NegROI: Click-Centric Uncertainty-Guided Refinement with Scene-Conditioned Negative Prompts for Robust Interactive 3D Segmentation

Shuheng Zhang and Feng Wu(✉)*

School of Computer Science and Technology,
University of Science and Technology of China, Hefei, Anhui, China.
zsh123456@mail.ustc.edu.cn, wufeng02@ustc.edu.cn

Abstract. Interactive 3D segmentation aims to extract object masks in point clouds with minimal user clicks. Despite recent progress, most existing approaches still struggle with (i) coarse voxel resolution that blurs fine boundaries under limited clicks and (ii) hard false positives caused by confusing background structures. These issues are exacerbated by density and scale shifts across datasets (e.g., dense RGB-D reconstructions vs. sparse LiDAR scans), where fixed refinement heuristics and purely click-driven decoding generalize poorly. To address them, we propose NegROI — a novel transformer-based interactive framework that couples click-centric multi-resolution refinement with scene-conditioned negative prompts. Given a coarse voxel prediction, it refines only a local Region Of Interest (ROI) around the current click on a finer grid and fuses refined logits back to the coarse mask. To improve robustness and efficiency, we introduce uncertainty-driven selective refinement that prioritizes ambiguous regions. Meanwhile, we model hard background patterns via a set of scene-conditioned negative prompts obtained by cross-attention over scene tokens. We further stabilize these prompts with a diversity regularizer. Finally, we propose boundary-aware hard negative mining to supervise negative-prompt attention toward boundary-proximal, high-confidence false positives. Our experiments on common benchmark datasets (i.e., ScanNet, S3DIS, and KITTI) demonstrate improved click efficiency and reduced false positives, with stronger cross-dataset robustness than the state-of-the-art baselines.

Keywords: Interactive 3D segmentation · Point clouds · Transformers · Prompting · Domain robustness

1 Introduction

Interactive segmentation reduces annotation cost by iteratively refining a mask with user feedback. In many 3D applications, interactive segmentation is particularly valuable because dense point-wise labels are often expensive to acquire.

* Feng Wu is the corresponding author.

However, point clouds exhibit large variability in density, noise, and scale. Therefore, interactive pipelines must respond reliably to sparse user clicks.

Modern interactive 3D segmentation methods [12, 25, 32, 34] commonly operate on voxel grids for efficiency and use transformer decoders to integrate click tokens with scene features. While effective, we observe two persistent failure modes: (1) *Boundary under-segmentation:* A single-resolution voxel representation tends to blur object boundaries, especially when the click budget is small. (2) *Hard false positives (FPs):* Even with clicks, confusing background structures can trigger high-confidence FPs. Both issues are amplified under distribution shifts (e.g., ScanNet-trained models tested on S3DIS or KITTI), where background appearance and sampling patterns differ substantially.

Against this background, we propose a novel method named *NegROI*, which targets both issues with a unified design. Firstly, we introduce *click-centric multi-resolution refinement*: after a coarse prediction, we refine only a local ROI around the *current* click on a finer grid and fuse the refined prediction back to the coarse mask. Secondly, we introduce *scene-conditioned negative prompts* that act as background prototypes learned by cross-attending to scene tokens, providing a structured way to suppress hard FPs. To improve robustness and efficiency, we add uncertainty-driven selective refinement, a diversity regularizer to prevent prompt collapse, and boundary-aware hard negative mining to directly supervise negative-prompt attention on boundary-adjacent confusing background.

Our main contributions are summarized as follows:

(1) We propose the *click-centric multi-resolution refinement* module that performs efficient fine-grid refinement inside a local ROI and fuses fine predictions back to coarse logits.
(2) We propose the *uncertainty-driven selective refinement* to focus refinement computation on ambiguous regions, improving efficiency and robustness.
(3) We propose the *scene-conditioned negative prompts* for explicit background modeling, augmented with a *prompt diversity regularizer*.
(4) We propose the *boundary-aware hard negative mining* to supervise negative-prompt attention toward boundary-proximal, high-confidence false positives.

All together, we advance the state-of-the-art for robust interactive 3D segmentation. The experimental results on common benchmark datasets (i.e., ScanNet, S3DIS, and KITTI) demonstrate im- proved click efficiency, reduced false positives, and cross-dataset robustness of our approach.

2 Related Work

In this section, we briefly review the related work on interactive 3D segmentation.

Interactive image segmentation. Early interactive segmentation methods leverage sparse user cues such as extreme points [17] and iterative refinement strategies [26, 27]. Recent work improves click efficiency by better modeling click locality and hard errors (e.g., [2, 15]), and by adding lightweight high-resolution

refinement heads for sharper boundaries [10]. Foundation models further broaden promptability in 2D, most notably Segment Anything (SAM) [9] and its video extension [23]. Our work adapts the "refine-where-needed" principle to 3D by performing click-centric local refinement on a fine voxel grid and explicitly suppressing hard false positives.

3D point cloud segmentation. Backbones for 3D understanding span point-based networks [20, 21, 30, 35] and efficient large-scale designs [6], as well as sparse convolutional approaches [3, 5, 29]. For 3D instance segmentation, grouping-based methods such as PointGroup [8] and SoftGroup [31] are widely used, while transformer-based mask prediction has shown strong performance in recent unified frameworks [11, 24]. NegROI is complementary to these backbones and focuses on interactive prompting, local refinement, and robust background modeling.

