Prompting Diffusion Models for Zero-Shot Instance Segmentation

İrem Zeynep Alagöz[1*] **Nils Morbitzer**[1,2*] **Andrea Ramazzina**[1,3]
Nassir Navab[1,2] **Federico Tombari**[1,2] **Stefano Gasperini**[1,2,4]

[1]Technical University of Munich (TUM) [2]Munich Center for Machine Learning (MCML)
[3]Mercedes-Benz AG [4]Visualais

Project Page: `https://prompt2seg.github.io`

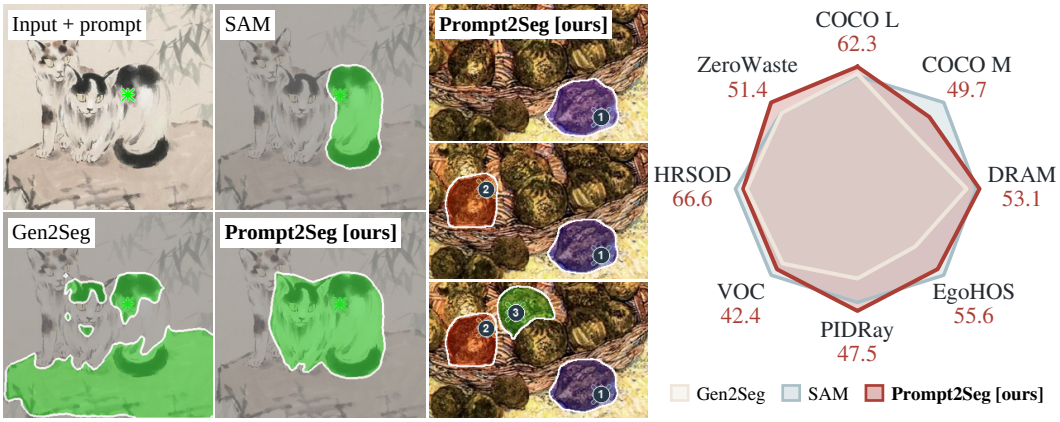


Figure 1: **Prompt2Seg** conditions diffusion on spatial prompts for semantics-aware segmentation. *Left:* Discriminative models [1] and prior work leveraging diffusion priors [2] can miss parts of an object that differ in appearance or blend into the background. *Middle:* Further clicks can select other instances interactively (or be used for refining the mask). *Right:* mIoU comparison on 7 benchmarks.

Abstract

Interactive instance segmentation enables users to identify and iteratively refine object masks, offering fine-grained control. Discriminative foundation models achieve strong zero-shot generalization across diverse natural-image domains at the cost of large-scale, task-specific mask annotations. On the other hand, early generative attempts demonstrate annotation-efficient fine-tuning but use prompts only during post-processing, yielding suboptimal segments. In this paper, we address this limitation with Prompt2Seg, a spatial prompt-conditioning branch that guides a frozen diffusion-based segmentation model. Our approach is prompted with 2D Gaussians as explicit input signals, and trains the conditioning branch to respond directly to user intent. Fine-tuned on a deliberately constrained set of synthetic data from Hypersim and Virtual KITTI 2, Prompt2Seg generalizes zero-shot to a wide range of real data, with diverse visual domains and unseen object types. Extensive experiments on seven datasets, including paintings, egocentric views, and X-ray data, show Prompt2Seg consistently outperforming the underlying diffusion segmentation backbone. Additionally, we show how Prompt2Seg can be prompted with uncertainty maps to identify unknown object instances.

*Equal contribution. Correspondence to: İrem Zeynep Alagöz (iremzeynep.alagoez@tum.de).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

In recent years, scene understanding has shifted from models trained and evaluated on fixed, domain-specific datasets [3, 4, 5] toward foundation models [6, 7, 1, 8, 9], which generalize to contexts never explicitly seen during training. In segmentation, discriminative models, such as SAM [1] and its successors [8, 9], have demonstrated strong zero-shot performance, enabling users to specify target instances through prompts such as clicks or bounding boxes.

This generalization comes at a high cost. For example, SAM was trained on SA-1B, containing approximately 11M images with over 1B masks [1]. Collecting annotations at this scale is highly resource-intensive. Moreover, large-scale training does not guarantee uniformly strong generalization across all image inputs and domains, such as X-ray security imagery (PIDRay [10]) or cluttered waste scenes (ZeroWaste [11]), where annotation data is scarce and expensive. This motivates seeking a way to incorporate diverse knowledge into the model while circumventing the need for extensive, dense, task-specific mask annotations.

Diffusion models offer such a source of knowledge. Trained to recover images from noisy observations, models such as Stable Diffusion [12], learn rich priors over visual appearance and structure without dense annotations. Despite being originally trained for image synthesis, their strong internal representations have been effectively repurposed for several downstream tasks, including estimating geometric attributes [13, 14, 15], pose [16, 17], or motion [18]. Since generating a realistic image requires understanding and modeling objects and scenes, these priors are particularly suitable for segmentation.

Recently, diffusion models have been adapted to segmentation, too, but existing methods offer limited control over which instance is segmented. Some are conditioned only on language [19, 20], which cannot distinguish between multiple instances of the same class. Gen2Seg [2] supports user clicks as prompts, but only after the image features have already been extracted. It computes the features independently of the prompt, and the click only selects among them. Thus, the model and the diffusion process have no knowledge of the user's intent while building its representation, missing the opportunity to leverage the highly informative user prompt. As shown in Figure 1, this can lead to masks that are incomplete or bleed into neighboring objects.

In this paper, we address this limitation with Prompt2Seg, a diffusion-based interactive instance segmentation model in which the user prompt guides the mask generation from the start. Given an image and a spatial prompt that identifies a target instance, Prompt2Seg uses them to condition the denoising process directly. By doing so, the generated mask highly depends on the user's intent rather than being extracted from features that are agnostic to the user's prompt. We achieve this by freezing the diffusion-based segmentation backbone and training the conditioning branch. Notably, Prompt2Seg is trained solely on synthetic images with synthetic mask annotations, removing the need to collect vast amounts of diverse, manually annotated data, yet it transfers well zero-shot to a wide range of real images. Moreover, we show how the prompt can be a user's click expanded into a 2D Gaussian or a dense confidence map, such as the uncertainty map from another model. To the best of our knowledge, Prompt2Seg is the first diffusion model explicitly trained for click-based interactive instance segmentation.

Our contributions can be summarized as follows:

- We present Prompt2Seg, a prompting mechanism for diffusion-based interactive instance segmentation using spatial prompts, represented as 2D Gaussians.
- We show that our conditioning strategy yields a strong zero-shot interactive segmentation model, especially on challenging domains, including artistic images (DRAM), X-ray data (PIDRay), and cluttered scenes (ZeroWaste), despite being trained only on synthetic data.
- We show how the same approach can be extended to extracting instances of unknown objects from the uncertainty maps of another model.

