Domain normalization for the k -th layer’s activations from the j -th domain, $\bar{\mathbf{x}}_j^k$, is expressed as:

$$\bar{\mathbf{x}}_j^k = \frac{\mathbf{x}_j^k - \mathbb{E}[\mathbf{x}_j^k]}{\sqrt{\text{Var}[\mathbf{x}_j^k] + \epsilon}} \cdot \gamma_j^k + \beta_j^k, \quad (1)$$

where γ_j^k and β_j^k are learned domain-specific scale and shift parameters. Compared with standard batch and dataset-specific normalizations, this domain-specific design allows the model to effectively adapt the normalization to the distinct statistics of each domain while sharing the core model weights across all domains.

3.2. Spatial-prompted Segmentation

Spatial Prompt Encoder. Given a spatial prompt point $\mathbf{p}_{sp} \in \mathbb{R}^3$, we identify its nearest neighbor in the point cloud: $i^* = \arg \min_i \|\mathbf{p}_{sp} - \mathbf{p}_i\|_2$, and retrieve the corresponding point embedding, $\mathbf{f}_{sem} = \mathbf{f}_{i^*}$, from the point-wise PTV3 embeddings $\mathbf{F}_{pc}$. We also compute positional encodings of the prompt point: $\mathbf{f}_{pos} = \phi(\mathbf{p}_{sp})$, where $\phi(\cdot)$ is the Fourier encoding function as in [34]. The final spatial prompt embedding combines semantic features and the positional encodings, which help preserve spatial information: $\mathbf{F}_{sp} = \Phi_{sp}([\mathbf{f}_{sem} \parallel \mathbf{f}_{pos}]) \in \mathbb{R}^D$ where $\Phi_{sp}(\cdot)$ is a learned projection function, and $[\cdot \parallel \cdot]$ is feature concatenation. Given P prompt points (clicks) on M objects, we get $\mathbf{F}_{sp} \in \mathbb{R}^{M \times P \times D}$.

Mask Decoder. The mask decoder module takes in two inputs: $\mathbf{F}_{pc} \in \mathbb{R}^{N \times D}$ for the point cloud embeddings, and $\mathbf{F}_{sp} \in \mathbb{R}^{M \times P \times D}$ for the prompt embeddings. Inspired by the design in SAM [1], we additionally introduce *three* task-specific learnable tokens per object. These tokens are responsible for predicting the final mask, its confidence score, and CLIP embeddings, respectively, resulting in $\mathbf{F}_{sp} \in \mathbb{R}^{M \times (P+3) \times D}$. Following the approach of SAM [1], we first tile the point embeddings to match the prompt dimension: $\tilde{\mathbf{F}}_{pc} \in \mathbb{R}^{M \times N \times D}$. The decoder then uses a series of transformer blocks to iteratively refine the embeddings. The process is designed to first incorporate contextual information from the point cloud into the prompt embeddings, and then use the refined prompts to condition

the point cloud embeddings. Within each block, the updates occur in the following sequence:

- Prompt Self-Attention: $\mathbf{Z}_1 = \Psi_1(\mathbf{F}_{\text{sp}}, \mathbf{F}_{\text{sp}})$.
- Prompt-to-point Cross-Attention: $\mathbf{Z}_2 = \Psi_2(\mathbf{Z}_1, \tilde{\mathbf{F}}_{\text{pc}})$.
- Feedforward Network (FFN): $\mathbf{Z}_{\text{sp}} = \Phi(\mathbf{Z}_2)$.
- Point-to-Prompt Cross-Attention: $\mathbf{Z}_{\text{pc}} = \Psi_3(\tilde{\mathbf{F}}_{\text{pc}}, \mathbf{Z}_{\text{sp}})$.

Here, each $\Psi(\cdot)$ denotes a multi-head attention block. Each of them takes two inputs, where the first denotes the query and the second indicates the key and value. Φ is a position-wise feedforward network. The final output of the entire decoder module is a set of refined prompt embeddings $\mathbf{Z}_{\text{sp}} \in \mathbb{R}^{M \times (P+3) \times D}$ and conditioned point cloud embeddings $\mathbf{Z}_{\text{pc}} \in \mathbb{R}^{M \times N \times D}$.

To generate predictions, we first separate the learnable token embeddings from the refined prompt embeddings $\mathbf{Z}_{\text{sp}}$. Recall that during prompt encoding, we appended three learnable tokens to each object’s prompt sequence. We extract the following embeddings accordingly: mask token embeddings $\mathbf{Z}_{\text{mask}} \in \mathbb{R}^{M \times 1 \times D}$, CLIP token embeddings $\mathbf{Z}_{\text{CLIP}} \in \mathbb{R}^{M \times 1 \times D}$, mask confidence score token embeddings $\mathbf{Z}_{\text{S}} \in \mathbb{R}^{M \times 1 \times D}$, and auxiliary mask token embeddings $\mathbf{Z}_{\text{aux}} \in \mathbb{R}^{M \times P \times D}$ and retain the original prompt point embeddings as auxiliary embeddings $\mathbf{Z}_{\text{aux}} \in \mathbb{R}^{M \times P \times D}$. The first three token embeddings are fed into dedicated prediction heads to generate mask, mask confidence scores, and CLIP embeddings, while the auxiliary embeddings are used for an additional supervision described in Sec. 3.4.

- **Mask Head.** We pass the mask token embeddings $\mathbf{Z}_{\text{mask}} \in \mathbb{R}^{M \times 1 \times D}$ through a small MLP and then compute the dot product between the mask token embeddings and point cloud embeddings to get M segmentation logits $\mathbf{L}_{\text{mask}} = \mathbf{Z}_{\text{pc}} \cdot (\mathbf{F}_{\text{mask}})^\top \in \mathbb{R}^{M \times N}$. The final segmentation masks can then be obtained by applying a sigmoid function σ to the logits: $\mathcal{M}_{\text{mask}} = \sigma(\mathbf{L}_{\text{mask}})$.
- **Confidence Score Head.** Mask confidence score predictions can be similarly obtained by passing the mask score token embeddings through a MLP: $\mathbf{L}_{\text{S}} = \Phi_{\text{S}}(\mathbf{Z}_{\text{S}})$ and applying a sigmoid function σ : $\mathcal{S} = \sigma(\mathbf{L}_{\text{S}})$.
- **CLIP Embedding Head.** Motivated by SAL [7], we learn to predict CLIP embeddings for each mask using a MLP Φ_{CLIP} : $\mathbf{L}_{\text{CLIP}} = \Phi_{\text{CLIP}}(\mathbf{Z}_{\text{CLIP}})$. We show in **Tab. 6** that this also improves segmentation accuracy.

3.3. Text-prompted Segmentation

Our text-prompted segmentation pipeline consists of two stages: (1) automatic generation of prompt points to produce mask proposals via our spatial-prompted segmentation module, and (2) matching of predicted CLIP embeddings from each mask to the input textual prompt.

Automatic Prompt Points Generation. To comprehensively segment the input point cloud without manual prompts, we employ an iterative coarse-to-fine strategy.

Starting with coarse voxelization, we use voxel centers as prompt points to generate initial mask proposals. We then identify unsegmented regions and generate new prompts using progressively smaller voxel sizes in the unsegmented regions, continuing for a fixed number of iterations. Finally, non-maximum suppression removes redundant overlapping masks. This approach ensures comprehensive coverage while avoiding redundant computation in well-segmented regions, striking a better balance between accuracy and efficiency compared to uniform grid sampling as in the 2D counterpart of SAM [1]. See § 6.3 for more details.

Text Prompt Encoder and Matching. To classify the generated mask proposals, we encode input text prompts using the CLIP text encoder [16]. During training, we use category names as text prompts to train SNAP. Specifically, we wrap category names in full sentences (*e.g.*, **a photo of a class_name**). At inference, the generated mask proposals are matched to text queries by comparing their predicted CLIP embeddings with encoded text prompts. This enables both panoptic segmentation using predefined vocabularies and open-vocabulary segmentation for novel object classes.

3.4. Training

Click Sampling Strategy. Following AGILE3D [3] and Interactive4D [5], we use an iterative click sampling strategy to simulate user clicks. Unlike their computationally intensive process of calculating and then ranking error regions for additional clicks, we adopt a simpler approach by randomly sampling clicks from the unsegmented regions of each object.

Training Losses. In line with prior works [1, 3, 5], we supervise our mask predictions using the ground truth annotated instance segmentation labels using Focal [35] and Dice [36] loss. Following [5], we also incorporate weights on these losses to modulate them based on proximity to the user clicks. Additionally, to encourage each click to independently yield a plausible mask prediction, we use an auxiliary mask loss ($\mathcal{L}_{\text{aux}}$), which adopts the auxiliary mask token embeddings (as defined in Sec. 3.2) to generate additional mask predictions. To improve the reliability of mask confidence score estimation, we also introduce a score prediction loss ($\mathcal{L}_{\text{score}}$). Specifically, the target score for each mask is defined as the Intersection-over-Union (IoU) between the predicted mask and its corresponding ground-truth mask. To supervise the predicted class labels, we use $\mathcal{L}_{\text{text}}$ to penalize incorrect alignment between the CLIP embeddings of predicted text tokens for each mask and the class vocabulary, using a cosine distance loss. Our overall loss function can be written as:

$$\mathcal{L}_{\text{SNAP}} = \mathcal{L}_{\text{focal}} + \mathcal{L}_{\text{dice}} + \mathcal{L}_{\text{aux}} + \mathcal{L}_{\text{score}} + \mathcal{L}_{\text{text}}. \quad (2)$$

See § 6.2 for additional details of each loss term.

Table 1. **In-distribution interactive point cloud segmentation.** SNAP-*dataset* refers to the model trained exclusively on the respective dataset. † denotes the method is evaluated by us.

Method	IoU@ k ↑				
	@1	@2	@3	@5	@10
Trained and evaluated on SemanticKITTI					
AGILE3D [3]	53.1	63.7	70	76.7	83.3
Interactive4D [5]	67.5	73.9	78.3	83.4	88.2
SNAP - KITTI	68.1	75.9	80.1	84.5	88.7
SNAP - C	71.5	78.1	81.0	86.0	90.1
Trained and evaluated on ScanNet20					
InterObject3D [2]	40.8	55.9	63.9	72.4	79.9
AGILE3D [3]	63.3	70.9	75.4	79.9	83.7
Point-SAM† [4]	52.7	69.6	75.9	80.6	83.3
SNAP - SN	68.6	74.2	78.4	82.1	84.6
SNAP - C	67.7	74.7	78.5	82.3	85.5

4. Experiments

4.1. Experimental Setup

Datasets. We train SNAP on 7 diverse datasets with ground-truth instance segmentation labels, which span three domains: (i) *indoor* scenes from ScanNet [11] and HM3D [12, 37]; (ii) *outdoor* driving sequences from SemanticKITTI [8, 20], nuScenes [9], and Pandaset [10]; and (iii) *aerial* point clouds from STPLS3D [13] and DALES [14, 23]. We evaluate zero-shot performance on held-out datasets from each domain: S3DIS [19] (full and crops), ScanNet++ [17], and Matterport3D [18] for indoor; Waymo [22] and KITTI-360 [21] (full, crops, and single scan) for outdoor; and UrbanBIS [24] for aerial. See § 7.2 for details about all the datasets.

Evaluation Metrics. Following conventions from [2–5], we evaluate spatial-prompted segmentation using IoU@ k , the average intersection over union (IoU) achieved with k clicks per object across all objects. To assess object category prediction capabilities, we use the mean Average Precision (mAP) metric following [25–27] and for panoptic segmentation, we use panoptic quality (PQ), segmentation quality (SQ), and recognition quality (RQ) as done in [7].

Model Variants. SNAP operates on XYZ coordinates only, without relying on any additional per-point attributes (*e.g.*, color, normals, intensity), as such features are not consistently available across datasets. As shown in Sec. 4.4, using only XYZ coordinates achieves performance comparable to models that leverage all available modalities, while avoiding the dependency on dataset-specific properties. We evaluate 6 SNAP variants to provide comprehensive comparisons. First, we train dataset-specific models: SNAP-KITTI on SemanticKITTI [8] and SNAP-SN on ScanNet [11]. Second, we train domain-specific mod-

els that leverage all available in-domain datasets: SNAP-Indoor, SNAP-Outdoor, and SNAP-Aerial. Finally, SNAP-C represents our complete model trained across all datasets, serving as our most generalizable variant. In Sec. 4.3, SNAP-C@auto refers to the automatic prompt points generation setting, in which the model is provided solely with the point cloud to enable a fair comparison with baseline methods.

4.2. Spatial-prompted Interactive Segmentation

To fairly evaluate SNAP against other purely interactive segmentation baselines like [3–5], we first evaluate the predicted instance mask quality without taking the class-prediction into account (class-agnostic), as typically done in prior methods on interactive point cloud segmentation.

In-distribution Evaluation. In Tab. 1 we compare SNAP’s performance on the SemanticKITTI [8] and ScanNet [11] datasets, where SNAP outperforms all prior baselines on both datasets. Notably, while AGILE3D [3] needs to be trained separately for both datasets, SNAP is evaluated with the same set of parameters for both datasets. This versatility makes SNAP a more practical segmentation tool, which can thus be evaluated on a broader set of datasets.

Zero-Shot Evaluation. To test the generalization capabilities of the SNAP, we evaluate on 9 *unseen* benchmarks covering indoor, outdoor, and aerial domains. As shown in Tab. 2, SNAP-C outperforms the baseline AGILE3D [3], Point-SAM [4] and Interactive4D [5] on 8 out of 9 benchmarks. Particularly in the 1-Click experiments, SNAP-C provides a **20.6%** average improvement across all benchmarks with a single set of parameters.

This positions SNAP-C as a practical, general-purpose tool for interactive segmentation, addressing the key limitation of existing domain-specific approaches.

4.3. Text-prompted Segmentation

For text-prompted segmentation, we evaluate SNAP’s effectiveness for panoptic segmentation and open-vocabulary segmentation. Our primary evaluation targets SNAP-C@auto; we additionally report SNAP-C@1 Click as an upper-bound reference, with the click sampled from the *ground-truth* masks to simulate ideal user input. SNAP-auto takes only the point cloud as input and outputs panoptic segmentation masks like the baselines we compare against.

Panoptic Segmentation. SAL [7] pioneered automated panoptic segmentation on outdoor datasets, making it our primary baseline for outdoor scenes. As shown in Tab. 3, SNAP-C@auto, which has the equivalent setting to SAL with only point clouds as input, performs robustly against it and improves PQ score by 3.4 points on SemanticKITTI while remaining comparable on nuScenes (37.9 PQ vs 38.4 from SAL). With single-click guidance, SNAP-C@1 Click

Table 2. **Zero-shot interactive point cloud segmentation.** We compare SNAP with the current methods for interactive segmentation across different domains on several unseen datasets. [†] denotes that the methods are evaluated by us.

	Dataset	Method	IoU@k $\uparrow$		
			@1	@3	@5
Indoor	ScanNet++	Point-SAM [†] [4]	28.6	56.3	62.9
		SNAP - C	52.0	67.3	73.2
		Δ	+23.4	+11.0	+10.3
	Matterport3D	Point-SAM [†] [4]	41.1	67.2	73.7
		SNAP - C	52.6	69.6	75.2
		Δ	+11.5	+2.4	+1.5
	S3DIS (crops)	AGILE3D [3]	58.7	77.4	83.6
		Point-SAM [†] [4]	45.9	77.6	84.6
		SNAP - C	56.6	73.8	80.9
	S3DIS (full)	Δ	-2.1	-3.6	-3.7
		Point-SAM [†] [4]	35.6	68.0	76.3
		SNAP - C	53.6	71.1	77.6
		Δ	+18.0	+3.1	+1.3
Outdoor	Waymo	Point-SAM [†] [4]	12.8	43.0	53.1
		Interactive 4D [†] [5]	7.2	7.3	7.5
		SNAP - C	69.8	82.3	86.6
		Δ	+57.0	+39.3	+33.5
	KITTI-360 (full)	Point-SAM [†] [4]	6.8	22.7	28.1
		SNAP - C	23.1	40.1	48.1
		Δ	+16.3	+17.4	+20.0
	KITTI-360 Single Scan	AGILE3D [3]	36.3	47.3	53.5
		Interactive 4D [5]	47.7	59.4	64.1
		SNAP - C	60.4	64.6	67.7
		Δ	+12.7	+5.2	+3.6
	KITTI-360 (crops)	AGILE3D [3]	34.8	42.7	44.4
		Point-SAM [4]	49.4	74.4	81.7
		SNAP - C	65.6	76.1	80.0
		Δ	+16.2	+1.7	-1.7
Aerial	UrbanBIS	Point-SAM [†] [4]	39.3	79.1	89.4
		SNAP - C	71.6	86.2	90.2
		Δ	+32.3	+7.1	+0.8

achieves even better results on both datasets.