Interactive and promptable 3D segmentation. Interactive 3D segmentation has advanced from early click-based pipelines to transformer-driven promptable systems. InterObject3D [12] and AGILE3D [34] demonstrate that attention-guided decoding can effectively integrate user clicks for object extraction in cluttered scenes. Easy3D [25] provides a strong and simple interactive baseline, while Point-SAM [37] extends the SAM-style prompting paradigm to point clouds. Compared to prior work, we explicitly model confusing background via scene-conditioned negative prompts and couple them with click-centric fine-grid refinement to improve boundary fidelity and cross-dataset robustness.

Open-vocabulary vision-language and 3D understanding. Open-vocabulary segmentation commonly builds on large-scale vision-language pretraining [7, 22]. Text-supervised grouping and segmentation emerge in models such as GroupViT [33], and open-vocabulary recognition can be bootstrapped from image-level supervision [36] or open-set detectors [16]. For segmentation, CLIP-adapted mask models provide a practical recipe for open-vocabulary transfer [13]. In 3D, OpenScene [19] studies open-vocabulary 3D scene understanding, while OpenMask3D [28] and Open3DIS [18] address open-vocabulary 3D instance segmentation by leveraging 2D foundation models and mask guidance. Our work targets the *interactive* setting and is compatible with open-vocabulary querying, while focusing on improved error localization and suppression of hard distractors.

Hard negatives and boundary-focused refinement. A recurring challenge in interactive segmentation is that errors are highly concentrated around ambiguous boundaries and confusing distractors. Prior 2D interactive methods improve click efficiency by explicitly focusing on hard regions [2, 15] and by adding lightweight high-resolution heads to sharpen boundaries [10]. In 3D interactive segmentation, clutter and object attachments further amplify boundary-proximal false positives [12, 34]. Our approach follows the same philosophy of *refine-where-needed* and *focus-on-hard-negatives*, but instantiates it with (i) boundary-aware

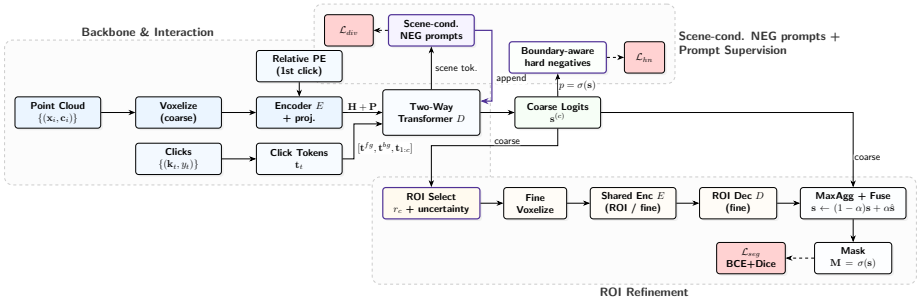

Fig. 1: NegROI overview (aligned with our implementation). We use a sparse voxel backbone and the Easy3D two-way transformer decoder to produce coarse logits. We augment decoding with scene-conditioned negative prompts supervised by $\mathcal{L}_{hn}$ and $\mathcal{L}_{div}$. For boundary accuracy, we refine a click-centric ROI on a fine grid using the shared encoder and an ROI decoder, then max-aggregate and fuse back to coarse logits.

hard negative mining that supervises negative-prompt attention, and (ii) click-centric fine-grid ROI refinement that improves boundary fidelity without dense high-resolution decoding.

3 Main Method

As aforementioned, we refer to our approach as *NegROI*, which combines scene-conditioned negative prompts with click-centric ROI refinement for robust interactive 3D segmentation.

3.1 Problem Statement

Given a point cloud $\mathcal{P} = \{(\mathbf{x}_i, \mathbf{c}_i)\}_{i=1}^{P}$ with coordinates $\mathbf{x}_i \in \mathbb{R}^3$ and color/features $\mathbf{c}_i$, we voxelize it into a coarse voxel set $\mathcal{V} = \{(\mathbf{v}_j, \mathbf{f}_j)\}_{j=1}^{V}$, where $\mathbf{v}_j \in \mathbb{Z}^3$ and $\mathbf{f}_j$ is a voxel feature. In interactive segmentation, at step c we are given clicks $\mathcal{K}_c = \{(\mathbf{k}_t, y_t)\}_{t=1}^{c}$ with $y_t \in \{\text{pos}, \text{neg}\}$. We predict voxel logits $\mathbf{s}^{(c)} \in \mathbb{R}^V$ and a binary mask by thresholding $\sigma(\mathbf{s}^{(c)})$.

Click coordinates and voxel indexing. We denote each click location $\mathbf{k}_t$ in *metric* coordinates, i.e., $\mathbf{k}_t \in \mathbb{R}^3$ in the same coordinate system as $\mathbf{x}_i$. For operations on the coarse voxel grid (voxel size v), we quantize $\mathbf{k}_t$ to a voxel coordinate $\tilde{\mathbf{k}}_t = \Pi_v(\mathbf{k}_t) \in \mathbb{Z}^3$ using the same voxelization rule as $\mathcal{V}$. We further denote by $j_t \in \{1, \ldots, V\}$ the index of the coarse voxel whose coordinate matches (or is nearest to) $\tilde{\mathbf{k}}_t$. Unless otherwise specified, we use $\mathbf{k}_t$ for metric coordinates and $(\tilde{\mathbf{k}}_t, j_t)$ for coarse-voxel coordinates/indexing.