Our goal is not to replace SAM, but to show that diffusion priors offer a competitive, annotation-efficient alternative in domains where discriminative foundation models struggle.

2 Related Work

2.1 Generative Models for Perceptual Tasks

Generative modeling has increasingly influenced visual perception, shifting from auxiliary training objectives toward direct inference mechanisms. Early works [21, 22] show that generative objectives, i.e. autoregressive prediction or masked reconstruction, are effective for learning transferable visual representations from unlabeled data. Subsequent works demonstrated that large-scale text-conditional generation models encode rich, reusable perceptual knowledge as a byproduct of learning to synthesize images from language: features extracted from generatively pretrained models transfer surprisingly well to discriminative tasks, often rivaling representations learned with explicit supervision [23, 24].

More recent approaches go a step further, repurposing generative models as direct inference engines for dense prediction; rather than extracting features, they generate the output itself through diffusion or autoregressive decoding. This paradigm has proven broadly applicable, spanning geometric estimation [13, 14, 15], pose [16, 17] motion [18], and, most recently, semantic tasks such as segmentation [2, 19]. Our work follows this direction, investigating whether the conditioning machinery of diffusion models can be repurposed for promptable instance segmentation, combining the inference-time flexibility of generative decoding with the broad visual understanding acquired during large-scale pretraining.

2.2 Diffusion Models for Segmentation

Diffusion models [25] have become the de-facto standard for high-quality image synthesis, achieving impressive performance in text-to-image synthesis [12] and controlled generation [26]. Beyond generation, the internal features of denoising U-Nets have proven surprisingly useful for perception. Baranchuk et al. [23] show that DDPM activations capture semantic information, thereby, enabling few-shot segmentation without any task-specific training. Since then, diffusion models have been explored as visual backbones for a range of perception tasks, with segmentation being among the more recent directions to gain traction [2, 19, 20]. Existing approaches can be grouped in four categories: (1) Utilizing diffusion model's strong frozen features for discriminative models [24, 27]; (2) generating pseudo labels to supervise discriminative models [28, 29, 30]; (3) inversion-based approaches that repurpose internal attention maps for zero-shot object localization [31, 32, 33], (4) or directly synthesizing segmentation masks in label [20] or RGB space [2].

Gen2Seg [2] shows that diffusion models trained on billions of images serve as strong priors for segmentation, enabling remarkable zero-shot generalization and label-efficient fine-tuning. However, they treat the model as a passive predictor, yet diffusion models are by nature interactive trained to respond to rich multimodal conditioning. We show that their conditioning mechanism can be repurposed for user-guided instance segmentation. This yields a promptable segmentation framework that inherits the strong generalization of generative pretraining.

2.3 Promptable Instance Segmentation

Interactive segmentation has been extensively studied through discriminative approaches, where models directly predict segmentation masks given user-provided prompts. Click-based methods such as SimpleClick [34] leverage strong vision encoders to produce accurate masks from sparse point inputs, while the Segment Anything Model (SAM) [1] and its successor SAM 2 [8] scale this paradigm to large-scale data, achieving class-agnostic segmentation from clicks, boxes, or masks. To incorporate semantic understanding, Grounded SAM [35] and SAM 3 [9] couple open-vocabulary detectors with SAM, enabling language-guided segmentation. Despite their strong performance, discriminative models require task-specific architectures and large amounts of annotated data to learn each new prompt modality. Generative models, by contrast, intrinsically develop rich visual representations during pretraining that can be unlocked with only lightweight adaptation [36], suggesting that a generative backbone is inherently better suited for few-shot generalization to new prompting conditions. Recent work has begun to exploit this property for segmentation: GS [20] reformulates language-driven segmentation as a label diffusion process, and LlamaSeg [19] treats it as autoregressive mask token generation conditioned on language instructions. Most closely related to our work, Gen2Seg [2] finetunes Stable Diffusion [12] for category-agnostic instance segmentation by casting it as an image-to-image translation problem, where instances are distinguished by color.

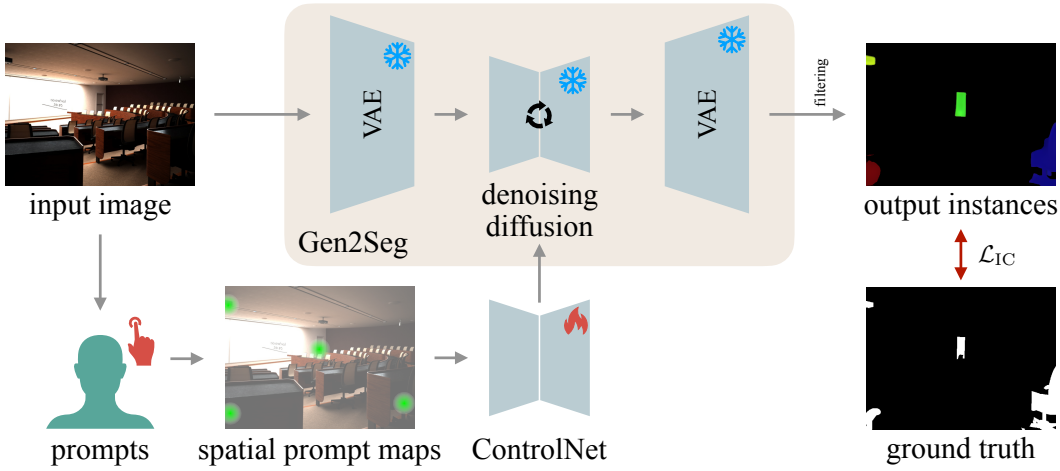


Figure 2: **Our Prompt2Seg framework**. Prompt2Seg segments any instance by guiding the diffusion process with a ControlNet conditioned on the user prompts as Gaussian maps.

However, its prompting mechanism reduces to post-hoc color-based instance retrieval; that is the model is never trained to incorporate spatial user prompts, which we show leads to consistently inferior performance across all zero-shot benchmarks (see Table 1). More broadly, existing generative approaches either rely exclusively on language conditioning or, like Gen2Seg, treat spatial interaction as an auxiliary mechanism. In contrast, our approach explicitly trains the model to condition on spatial prompts; thereby, integrating interactivity as a core component of the generative process and leveraging generative pretraining's few-shot adaptability for interactive instance segmentation.

3 Method

We introduce Prompt2Seg, a spatially prompted diffusion model for category-agnostic interactive instance segmentation. Given an image and a sparse spatial prompt, Prompt2Seg predicts the mask of the prompted object. The central idea is to preserve the visual priors of a pretrained diffusion model while adding an explicit spatial control interface to guide the segmentation. To this end, we train only a lightweight conditioning branch that injects prompt-dependent residuals into the frozen backbone.