Open-Vocabulary Segmentation. There exist limited number of approaches for direct alignment between 3D point clouds and CLIP [16] text embeddings without intermediate image representations, we therefore evaluate against three non-interactive instance segmentation baselines. First, OpenScene (3D Distill) [25] represents the most comparable approach to ours as it only uses the predicted per point CLIP embeddings during inference without requiring images. Second, OpenIns3D [26] generates class-agnostic instance masks and render multiview images of the input point cloud for CLIP [16] based segmentation. Third, OpenMask3D [27] and SAI3D [28] leverage the original

Table 3. **Panoptic segmentation on outdoor Lidar datasets.** We compare the SNAP-C @ auto variant with SAL [7] for panoptic segmentation on outdoor lidar datasets. We also show SNAP-C @ 1 Click results as a reference that represents the upper-bound for SNAP-C @ auto. Note that SNAP-C @ auto represents the equivalent setting to SAL [7] since the only input is raw point clouds.

Method	PQ	RQ	SQ	PQ _{Th}	PQ _{St}
Evaluation $\rightarrow$ SemanticKITTI					
SAL [7]	24.8	32.3	66.8	17.4	30.2
SNAP - C @ auto	28.2	34.1	78.6	20.7	34.4
SNAP - C @ 1 Click	40.1	47.3	81.6	30.7	46.9
Evaluation $\rightarrow$ nuScenes					
SAL [7]	38.4	47.8	77.2	47.5	29.2
SNAP - C @ auto	37.9	47.1	82.2	33.1	52.4
SNAP - C @ 1 Click	47.2	56.1	83.9	40.3	56.2

2D images from the dataset to perform open-vocabulary instance segmentation.

Results are summarized in **Tab. 4**, where baseline methods are trained on ScanNet200 [11]. SNAP-C demonstrates strong performance across 3 zero-shot datasets without using image embeddings. On Matterport3D [18], SNAP-C@auto achieves 16.5 AP, substantially outperforming baselines which rely on images and CLIP image embeddings. This is largely attributed to the large overlap between semantic classes of Matterport3D [18] and ScanNet200 [11]. On Replica [38] and S3DIS [19], SNAP-C@auto shows competitive results against baselines that have access to visual information and surpasses the image-free OpenScene(3D Distill) [25]. We also include qualitative results on the ScanNet++ [17] dataset in **Fig. 5**.

In summary, unlike methods that need RGB images or pre-computed CLIP embeddings, SNAP-C@auto performs open-vocabulary segmentation directly on point clouds and achieves competitive results against image-based methods. The image-free approach also offers practical advantages for handling LiDAR-only datasets, synthetic data, or legacy datasets where corresponding images were never collected.

4.4. Ablation Studies

Effect of Input Properties. We conducted an ablation study to evaluate the contribution of different input properties to our model’s performance, with results summarized in **Tab. 5**. The experiments reveal that SNAP relies predominantly on geometric features. On the ScanNet20 dataset, the inclusion of surface normals (N) consistently improves performance across all metrics, boosting IoU@1 from 67.4 to 68.3. In contrast, color information (C) appears to be largely redundant; its inclusion provides no discernible benefit when normals are present and only a marginal gain otherwise. This trend is consistent across

Table 4. **Open-Vocabulary Segmentation.** We evaluate the SNAP-C @ auto variant against image-based and image-free open vocabulary segmentation models. We also show SNAP-C @ 1 Click results as a reference that represents the upper-bound for SNAP-C @ auto. Note that SNAP-C @ auto represents the equivalent setting to baseline methods since the only input is raw point clouds. [†]Uses RGB images, [‡]Uses CLIP image embeddings from rendered point cloud views.

Model	Uses Img [†]	CLIP IE [‡]	AP	AP ₅₀	AP ₂₅
Zero Shot - Matterport3D					
OpenMask3D [27]	✓	✓	7.7	13.9	20.3
SAI3D [28]	✓	✓	8.9	15.3	20.9
SNAP - C @ auto	✗	✗	16.5	25.2	30.7
SNAP - C @ 1 click	✗	✗	18.3	28.2	34.7
Zero Shot - Replica					
OpenMask3D [27]	✓	✓	13.1	18.4	24.2
OpenIns3D [26]	✗	✓	13.6	18.0	19.7
OpenScene(3D Distill) [25]	✗	✗	8.2	10.5	12.6
SNAP - C @ auto	✗	✗	10.1	11.7	13.8
SNAP - C @ 1 click	✗	✗	11.7	14.5	15.4
Zero Shot - S3DIS					
OpenIns3D [26]	✗	✓	21.1	28.3	29.5
OpenScene(3D Distill) [25]	✗	✗	15.2	21.5	23.7
SNAP - C @ auto	✗	✗	16.1	23.8	30.9
SNAP - C @ 1 click	✗	✗	17.6	25.5	33.6



“Segment bar table.”



“Segment cooking stove.”

Figure 5. **Qualitative segmentation results of open-set scene understanding on the ScanNet++ Dataset.** Given a text prompt in the format of “Segment {open-set vocabulary}”, our SNAP model finds the corresponding masks ■ in the scenes.

other datasets, where Lidar intensity (S) on SemanticKITTI and color on STPLS3D offer only minor improvements to IoU@1. While surface normals improve performance, computing accurate normals is challenging for real-world point clouds. Therefore, SNAP operates solely on XYZ coordinates to ensure applicability across diverse data sources.

Effect of Loss functions on Mask Accuracy. SNAP uses a combination of loss functions to guide the network. To check the effectiveness of each module, we run ablations by

Table 5. **Input Properties Ablations.** X - XYZ coordinates, C - Color, N- Normals, S-Strength / Lidar intensity. - marks property data not available.

Dataset	X	C	N	S	IoU@1	IoU@5	IoU@10
SemanticKITTI	✓	-	-	✓	68.2	84.5	88.7
SemanticKITTI	✓	-	-	✗	68.1	84.5	88.7
ScanNet20	✓	✓	✓	-	69.6	83.0	85.5
ScanNet20	✓	✓	✗	-	68.8	82.1	84.6
ScanNet20	✓	✗	✓	-	69.3	83.0	85.5
ScanNet20	✓	✗	✗	-	68.6	82.1	84.6
STPLS3D	✓	✓	-	-	67.9	78.9	83.9
STPLS3D	✓	✗	-	-	67.2	78.9	83.9

Table 6. **Loss Function Ablations.** We report IoU@*k* with different combinations of Focal, Dice, Auxiliary, Weighted, and Text losses on the ScanNet20 dataset.

Component					IoU@ <i>k</i> ↑		
Dice	Focal	Aux	Weighted	Text	@1	@5	@10
✓	✓				66.7	80.5	82.4
✓	✓	✓			67.3	81.0	83.2
✓	✓	✓	✓		67.5	81.7	83.7
✓	✓	✓	✓	✓	68.6	82.1	84.6

adding each component sequentially. The results are summarized in **Tab. 6**. Results indicate that auxiliary loss and weighted loss improve performance at a higher number of clicks. Including text classification loss leads to the highest improvements and indicates that semantic understanding is important for segmentation tasks.

Effect of Normalization Strategies. We evaluate 3 normalization strategies to validate our domain-based design choice: (1) a single model with standard batch normalization across all data, (2) dataset-specific normalization with individual layers for each training dataset, and (3) our proposed domain normalization with separate layers for indoor, outdoor, and aerial domains. As shown in **Tab. 7**, batch normalization shows reasonable in-distribution performance (64.3 IoU@1 on SemanticKITTI) but tends to underperform on zero-shot datasets. Dataset normalization can achieve slightly higher scores on in-distribution sets (e.g., +2.1 on ScanNet20), but this comes at the cost of limited generalization capability. Domain-specific normalization performs the best in zero-shot generalization, suggesting potential benefits to domain-level grouping over dataset-specific parameters. More importantly, this simple domain-norm translates into a simple domain-type checkbox selection, making it highly user-friendly.

Effect of Progressively Adding Datasets. To verify the effectiveness of scaling up training data and grouping them into domains, we evaluate various SNAP variants on in-distribution and unseen datasets. As shown in **Tab. 8**, scal-

Table 7. **Domain Normalization Ablation.** We compare 3 normalization strategies for training a multi-dataset model: Batch Norm, Dataset Norm, and Domain Norm. The applied normalization for Dataset Norm and Domain Norm is provided in *teal* beneath each result.

Model	IoU@k $\uparrow$											
	In-Distribution						Zero-Shot					
	SemanticKITTI		ScanNet20		STPLS3D		Waymo		ScanNet++		UrbanBIS	
	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5
Batch Norm	64.3	83.9	63.5	81.2	54.2	68.4	52.1	64.3	38.1	56.2	31.3	61.1
Dataset Norm	72.0	86.1	69.5	83.8	67.9	79.0	56.2	76.3	48.5	67.3	61.8	77.8
<i>Norm used</i>	<i>KITTI</i>		<i>ScanNet</i>		<i>STPLS3D</i>		<i>KITTI</i>		<i>ScanNet</i>		<i>STPLS3D</i>	
Domain Norm	71.5	86.0	67.7	82.3	67.8	80.4	69.8	86.6	52	73.2	71.6	90.2
<i>Norm used</i>	<i>Outdoor</i>		<i>Indoor</i>		<i>Aerial</i>		<i>Outdoor</i>		<i>Indoor</i>		<i>Aerial</i>	

Table 8. **Effect of Adding Datasets.** We progressively evaluate baseline models trained on single datasets (*SNAP-SN* or *SNAP-KITTI*) and baseline models trained on specific domains *SNAP-Indoor*, *SNAP-Outdoor*, and *SNAP-Aerial* against *SNAP-C*, the complete model trained on all data. The gray-shaded cells denote evaluations performed on out-of-distribution datasets.

Model	IoU@k $\uparrow$															
	In-Distribution								Zero-Shot							
	SemanticKITTI		ScanNet20		STPLS3D		Matterport3D		S3DIS Full		KITTI-360 Full		Waymo		UrbanBIS	
	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5
SNAP - SN	3.1	5.9	68.6	82.1	2.5	4.1	53.4	71.3	51.4	70.0	0.2	0.5	0.7	3.1	6.3	16.8
SNAP - KITTI	68.1	84.5	13.8	33.6	24.6	52.9	16	33.3	11.6	29.1	6.7	29.7	48.2	66.3	43.1	71.3
SNAP - Indoor	3.1	15.3	66.0	81.3	5.9	18.6	49.9	74.2	51.9	76.9	0.2	2.1	1.1	6.5	3.4	33.6
SNAP - Outdoor	71.3	85.7	14.9	43.5	36.4	57.9	15.1	39.2	14.9	43.6	18.3	44.6	68.5	86.0	45.9	78
SNAP - Aerial	27.1	57.9	6.3	19.1	65.8	79.1	9.4	21.2	7.3	20.6	5.6	19.4	25.1	60.4	74.2	86.9
SNAP - C	71.5	86.0	67.7	82.3	67.8	80.4	52.6	75.2	53.6	77.6	23.1	48.1	69.8	86.6	71.6	90.2

Table 9. **Comparison of Automatic Prompt Points Generation Strategies.** We evaluate our iterative prompting approach against four baseline methods on ScanNet200. Best results are highlighted in **bold**, and second best results are underlined.

Approach	AP	AP ₅₀	AP ₂₅	Head mAP	Common mAP	Tail mAP	Time (s)
Uniform grid	39.1	58.5	68.8	39.8	40.6	36.3	1421
FPS	36.2	54.5	65.0	<u>38.6</u>	38.4	30.4	702
HDBSCAN	4.7	6.5	7.2	6.5	5.4	1.7	714
Ours	<u>38.7</u>	<u>58.4</u>	69.2	38.2	<u>39.7</u>	38.1	461

ing up the training set leads to consistent improvements in performance over both in-distribution as well as unseen datasets. SNAP-C consistently matches or surpasses the performances of single dataset or single domain models, suggesting that the proposed domain normalization is largely able to reduce the effects of any negative transfer from cross-domain datasets.

Design Choices for Automatic Prompt Points Generation. To generate prompt points automatically, we initially considered uniform grid sampling following SAM [1]. However, this approach faces a fundamental trade-off: large voxel sizes miss small objects while small voxels yield computationally prohibitive numbers of points. Alternative ap-

proaches like Farthest Point Sampling and HDBSCAN [39] require scene-specific tuning and generally undersample dense regions while oversampling sparse regions. Our iterative approach addresses these limitations and maintains uniform sampling independent of local point density. As shown in **Tab. 9**, this strategy achieves comparable coverage to naive grid sampling (39.1 AP vs 38.7 AP) with significantly fewer points, reducing computation time by **68%**. Notably, our approach excels on tail classes (38.1 mAP), outperforming all baselines.

5. Conclusion

In this paper, we introduced SNAP, a unified model for flexible promptable point cloud segmentation that is compatible with both spatial and textual prompts. By training on heterogeneous datasets with cross-domain normalization SNAP demonstrates state-of-the-art performance across a wide range of datasets. We believe that this unified model will come across as a handy tool for users, moving away from the need for dataset or domain specific models. For future work, following the promising trend found in this paper, investigating the effect of scaling up further via self-supervised or weakly-supervised learning on unlabeled data is an appealing direction.

References

- [1] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo et al., “Segment anything,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1, 2, 3, 4, 8
- [2] T. Kontogianni, E. Celikkan, S. Tang, and K. Schindler, “Interactive object segmentation in 3d point clouds,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2891–2897. 1, 2, 5, 12
- [3] Y. Yue, S. Mahadevan, J. Schult, F. Engelmann, B. Leibe, K. Schindler, and T. Kontogianni, “Agile3d: Attention guided interactive multi-object 3d segmentation,” *arXiv:2306.00977*, 2024. [Online]. Available: <https://arxiv.org/abs/2306.00977> 2, 4, 5, 6, 3, 12
- [4] Y. Zhou, J. Gu, T. Y. Chiang, F. Xiang, and H. Su, “Point-sam: Promptable 3d segmentation model for point clouds,” *arXiv:2406.17741*, 2024. [Online]. Available: <https://arxiv.org/abs/2406.17741> 1, 2, 5, 6, 3, 12
- [5] I. Fradlin, I. E. Zulfikar, K. Yilmaz, T. Kontogianni, and B. Leibe, “Interactive4d: Interactive 4d lidar segmentation,” *arXiv:2410.08206*, 2024. [Online]. Available: <https://arxiv.org/abs/2410.08206> 1, 2, 4, 5, 6, 12
- [6] X. Wu, Z. Tian, and et al., “Towards large-scale 3d representation learning with multi-dataset point prompt training,” in *CVPR*, 2024. 1, 3
- [7] A. Ošep, T. Meinhardt, F. Ferroni, N. Peri, D. Ramanan, and L. Leal-Taixé, “Better call sal: Towards learning to segment anything in lidar,” *arXiv:2403.13129*, 2024. [Online]. Available: <https://arxiv.org/abs/2403.13129> 2, 4, 5, 6
- [8] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, “SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences,” in *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2019. 2, 5, 3, 6
- [9] W. K. Fong, R. Mohan, J. V. Hurtado, L. Zhou, H. Caesar, O. Beijbom, and A. Valada, “Panoptic nusenes: A large-scale benchmark for lidar panoptic segmentation and tracking,” *arXiv preprint arXiv:2109.03805*, 2021. 5, 3
- [10] P. Xiao, Z. Shao, S. Hao, Z. Zhang, X. Chai, J. Jiao, Z. Li, J. Wu, K. Sun, K. Jiang, Y. Wang, and D. Yang, “Pandaset: Advanced sensor suite dataset for autonomous driving,” 2021. [Online]. Available: <https://arxiv.org/abs/2112.12610> 2, 5, 3
- [11] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, “ScanNet: Richly-annotated 3d reconstructions of indoor scenes,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5828–5839. 2, 5, 6, 3
- [12] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. M. Turner, E. Underlander, W. Galuba, A. Westbury, A. X. Chang, M. Savva, Y. Zhao, and D. Batra, “Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI,” in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. [Online]. Available: <https://arxiv.org/abs/2109.08238> 5, 3
- [13] M. Chen, Q. Hu, Z. Yu, H. THOMAS, A. Feng, Y. Hou, K. McCullough, F. Ren, and L. Soibelman, “Stpls3d: A large-scale synthetic and real aerial photogrammetry 3d point cloud dataset,” in *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. [Online]. Available: <https://bmvc2022.mpi-inf.mpg.de/0429.pdf> 2, 5, 4
- [14] N. M. Singer and V. K. Asari, “Dales objects: A large scale benchmark dataset for instance segmentation in aerial lidar,” *IEEE Access*, pp. 1–1, 2021. 2, 5, 4
- [15] X. Wu, Z. Tian, X. Wen, B. Peng, X. Liu, K. Yu, and H. Zhao, “Towards large-scale 3d representation learning with multi-dataset point prompt training,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 551–19 562. 2, 4
- [16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” 2021. [Online]. Available: <https://arxiv.org/abs/2103.00020> 2, 4, 6
- [17] C. Yeshwanth, Y.-C. Liu, M. Nießner, and A. Dai, “ScanNet++: A high-fidelity dataset of 3d indoor scenes,” in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 2, 5, 6, 3
- [18] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, “Matterport3d: Learning from rgb-d data in indoor environments,” *International Conference on 3D Vision (3DV)*, 2017. 5, 6, 3
- [19] Stanford Doerr School of Sustainability, “Stanford Large-Scale 3D Indoor Spaces Dataset (S3DIS),” Jun. 2024. [Online]. Available: <https://redvis.com/datasets/9q3m-9w5pa1a2h?v=1.0> 2, 5, 6, 3
- [20] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite,” in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3354–3361. 2, 5, 3
- [21] Y. Liao, J. Xie, and A. Geiger, “KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d,” *arXiv preprint arXiv:2109.13410*, 2021. 5, 3, 4
- [22] P. Sun, H. Kretschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, “Scalability in perception for autonomous driving: Waymo open dataset,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 5, 4
- [23] N. Varney, V. K. Asari, and Q. Graehling, “Dales: A large-scale aerial lidar data set for semantic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 186–187. 2, 5
- [24] G. Yang, F. Xue, Q. Zhang, K. Xie, C.-W. Fu, and H. Huang, “Urbanbis: a large-scale benchmark for fine-grained urban building instance segmentation,” in *ACM SIGGRAPH*, 2023, pp. 16:1–16:11. 2, 5, 4, 6, 11