3.2 Framework Overview

As illustrated in Fig. 1, *NegROI* consists of three coupled parts: *(i) Backbone & interaction*, which encodes the coarse voxel grid and injects click tokens into a two-way transformer decoder; *(ii) Scene-conditioned negative prompts*, which explicitly model confusing background via K negative prototypes and supervises their attention on boundary hard negatives; and *(iii) ROI refinement & output*, which refines the prediction in a click-centered region on a fine grid and fuses it back to the coarse logits.

3.3 Backbone and Interaction Tokens

Coarse voxel encoding. A voxel encoder $E(\cdot)$ maps voxelized inputs to scene tokens as below:

$$\mathbf{H} = E(\mathcal{V}) \in \mathbb{R}^{V \times F}. \tag{1}$$

Relative positional encoding. We use the quantized first click $\tilde{\mathbf{k}}_1$ as the coordinate origin on the coarse voxel grid. For each voxel j,

$$\Delta\mathbf{v}_j = \mathbf{v}_j - \tilde{\mathbf{k}}_1, \qquad \mathbf{P}_j = W_p\, \phi(\Delta\mathbf{v}_j / S) \in \mathbb{R}^F, \tag{2}$$

where $\phi(\cdot)$ is harmonic encoding. We set S to the per-scene voxel-grid scale, e.g.,

$$S = \max\{(v_{\max} - v_{\min})_x,\, (v_{\max} - v_{\min})_y,\, (v_{\max} - v_{\min})_z\}, \tag{3}$$

with $v_{\min}, v_{\max} \in \mathbb{Z}^3$ being the min/max voxel coordinates of the scene.

Click tokens. We construct interaction tokens consisting of two learnable mask tokens $\mathbf{t}^{fg}, \mathbf{t}^{bg} \in \mathbb{R}^F$ and c click tokens. For click t, we map it to the coarse voxel index j_t (defined in Sec. 3) and form

$$\mathbf{t}_t = \mathbf{P}_{j_t} + \mathbf{e}(y_t), \tag{4}$$

where $\mathbf{e}(\cdot)$ is a learned embedding for click type. The base token sequence is

$$\mathbf{T}^{(c)}_{\text{base}} = \left[\mathbf{t}^{fg}, \mathbf{t}^{bg}, \mathbf{t}_1, \ldots, \mathbf{t}_c\right] \in \mathbb{R}^{(2+c)\times F}. \tag{5}$$

3.4 Two-Way Transformer Decoder

We adopt the two-way transformer decoder from Easy3D [25], which exchanges information between scene tokens and interaction tokens:

$$(\tilde{\mathbf{H}}, \tilde{\mathbf{T}}) = D(\mathbf{H}, \mathbf{P}, \mathbf{T}). \tag{6}$$

Coarse voxel logits are obtained via a foreground/background token difference:

$$s^{(c)}_j = \langle \tilde{\mathbf{t}}^{fg}, \tilde{\mathbf{h}}_j \rangle - \langle \tilde{\mathbf{t}}^{bg}, \tilde{\mathbf{h}}_j \rangle. \tag{7}$$

3.5 Scene-Conditioned Negative Prompts and Supervision

Hard false positives often originate from confusing background structures. To explicitly model such distractors, we introduce K learnable negative prompt queries $\{\mathbf{q}_k\}_{k=1}^{K}$ and compute scene-conditioned prototypes by cross-attending to the scene tokens:

$$\mathbf{p}_k^{(c)} = \mathrm{Attn}(\mathbf{q}_k,\ \mathbf{H}+\mathbf{P},\ \mathbf{H}+\mathbf{P}) \in \mathbb{R}^F, \qquad \hat{\mathbf{p}}_k^{(c)} = \mathrm{MLP}(\mathbf{p}_k^{(c)}) + \mathbf{e}(\mathrm{neg}). \tag{8}$$

We append $\{\hat{\mathbf{p}}_k^{(c)}\}$ to the interaction tokens before decoding:

$$\mathbf{T}^{(c)} = \left[\mathbf{T}_{\mathrm{base}}^{(c)}, \hat{\mathbf{p}}_1^{(c)}, \ldots, \hat{\mathbf{p}}_K^{(c)}\right]. \tag{9}$$

Prompt diversity regularization. To avoid prompt collapse, we encourage diversity among negative prototypes at each click step by penalizing off-diagonal cosine similarity:

$$\mathcal{L}_{\mathrm{div}}^{(c)} = \frac{1}{K(K-1)} \sum_{i \neq j} \left(\cos(\mathbf{p}_i^{(c)}, \mathbf{p}_j^{(c)})\right)^2. \tag{10}$$

where $\{\mathbf{p}_k^{(c)}\}_{k=1}^{K}$ are the scene-conditioned negative prototypes computed at click step c.

Boundary-aware hard negative supervision. Let $A^{(c)} \in \mathbb{R}^{K \times V}$ denote the attention weights produced when generating negative prototypes (averaged across heads), and $\bar{A}^{(c)} = \frac{1}{K}\sum_k A_k^{(c)}$. We supervise $\bar{A}^{(c)}$ to focus on hard background near object boundaries. Specifically, we define boundary-adjacent background candidates

$$\mathcal{B} = \{j \mid y_j = 0,\ \exists j' \in \mathcal{N}(j) \text{ s.t. } y_{j'} = 1\}, \tag{11}$$

where $\mathcal{N}(j)$ is the 6-neighborhood in the voxel grid and y_j denotes the ground-truth voxel label. Among candidates in $\mathcal{B}$ (falling back to all background if $\mathcal{B}$ is too small), we select the top-k voxels with highest predicted foreground probability $p_j = \sigma(s_j^{(c)})$ to form $\mathcal{H}$ and define a uniform target distribution $t_j = \frac{1}{|\mathcal{H}|}\mathbf{1}[j \in \mathcal{H}]$. We minimize cross-entropy:

$$\mathcal{L}_{\mathrm{hn}}^{(c)} = -\sum_{j=1}^{V} t_j \log\left(\bar{A}_j^{(c)} + \epsilon\right). \tag{12}$$

3.6 Click-Centric ROI Refinement and Output

Coarse voxels may blur thin structures and boundaries. We therefore refine locally around the *current click* on a finer grid and fuse the refined logits back to the coarse prediction, as shown in Fig. 1.