Let $x \in \mathbb{R}^{3\times H\times W}$ be an input image and let $c \in \mathbb{R}^{C\times H\times W}$ be a spatial condition derived from the sparse user prompts. Prompt2Seg predicts an RGB instance-coloring map

$$y = f_{\theta,\phi}(x, c), \qquad y \in \mathbb{R}^{3\times H\times W}, \tag{1}$$

where θ denotes the frozen diffusion segmentation backbone and ϕ denotes the trainable prompt-conditioning branch.

3.1 Pretrained Diffusion Segmentation Backbone

We build on Gen2Seg [2], which repurposes a Stable-Diffusion-initialized image-to-image model for instance segmentation. Gen2Seg predicts an RGB instance-coloring map for the entire image: pixels belonging to the same region are encouraged to share a color, different regions are encouraged to be separated in color space, and the background is mapped to black. A target instance is then recovered only after prediction by post-processing this full-image representation, e.g., by thresholding distances in the predicted color space. This can introduce ambiguities when nearby instances receive similar colors. In contrast, Prompt2Seg keeps the same RGB output representation but changes its semantics: the prompt selects the target instance before decoding, so the model directly predicts a single prompt-conditioned foreground region against a black background. Thus, Prompt2Seg avoids full-image instance selection at post-processing time while leaving the pretrained diffusion segmentation backbone and output loss unchanged.

3.2 Spatial Prompt Representation

A point prompt is a pixel coordinate within the pixels of the target object. We transform this point into an image-like continuous spatial map that can be consumed by a convolutional conditioning branch. Let $\mathcal{P}^+ = \{r_i\}_{i=1}^P$ be the set of positive prompts, where $r_i = (u_i, v_i)$ is an image coordinate inside the target object. For a pixel coordinate q, we define a Gaussian prompt map

$$G_\sigma(q; r_i) = \exp\left(-\frac{\|q - r_i\|_2^2}{2\sigma^2}\right), \tag{2}$$

where σ controls the spatial extent of the prompt. Multiple positive prompts are aggregated by a pixel-wise maximum,

$$h_\sigma^+(q) = \max_{r_i \in \mathcal{P}^+} G_\sigma(q; r_i). \tag{3}$$

This aggregation keeps the condition's dimensionality fixed while allowing multiple clicks.

Our main model uses a two-scale Gaussian prompt encoding. We construct

$$h_s^+(q) = h_{\sigma_s}^+(q), \qquad h_l^+(q) = h_{\sigma_l}^+(q), \qquad \sigma_s < \sigma_l. \tag{4}$$

Given the effectiveness of our dual-scale design choice shown in Table 8, we hypothesize that the two scales provide complementary cues by emphasizing the click location at the narrower scale and its surrounding neighborhood at the broader scale. The smaller bandwidth σ_s preserves precise localization at the click, whereas the larger bandwidth σ_l increases the spatial support of the prompt signal. The final spatial condition is a three-channel tensor

$$c = \Psi(h_s^+, h_s^+, h_l^+) \in \mathbb{R}^{3 \times H \times W}, \tag{5}$$

where Ψ maps the two-scale prompt representation to the channel layout expected by the conditioning branch. The prompt is provided as an explicit spatial signal before prediction, rather than being used as a post-processing query on an already generated instance map, as in Gen2Seg. By doing so, our model learns to directly focus on generating of the prompted object segmentation.

While the description above focuses on a 2D Gaussian representation, our prompt can differ from the sparse clicks paradigm. This is useful when the region of interest is produced by another model (e.g., a dense confidence or uncertainty map) rather than by a user.

3.3 Prompt Injection with a ControlNet Branch

We inject spatial prompts through a ControlNet-style branch [26]. Since Gen2Seg uses a single-step formulation, all predictions are evaluated at a fixed timestep $t^\star$. Let z_x denote the VAE latent of the input image, h the encoder hidden states, and c the spatial conditioning input. The trainable ControlNet branch predicts down-block and mid-block residuals,

$$\left(\Delta^{(1)}, \ldots, \Delta^{(L)}, \Delta^{\text{mid}}\right) = g_\phi(z_x, t^\star, h, c). \tag{6}$$

These residuals are injected into the frozen Gen2Seg U-Net via its additional residual inputs:

$$\hat{u} = f_\theta\left(z_x, t^\star, h; \Delta^{(1)}, \ldots, \Delta^{(L)}, \Delta^{\text{mid}}\right), \tag{7}$$

where $\hat{u}$ is the diffusion-model prediction. The Gen2Seg scheduler maps this prediction to a segmentation latent,

$$\hat{z}_y = \mathcal{S}_{t^\star}(z_x, \hat{u}), \qquad \hat{y} = D_{\text{VAE}}(\hat{z}_y), \tag{8}$$

where $\hat{y}$ is the decoded RGB instance-coloring map.

During training, the VAE, text encoder, and denoising U-Net remain frozen, while the ControlNet parameters ϕ are optimized.

3.4 Prompt-Conditioned Supervision

We inherit the instance-coloring supervision of Gen2Seg. For a set of supervised regions $\mathcal{S}$, the Gen2Seg objective encourages pixels within each region to share a color, encourages different regions to be separated in color space, and fixes the background mean to black. We use the original objective

$$\mathcal{L}_{\text{IC}} = \mathcal{L}_{\text{var}} + \lambda_{\text{sep}}\mathcal{L}_{\text{sep}} + \lambda_{\text{mean}}\mathcal{L}_{\text{mean}}, \tag{9}$$

with the same definitions and hyperparameters as Gen2Seg. We include the full loss definition in the appendix for completeness.

Instead of coloring all instances in an image [2], Prompt2Seg is trained to predict only the instance specified by the spatial prompt. For each training image, we sample one or more target instances, construct the spatial condition from prompts inside those instances, and keep only the selected instances as foreground in the target instance-coloring map. All non-target pixels are assigned to the black background. Thus, the supervision matches the interactive task: the correct output is not an all-instance segmentation, but the mask of the prompted target.

This prompt-conditioned target construction is category-agnostic. The model receives no semantic class labels, and the prompt tensor contains only spatial information. Therefore, the trainable branch must learn to associate the prompt location with image evidence and object boundaries rather than memorizing a fixed label space. This design allows us to train on a deliberately constrained set of synthetic categories and evaluate whether the frozen diffusion prior generalizes to unseen categories and domains.

3.5 Inference

At inference time, the model receives an image and a spatial condition, constructed from the user point prompts or from the confidence or uncertainty map.

The final binary mask is obtained by thresholding the distance from black (i.e., background) in RGB embedding space:

$$\hat{m}(q) = \mathbf{1}\left[\|y_q - \mathbf{0}\|_2 > \tau\right], \tag{10}$$

where τ is selected on a validation set and kept fixed during evaluation. In the interactive setting, $\hat{m}$ is the mask of the prompted target object.