- [25] S. Peng, K. Genova, C. M. Jiang, A. Tagliasacchi, M. Pollefeys, and T. Funkhouser, “Openscene: 3d scene understanding with open vocabularies,” in *CVPR*, 2023. 2, 5, 6, 7
- [26] Z. Huang, X. Wu, X. Chen, H. Zhao, L. Zhu, and J. Lasenby, “Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation,” 2024. [Online]. Available: <https://arxiv.org/abs/2309.00616> 2, 6, 7
- [27] A. Takmaz, E. Fedele, R. W. Sumner, M. Pollefeys, F. Tombari, and F. Engelmann, “Openmask3d: Open-vocabulary 3d instance segmentation,” *arXiv preprint arXiv:2306.13631*, 2023. 5, 6, 7
- [28] Y. Yin, Y. Liu, Y. Xiao, D. Cohen-Or, J. Huang, and B. Chen, “Sai3d: Segment any instance in 3d scenes,” 2024. [Online]. Available: <https://arxiv.org/abs/2312.11557> 2, 6, 7
- [29] W. Roh, H. Jung, G. Nam, J. Yeom, H. Park, S. H. Yoon, and S. Kim, “Edge-aware 3d instance segmentation network with intelligent semantic prior,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 20 644–20 653. 2, 3, 5, 11
- [30] Y. Yang, X. Wu, T. He, H. Zhao, and X. Liu, “Sam3d: Segment anything in 3d scenes,” *arXiv:2306.03908*, 2023. [Online]. Available: <https://arxiv.org/abs/2306.03908> 2
- [31] M. Xu, X. Yin, L. Qiu, Y. Liu, X. Tong, and X. Han, “Sampro3d: Locating sam prompts in 3d for zero-shot instance segmentation,” *International Conference on 3D Vision (3DV)*, 2015. 2
- [32] R. Huang, S. Peng, A. Takmaz, F. Tombari, M. Pollefeys, S. Song, G. Huang, and F. Engelmann, “Segment3d: Learning fine-grained class-agnostic 3d segmentation without manual labels,” *arXiv:2312.17232*, 2023. [Online]. Available: <https://arxiv.org/abs/2312.17232> 2
- [33] X. Wu, L. Jiang, P.-S. Wang, Z. Liu, X. Liu, Y. Qiao, W. Ouyang, T. He, and H. Zhao, “Point transformer v3: Simpler, faster, stronger,” in *CVPR*, 2024. 2, 5
- [34] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, “Fourier features let networks learn high frequency functions in low dimensional domains,” *Advances in neural information processing systems*, vol. 33, pp. 7537–7547, 2020. 3
- [35] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” 2018. [Online]. Available: <https://arxiv.org/abs/1708.02002> 4
- [36] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, *Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations*. Springer International Publishing, 2017, p. 240–248. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-67558-9_28 4
- [37] K. Yadav, R. Ramrakhya, S. K. Ramakrishnan, T. Gervet, J. Turner, A. Gokaslan, N. Maestre, A. X. Chang, D. Batra, M. Savva *et al.*, “Habitat-matterport 3d semantics dataset,” *arXiv preprint arXiv:2210.05633*, 2022. [Online]. Available: <https://arxiv.org/abs/2210.05633> 5
- [38] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, A. Clarkson, M. Yan, B. Budge, Y. Yan, X. Pan, J. Yon, Y. Zou, K. Leon, N. Carter, J. Briales, T. Gillingham, E. Mueggler, L. Pesqueira, M. Savva, D. Batra, H. M. Strasdat, R. D. Nardi, M. Goesele, S. Lovegrove, and R. Newcombe, “The replica dataset: A digital replica of indoor spaces,” 2019. [Online]. Available: <https://arxiv.org/abs/1906.05797> 6
- [39] L. McInnes, J. Healy, S. Astels *et al.*, “hdbscan: Hierarchical density based clustering,” *J. Open Source Softw.*, vol. 2, no. 11, p. 205, 2017. 8, 3, 4
- [40] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, “Panoptic segmentation,” 2019. [Online]. Available: <https://arxiv.org/abs/1801.00868> 2
- [41] C. Choy, J. Gwak, and S. Savarese, “4d spatio-temporal convnets: Minkowski convolutional neural networks,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3070–3079. 5
- [42] J. Schult, F. Engelmann, A. Hermans, O. Litany, S. Tang, and B. Leibe, “Mask3D: Mask Transformer for 3D Semantic Instance Segmentation,” 2023. 11
- [43] K. Yilmaz, J. Schult, A. Nekrasov, and B. Leibe, “Mask4former: Mask transformer for 4d panoptic segmentation,” 2024. [Online]. Available: <https://arxiv.org/abs/2309.16133> 6, 11
- [44] L. Yao, Y. Wang, C. Yawen, M. Liu, and L.-P. Chau, “Lassm: Efficient semantic-spatial query decoding via local aggregation and state space models for 3d instance segmentation,” *xxx*, 2025. 6, 11
- [45] A. Jain, P. Katara, N. Gkanatsios, A. W. Harley, G. Sarch, K. Aggarwal, V. Chaudhary, and K. Fragkiadaki, “Odin: a single model for 2d and 3d segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3564–3574. 6, 11

SNAP: Towards Segmenting Anything in Any Point Cloud

Supplementary Material

This supplementary document is structured as follows:

- **Model Details**
 - Detailed Model Architecture
 - Details on Loss functions
 - Auto Prompt Generation
- **Implementation Details**
 - Details on Evaluation Metrics
 - Dataset Details
 - HDBSCAN Details
 - Training Details
- **Additional Ablations**
 - Backbone Ablation
 - Click-Strategy Ablation
 - Cross-Domain Input Ablation
- **Additional Quantitative Results**
 - Timing and Memory Comparison
 - Class Agnostic Interactive Segmentation against Non-Interactive Fully Supervised Methods
 - Interactive segmentation results with all model variants
- **Additional Qualitative results**

6. Model Details

6.1. Model Architecture

To complement the simplified pipeline presented in the main paper, we provide a more detailed illustration of SNAP in **Fig. 6**. Specifically, the Mask Decoder is decomposed into a Prompt-Point Attention module and three dedicated MLP heads that process the mask, score, and CLIP tokens, together with the spatial prompt embeddings. For completeness, we also indicate the corresponding supervision signals, showing how each type of prediction contributes to the overall training objective through its associated loss functions.

6.2. Details on Loss Functions

Focal and Dice Loss. Following Interactive4D [5], we apply a distance-based, click-localized weight to the point-wise Focal loss and Dice loss terms ($\mathcal{L}_{\text{focal}}$ and $\mathcal{L}_{\text{dice}}$). Concretely, for each point $\mathbf{p}_i \in \mathbf{P}$, we compute its normalized distance to its nearest spatial prompt point $\mathbf{p}_{\text{sp}}^*$: $d_i = \text{Dist}(\mathbf{p}_i, \mathbf{p}_{\text{sp}}^*)$. If d_i is below a threshold τ_d , the weight is defined to decay linearly from w_{max} to w_{min} as the distance increases; otherwise, the weight is set to w_{min} . Formally, the weight of each point is defined as:

$$w(\mathbf{p}_i) = \begin{cases} w_{\text{max}} - (w_{\text{max}} - w_{\text{min}}) d_i, & d_i < \tau_d, \\ w_{\text{min}}, & \text{otherwise.} \end{cases} \quad (3)$$

In our implementation, we set $w_{\text{max}} = 2$, $w_{\text{min}} = 1$, and $\tau = 0.5$. This weighting strategy increases the contribution of points closer to spatial prompts (clicks), encouraging the model to focus its supervision around click regions while preserving global mask consistency.

Auxiliary Loss. In addition to the final mask prediction loss described above, we strengthen supervision by leveraging the spatial prompt embeddings corresponding to individual clicks. The intuition is that each click should independently guide a plausible mask prediction, rather than only contributing through the aggregated mask token. To achieve this, we treat the P prompt embeddings extracted from $\mathbf{Z}_{\text{sp}} \in \mathbb{R}^{M \times (P+3) \times D}$ as *auxiliary mask tokens* and feed them through the same mask head described in the method section. This yields $M \times P$ auxiliary mask predictions, which are supervised using standard point-wise focal and Dice loss terms. The resulting auxiliary loss $\mathcal{L}_{\text{aux}}$ encourages individual clicks to directly align with the segmentation masks, thereby providing more fine-grained supervision.

Confidence Score Loss. To improve the reliability of mask confidence score estimation, we supervise this score prediction process with $\mathcal{L}_{\text{score}}$. Intuitively, this score should reflect the quality of the predicted mask, which we approximate by its intersection-over-union (IoU) with its corresponding ground-truth mask. Concretely, given the predicted mask $\mathcal{M}_i$, we first obtain a binarized mask by thresholding it: $\mathcal{M}_i > \tau$. The IoU between this mask and its ground-truth counterpart $\mathcal{M}_i^*$ is then computed and used as the regression target: $\mathcal{S}_i^* = \text{IoU}(\mathcal{M}_i > \tau, \mathcal{M}_i^*)$. Finally, we formulate the score loss as a mean squared error (MSE) between the predicted score and the IoU target:

$$\mathcal{L}_{\text{score}} = \frac{1}{M} \sum_{i=1}^M (\mathcal{S}_i - \mathcal{S}_i^*)^2, \quad (4)$$

where $\mathcal{S}_i$ denotes the predicted score for the i -th mask.

Text Loss. To supervise the predicted CLIP tokens $\mathbf{L}_{\text{CLIP}}$, we follow a prototype-based classification scheme against the CLIP text vocabulary embeddings $\mathbf{T} \in \mathbb{R}^{C \times D_{\text{CLIP}}}$. After L2 normalization, cosine similarities between $\mathbf{L}_{\text{CLIP}}$ and $\mathbf{T}$ yield logits $\mathbf{Z} = \mathbf{L}_{\text{CLIP}} \cdot \mathbf{T}^\top \in \mathbb{R}^{M \times C}$. For the i -th sample with ground-truth label y , let $p_i = \text{softmax}(\mathbf{Z}_i)_y$ denote the probability of the correct class. We then apply a focal loss with focusing parameter $\gamma = 2.0$ and no additional class re-weighting, which gives the text loss:

$$\mathcal{L}_{\text{text}} = \frac{1}{M} \sum_{i=1}^M (1 - p_i)^\gamma (-\log p_i). \quad (5)$$

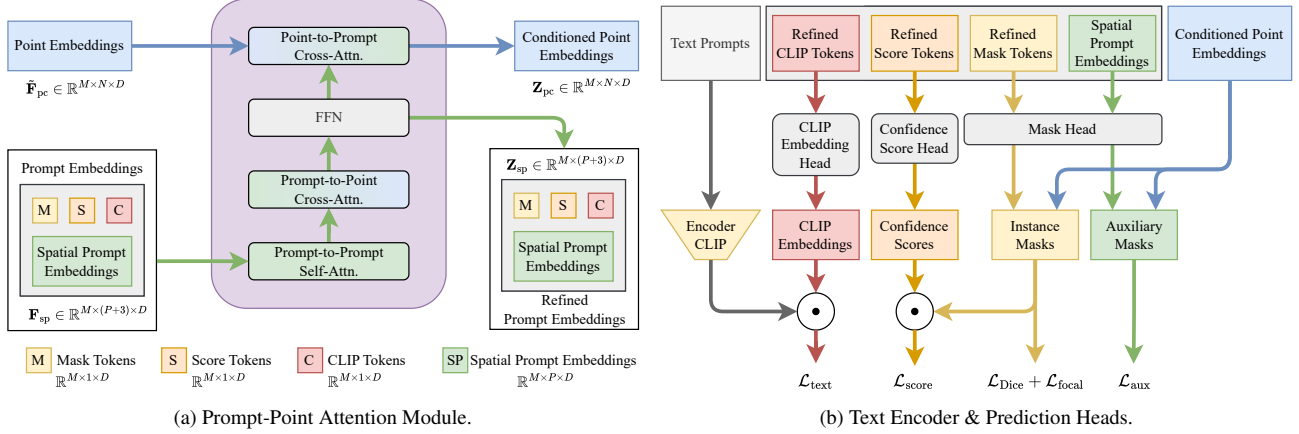


Figure 6. **Detailed architecture of SNAP.** (a) The Prompt-Point Attention module refines both point and prompt embeddings within the Mask Decoder. (b) The refined embeddings are then fed into several lightweight prediction heads for mask, confidence score, and CLIP embedding predictions. For completeness, we also indicate the external CLIP Text Encoder for processing text prompts and the supervision signals associated with each branch.

6.3. Auto Prompt Generation

Given an input point cloud $\mathcal{P}$ with N points, let $\mathcal{F}$ denote the segmentation model and d the scene domain (outdoor / indoor / aerial). We define v_0 as the initial domain-specific voxel size, $K_{\max}$ as the maximum number of iterations, τ_s as the predicted confidence score threshold, and τ_{nms} as the NMS IoU threshold. The algorithm iteratively generates prompt points, segments objects of decreasing scales, and refines the results across iterations, as illustrated in **Fig. 7**. Finally, it outputs a set of masks $\mathcal{M}$, their corresponding text embeddings $\mathcal{T}$, and confidence scores $\mathcal{S}$. The complete procedure is summarized in **Alg. 1**.

7. Implementation Details

7.1. Details on Evaluation Metrics

IoU@k. Following conventions from [2, 3, 5], we evaluate using IoU@ k , the average intersection over union (IoU) achieved with k clicks per object, averaged across all objects.

Average Precision. In our comparisons against the baselines for open-vocabulary segmentation, we use the Average Precision metric defined as follows -

$$\text{mAP} = \frac{1}{C \times 10} \sum_{c=1}^C \sum_{\tau=0.5}^{0.95} \text{AP}_{\tau}^c \quad (\text{step size} = 0.05) \quad (6)$$

where C is the number of classes.

Panoptic Segmentation metrics. To assess panoptic quality, we utilize the Panoptic Segmentation metrics as defined in [40] and as used by SAL [7]. Specifically, PQ is Panoptic Quality, SQ is Segmentation Quality and RQ is Recognition Quality. TP, FP, FN represent True Positives, False Positives and False Negatives respectively. For class-aware segmentation, A prediction is counted as a TP if it has a IoU > 0.5

Algorithm 1 Mask generation with iterative prompting algorithm.