ROI selection with adaptive radius. Let the current click center be $\mathbf{k}_c$ (metric coordinates). We select ROI points

$$\mathcal{P}_{roi} = \{i \mid \|\mathbf{x}_i - \mathbf{k}_c\| \le r_c\}, \tag{13}$$

where r_c is an adaptive radius predicted from click context and local density (Sec. 3.6). We voxelize $\mathcal{P}_{roi}$ at a finer resolution $v_f = v/\eta$ (fine scale η) to obtain fine voxels $\mathcal{V}_f$ and fine scene tokens $\mathbf{H}_f \in \mathbb{R}^{V_f \times F}$.

Shared encoder and ROI decoder. We reuse the same encoder E to embed the fine ROI voxels (weight sharing with the coarse branch). We then apply the same Easy3D-style two-way decoder D to the ROI tokens (again weight-shared), producing fine logits $\mathbf{s}_f^{(c)}$. Inside the ROI branch, we optionally build *ROI-conditioned* negative prompts by cross-attending learnable NEG queries to the ROI tokens, and append them to the ROI interaction tokens ("ROI-cond. NEG" in Fig. 1).

Fine-to-coarse max aggregation and fusion. Each fine voxel maps to a coarse voxel by $\lfloor \mathbf{u}/\eta \rfloor$. We aggregate by max:

$$\hat{s}_j^{(c)} = \max_{u \in \mathcal{M}(j)} s_{f,u}^{(c)}, \tag{14}$$

and fuse with the coarse logits using a residual update:

$$s_j^{(c)} \leftarrow (1-\alpha)\, s_j^{(c)} + \alpha\, \hat{s}_j^{(c)}. \tag{15}$$

Adaptive Radius Prediction. We begin with a decayed base radius $r_c^{\text{base}} = r_0\, \gamma^{c-1}$. We estimate local density around the click using kNN distances and compute mean/median (μ, m). We predict a log-scale adjustment

$$\Delta_c = \text{MLP}\big([\mathbf{t}_c, \log \mu, \log m, \log r_c^{\text{base}}, c/C]\big), \qquad r_c = \text{clip}\big(r_c^{\text{base}} \exp(\Delta_c)\big), \tag{16}$$

where clip clamps to $[r_{\min}, r_{\max}]$. If the ROI contains too few points, we enlarge r_c (bounded) a few times to ensure stable refinement.

Uncertainty-Driven Selective Refinement. We prioritize ambiguous regions inside the ROI using coarse uncertainty:

$$u_j = 1 - 2\left|\sigma(s_j^{(c)}) - 0.5\right| \in [0, 1]. \tag{17}$$

Each ROI point is mapped to its coarse voxel index $v(i)$; we keep points with $u_{v(i)} \ge \tau$ (fallback to the unfiltered ROI if too few points remain), focusing fine decoding on boundary and hard regions.

3.7 Training Objective

At click step c, we use the segmentation loss

$$\mathcal{L}_{\text{seg}}^{(c)} = \mathcal{L}_{\text{bce}}(\mathbf{s}^{(c)}, \mathbf{y}) + \mathcal{L}_{\text{dice}}(\mathbf{s}^{(c)}, \mathbf{y}), \tag{18}$$

and add scene-conditioned negative prompts regularizers:

$$\mathcal{L} = \frac{1}{C} \sum_{c=1}^{C} \left(\mathcal{L}_{\text{seg}}^{(c)} + \lambda_{\text{hn}} \mathcal{L}_{\text{hn}}^{(c)} + \lambda_{\text{div}} \mathcal{L}_{\text{div}}^{(c)} \right). \tag{19}$$

Click simulation. During training, we simulate interactions using an oracle policy that places the next click at informative error regions between prediction and ground truth.

Implementation note (DDP stability). In our implementation, adaptive radii affect ROI selection via a hard threshold. To avoid unused-parameter issues in DDP, we optionally include a tiny regularizer on Δ_c with detached inputs:

$$\mathcal{L}_{\text{rad}} = \lambda_{\text{rad}} \|\Delta_c\|_2^2. \tag{20}$$

4 Experiments

In this section, we conducted extensive experiments on several common benchmarks to empirically evaluate the advantages of our method by comparing with the leading interactive and non-interactive baselines. We also did ablation studies to show the effectiveness of our every modules.

4.1 Benchmarks

We evaluate interactive 3D instance segmentation on three widely-used benchmarks: **ScanNet40** [4], **S3DIS** [1], and **KITTI-360** [14]. ScanNet40 and S3DIS represent cluttered indoor environments with frequent object attachments and ambiguous boundaries, while KITTI-360 features large-scale outdoor scans with sparse geometry and long-range context. Unless otherwise stated, we follow the official dataset splits and evaluation protocols adopted in prior interactive 3D segmentation works (e.g., [12,34]).