4 Experiments and Results

Datasets & Metrics **Training:** Following Gen2Seg [2], we used only two synthetic datasets for fine-tuning: Hypersim [37] and Virtual KITTI 2 [38]. Hypersim covers a wide range of indoor environments, bathrooms, bedrooms, libraries, kitchens, and more, while Virtual Kitti 2 provides photorealistic outdoor driving scenes. We also followed their filtering procedure. Despite their photorealistic quality and scene variety, both datasets' combined category coverage is inherently limited by their domain focus, lacking annotations for many common everyday object categories (e.g., animals, people). These characteristics make Gen2Seg's training set a deliberately constrained testbed to cleanly isolate and evaluate our approach's zero-shot generalization capacity to object categories and visual domains entirely absent from training. **Evaluation:** Once fine-tuned on the limited category set described above, we evaluate zero-shot generalization for interactive instance segmentation using mIoU on seven benchmarks spanning diverse domains and object types. Following Gen2Seg [2] and SAM [1], we adopt four benchmarks, as a direct basis for comparison: COCO [39] (validation set, but categories seen during fine-tuning are excluded), DRAM [40] (artistic imagery), EgoHOS [41] (egocentric scenes), and PIDRay [10] (luggage X-rays). We further include Pascal VOC [42] and HRSOD [43] to broaden evaluation coverage to natural images with diverse object categories and high-resolution salient objects, respectively. Together, these benchmarks span a wide range of visual styles and object categories absent from our training data, providing a rigorous test of generalization. We follow the evaluation splits used by Gen2Seg for DRAM and COCO. For EgoHOS, we report results on both the in-domain and out-of-domain test splits. For the remaining datasets, we use the standard evaluation splits: VOC2012 val, HRSOD val, ZeroWaste test, and the PIDRay easy test.

Baselines We compare against three main baselines. Gen2Seg [2], most closely related to our work, is a generative approach that fine-tunes Stable Diffusion for category-agnostic instance segmentation by casting it as an image-to-image translation problem, where instances are distinguished by color. Our method directly extends this framework by introducing principled spatial prompting, allowing us to isolate the contribution of explicit prompt conditioning over Gen2Seg's post-hoc color-based instance retrieval. Beyond Gen2Seg's primary Stable Diffusion-based model, we also report results for its MAE-based variant. SimpleClick [34] is a state-of-the-art promptable segmenter built on a MAE-pretrained ViTDet backbone. To ensure a fair comparison, we fine-tune it on the same synthetic dataset as ours using its official training code, following the protocol of Gen2Seg. This

Table 1: **Zero-shot quantitative evaluation** on 7 different datasets. Diffusion-based models are reported in the lower block. †: ViT-H model trained on SA-1B [1], unlike the others trained only on synthetic data from [37, 38]. P2Seg is short for Prompt2Seg.

	COCO								
Method	S	M	L	DRAM	EgoHOS	PIDRay	VOC	HRSOD	ZeroWaste
SAM †	57.0	59.7	57.6	51.6	59.8	44.2	45.5	71.2	45.4
SimpleClick	34.6	45.2	56.5	48.7	51.8	14.1	42.3	74.2	45.3
MAE-H	4.3	27.2	59.9	48.7	33.3	25.4	37.0	66.4	36.9
Gen2Seg	8.3	38.4	58.7	47.5	40.0	34.9	39.4	64.5	45.2
P2Seg-Concat	4.3	26.1	53.7	47.0	37.3	24.0	32.1	53.5	39.4
P2Seg-Adapter	6.9	37.2	60.4	48.3	46.2	41.7	39.0	64.4	45.3
P2Seg	**12.1**	**49.7**	**62.3**	**53.1**	**55.6**	**47.5**	**42.4**	**66.6**	**51.4**

makes SimpleClick a natural discriminative baseline for assessing whether standard architectures can generalize beyond their supervised categories. SAM [1] is used off-the-shelf (ViT-H backbone) as upper-bound reference, since it was supervised on SA-1B [1] with over one billion annotated masks. For each prompt, we evaluate only the highest-confidence mask predicted by SAM.

In addition, we introduce two simpler alternatives to the proposed ControlNet-style conditioning: P2Seg-Concat, with early input fusion, and P2Seg-Adapter, with a lightweight multi-scale feature injection. These approaches show whether simpler conditioning mechanisms suffice to incorporate spatial prompts, or whether the higher-capacity ControlNet branch offers additional benefits. *P2Seg-Concat* resizes the three-channel spatial prompt map to the image-latent resolution, concatenates it with the four-channel Gen2Seg latent, and expands the U-Net input convolution from four to seven channels. The pretrained weights corresponding to the image-latent channels are preserved, while the weights for the three additional prompt channels are zero-initialized. Only the expanded input convolution and the first U-Net down block are trained. *P2Seg-Adapter* uses a lightweight, T2I-Adapter-style four-scale branch. The prompt is processed through residual stages, projected to the corresponding U-Net feature dimensions using zero-initialized convolutions, and injected as intrablock residuals into the U-Net down blocks. Only the adapter branch is trained.

Implementation Details The implementation details are provided in the supplementary material.

4.1 Zero-Shot Evaluation

Table 1 reports zero-shot performance on seven benchmarks. Following Gen2Seg [2], we additionally split COCO [39] by instance size to examine how performance varies with object scale.

Prompt2Seg consistently outperforms SimpleClick [34] and all Gen2Seg variants across all datasets, confirming that neither discriminative pretraining nor post-hoc color-based prompting generalizes as effectively as our approach. Compared to SAM [1], trained on the large-scale, human-annotated SA-1B dataset, the picture is more nuanced. SAM performs better on benchmarks that closely resemble its training distribution, namely COCO [39], EgoHOS [41], Pascal VOC [42], and HRSOD [43]. In contrast, Prompt2Seg surpasses SAM on DRAM [40], PIDRay [10], and ZeroWaste [11], datasets that are more long-tail in nature, covering artistic imagery, X-ray scans, and cluttered waste scenes with ambiguous object boundaries. This pattern is consistent with generative priors being more robust to domain shift, having been exposed to a substantially broader visual distribution during pretraining than discriminative models trained on curated segmentation data. Interestingly, for COCO, Prompt2Seg outperforms SAM on large objects, while SAM gains the upper hand on medium-sized instances. Furthermore, our method exposes substantial limitations on small objects, a weakness shared by Gen2Seg, suggesting it may be a fundamental challenge for diffusion-based approaches rather than specific to our prompting mechanism.