Input: Point cloud $\mathcal{P} = \{p_i\}_{i=1}^N$, segmentation model $\mathcal{F}$, scene domain d , initial domain-specific voxel size v_0 , maximum number of iterations $K_{\max}$, confidence score threshold τ_s , NMS IoU threshold τ_{nms}

Output: Set of masks $\mathcal{M}$, text embeddings $\mathcal{T}$, confidence scores $\mathcal{S}$

```

1:  $v_0 \leftarrow$  domain-specific voxel size
2:  $\mathcal{C} \leftarrow \{0\}^N$  ▷ Coverage mask
3:  $\mathcal{M}, \mathcal{T}, \mathcal{S} \leftarrow \emptyset$ 
4: for  $k = 0$  to  $K_{\max} - 1$  do
5:    $\mathcal{U} \leftarrow \{p_i : \mathcal{C}_i = 0\}$  ▷ Get uncovered points
6:    $v_k \leftarrow v_0 / 2^k$  ▷ Halve voxel size
7:    $\mathcal{Q} \leftarrow \text{VoxelDownsample}(\mathcal{U}, v_k)$  ▷ Generate prompts
8:    $\mathcal{M}^k, \mathcal{T}^k, \mathcal{S}^k \leftarrow \mathcal{F}(\mathcal{P}, \mathcal{Q})$  ▷ Run model
9:   for  $j = 1$  to  $|\mathcal{M}^k|$  do
10:    if  $\mathcal{S}_j^k \geq \tau_s$  then
11:       $\mathcal{M} \leftarrow \mathcal{M} \cup \{\mathcal{M}_j^k\}$ 
12:       $\mathcal{T} \leftarrow \mathcal{T} \cup \{\mathcal{T}_j^k\}$ 
13:       $\mathcal{S} \leftarrow \mathcal{S} \cup \{\mathcal{S}_j^k\}$ 
14:       $\mathcal{C} \leftarrow \mathcal{C} \vee \mathcal{M}_j^k$  ▷ Update coverage
15:    end if
16:  end for
17: end for
18:  $\mathcal{M}, \mathcal{T}, \mathcal{S} \leftarrow \text{NMS}(\mathcal{M}, \mathcal{T}, \mathcal{S}, \tau_{nms})$ 
19: return  $\mathcal{M}, \mathcal{T}, \mathcal{S}$ 

```

and the correct label. For class-agnostic segmentation, we assume class predictions are correct and TP is counted if IoU > 0.5.

Table 10. Summary of Dataset Statistics.

Training Datasets				
Dataset	Train	Val	Domain	Sensor Type
SemanticKITTI	19,130	4,071	Outdoor	HDL-64 LiDAR
nuScenes	28,130	6,019	Outdoor	32-beam LiDAR
PandaSet	2,000	400	Outdoor	Pandar64
ScanNet	1,201	312	Indoor	RGBD Camera
HM3D	1805	481	Indoor	RGBD camera
STPLS3D	3,395	500	Aerial	Photogrammetry
DALES	2,900	1,100	Aerial	Aerial LiDAR
Total	58,561	12,883		
Zero-Shot Validation Datasets				
Dataset	Val	Domain	Sensor Type	
Waymo	5,976	Outdoor	Proprietary LiDAR	
KITTI-360 SS	13,440	Outdoor	HDL-64 LiDAR	
KITTI-360 Full	61	Outdoor	HDL-64 LiDAR	
KITTI-360 Crops	3,421	Outdoor	HDL-64 LiDAR	
Matterport3D	233	Indoor	RGBD Camera	
ScanNet++	178	Indoor	RGBD Camera	
S3DIS Crops	2,330	Indoor	RGBD Camera	
S3DIS Full	68	Indoor	RGBD Camera	
UrbanBIS	46	Aerial	Photogrammetry	
Total	25,753			

$$PQ = \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{Recognition Quality (RQ)}} \times \underbrace{\frac{\sum_{(p,g) \in TP} \text{IoU}(p,g)}{|TP|}}_{\text{Segmentation Quality (SQ)}} \quad (7)$$

7.2. Dataset Details

We provide a summary of the dataset statistics in **Tab. 10**. Samples from each dataset illustrating the various domains are visualized in **Fig. 8, 9** and **10**.

7.2.1 Indoor Datasets

ScanNetV2 [11] is a richly annotated dataset of 3D indoor scenes, covering a wide variety of scenes including offices, rooms, hotels etc. It provides semantic segmentation masks for 200 fine-grained classes, known as ScanNet200, and 20 coarser classes known as ScanNet20. We evaluate on both the benchmarks.

Habitat Matterport 3D [12] is a large scale annotated dataset for 3D indoor scenes which covers 216 3D spaces and 3100 rooms within these spaces. It provides instance annotation for 40 categories. After processing, it provides us with 1805 samples for training and 481 samples for validation.

ScanNet++ [17] comes with high-fidelity 3D mask annotations including smaller objects which are not well labeled in

the ScanNet datasets. It includes high-resolution 3D scans captured at sub-millimeter precision and annotated comprehensively, covering objects of varying sizes.

Matterport3D [18] is a collection of 90 high-quality 3D reconstructions of indoor environments with instance annotations for 21 object categories. After processing, it provides us with 233 samples for validation.

S3DIS Full [19] is a collection of 6 large scale scenes covering 271 rooms. It provides annotations for 13 semantic classes. Following prior works [3, 4, 29] in instance segmentation, we use *Area_5* for evaluation which contains 68 samples for validation.

S3DIS Crops is proposed by AGILE3D [3] in their evaluation setting, cropping the validation samples from the original S3DIS dataset around each instance into $3\text{m} \times 3\text{m}$ blocks. They provide the processed data on their github [here](#). We call this dataset variant S3DIS Crops.

7.2.2 Outdoor Datasets

Outdoor datasets include two types of classes: *things*, which have instance labels, and *stuff*, which do not. This distinction can hinder the effectiveness of a promptable segmentation model. To address this, we use an off-the-shelf clustering algorithm, HDBSCAN [39], to provide us with pseudo instance labels for the *stuff* classes, enabling instance-wise promptable training on these categories. Details about HDBSCAN are provided in § 7.3.

SemanticKITTI [8] is derived from the KITTI Odometry [20] datasets. Each point in the dataset is densely labeled with one of $C = 19$ classes divided into *things* (with instance labels) and *stuff* (without instance labels) classes. We run HDBSCAN [39] to generate instance labels for the *stuff* classes.

nuScenes [9] is a comprehensive dataset that includes over 1000 diverse driving records, each lasting around 20 seconds. The LiDAR data from nuScenes is densely annotated with $C = 32$ classes, again divided into *things* and *stuff* classes. We run HDBSCAN [39] to generate instance labels for the *stuff* classes.

PandaSet [10] is an autonomous driving dataset featuring 103 driving sequences lasting about 8 seconds each. We only use a subset of 39 out of 103 scenes as the original dataset download links have expired. The dataset used for our training can be found at [Kaggle](#). It provides semantic annotations for 37 classes divided into 28 *things* and 9 *stuff* classes. We run HDBSCAN [39] to generate instance labels for the *stuff* classes.

KITTI-360 Full [21] is a large-scale outdoor driving dataset which provides 360-degree annotations on point clouds, including bounding boxes, semantic, and instance annotations. The original dataset only provides labels for down-sampled superimposed point clouds. We call this

original version KITTI-360 Full. This dataset provides annotations on 37 semantic classes in 19 object categories.

KITTI-360 Single Scan is derived from the KITTI-360 [21] Full dataset by following Interactive4D [5], where we applied a nearest-neighbor algorithm to propagate labels to individual points in individual scans. We use publicly available scripts for this purpose (Sanchez, 2021). We call this derived version KITTI-360 Single Scan. This also contains annotations for 37 semantic classes in 19 object categories. Since SNAP is trained using HDBSCAN-based instance labels for *stuff* classes, it would constitute a different evaluation setting if we evaluate on all classes. To keep evaluations fair, we only evaluate on the 11 *things* classes and compare against baselines.

KITTI-360 Crops is also derived from KITTI-360 [21] Full dataset. Specifically, to keep consistent with prior works like AGILE3D [3], which evaluated their indoor models on this variant, we also use the cropped version from their provided list of preprocessed scenes. This preprocessing includes dividing the original superimposed point clouds into smaller $3\text{m} \times 3\text{m}$ chunks centered around the object instance.

Waymo [22] is a large-scale outdoor driving dataset which provides semantic labels across 23 classes but does not provide any instance annotations. However, Waymo does provide bounding boxes for 4 classes, including *vehicle*, *cyclist*, *sign*, and *pedestrian*. We use the combination of bounding boxes and semantic labels to generate instance labels for these 4 classes. After preprocessing on the entire validation set of Waymo, we get 5,976 samples for validation. Our preprocessed dataset will be released for reproducibility.

7.2.3 Aerial Datasets

STPLS3D [13] is a large-scale photogrammetry point cloud dataset covering approximately 16km^2 of urban and rural areas in Malaysia. Released in 2020, it contains over 2 billion labeled points across 25 scenes with annotations for 14 semantic classes. To keep computational demands tractable, we crop the point clouds to $50\text{m} \times 50\text{m}$ blocks. The dataset is generated from aerial imagery using photogrammetric techniques, providing dense colored point clouds.

DALES [14] is a large-scale aerial LiDAR dataset covering 10km^2 of diverse landscapes, including urban, suburban, rural, and forested areas. It contains over 505 million points manually annotated with 8 semantic classes. Since DALES does not provide instance annotations, we again employ HDBSCAN [39] to generate instance labels for training and validation. We also crop the point clouds to $50\text{m} \times 50\text{m}$ blocks. The dataset provides high-density aerial LiDAR data (50 points per m^2) captured at varying altitudes, making it particularly challenging due to large variations in point density and object scales.

UrbanBIS [24] is a dataset for large-scale 3D urban understanding, supporting practical urban-level semantic and building-level instance segmentation. UrbanBIS comprises six real urban scenes, with 2.5 billion points, covering a vast area of 10.78km^2 and 3,370 buildings, captured by 113,346 views of aerial photogrammetry. It provides annotations on 6 scenes, out of which we evaluate on the *Yingrenshi* test scenes. After cropping to $50\text{m} \times 50\text{m}$ blocks, this provides us with 46 validation samples.

7.3. HDBSCAN Details

For the outdoor datasets used to train SNAP-C, the datasets include two types of classes: *things*, which have instance labels and *stuff* which do not have instance labels. From a promptable segmentation perspective, instance labels from *things* classes fit in directly. However, *stuff* includes classes such as vegetation, roads, buildings, *etc.*, and is assigned a single label for all of them. The objects from these classes can be far away from each other and thus using one label directly is counterproductive in training a promptable segmentation model. To solve this issue, we propose to preprocess the datasets with HDBSCAN [39]. Specifically, we first take all the points belonging to a *stuff* class, and apply clustering on it. This helps in making multiple clusters from single class labels, which can then be used for promptable segmentation training.

7.4. Training and Inference Details

All SNAP variants are trained for 100 epochs using $8 \times \text{NVIDIA A6000 GPUs}$. We train with a batch size of 1, where each batch corresponds to a single point cloud, from which 32 objects are randomly sampled for supervision. During training, we set the maximum click budget to 10. In each iteration, the number of clicks is randomly sampled between 1 and 10, ensuring that the model is consistently exposed to varying levels of user interaction. We use mixed-precision training to speed up both the training and evaluation process. We employ a round-robin style multi-dataset dataloader that repeats smaller datasets multiple times to keep the sample count similar to large datasets. During training, this dataloader provides a point cloud from one of the datasets at each training iteration, with each batch containing samples from a single dataset. To ensure proper routing through the correct normalization layer, we follow [15] and attach a *domain* variable to each point cloud. During training, this approach allows the network to route the data to the correct normalization layer. For interactive inference, this functionality translates into a simple domain-type checkbox selection, making it highly user-friendly.

Table 11. **Backbone Ablation on the ScanNet dataset.** Note that memory and time statistics are reported for 1-Click experiments.

Backbone	IoU@ k $\uparrow$			Memory	Time
	@1	@5	@10	@1 click	@1 click
AGILE3D [3]	63.3	79.9	83.7	1.2 GB	203 ms
Minkowski [41]	68.4	82.2	83.4	1.8 GB	213 ms
PTv3 [33]	68.6	82.1	84.6	1.3 GB	197 ms

Table 12. **Inference Click Strategy Ablations.** We evaluate different click strategies on the ScanNet20 dataset. Random Sampling represents all click points sampled randomly on the target object. Iterative sampling represents additional click points sampled in the unsegmented region from previous click mask.

Strategy	IoU@ k $\uparrow$		Time (ms)
Random Sampling	@1	68.6	170
	@5	79.4	178
	@10	80.5	185
Iterative Sampling	@1	68.6	170
	@5	82.3	190
	@10	85.5	211

8. Additional Ablations

8.1. Effect of Backbone Architecture

We use the PTv3 [33] backbone for feature extraction, but a natural question to ask is, ‘‘How is the model performance affected if we use a different backbone?’’ To answer this, we compare PTv3 [33] with the Minkowski Res16UNet34C[41] backbone, which has been employed by [2, 3, 5]. The comparison, conducted on the ScanNet dataset, is summarized in **Tab. 11**. We observe consistent improvement in both PTv3 and Minkowski Engine backbones against AGILE3D [3], showing that our approach is equally applicable across both recent transformer-based as well as the common sparse-convolution-based backbones. To compute the memory and timing requirements, we use a random uniform point cloud with 100,000 points on all methods.

8.2. Effect of Click Strategy

We evaluate two click strategies during inference: (1) Random-sampling and (2) Iterative Refinement. The results are shown in **Tab. 12**, with timing measurements obtained by running the evaluation on a single NVIDIA RTX 3090 GPU. While Iterative Refinement performs much better than random sampling, it also runs slower in comparison. The random sampling strategy is especially helpful for users when trying to segment objects in the scene, because users can give multiple clicks at the beginning (which is equivalent to random sampling) to get a high-quality mask and later use refinement clicks to further improve the mask

quality. In SNAP, we provide the flexibility to use both approaches during inference. To compute inference time, we used a point cloud from the ScanNet dataset with about 80,000 points.

8.3. Cross-Domain Input Ablation

To determine the effect of passing the wrong *domain input* when running zero-shot evaluations, we evaluate different domain settings of SNAP-C on 3 in-distribution and 6 zero-shot datasets. As demonstrated in **Tab. 13**, the correct domain input is crucial for getting good performance from the model on both in-distribution and zero-shot datasets. Moreover, while using the outdoor domain on indoor scenes completely disrupts performance, using the outdoor domain on aerial scenes still yields reasonable segmentation results, and vice versa. This is expected because aerial LiDAR captures are often collected over outdoor environments, which introduces partial similarities between aerial and outdoor domains while still retaining distinct characteristics.

9. Additional Quantitative Results

9.1. Timing and Memory Consumption Comparison

We compare the computational efficiency of SNAP with other interactive segmentation methods in **Tab. 14**, reporting inference time and memory consumption on an RTX 3090 GPU for single-object segmentation with 1 click. The results indicate that SNAP maintains competitive efficiency across both time and memory. For this test, we use the same uniform random point cloud with 100,000 points for all methods.

9.2. Class-Agnostic Interactive Segmentation against Non-interactive Fully-Supervised Methods

We evaluate our interactive model variants against state-of-the-art *non-interactive baselines* on both in-distribution and zero-shot datasets as a sanity check for their effectiveness. Since SNAP variants benefit from click supervision on all objects in the scene, they are expected to outperform non-interactive instance segmentation methods. As shown in **Tab. 15**, all SNAP variants achieve substantial gains over the current SOTA method EASE[29] on the ScanNet200 [11] benchmark, surpassing it by a 19.3 point margin with a single click and further improving as the number of clicks increases (5 and 10). This large advantage comes from the fact that predicting 200 categories, especially the long-tailed proportion, is inherently difficult for non-interactive methods, whereas SNAP benefits from click guidance that helps disambiguate object boundaries, leading to a much stronger performance. On the aerial STPLS3D [13] dataset with 14 semantic classes,

Table 13. **Effect of Cross-Domain Selection in Domain Normalization.** Evaluation results of SNAP-C when applying different domain types (Indoor, Outdoor, Aerial) in Domain Norm. The domain used for normalization is indicated in *teal* beneath each result.

Model	IoU @ k															
	In-Distribution								Zero-Shot							
	SemanticKITTI		ScanNet20		STPLS3D		Matterport3D		S3DIS Full		KITTI-360 Full		Waymo		UrbanBIS	
	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5
SNAP - C	71.5	86.0	19.2	52.3	41.4	66.9	17.4	45.6	15.7	51.3	23.1	48.1	69.8	86.6	55.2	80.1
<i>Norm used</i>	<i>Outdoor</i>		<i>Outdoor</i>		<i>Outdoor</i>		<i>Outdoor</i>		<i>Outdoor</i>		<i>Outdoor</i>		<i>Outdoor</i>		<i>Outdoor</i>	
SNAP - C	7.2	19.1	67.7	82.3	4.9	6.6	52.6	75.2	53.6	77.6	2.4	4.1	1.2	7.8	19.9	27.9
<i>Norm used</i>	<i>Indoor</i>		<i>Indoor</i>		<i>Indoor</i>		<i>Indoor</i>		<i>Indoor</i>		<i>Indoor</i>		<i>Indoor</i>		<i>Indoor</i>	
SNAP - C	27.1	57.9	11	22.1	67.8	80.4	12.6	22.3	8.2	25.6	6.8	28.3	25.1	60.4	71.6	90.2
<i>Norm used</i>	<i>Aerial</i>		<i>Aerial</i>		<i>Aerial</i>		<i>Aerial</i>		<i>Aerial</i>		<i>Aerial</i>		<i>Aerial</i>		<i>Aerial</i>	

Table 14. **Model Efficiency Comparison Results.** We compare the timing and memory consumption on an RTX 3090 GPU when performing single-object segmentation with 1 click.