4.2 Baselines

We compare against representative interactive 3D segmentation baselines, including click-conditioned transformer decoders without ROI refinement, multi-resolution refinement methods without explicit background prompting, and prompt/query-based interactive approaches (e.g., [12,34,37]). When official implementations are available, we run them under the same evaluation protocol. Otherwise, we follow the paper descriptions and ensure identical click simulation, data preprocessing, and evaluation metrics. We also report Easy3D [25] as a strong promptable interactive baseline in our setting.

4.3 Interactive Evaluation Protocol

Click simulation. To ensure fair and repeatable evaluation, we adopt a standard simulated-click protocol. For each target instance, the first click is sampled inside the ground-truth region. At each subsequent step, we update the mask prediction and place the next click on the largest error region: a positive click on the largest false-negative component if under-segmentation dominates, otherwise a negative click on the largest false-positive component. This policy approximates an oracle user that always clicks the most informative mistake and is commonly used for benchmarking interactive methods.

Target specification (prompts). Our primary setting is class-agnostic interactive instance segmentation, where the target instance is specified by clicks only. If a baseline requires an additional query signal (e.g., an open-vocabulary text prompt), we use simple class-name prompts with a consistent template across methods.

Metrics. We report point-wise IoU (%) under click budgets $k \in \{1, 2, 3, 5, 10\}$, denoted as IoU@k. All methods are evaluated with the same click simulator and the same preprocessing on each dataset to isolate algorithmic differences.

4.4 Implementation Details

Backbone and voxelization. Our model uses a sparse voxel encoder and a two-way transformer decoder. We follow dataset-standard sparse voxel preprocessing and keep voxelization consistent between training and testing. To improve boundary accuracy without incurring dense high-resolution costs, our refinement branch re-voxelizes only within selected ROIs using a finer resolution than the coarse grid.

Scene-conditioned negative prompts and click-centric ROI refinement. NegROI combines *scene-conditioned negative prompts* with *click-centric ROI refinement* (Fig. 1). At each click step, we first decode coarse mask logits using click tokens augmented by K scene-conditioned negative prototypes. We then refine boundaries with a click-centric ROI branch: we select an ROI centered at the current click using an adaptive radius r_c and a coarse uncertainty signal, re-voxelize the ROI on a finer grid, and run the *shared encoder* E together with an *ROI decoder* to produce fine logits. Finally, we *max-aggregate* fine logits to the coarse grid and *fuse* them back with weight α. For prompt learning, we supervise negative-prompt attention via a boundary-aware hard-negative loss $\mathcal{L}_{hn}$ and regularize prompt collapse with the diversity loss $\mathcal{L}_{div}$. We tune these hyperparameters on a ScanNet40 validation split and keep them fixed across datasets.

Hyperparameters. Unless otherwise stated, we use K=8 negative prompts, fine scale η=2 (i.e., v_f=v/η), uncertainty threshold τ=0.20, and fusion weight α=0.7. For the click-centric ROI, we set the base radius r_0=1.0m with decay γ=0.9,

and clamp the adaptive radius to $[r_{\min}, r_{\max}]$=[0.35r_0, 3.0r_0]. We estimate local density with kNN using k=32, and mine hard negatives with top-k background voxels using k=256. All these hyperparameters are tuned once on a ScanNet40 validation split and then fixed for all datasets.

Training. We train end-to-end using the segmentation loss on the fused logits and auxiliary objectives: $\mathcal{L}_{seg} = \mathcal{L}_{bce} + \mathcal{L}_{dice}$, plus the prompt regularizers $\lambda_{hn}\mathcal{L}_{hn}$ and $\lambda_{div}\mathcal{L}_{div}$. We simulate up to C clicks per query during training using the oracle policy described above.

4.5 Main Results

Table 1 reports the main comparison under a unified *train-on-ScanNet40* setting. We train all methods on ScanNet40 and evaluate on three test benchmarks: **ScanNet40** (in-domain) and **S3DIS**/**KITTI-360** (out-of-domain). This protocol directly measures cross-dataset robustness under substantial domain shifts in scene scale, point density, and sensor characteristics.

In-domain performance. On ScanNet40, NegROI consistently improves IoU across click budgets, with the largest gains under sparse interactions (IoU@1–IoU@3). These improvements align with our design goals: scene-conditioned negative prompts reduce hard distractors and attached clutter early, while the click-centric fine-grid ROI branch sharpens boundaries through re-voxelization and ROI decoding, followed by max-aggregation and fusion back to coarse logits.

Out-of-domain generalization. When transferring to S3DIS and KITTI-360 without any target-domain training, NegROI maintains strong performance and outperforms prior baselines across most click budgets. The gains are particularly pronounced at low click counts, where boundary ambiguity and hard distractors dominate. This behavior is consistent with (i) boundary-aware hard-negative prompt supervision ($\mathcal{L}_{hn}$), which suppresses boundary-proximal false positives, and (ii) click-centric ROI refinement, which allocates fine-grained computation to the most informative region indicated by the click.