Crucially, despite being fine-tuned solely on synthetic data, Prompt2Seg closes the gap to SAM on EgoHOS, VOC, and HRSOD, where SAM leads by only 3-5 IoU points, while the gap widens substantially on COCO_S and COCO_M. This demonstrates that principled spatial prompting, combined

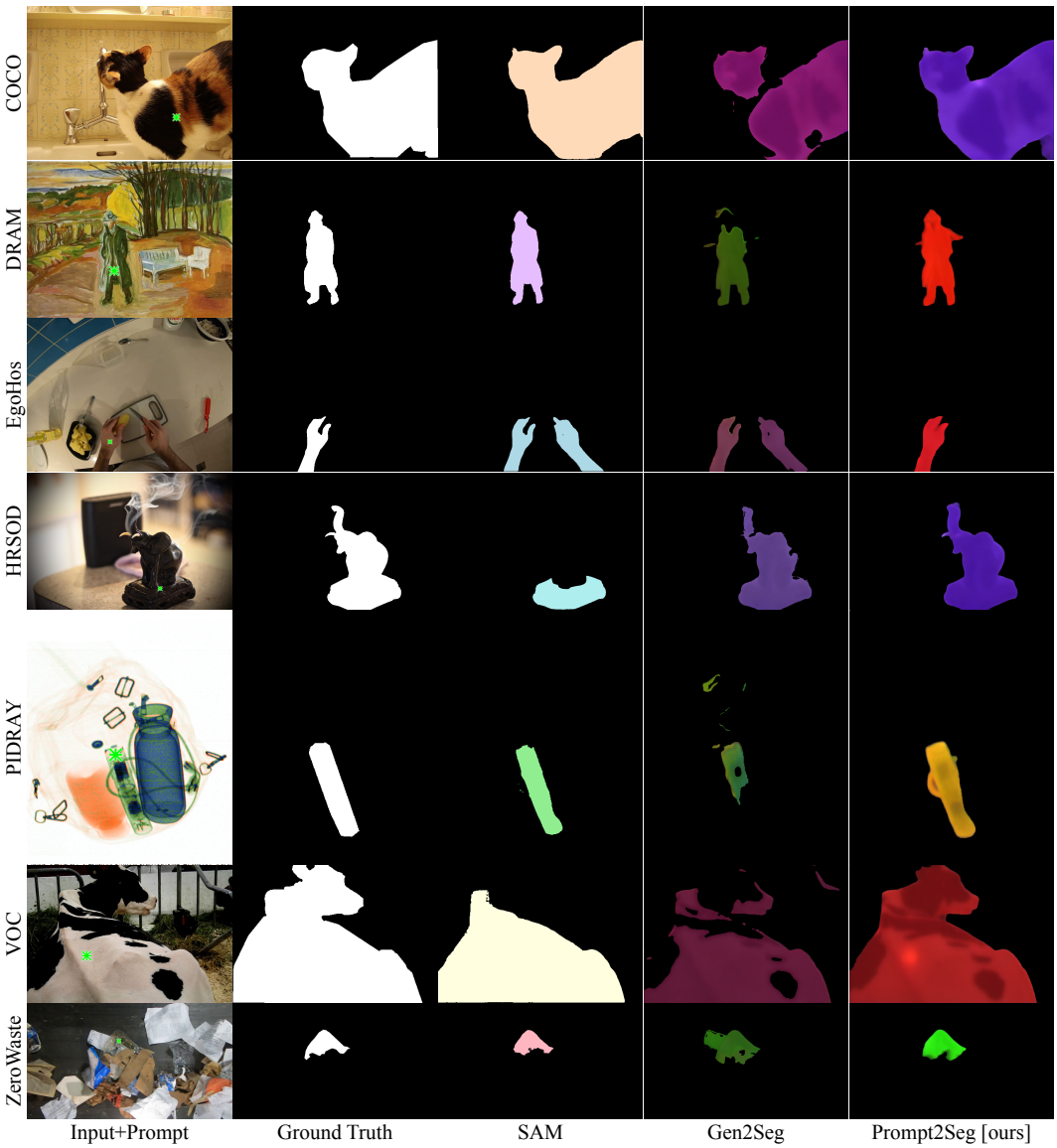


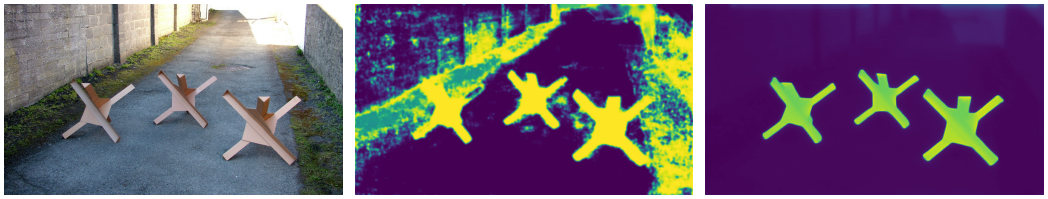
Figure 3: **Qualitative Results** from various datasets (left) comparing Prompt2Seg with SAM and Gen2Seg. SAM masks are binary, unlike the others, which output features, shown here masked.

with the rich priors of generative pretraining, offers a compelling and low-effort path to generalizable interactive instance segmentation.

4.2 Qualitative Results

Figure 3 qualitatively compare Gen2Seg [2] and SAM [1] with our Prompt2Seg across seven datasets. The qualitative results confirm the quantitative findings, suggesting the effectiveness of our introduced prompting mechanism over Gen2Seg's post-selection process.

While Gen2Seg's extracted instance masks look competitive to ours for HRSOD [43], it tends to underperform on instances with complex or irregular boundaries, such as PIDRay [10] or ZeroWaste [11]. In addition, Gen2Seg tends to struggle with objects having contrastive colors. For example, it assigns the cat's (COCO [39]) and cow's (VOC [42]) black patterns to the background. Interestingly, for EgoHos [41], Gen2Seg and SAM cluster both visible hands together. This could stem from semantic similarity between arms and the ambiguity arising from both arms belonging to the same person.



Input image | Prompt (uncertainty map) | Prompt2Seg output

Figure 4: **Qualitative results on RoadAnomaly [46].** Conditioned on the source uncertainty map, Prompt2Seg suppresses noisy background responses while preserving the true anomalous objects.

Nevertheless, precise separation of individual arms is often critical in egocentric settings, where our spatial prompting mechanism seems to effectively disambiguate both arms.

4.3 Unknown Instances from Uncertainty Maps

Prompt2Seg is not limited to sparse prompts such as points (e.g., user clicks); it can also use dense maps as spatial conditioning signals, such as the outputs of another model. We evaluate this setting on RoadAnomaly [46] using RbA [44], an anomaly segmentation model, and Mask2Former [45], a closed-set semantic segmentation model. Prompt2Seg is conditioned on RbA anomaly-score maps or on uncertainty maps derived from Mask2Former predictions using the RbA scoring formulation.

Table 2: **RoadAnomaly.** P2Seg is applied on RbA or Mask2Former outputs.