Method	Model Size (M)	Memory (GB)	Inference Time (ms)
AGILE3D	39.3	1.20	203
Point-SAM	311.0	3.70	287
Interactive4D	39.3	1.05	200
SNAP	49.6	1.27	197

SNAP-C@1 Click shows slightly better performance than the current SOTA methods, and as expected, this continues to improve with additional clicks. On the outdoor SemanticKITTI [8] dataset, SNAP variants show very strong performance, significantly outperforming the SOTA Mask4Former [43] with 1-Click.

When evaluating zero-shot on unseen datasets like ScanNet++ [17] and Matterport3D [18], SNAP outperforms these method by 8.6 points on ScanNet++ and 14.5 points on Matterport3D. Notably, both the baseline methods LaSSM [44] and ODIN [45] were trained on ScanNet++ and Matterport3D datasets respectively. Further on the aerial UrbanBIS [24] dataset, SNAP again outperforms the B-Seg [24] baseline which was trained on the dataset. With additional clicks, this performance continues to improve.

9.3. Interactive segmentation results with all model variants

Tab. 16 presents results for all SNAP variants on in-distribution datasets. For datasets lacking established baselines, we compare against zero-shot results from single-dataset models and in-distribution results from single-domain models. SNAP-C achieves the best 1-click performance on 4/7 datasets, best 3-click results on 6/7 datasets,

and optimal performance across all datasets for higher click counts, demonstrating effective performance with a unified model.

Tab. 17 evaluates all SNAP variants on unseen datasets. SNAP-C outperforms baselines on 6/9 datasets for 1-click performance and maintains strong performance across different click counts (7/9 for 3-click, 6/9 for 5-click, 7/9 for 7-click, and 7/9 for 10-click). This demonstrates robust generalization across diverse domains. Notably, while SNAP-C may not achieve state-of-the-art performance on every individual dataset, it is the only method that operates across all domains with a single set of weights, unlike approaches such as AGILE3D [3] that require separate models for different scene types.

10. Additional Qualitative Results

We provide additional qualitative results, showcasing the performance of our SNAP model across different tasks. **Fig. 11** demonstrates the model’s capability in open-vocabulary scene understanding on the ScanNet++ dataset by using arbitrary queries involving object categories that are not present during training. **Fig. 12-20** present point-based segmentation results on the ScanNet, SemanticKITTI, and STPLS3D datasets, respectively. For each domain, we compare the ground truth masks with our model’s outputs under 1-click, 5-click, and 10-click interaction settings, reporting the corresponding mean IoU values. These results demonstrate the effectiveness of our approach in diverse environments, emphasizing the flexibility and robustness of our method across diverse segmentation challenges.

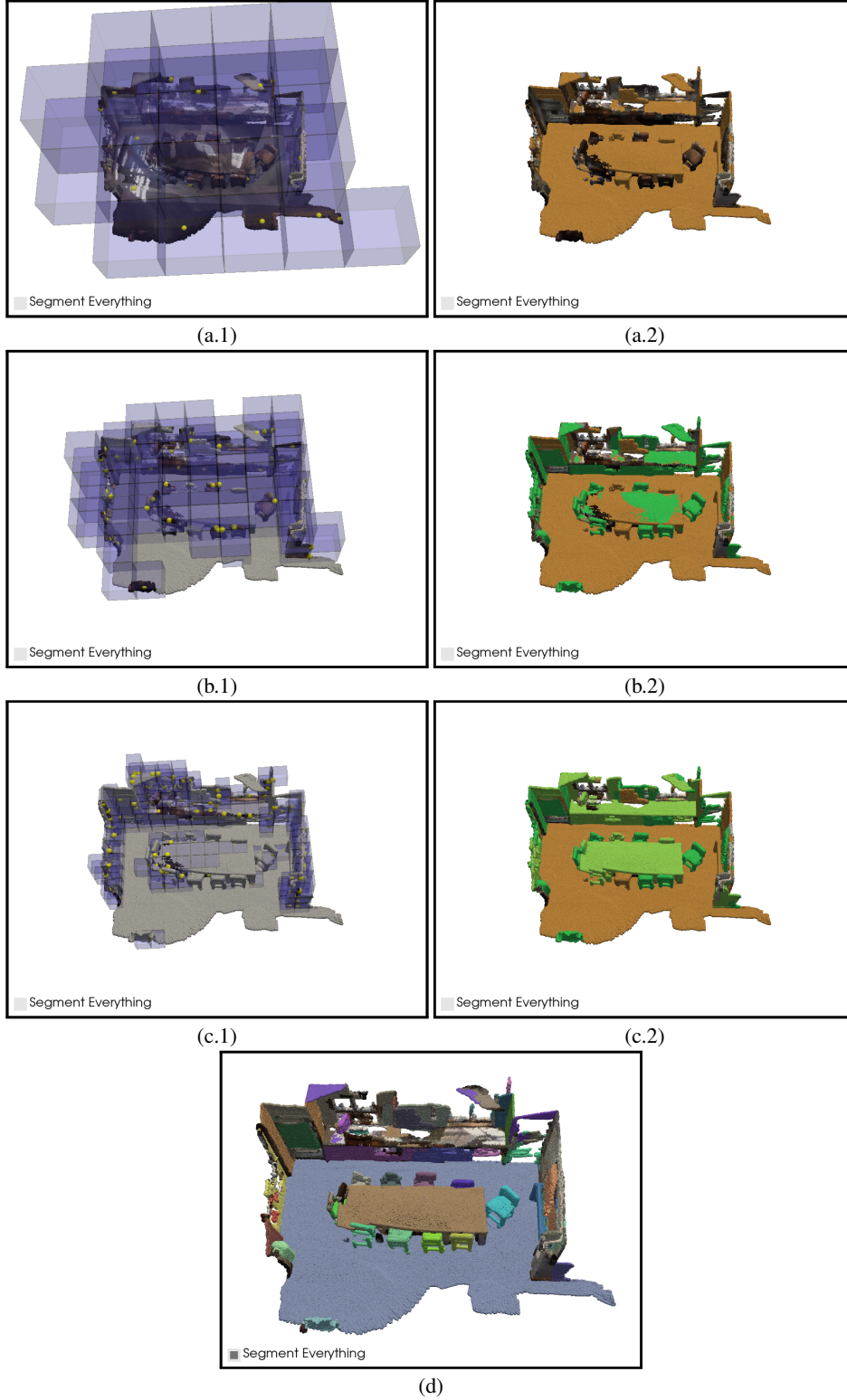


Figure 7. **Visualization of the Iterative Prompting algorithm.** (a.1): Generate prompt points with a large voxel size to segment out large objects in the scene, (a.2): All the points segmented after first iteration (Yellow color). (b.1): Reduce the voxel size and generate prompt points on the unsegmented points. (b.2): All points segmented after Iteration 1 (yellow) and 2(dark green). (c.1): Reduce voxel size again and repeat. (c.2): All points segmented after Iteration 1 (yellow), 2(dark green) and 3 (light green). (d): Final instance masks after Non Maximum Suppression.

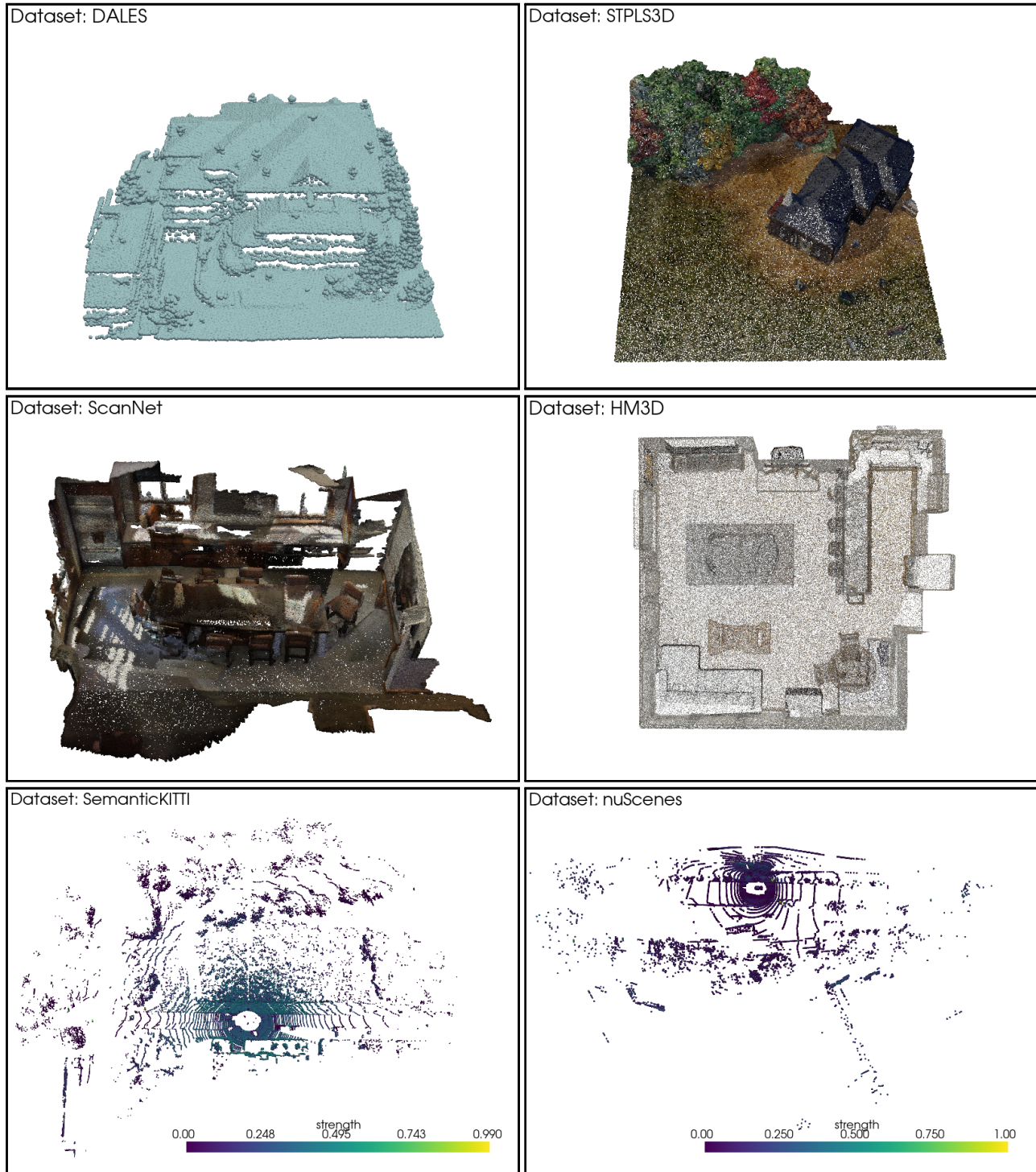


Figure 8. **Samples from Training Datasets.** Here we present samples taken from the training datasets to showcase the difference in scale, point density and scene types. Note that the dataset name is displayed in each figure. From the top - DALES and STPLS3D are aerial datasets, ScanNet and HM3D are indoor scene datasets and SemanticKITTI and nuScenes are outdoor scene datasets. HM3D provides full room scenes, point size has been reduced for better understanding of the scene.

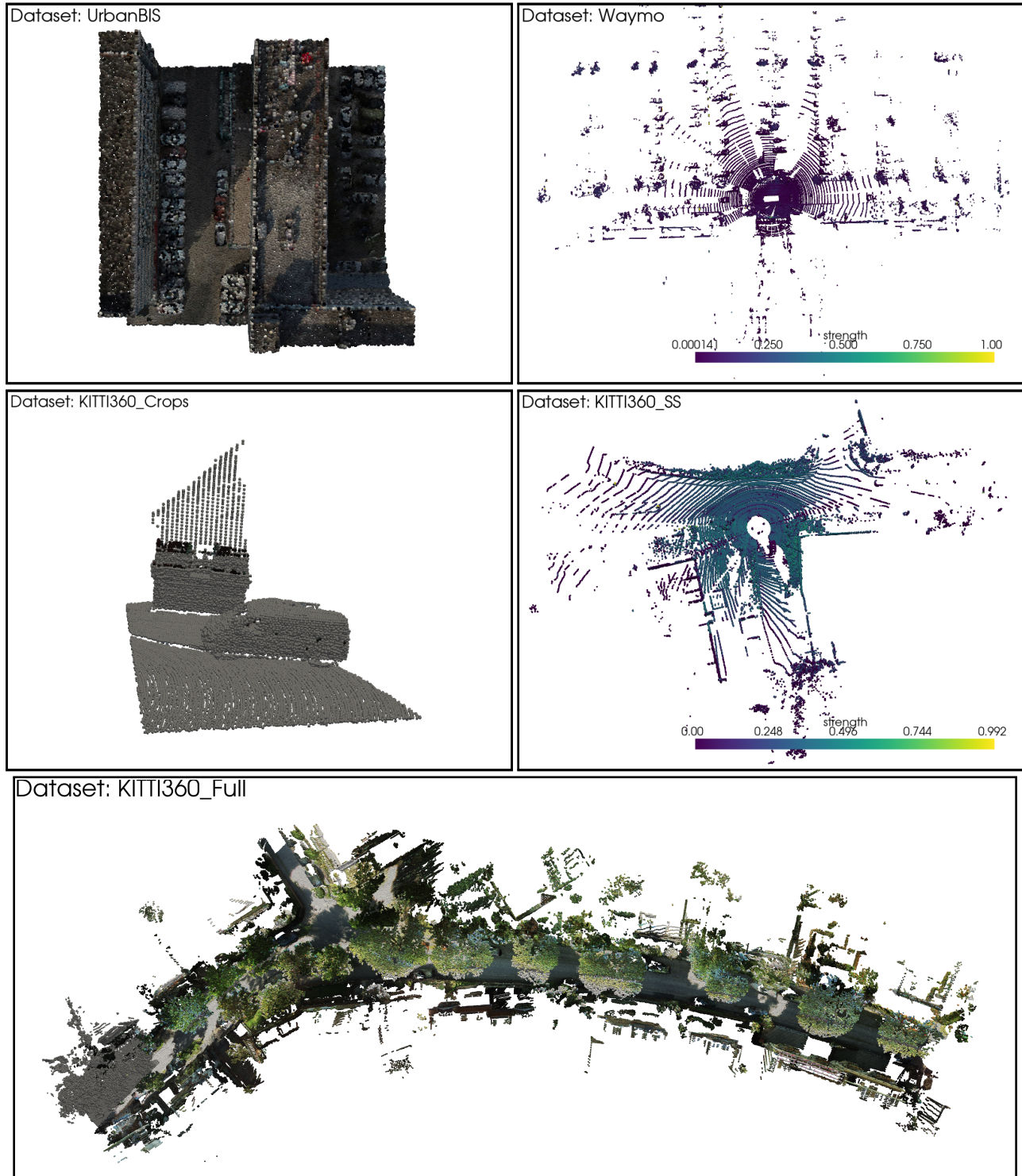


Figure 9. **Samples from Validation Datasets.** We evaluate SNAP on a variety of datasets. From the top - UrbanBIS is an Aerial scene; Waymo, KITTI-360 Crops, KITTI-360 Single Scan, and KITTI-360 Full are outdoor scene datasets. Note the difference in the variants of KITTI-360. KITTI-360 Crops particularly represents small-scale dense scenes generally found in indoor point clouds, while KITTI-360 Single Scan shows a traditional point cloud collected with a LiDAR sensor, and KITTI-360 Full shows an aggregated point cloud map built by combining multiple LiDAR scans.