4.6 Comparison with Non-Interactive Methods

Although NegROI is designed for interactive refinement, we further evaluate its *single-click* performance under the ScanNet20 protocol commonly used for assessing generalization to unseen categories. Specifically, we train models *only* on ScanNet20 and test in two settings: ScanNet20 → ScanNet20 (only seen objects) and ScanNet20 → ScanNet40 (seen + unseen objects). Table 2 compares NegROI with representative non-interactive/open-vocabulary 3D instance segmentation baselines, including Mask3D [24], AGILE3D [34], and Easy3D [25]. NegROI improves mAP as well as AP_{50}/AP_{25} in both the in-domain (seen-only) and the seen + unseen transfer setting, indicating that our scene-conditioned negative prompting and click-centric refinement also benefit early (single-click)

Table 1: Main results under a unified train-on-ScanNet40 protocol. All methods are trained on ScanNet40 and tested on ScanNet40 (in-domain), S3DIS and KITTI-360 (out-of-domain). Numbers denote point-wise IoU (%) under click budgets $k \in \{1, 2, 3, 5, 10\}$ (IoU@k). Best results are in bold.

Dataset	Method	IoU@1	IoU@2	IoU@3	IoU@5	IoU@10
ScanNet40	InterObject3D [12]	40.8	55.9	63.9	67.6	77.6
	AGILE3D [34]	63.0	70.6	75.1	79.7	83.5
	Easy3D [25]	68.2	74.6	77.3	79.6	81.7
	NegROI (Ours)	**72.1**	**78.1**	**80.9**	**82.3**	**83.6**
S3DIS	InterObject3D [12]	38.5	54.0	62.5	72.4	79.9
	AGILE3D [34]	58.5	70.7	77.4	83.6	**88.3**
	Point-SAM [37]	38.8	n/a	67.1	72.2	80.6
	Easy3D [25]	65.7	76.0	80.8	84.9	87.8
	NegROI (Ours)	**66.9**	**77.8**	**82.2**	**85.4**	87.8
KITTI-360	InterObject3D [12]	2.0	5.1	8.5	72.4	83.6
	AGILE3D [34]	34.8	40.7	42.7	44.4	49.6
	Point-SAM [37]	44.0	n/a	67.1	72.2	80.8
	Easy3D [25]	46.3	58.7	66.7	76.2	83.6
	NegROI (Ours)	**47.5**	**62.6**	**71.4**	**80.1**	**87.7**

predictions and enhance robustness beyond the ScanNet40-trained interactive evaluation in Table 1.

4.7 Ablation Studies

We conduct ablations to quantify the contribution of each component and the proposed regularizers.

Component ablation. We progressively enable (i) scene-conditioned negative prompts, (ii) click-centric fine-grid ROI refinement, (iii) uncertainty-driven ROI selection, (iv) boundary-aware hard negative supervision for prompt attention ($\mathcal{L}_{hn}$), and (v) prompt diversity regularization ($\mathcal{L}_{div}$). Table 3 summarizes the results on ScanNet40 and S3DIS.

Scene-conditioned negative prompts boost performance under sparse clicks (e.g., IoU@1: 67.0→69.2 on ScanNet40 and 62.5→64.2 on S3DIS). Adding click-centric fine-grid ROI refinement further improves boundary quality (e.g., ScanNet40 IoU@5: 80.4→81.6). Uncertainty-driven ROI selection brings additional

Table 2: Single-click evaluation for models trained only on ScanNet20. We report results under the ScanNet20 → ScanNet20 setting (only seen objects) and the ScanNet20 → ScanNet40 setting (seen + unseen objects).

Setting	Method	mAP	AP_{50}	AP_{25}
ScanNet20 → ScanNet20	Mask3D [24]	51.5	77.0	90.2
	AGILE3D [34]	53.5	75.6	91.3
	Easy3D [25]	56.1	79.5	93.1
	NegROI (Ours)	**59.6**	**82.3**	**95.2**
ScanNet20 → ScanNet40	Mask3D [24]	5.3	13.1	24.7
	AGILE3D [34]	24.8	45.7	72.4
	Easy3D [25]	39.2	64.6	85.5
	NegROI (Ours)	**43.0**	**67.9**	**87.8**

gains (e.g., ScanNet40 IoU@1: 70.9→71.8). Overall, combining the proposed regularizers yields the best full model on both datasets (ScanNet40: 72.1/82.3/83.6; S3DIS: 66.9/85.4/87.8 for IoU@1/5/10).

Loss ablation. We further examine the effect of the proposed prompt regularizers. As shown in Table 3, adding boundary-aware hard negatives ($\mathcal{L}_{hn}$) and prompt diversity ($\mathcal{L}_{div}$) improves the overall results, with the full model achieving the best performance across datasets and click budgets. In particular, $\mathcal{L}_{div}$ provides consistent gains when combined with $\mathcal{L}_{hn}$, suggesting that preventing prompt collapse helps stabilize negative guidance during refinement.

4.8 Qualitative Results

We visualize representative cases across datasets in Fig. 2 to illustrate how scene-conditioned negative prompts and click-centric ROI refinement improve interactive segmentation. Compared to Easy3D [25], **NegROI** better suppresses attached or visually similar distractors and recovers sharper object boundaries under sparse clicks (e.g., 1–3 clicks), consistently yielding higher IoU@k.

Boundary-proximal false positives. Figure 3 visualizes *boundary-band false positives* (Band-FP): Gray denotes a narrow ground-truth boundary band (darker is closer to the boundary) and red marks false-positive predictions inside the band. Across ScanNet40, S3DIS, and KITTI-360, NegROI consistently shows fewer boundary-proximal false positives than Easy3D [25], and the red points diminish faster from Click 1 to Click 3. This trend is consistent with our design: the boundary-aware hard negative loss explicitly guides the negative prompts

Table 3: Component ablation on ScanNet40 and S3DIS. Numbers denote point-wise IoU (%) under click budgets $k \in \{1, 5, 10\}$.