Method	AP ↑	FPR ↓
RbA [44]	85.42	6.92
RbA + **P2Seg**	**85.83**	**2.25**
Mask2Former [45]	78.45	11.83
M2F + **P2Seg**	**88.41**	**4.95**

Table 2 shows that Prompt2Seg improves both source signals, indicating that it does not simply reproduce the conditioning map but refines noisy dense cues into more reliable anomaly masks.

Figure 4 illustrates this behavior qualitatively: the source map captures the anomalous objects but contains substantial background noise, while Prompt2Seg suppresses spurious activations and produces a cleaner, more localized prediction. These results show that Prompt2Seg can serve as a dense-prompt refinement module beyond sparse spatial prompts, opening new opportunities to refine the otherwise jittery and imprecise uncertainty maps.

4.4 Edge Detection Accuracy

Table 3 reports a comparison with Gen2Seg on the quality of the segmentation boundaries. Despite using all the Gen2Seg model parameters, ours delivers superior mask boundaries by directly conditioning the diffusion process via the spatial prompt.

Table 3: **Masks boundary evaluation.** Comparison with Gen2Seg on COCO_L.

Method	Boundary F	Precision	Recall	Boundary IoU
Gen2Seg	47.09	42.62	**52.61**	19.16
P2Seg	**52.63**	**56.43**	49.31	**21.82**

4.5 Multi-Click Setting

We evaluate Prompt2Seg in a multi-click setting using up to 4 clicks on two datasets, namely COCO and DRAM. We simulate the interactivity as follows: the first click is placed at the ground-truth distance-transform center, while each subsequent click (up to 3 more) targets the largest remaining false-negative region. We report mIoU after each number of clicks. As shown in Table 4, Prompt2Seg improves consistently on both datasets as more spatial prompts are provided, with the largest gain after the second prompt. For small objects, however, performance remains largely unchanged. This may result from the ambiguity introduced by Gaussian prompts that extend beyond object boundaries (which we do not account for), the limited spatial resolution of latent diffusion features, or Stable Diffusion's pre-training bias toward larger, more salient objects. While Gen2Seg also improves with subsequent clicks, Prompt2Seg maintains superior performance across all settings.

Table 4: **Multiple iterative clicks** as prompt. Comparison with Gen2Seg on COCO and DRAM. Best per row in **bold**; best per dataset underlined. The rightmost column shows the gain from interactivity.

Dataset	Method	1st Prompt	2nd Prompts	3rd Prompts	4th Prompts	Gain ($1 \rightarrow 4$)
COCO_S	Gen2Seg	8.23	**8.25**	**8.25**	**8.25**	+0.02
	Prompt2Seg	11.84	11.87	**11.88**	**11.88**	+0.04
COCO_M	Gen2Seg	38.42	39.63	39.74	**39.75**	+1.33
	Prompt2Seg	49.44	51.01	51.23	**51.30**	+1.86
COCO_L	Gen2Seg	58.77	67.53	70.01	**70.84**	+12.07
	Prompt2Seg	62.94	71.92	74.77	**75.82**	+12.88
COCO	Gen2Seg	31.63	34.31	34.98	**35.20**	+3.57
	Prompt2Seg	37.93	40.79	41.61	**41.90**	+3.97
DRAM	Gen2Seg	47.49	57.53	60.90	**62.23**	+14.74
	Prompt2Seg	53.11	62.47	65.35	**66.83**	+13.72

4.6 Non-Idealized-Click Setting

In the other experiments in this paper, the prompt is placed at the distance-transform maximum of the ground-truth object, i.e., its most central point. To simulate realistic user imprecision, we sample a 2D Gaussian displacement around this ideal location, with $\sigma \in \{0, 5, 10, 15\}$ pixels. The perturbed click is constrained to remain inside the target ground-truth mask. We use a fixed annotation-specific random seed to ensure that all compared methods receive exactly the same perturbed click.

Table 5 shows that all three methods appear robust under the evaluated non-ideal click protocol, with performance (mIoU) decreasing only modestly as the click displacement increases from $\sigma = 0$ to $\sigma = 15$: by 0.9 for SAM and 1.6 for both Gen2Seg and Prompt2Seg. Although the diffusion-based methods degrade slightly more than SAM under the described non-idealized click setting, Prompt2Seg retains the highest absolute performance across all evaluated click displacements.

Table 5: **Non-idealized clicks** as prompt to simulate user imprecision. Comparison to SAM and Gen2Seg on DRAM.

Method	σ=0	σ=5	σ=10	σ=15
SAM	51.6	51.8	51.3	50.7
Gen2Seg	47.5	47.3	46.9	45.9
Prompt2Seg	**53.1**	**52.6**	**52.3**	**51.5**

5 Limitations and Conclusion

Limitations and Future Work: Despite its overall effectiveness, as shown in Table 1 on COCO_S, our model struggles with small objects. We inherit this issue from Gen2Seg [2], which we incorporated frozen and still outperformed on small instances, too. This could be due to the inherent resolution bottleneck of latent diffusion models, where aggressive spatial compression limits the model's ability to represent fine-grained details [12]. Then, although Prompt2Seg narrows the gap to discriminative models on natural images and even overtakes SAM in multiple settings, it still lags behind SAM due to SAM's large amount of task-specific, annotated training data. Future work could explore how to further leverage the diffusion priors. Moreover, we show failure cases in Figure 5, where, with a single spatial prompt, the model cannot always understand the desired granularity. Training with multiple positive prompts on the same instance or negative prompts on other objects could help here. Additionally, hierarchical solutions could be explored to output masks at different granularities.

Conclusion: We presented the simple and effective Prompt2Seg, a generalizable spatial prompting framework for diffusion-based interactive instance segmentation. Prompt2Seg enables interactive click-based segmentation powered by a diffusion model. With our experiments across seven diverse datasets, we demonstrated that conditioning based on spatial prompting significantly improves over post-hoc selection strategies for diffusion-based models, narrowing the gap and even outperforming strong supervised discriminative models such as SAM [1] in multiple settings while relying solely on synthetic training data. Furthermore, we showed how our method can effectively adapt to take an uncertainty map as conditioning input instead of the user's clicks, leading to refined uncertainty masks over OOD objects.

Acknowledgements

This project was supported by the Mercedes-Benz Group AG. In addition, Stefano Gasperini was supported by the German Federal Ministry of Research, Technology and Space (BMFTR) through the Robotics Institute Germany (RIG).

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 (ICCV)*, pp. 4015–4026, 2023.

[2] O. Khangaonkar and H. Pirsiavash, "gen2seg: Generative models enable generalizable instance segmentation," in *International Conference on Learning Representations*, 2026.

[3] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2961–2969, 2017.

[4] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision*, pp. 833–851, Springer, 2018.

[5] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2881–2890, 2017.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, pp. 8748–8763, PMLR, 2021.

[7] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, *et al.*, "Dinov3," *arXiv preprint arXiv:2508.10104*, 2025.