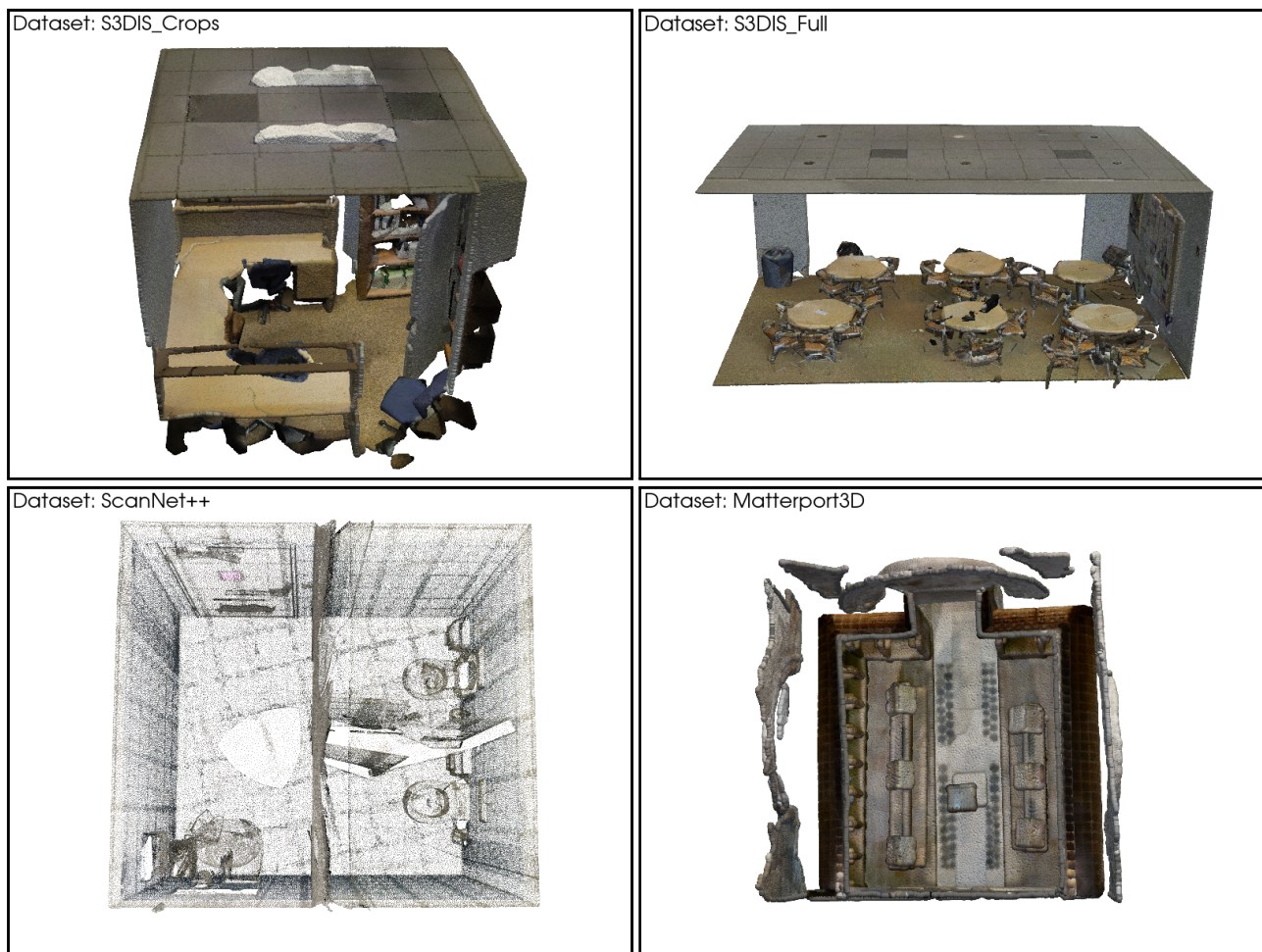


Figure 10. **Samples from Validation Datasets.** Here we show the examples from Indoor datasets used for validation. Note that S3DIS Crops is a cropped version of the full S3DIS point clouds, therefore some objects appear truncated. ScanNet++ provides full room scenes, point size has been reduced for better understanding of the scene.

Table 15. **Class Agnostic Instance Segmentation Comparison against Non-Interactive Fully-Supervised Methods.** We compare SNAP against state-of-the-art baselines for in-distribution and zero-shot datasets on the instance segmentation task in a class-agnostic fashion.

In-distribution Evaluation									
ScanNet200 Validation Set									
Method	mAP			mAP50			mAP25		
EASE [29]	29.9			38.8			44.7		
	@1	@5	@10	@1	@5	@10	@1	@5	@10
SNAP - SN200	47.9	68.1	73	69.7	89.6	92.4	84.2	97.3	98.8
SNAP - Indoor	45.2	66.7	73.3	69.3	90.5	95.3	84.8	98.8	99.7
SNAP - C	49.2	69.8	77.5	73.2	91.6	95.9	87.7	99.1	99.8
STPLS3D Validation Set									
Method	mAP			mAP50			mAP25		
Mask3D [42]	57.3			74.3			81.6		
	@1	@5	@10	@1	@5	@10	@1	@5	@10
SNAP - Aerial	56.2	72.9	80.7	74.4	88.9	94	86.5	97.1	98.6
SNAP - C	58.3	75.7	84.4	76.7	91.1	95.3	88.8	98.0	99.1
Semantic KITTI Validation Set									
Method	PQ			SQ			RQ		
Mask4Former [43]	61.7			81			71.4		
	@1	@5	@10	@1	@5	@10	@1	@5	@10
SNAP - KITTI	68.6	83.4	87.6	80.9	84.9	88.9	82.7	91.8	97.5
SNAP - Outdoor	69.9	85.1	90.1	82.3	87.9	91.3	84.1	96.3	98.3
SNAP - C	71.1	86.5	90.7	82.7	88.7	91.7	84.8	97.4	98.4
Zero-Shot Evaluation									
ScanNet++ Validation Set									
Method	mAP			mAP50			mAP25		
LaSSM [44]	29.1			43.5			51.6		
	@1	@5	@10	@1	@5	@10	@1	@5	@10
SNAP - SN200	32.9	52.1	58.1	49.1	71.9	77.3	62.9	86.1	89.6
SNAP - Indoor	35.9	59.8	66.5	55.5	84.9	89.4	73.3	97.3	98.6
SNAP - C	37.7	60.4	70.1	55.4	83.0	90.9	73.2	94.6	98.1
Matterport3D Validation Set									
Method	mAP			mAP50			mAP25		
ODIN [45]	24.7			–			63.8		
	@1	@5	@10	@1	@5	@10	@1	@5	@10
SNAP - SN200	42.7	62.3	68.8	64.4	85.6	90.9	77.5	95.9	97.1
SNAP - Indoor	36.5	63	70.7	59.2	90.3	93.3	74.2	98.1	98.8
SNAP - C	39.2	64.6	74.3	59.4	88.7	94.5	77.3	96.9	98.7
UrbanBIS Validation Set									
Method	mAP			mAP50			mAP25		
B-Seg [24]	62.1			70			73.9		
	@1	@5	@10	@1	@5	@10	@1	@5	@10
SNAP - Aerial	62.9	84.2	91.1	85.6	95.5	98.2	96.4	99.1	98.2
SNAP - C	62.2	89.1	94.9	84.1	100	100	94.6	100	100

Table 16. **In-distribution Interactive Point Cloud Segmentation Results.** * indicates models not trained on the evaluation dataset and † denotes that the methods are evaluated by us.

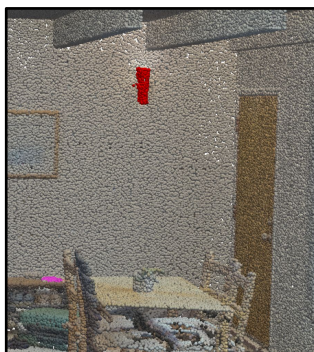
Domain	Dataset	Method	IoU@k				
			@1	@3	@5	@7	@10
Outdoor	SemanticKITTI	AGILE3D [3]	53.1	70	76.7	-	83
		Interactive4D [5]	67.5	78.3	83.4	-	88.2
		SNAP-KITTI	68.1	80.1	84.5	87.5	88.7
		SNAP-Outdoor	71.3	81.9	85.7	87.7	89.3
		SNAP-C	71.5	81.9	86	88.1	90
	nuScenes	AGILE3D* [3]	32.4	47.1	56.4	-	68.4
		Interactive4D* [5]	45.5	57.2	64.6	-	74.3
		SNAP-KITTI*	50.2	64.3	71.3	74.6	76.9
		SNAP-Outdoor	72.4	83.1	88.1	90.2	91.2
		SNAP-C	72.2	83.3	88.1	90.3	92.2
	Pandaset	SNAP-KITTI	17	29.7	34.2	35.8	36.7
		SNAP-Outdoor	60.5	74.8	80.1	82.1	84.3
		SNAP-C	56.3	74.6	80.2	82.6	84.4
Indoor	ScanNet20	InterObject3D [2]	40.8	63.9	72.4	-	79.9
		AGILE3D [3]	63.3	75.4	79.9	-	83.7
		Point-SAM† [4]	52.7	75.9	80.6	82.9	83.3
		SNAP-SN	68.6	78.4	82.1	83.4	84.6
		SNAP-Indoor	66	77.6	81.3	83	84
		SNAP-C	67.7	78.5	82.3	84.1	85.5
	HM3D	SNAP-SN	38.7	52.4	58.6	61.4	63.3
		SNAP-Indoor	47.1	65.2	71.2	74.0	75.9
		SNAP-C	50	66.7	72.9	76.1	78.7
	Aerial	STPLS3D	65.8	74.5	79.1	81.6	83.6
		SNAP-C	67.8	75.5	80.4	83.3	85.8
Aerial	DALES	SNAP-Aerial	60.7	72.5	76.8	78.7	80
		SNAP-C	61.6	74	78.2	80.4	82.3

Table 17. **Zero-Shot Interactive Point Cloud Segmentation Results.** † denotes that the methods are evaluated by us.

Domain	Dataset	Method	IoU@k				
			@1	@3	@5	@7	@10
Outdoor	Waymo	Interactive4D	7.2	7.3	7.5	7.7	7.9
		Point-SAM	12.8	43	53.1	57.4	60.2
		SNAP-KITTI	48.2	61.8	66.3	68.4	70
		SNAP-Outdoor	68.5	81.8	86	87.4	88.3
		SNAP-C	69.8	82.3	86.6	88.2	89.3
	KITTI 360 Full	Point-SAM	6.8	22.7	28.1	30.7	32.5
		SNAP-KITTI	6.7	25.9	29.7	31.2	32.4
		SNAP-Outdoor	18.3	35.2	44.6	47.3	50.2
		SNAP-C	23.1	40.1	48.1	51.7	54.2
	KITTI 360 Single Scan	AGILE3D	36.3	47.3	53.5	-	63.3
		Interactive4D	47.7	59.4	64.1	-	70
		SNAP-KITTI	54.4	60.9	63.9	65.5	66.8
		SNAP-Outdoor	59.8	63.3	65.9	67.5	69.1
		SNAP-C	60.4	64.6	67.7	70.1	72.6
	KITTI 360 Crops	AGILE3D	34.8	42.7	44.4	45.8	49.6
		Point-SAM	49.4	74.4	81.7	84.3	85.8
		SNAP-KITTI	56.1	68.8	72.8	74.3	75.3
		SNAP-Outdoor	56.9	70.6	75.6	78.1	80.3
		SNAP-SN	54.5	68.6	74.1	76.8	78.6
		SNAP-Indoor	54.9	70.2	76.6	79.3	80.4
		SNAP-C	65.6	76.1	80	82.1	83.6
Indoor	ScanNet++	Point-SAM†	28.6	56.3	62.9	65.5	67.2
		SNAP-SN	45.5	59.9	65.3	67.8	69.5
		SNAP-Indoor	51.5	67.9	73.4	75.9	77.6
		SNAP-C	52	67.3	73.2	76.3	78.6
	Matterport3D	Point-SAM†	41.1	67.2	73.7	76.2	77.9
		SNAP-SN	53.4	66.6	71.3	73.7	75.3
		SNAP-Indoor	49.9	68.4	74.2	76.4	78.3
		SNAP-C	52.6	69.6	75.2	78.2	80.5
	S3DIS Crops	AGILE3D	58.7	77.4	83.6	86.4	88.5
		Point-SAM	45.9	77.6	84.6	86.9	88.4
		SNAP-SN	55.8	68.7	74.1	77.2	79.4
		SNAP-Indoor	54.8	73.5	80.5	83.7	85.9
		SNAP-C	56.6	73.8	80.9	84.4	87
	S3DIS Full	Point-SAM†	35.6	68	76.3	78.9	80.6
		SNAP-SN	51.4	64.3	70	72.6	74.8
		SNAP-Indoor	51.9	70.1	76.9	79.9	81.9
		SNAP-C	53.6	71.1	77.6	80.8	83.2
Aerial	UrbanBIS	Point-SAM	39.3	79.1	89.4	92.7	94.3
		SNAP-Aerial	74.2	83.2	86.9	89.8	90.6
		SNAP-C	71.6	86.2	90.2	92.8	94.7



"Segment bar table."



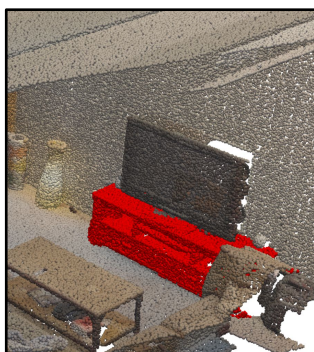
"Segment light."



"Segment swivel chair."



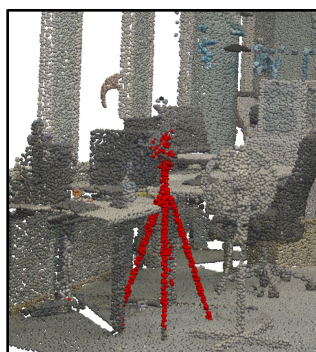
"Segment kitchen mat."



"Segment TV stand."



"Segment footrest."



"Segment tripod."



"Segment utility stand."



"Segment desktop."

Figure 11. **Additional qualitative segmentation results of open-set scene understanding on the ScanNet++ Dataset.** Given a text prompt in the format of "*Segment {open-set vocabulary}*", our SNAP model finds the corresponding masks ■ in the scenes.

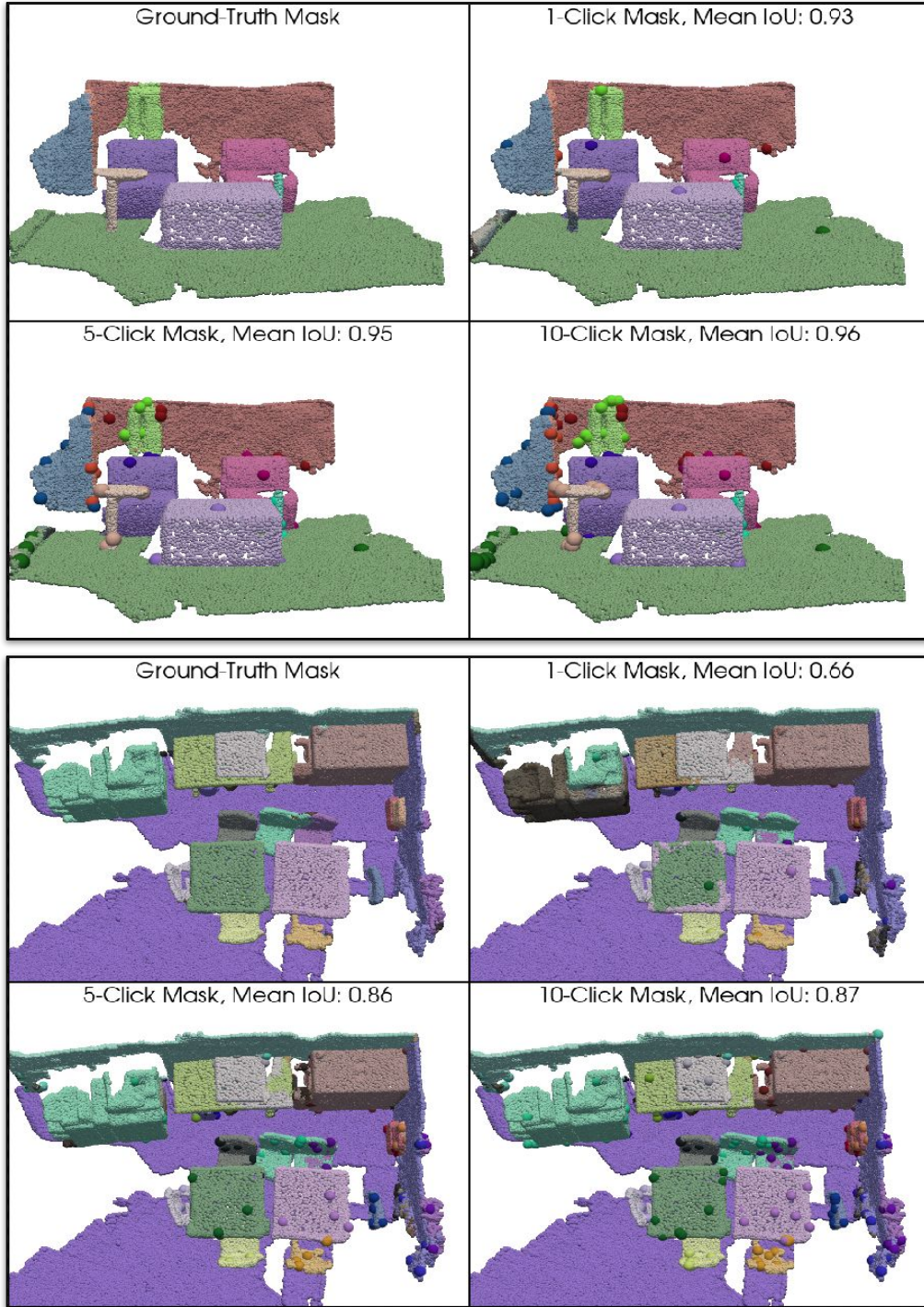


Figure 12. **Additional qualitative results for point-based segmentation on the ScanNet dataset.** For each block, we show the ground truth masks alongside our segmentation results for 1-click, 5-click, and 10-click interactions, including the corresponding mean IoU values. Points with the same color represent the same object, while clicks are highlighted using darker colors and larger spheres for better visibility.