Variant	ScanNet40			S3DIS		
	IoU@1	IoU@5	IoU@10	IoU@1	IoU@5	IoU@10
Base (no ROI, no NEG prompts)	67.0	78.8	80.8	62.5	80.8	83.8
+ Scene-cond. NEG prompts	69.2	80.4	82.3	64.2	83.0	85.7
+ ROI refinement (fine-grid)	70.9	81.6	83.0	65.8	84.1	86.9
+ Uncertainty-driven ROI selection	71.8	81.9	82.8	66.5	85.0	87.4
+ Boundary hard negatives ($\mathcal{L}_{hn}$)	71.2	82.1	83.1	65.6	85.2	87.1
+ Diversity ($\mathcal{L}_{div}$) (full)	**72.1**	**82.3**	**83.6**	**66.9**	**85.4**	**87.8**

toward boundary-adjacent confusers (Sec. 3.5), while click-centric ROI refinement sharpens local decision boundaries on a finer grid and fuses them back to the coarse mask (Sec. 3.6).

5 Conclusions

We presented a robust interactive 3D segmentation framework that couples click-centric multi-resolution ROI refinement with scene-conditioned negative prompts. NegROI refines boundaries efficiently near the current click, focuses computation on uncertain regions, and explicitly suppresses hard false positives via boundary-aware negative prompt supervision and diversity regularization. Experiments on ScanNet, S3DIS, and KITTI-style point clouds indicate improved click efficiency and stronger cross-dataset robustness. Future work includes fully differentiable ROI selection and stronger test-time adaptation driven solely by user clicks.

Acknowledgments

We thank all reviewers and area chairs for their insightful comments and valuable suggestions. We also thank Yueyang Wen for the fruitful discussions and sharing source codes. This work was supported in part by Major Research Plan of the National Natural Science Foundation of China (92048301), Anhui Provincial Major Research and Development Plan (202004H07020008), and Anhui Provincial Natural Science Foundation (2208085MF172).

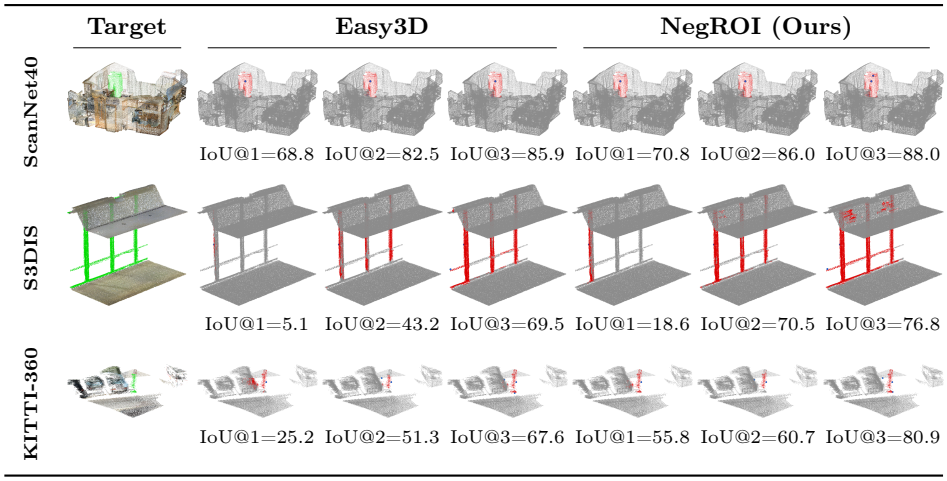

Fig. 2: Qualitative comparison across datasets. We compare Easy3D [25] and *NegROI* under the same simulated-click protocol on ScanNet40, S3DIS, and KITTI-360. Each row shows the target instance (green) and predicted masks (red) after 1/2/3 clicks, with the corresponding IoU@1/2/3 reported below each prediction.

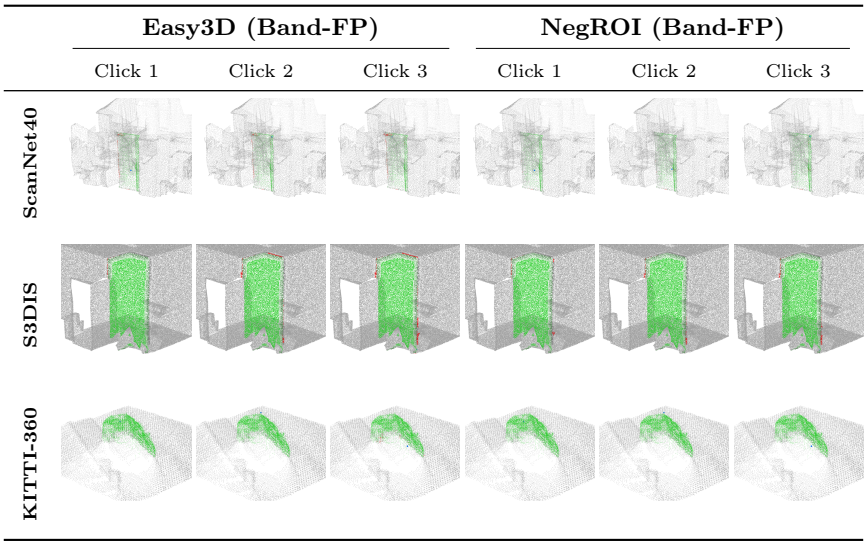

Fig. 3: Boundary-band false positive (Band-FP) visualization. Gray indicates the ground-truth boundary band (darker near the boundary); red highlights false-positive predictions within the band. Compared to Easy3D [25], **NegROI** produces fewer boundary-proximal false positives under the same simulated clicks across ScanNet40, S3DIS, and KITTI-360.