[8] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, *et al.*, "Sam 2: Segment anything in images and videos," in *International Conference on Learning Representations*, 2025.

[9] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, *et al.*, "Sam 3: Segment anything with concepts," *International Conference on Learning Representations*, 2026.

[10] L. Zhang, L. Jiang, R. Ji, and H. Fan, "Pidray: A large-scale x-ray benchmark for real-world prohibited item detection," *International Journal of Computer Vision*, vol. 131, no. 12, pp. 3170–3192, 2023.

[11] D. Bashkirova, M. Abdelfattah, Z. Zhu, J. Akl, F. Alladkani, P. Hu, V. Ablavsky, B. Calli, S. A. Bargal, and K. Saenko, "Zerowaste dataset: Towards deformable object segmentation in cluttered scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21147–21157, 2022.

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

[13] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, "Repurposing diffusion-based image generators for monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9492–9502, 2024.

[14] X. Fu, W. Yin, M. Hu, K. Wang, Y. Ma, P. Tan, S. Shen, D. Lin, and X. Long, "Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image," in *Proceedings of the European Conference on Computer Vision*, pp. 241–258, Springer, 2024.

[15] C. Ye, L. Qiu, X. Gu, Q. Zuo, Y. Wu, Z. Dong, L. Bo, Y. Xiu, and X. Han, "Stablenormal: Reducing diffusion variance for stable and sharp normal," *ACM Transactions on Graphics (ToG)*, vol. 43, no. 6, pp. 1–18, 2024.

[16] J. Gong, L. G. Foo, Z. Fan, Q. Ke, H. Rahmani, and J. Liu, "Diffpose: Toward more reliable 3d pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13041–13051, 2023.

[17] J. Wang, C. Rupprecht, and D. Novotny, "Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9773–9783, 2023.

[18] A. Luo, X. Li, F. Yang, J. Liu, H. Fan, and S. Liu, "Flowdiffuser: Advancing optical flow estimation with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19167–19176, 2024.

[19] J. Deng, T. Weng, T. Yang, W. Luo, Z. Li, and W. Jiang, "Llamaseg: image segmentation via autoregressive mask generation," in *European Conference on Computer Vision*, pp. 523–541, Springer, 2026.

[20] Y. Chen, S. Chen, L. Lin, and G. Wang, "Gs: Generative segmentation via label diffusion," *arXiv preprint arXiv:2508.20020*, 2025.

[21] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *International conference on machine learning*, pp. 1691–1703, PMLR, 2020.

[22] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

[23] D. Baranchuk, I. Rubachev, A. Voynov, V. Khrulkov, and A. Babenko, "Label-efficient semantic segmentation with diffusion models," in *International Conference on Learning Representations*, 2022.

[24] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu, "Unleashing text-to-image diffusion models for visual perception," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5729–5739, 2023.

[25] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[26] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, 2023.

[27] Z. Zhu, X. Feng, D. Chen, J. Yuan, C. Qiao, and G. Hua, "Exploring pre-trained text-to-video diffusion models for referring video object segmentation," in *Proceedings of the European Conference on Computer Vision*, pp. 452–469, Springer, 2024.

[28] Q. Nguyen, T. Vu, A. Tran, and K. Nguyen, "Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 76872–76892, 2023.

[29] C. Ma, Y. Yang, C. Ju, F. Zhang, J. Liu, Y. Wang, Y. Zhang, and Y. Wang, "Diffusionseg: Adapting diffusion towards unsupervised object discovery," *arXiv preprint arXiv:2303.09813*, 2023.

[30] W. Wu, Y. Zhao, M. Z. Shou, H. Zhou, and C. Shen, "Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1206–1217, 2023.

[31] D. Yang, R. Dong, J. Ji, Y. Ma, H. Wang, X. Sun, and R. Ji, "Exploring phrase-level grounding with text-to-image diffusion model," in *Proceedings of the European Conference on Computer Vision*, pp. 161–180, Springer, 2024.

[32] X. Liu, S. Huang, Y. Kang, H. Chen, and D. Wang, "Vgdiffzero: Text-to-image diffusion models can be zero-shot visual grounders," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2765–2769, IEEE, 2024.

[33] M. Ni, Y. Zhang, K. Feng, X. Li, Y. Guo, and W. Zuo, "Ref-diff: Zero-shot referring image segmentation with generative models," *Science China Information Sciences*, vol. 69, no. 8, p. 182104, 2026.

[34] Q. Liu, Z. Xu, G. Bertasius, and M. Niethammer, "Simpleclick: Interactive image segmentation with simple vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22290–22300, 2023.

[35] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, *et al.*, "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024.

[36] V. Gabeur, S. Long, S. Peng, P. Voigtlaender, S. Sun, Y. Bao, K. Truong, Z. Wang, W. Zhou, J. T. Barron, *et al.*, "Image generators are generalist vision learners," *arXiv preprint arXiv:2604.20329*, 2026.

[37] M. Roberts, J. Ramapuram, A. Ranjan, A. Kumar, M. A. Bautista, N. Paczan, R. Webb, and J. M. Susskind, "Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10912–10922, 2021.

[38] Y. Cabon, N. Murray, and M. Humenberger, "Virtual kitti 2," *arXiv preprint arXiv:2001.10773*, 2020.

[39] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision*, pp. 740–755, Springer, 2014.

[40] N. Cohen, Y. Newman, and A. Shamir, "Semantic segmentation in art paintings," in *Computer graphics forum*, vol. 41, pp. 261–275, Wiley Online Library, 2022.

[41] L. Zhang, S. Zhou, S. Stent, and J. Shi, "Fine-grained egocentric hand-object segmentation: Dataset, model, and applications," in *Proceedings of the European Conference on Computer Vision*, pp. 127–145, Springer, 2022.

[42] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, June 2010.

[43] Y. Zeng, P. Zhang, J. Zhang, Z. Lin, and H. Lu, "Towards high-resolution salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7234–7243, 2019.

[44] N. Nayal, M. Yavuz, J. F. Henriques, and F. Güney, "Rba: Segmenting unknown regions rejected by all," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 711–722, 2023.

[45] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1290–1299, 2022.

[46] K. Lis, K. Nakka, P. Fua, and M. Salzmann, "Detecting the unexpected via image resynthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2152–2161, 2019.

A Implementation Details

Backbone and conditioning branch. We initialize the segmentation backbone from the official Gen2Seg Stable-Diffusion checkpoint [2]. The VAE, CLIP text encoder, and denoising U-Net are kept frozen throughout training. We initialize the ControlNet branch from the frozen U-Net weights and train only the ControlNet parameters. All experiments use an empty text prompt, so the model receives task-specific information only through the spatial condition.