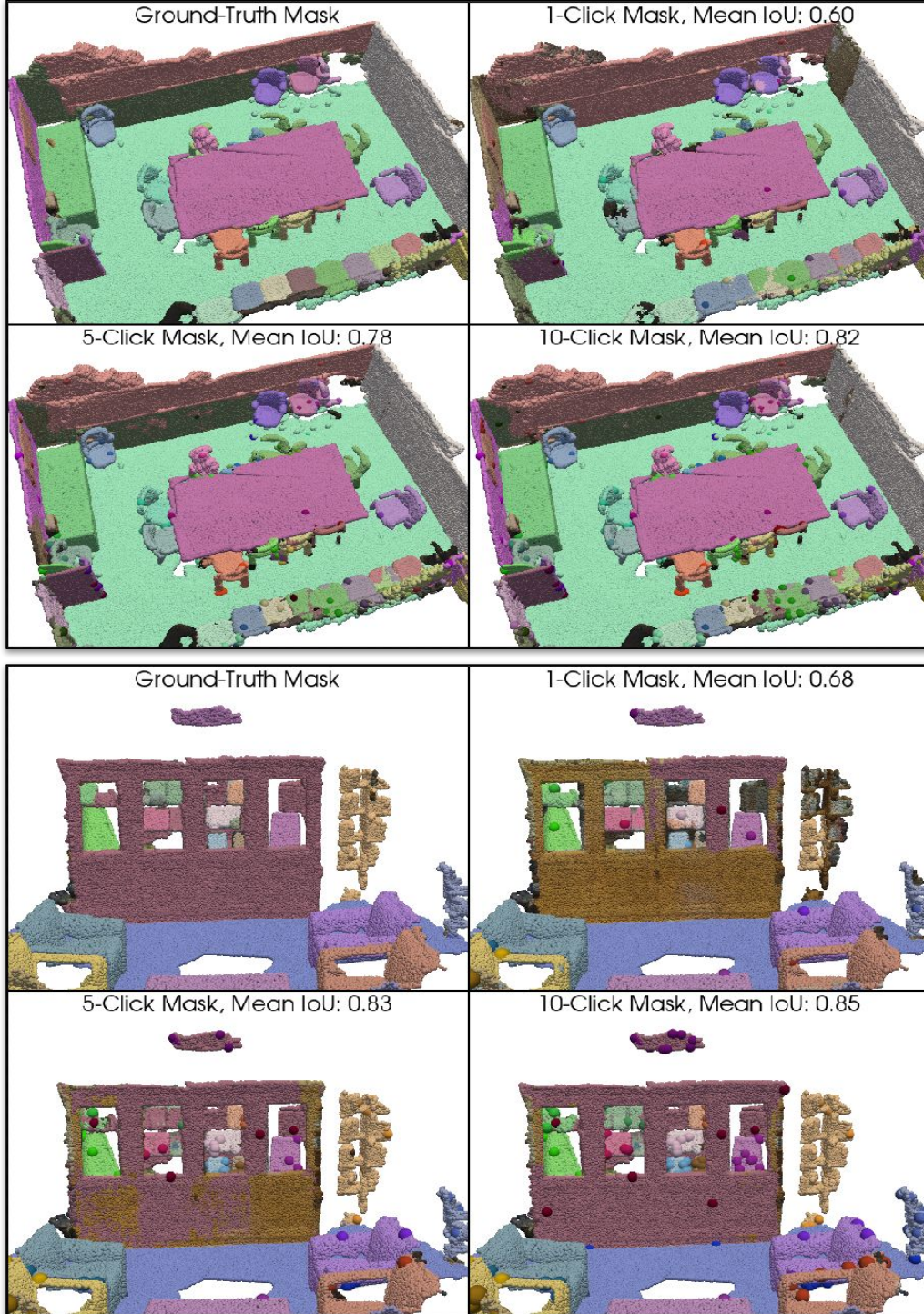


Figure 13. **Additional qualitative results for point-based segmentation on the ScanNet dataset.** For each block, we show the ground truth masks alongside our segmentation results for 1-click, 5-click, and 10-click interactions, including the corresponding mean IoU values. Points with the same color represent the same object, while clicks are highlighted using darker colors and larger spheres for better visibility.

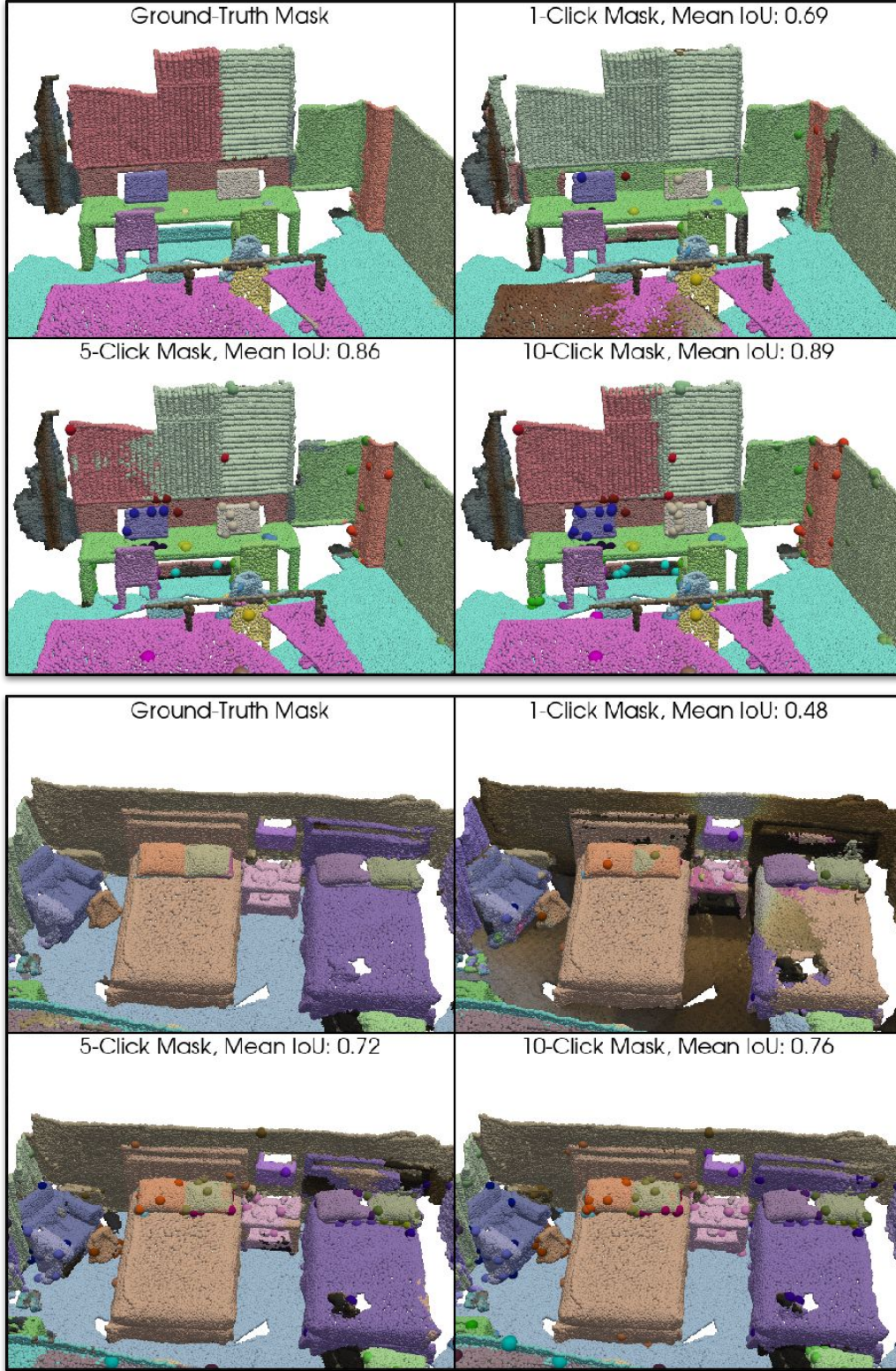


Figure 14. **Additional qualitative results for point-based segmentation on the ScanNet dataset.** For each block, we show the ground truth masks alongside our segmentation results for 1-click, 5-click, and 10-click interactions, including the corresponding mean IoU values. Points with the same color represent the same object, while clicks are highlighted using darker colors and larger spheres for better visibility.

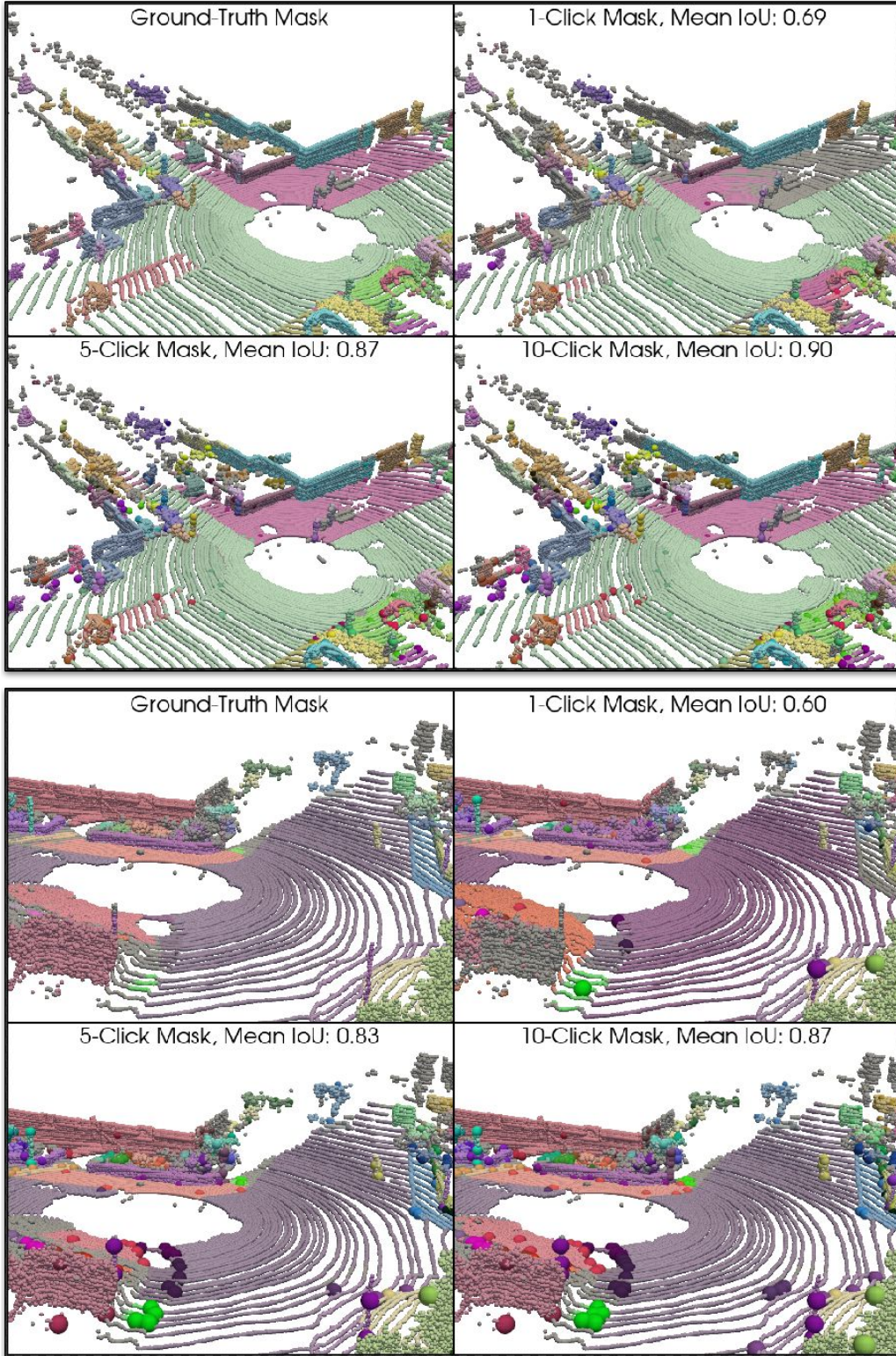


Figure 15. **Additional qualitative results for point-based segmentation on the SemanticKITTI dataset.** For each block, we show the ground truth masks alongside our segmentation results for 1-click, 5-click, and 10-click interactions, including the corresponding mean IoU values. Points with the same color represent the same object, while clicks are highlighted using darker colors and larger spheres for better visibility.

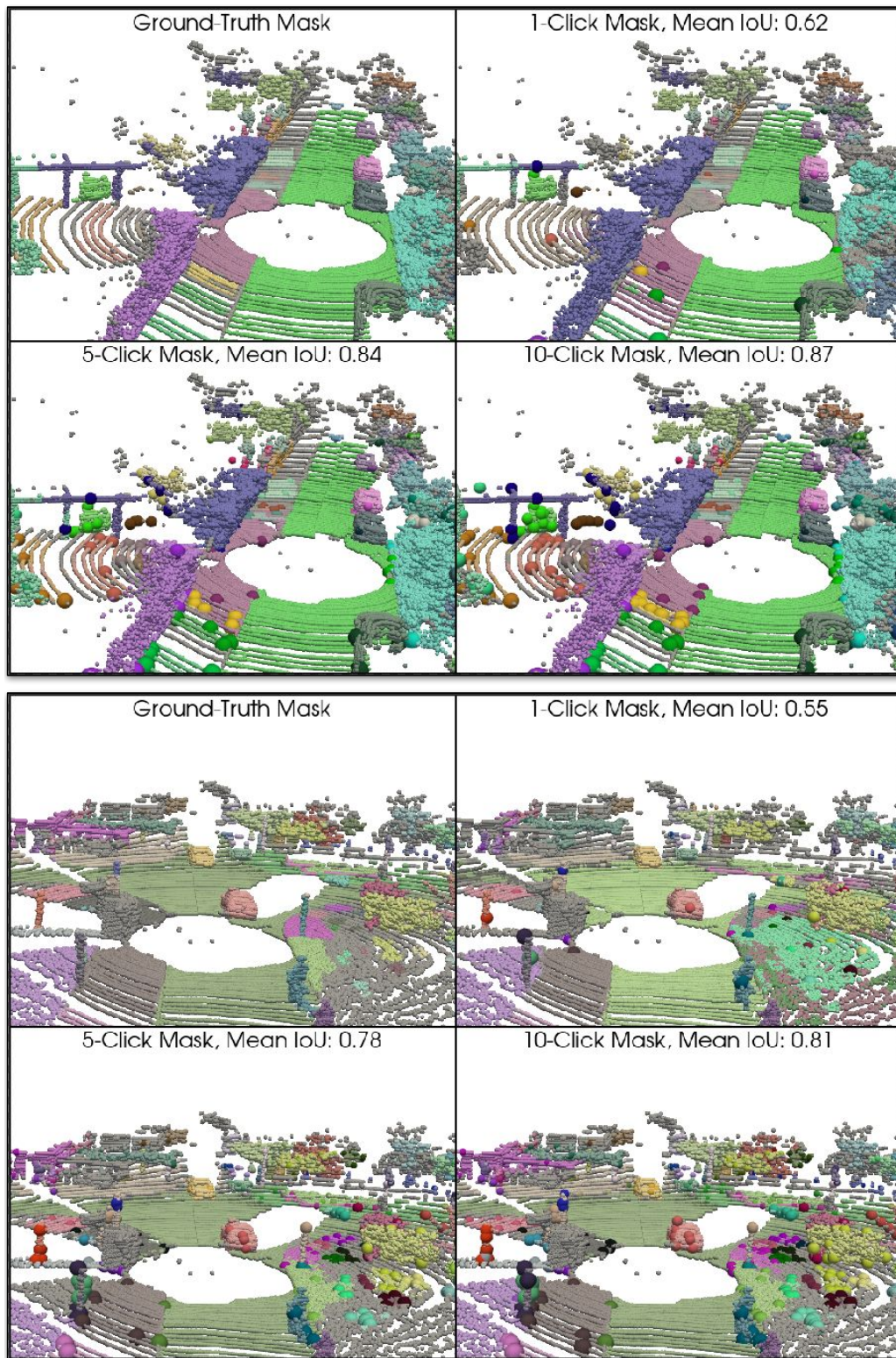


Figure 16. **Additional qualitative results for point-based segmentation on the SemanticKITTI dataset.** For each block, we show the ground truth masks alongside our segmentation results for 1-click, 5-click, and 10-click interactions, including the corresponding mean IoU values. Points with the same color represent the same object, while clicks are highlighted using darker colors and larger spheres for better visibility.