References

1. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3d semantic parsing of large-scale indoor spaces. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 1534–1543 (2016)
2. Chen, X., Zhao, Z., Zhang, Y., Duan, M., Qi, D., Zhao, H.: Focalclick: Towards practical interactive image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 1300–1309 (2022)
3. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 3075–3084 (2019)
4. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 5828–5839 (2017)
5. Graham, B., Engelcke, M., Van Der Maaten, L.: 3d semantic segmentation with submanifold sparse convolutional networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 9224–9232 (2018)
6. Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A.: Randla-net: Efficient semantic segmentation of large-scale point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 11108–11117 (2020)
7. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: Proceedings of the international conference on machine learning (ICML). pp. 4904–4916 (2021)
8. Jiang, L., Zhao, H., Shi, S., Liu, S., Fu, C.W., Jia, J.: Pointgroup: Dual-set point grouping for 3d instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition (CVPR). pp. 4867–4876 (2020)
9. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV). pp. 4015–4026 (2023)
10. Kirillov, A., Wu, Y., He, K., Girshick, R.: Pointrend: Image segmentation as rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 9799–9808 (2020)
11. Kolodiazhnyi, M., Vorontsova, A., Konushin, A., Rukhovich, D.: Oneformer3d: One transformer for unified point cloud segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20943–20953 (2024)
12. Kontogianni, T., Celikkan, E., Tang, S., Schindler, K.: Interactive object segmentation in 3d point clouds. In: Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 2891–2897 (2023)
13. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 7061–7070 (2023)
14. Liao, Y., Xie, J., Geiger, A.: Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 3292–3310 (2022)

15. Liu, Q., Xu, Z., Bertasius, G., Niethammer, M.: Simpleclick: Interactive image segmentation with simple vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22290–22300 (2023)
16. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: Proceedings of the European conference on computer vision (ECCV). pp. 38–55 (2024)
17. Maninis, K.K., Caelles, S., Pont-Tuset, J., Van Gool, L.: Deep extreme cut: From extreme points to object segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 616–625 (2018)
18. Nguyen, P., Ngo, T.D., Kalogerakis, E., Gan, C., Tran, A., Pham, C., Nguyen, K.: Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 4018–4028 (2024)
19. Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T., et al.: Openscene: 3d scene understanding with open vocabularies. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 815–824 (2023)
20. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 652–660 (2017)
21. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems (NeurIPS) **30** (2017)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the international conference on machine learning (ICML). pp. 8748–8763 (2021)
23. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. In: Proceedings of the International Conference on Learning Representations (ICLR). vol. 2025, pp. 28085–28128 (2025)
24. Schult, J., Engelmann, F., Hermans, A., Litany, O., Tang, S., Leibe, B.: Mask3d: Mask transformer for 3d semantic instance segmentation. In: Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 8216–8223. IEEE (2023)
25. Simonelli, A., Müller, N., Kontschieder, P.: Easy3d: A simple yet effective method for 3d interactive segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 24707–24716 (2025)
26. Sofiiuk, K., Petrov, I., Barinova, O., Konushin, A.: f-brs: Rethinking backpropagating refinement for interactive segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 8623–8632 (2020)
27. Sofiiuk, K., Petrov, I.A., Konushin, A.: Reviving iterative training with mask guidance for interactive segmentation. In: Proceedings of the 2022 IEEE international conference on image processing (ICIP). pp. 3141–3145. IEEE (2022)
28. Takmaz, A., Fedele, E., Sumner, R.W., Pollefeys, M., Tombari, F., Engelmann, F.: Openmask3d: Open-vocabulary 3d instance segmentation. arXiv preprint arXiv:2306.13631 (2023)

29. Tang, H., Liu, Z., Zhao, S., Lin, Y., Lin, J., Wang, H., Han, S.: Searching efficient 3d architectures with sparse point-voxel convolution. In: Proceedings of the European conference on computer vision (ECCV). pp. 685–702 (2020)
30. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: Kpconv: Flexible and deformable convolution for point clouds. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV). pp. 6411–6420 (2019)
31. Vu, T., Kim, K., Luu, T.M., Nguyen, T., Yoo, C.D.: Softgroup for 3d instance segmentation on point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 2708–2717 (2022)
32. Wen, Y., Hou, Y., Zhang, S., Wu, F.: Clickenhance: Efficient 3d interactive segmentation with click-specific encoder and contrastive learning. IEEE Transactions on Multimedia (IEEE TMM) pp. 1–12 (2026)
33. Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 18134–18144 (2022)
34. Yue, Y., Mahadevan, S., Schult, J., Engelmann, F., Leibe, B., Schindler, K., Kontogianni, T.: Agile3d: Attention guided interactive multi-object 3d segmentation. In: Proceedings of the international conference on learning representations (ICLR). vol. 2024, pp. 11391–11411 (2024)
35. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV). pp. 16259–16268 (2021)
36. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: Proceedings of the European conference on computer vision (ECCV). pp. 350–368 (2022)
37. Zhou, Y., Gu, J., Chiang, T.Y., Xiang, F., Su, H.: Point-sam: Promptable 3d segmentation model for point clouds. arXiv preprint arXiv:2406.17741 (2024)