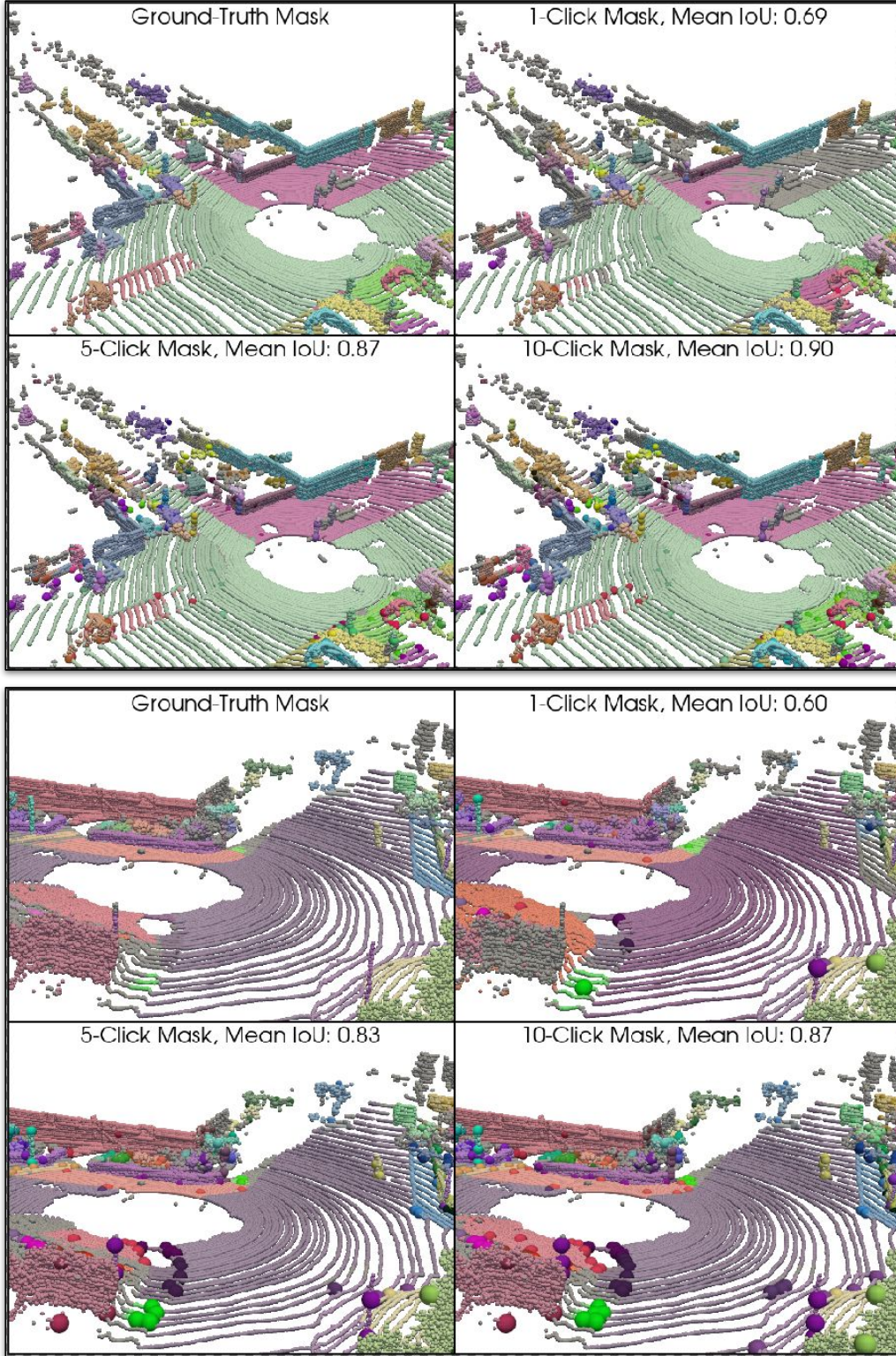


Figure 17. **Additional qualitative results for point-based segmentation on the SemanticKITTI dataset.** For each block, we show the ground truth masks alongside our segmentation results for 1-click, 5-click, and 10-click interactions, including the corresponding mean IoU values. Points with the same color represent the same object, while clicks are highlighted using darker colors and larger spheres for better visibility.

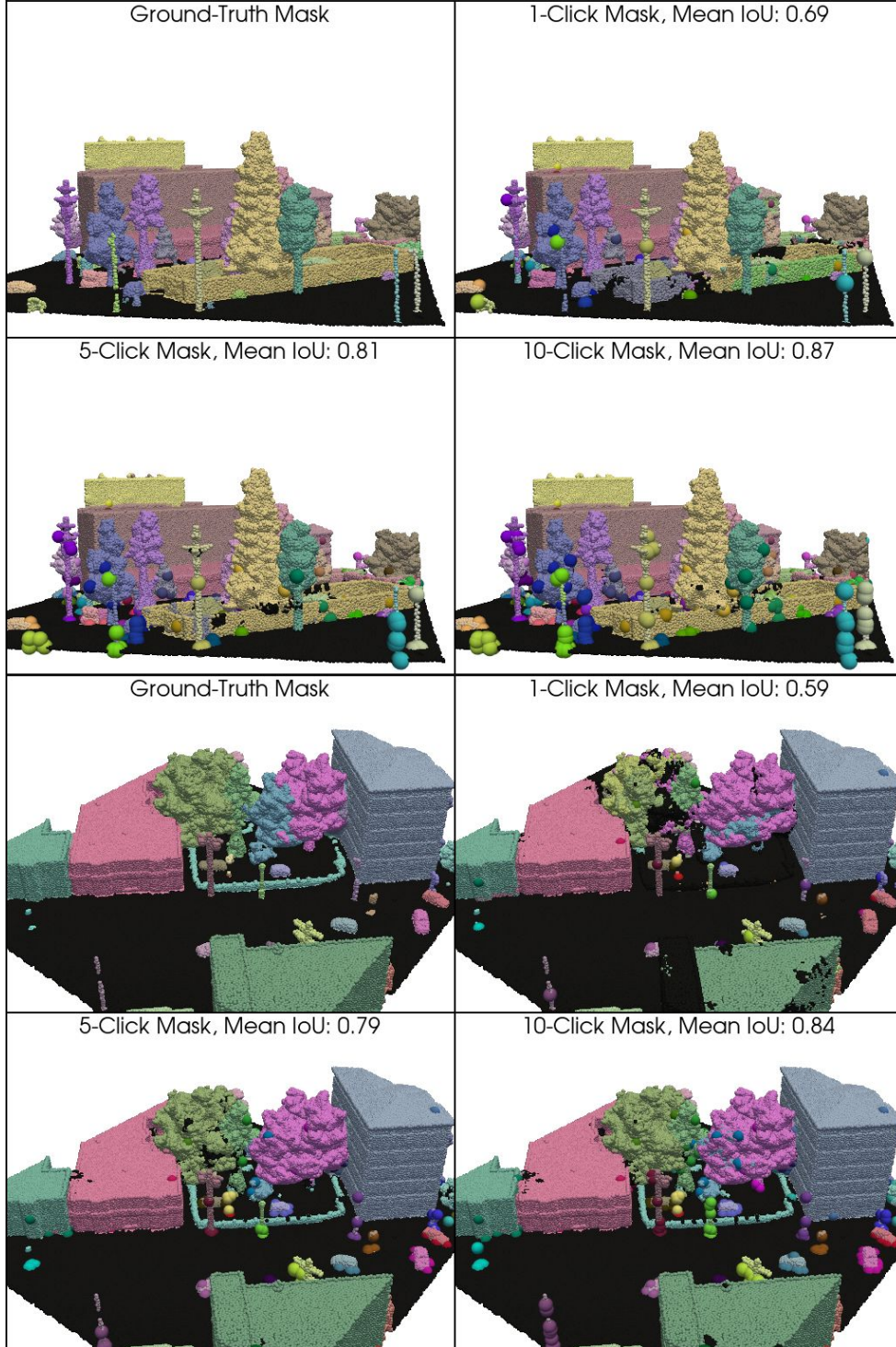


Figure 18. **Additional qualitative results for point-based segmentation on the STPLS3D dataset.** For each block, we show the ground truth masks alongside our segmentation results for 1-click, 5-click, and 10-click interactions, including the corresponding mean IoU values. Points with the same color represent the same object, while clicks are highlighted using darker colors and larger spheres for better visibility.

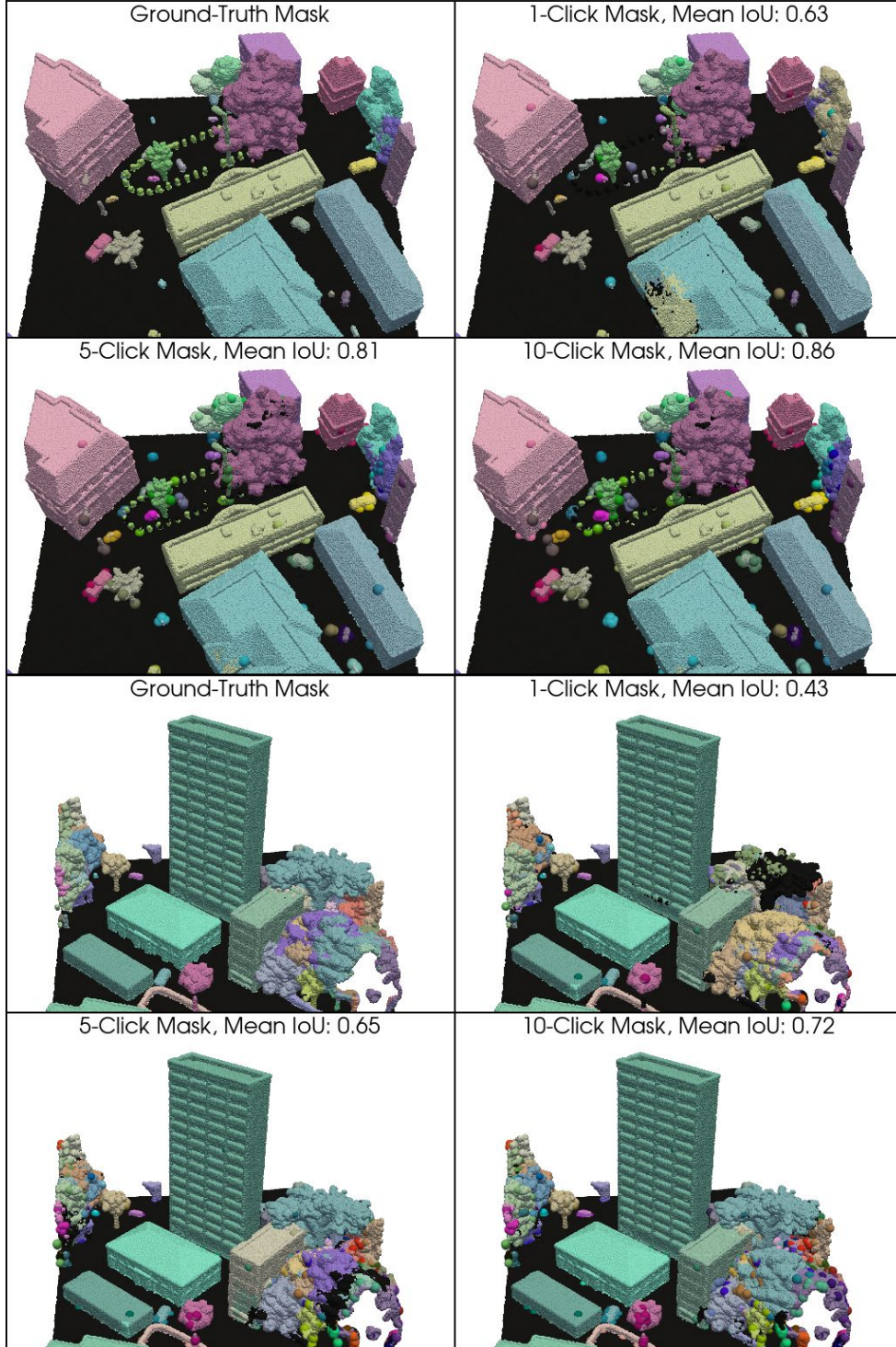


Figure 19. **Additional qualitative results for point-based segmentation on the STPLS3D dataset.** For each block, we show the ground truth masks alongside our segmentation results for 1-click, 5-click, and 10-click interactions, including the corresponding mean IoU values. Points with the same color represent the same object, while clicks are highlighted using darker colors and larger spheres for better visibility.

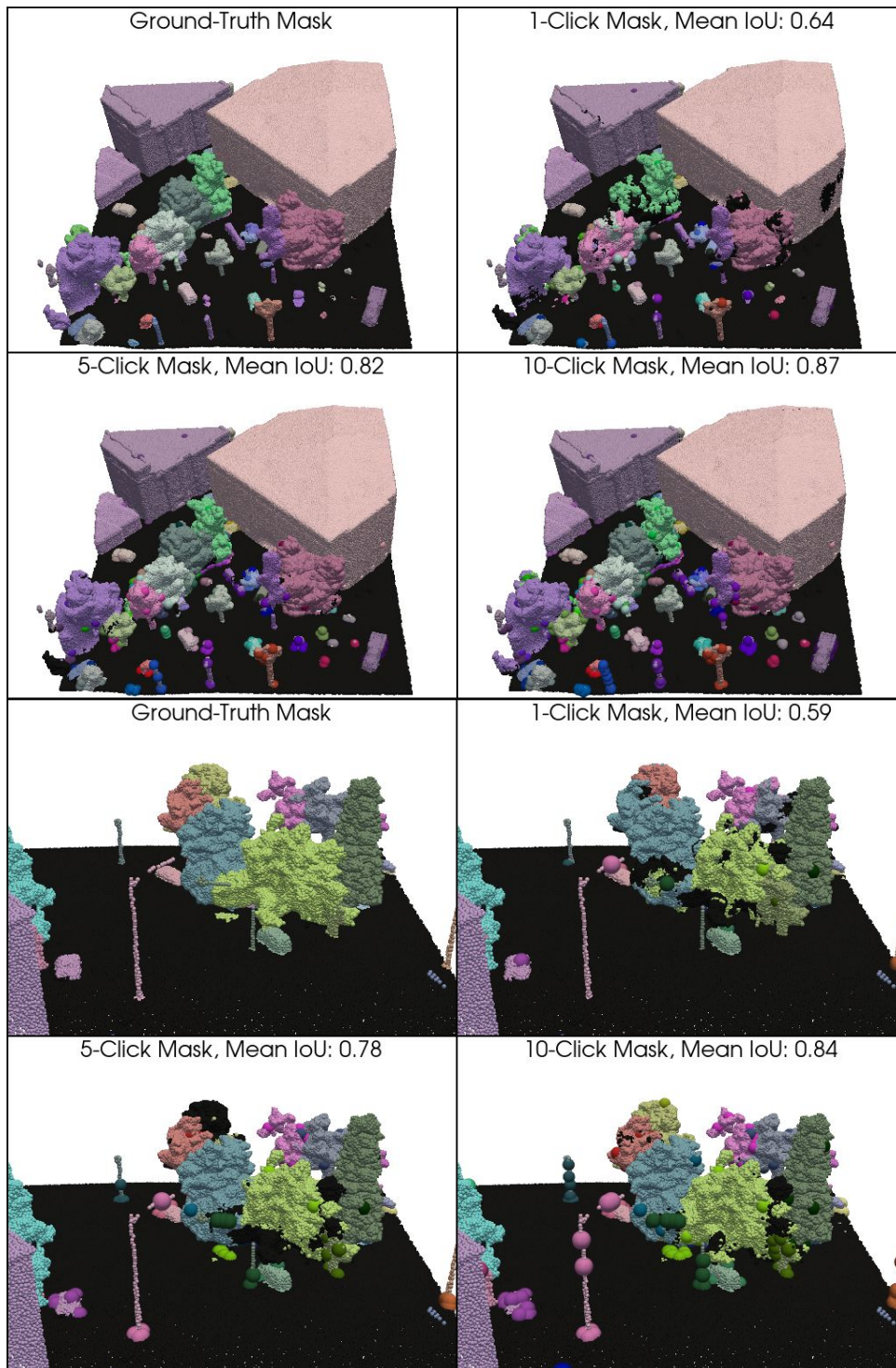


Figure 20. **Additional qualitative results for point-based segmentation on the STPLS3D dataset.** For each block, we show the ground truth masks alongside our segmentation results for 1-click, 5-click, and 10-click interactions, including the corresponding mean IoU values. Points with the same color represent the same object, while clicks are highlighted using darker colors and larger spheres for better visibility.