Prompt construction. For each selected target instance, we sample a positive point from the interior of the ground-truth mask using the maximum of the Euclidean distance transform. This avoids ambiguous boundary clicks during training and provides a stable one-click supervision signal. The prompt is converted into Gaussian heatmaps at two spatial scales. Unless otherwise stated, we use $\sigma_s = 8$ and $\sigma_l = 16$ pixels, which performed best in our prompt-encoding ablation. Multiple prompts, when present, are combined using a pixelwise maximum.

Prompt-conditioned targets. Training targets are constructed from instance-colored ground-truth masks. For each image, we randomly select valid target instances above a minimum area threshold and set all non-target pixels to black. Thus, the target instance-coloring map contains only the prompted object or objects as foreground, while other annotated objects are treated as background. This converts the original all-instance supervision into a prompt-conditioned target-only prediction task.

Mask extraction. The decoded output is an RGB instance-coloring map in the range $[0, 255]$. We convert it to a binary mask by thresholding the Euclidean distance of each pixel from black. The threshold is selected with the ablations and then kept fixed for all test datasets. For the reported interactive segmentation experiments, we use the same one-click protocol across all datasets.

Dense-prompt extension. For road anomaly experiments, the spatial condition is derived from a frozen uncertainty critic instead of user clicks. We compute the anomaly response from the critic logits, robustly normalize it using image-wise percentiles, smooth the response with local averaging, retain high-response regions, and map the resulting dense condition to $[-1, 1]$. The dense map is then provided to the same ControlNet conditioning interface. No change is made to the frozen diffusion backbone or to the instance-coloring objective.

B Ablation Studies

B.1 Instance Count

Table 6: Ablation on the number of target instances used during training

# Instances	COCO_S	COCO_M	COCO_L
1	9.26	43.98	64.37
4	**9.95**	**47.99**	64.09
8	7.86	43.67	**65.04**

Table 6 ablates the number of target instances used during training. Using four instances gives the best performance on COCO_S and COCO_M, while eight instances slightly improves COCO_L but degrades smaller scales. We therefore use four instances as the default, as it provides the best overall trade-off.

B.2 Gaussian Size

Table 7 ablates the standard deviation of the Gaussian point-prompt encoding. Very narrow prompts provide insufficient spatial context, while overly broad prompts reduce localization precision. A moderate value of $\sigma_s = 8$ px performs best across object scales.

Table 7: Ablation on the standard deviation of the Gaussian point-prompt encoding.

Gaussian σ_s (px)	COCO$_S$	COCO$_M$	COCO$_L$
1	0.7	5.2	27.6
8	**10.6**	**43.3**	**60.7**
16	9.3	41.0	60.4

Table 8: Ablation of single- versus dual-scale Gaussian point-prompt encodings. The object-area threshold for training filtering is fixed to 300 pixels for all runs.

Encoding	COCO$_S$	COCO$_M$	COCO$_L$
$\sigma = 8$ px	10.0	48.0	64.1
$\sigma \in \{8, 16\}$ px	12.1	49.7	62.3

Table 8 compares single- and dual-scale Gaussian encodings. Using two scales, $\sigma \in \{8, 16\}$ px, improves COCO$_S$ and COCO$_M$, indicating that multi-scale prompt encoding provides a more robust conditioning signal. We use the dual-scale encoding as the default.

B.3 Area Filtering

Table 9: Ablation on the minimum object-area threshold used for filtering training instances. A threshold of 0 indicates no area filtering.

px^2	COCO$_S$	COCO$_M$	COCO$_L$
0	**10.6**	43.3	60.7
100	10.4	44.4	59.4
300	10.0	**48.0**	**64.1**
400	10.3	42.2	58.8
500	9.4	44.2	58.8

Table 9 evaluates the minimum object-area threshold used to filter training instances. Since only a few target instances are sampled per image, very small objects often provide weak conditioning signals and are difficult to segment reliably. Filtering instances below 300 px^2 improves COCO$_M$ and COCO$_L$ while maintaining competitive COCO$_S$ performance; therefore, we use 300 px^2 as the default.

B.4 Magnitude Thresholding

Table 10: Ablation on the magnitude thresholding τ.

τ	COCO$_S$	COCO$_M$	COCO$_L$
50	7.7	38.8	61.3
75	9.0	43.1	63.6
100	10.1	45.5	**63.3**
150	**12.1**	**49.7**	62.34

Table 10 ablates the magnitude threshold τ used during post-processing. Increasing τ suppresses low-confidence regions and improves COCO$_S$ and COCO$_M$, with the best results at $\tau = 150$. COCO$_L$ peaks at $\tau = 100$, but the difference is small; therefore, we use $\tau = 150$ as the default.

B.5 Failure Cases

Figure 5 shows failure cases where our model misunderstood the prompt and either segmented only part of the object of interest or a different object. Additional prompts could help.



Figure 5: Additional qualitative failure cases. For each example, the left image shows the input with the point prompt and target boundary overlay, while the right image shows the predicted output. The model captures the approximate target region but fails to precisely follow object boundaries or separates only part of the intended instance.

C Additional Quantitative Results

C.1 Inference Latency

Table 11: **Inference latency** in milliseconds for 1-Click setting of Prompt2Seg, Gen2Seg, and SAM.

Method	Full	Prompt only
SAM	207.1	3.7
Gen2Seg	138.2	1.2
Prompt2Seg	150.0	150.0

Table 11 compares Prompt2Seg's inference latency with Gen2Seg and SAM. SAM encodes each input image only once; subsequent prompts are processed using the cached image embedding together with its prompt encoder and mask decoder. Gen2Seg likewise extracts image features once and handles additional prompts through a post-processing procedure, without rerunning the diffusion backbone. In contrast, Prompt2Seg injects the spatial prompt directly into the model input and therefore requires a full forward pass whenever either the image or the prompt changes. This increases its latency in interactive settings, particularly in multi-click settings, in favor of superior segmentation outputs. We used a single NVIDIA H200 NVL for measuring the inference time.

C.2 Comparison to additional discriminative models

Table 12: **Zero-shot quantitative evaluation** of additional discriminative, interactive segmentation models across DRAM, COCO, and PIDRay.

Method	DRAM	COCO_S	COCO_M	COCO_L	PIDRay
SAM	51.6	57.0	59.7	57.6	**44.2**
SAM 3	56.9	**64.0**	**69.9**	**68.0**	-
SimpleClick	48.7	34.6	45.2	56.5	14.1
SimpleClick-DINOv3	**62.1**	26.6	42.3	63.2	37.3
Gen2Seg	47.5	8.3	38.4	58.7	34.9
Prompt2Seg	**53.1**	**12.1**	**49.7**	**62.3**	**47.5**

For reference, Table 12 reports two additional discriminative segmentation models, SimpleClick [34] with DINOv3 [7] backbone and SAM 3 [9]. SimpleClick, with the powerful DINOv3 ViT-H+/16 backbone, is trained on the same Hypersim/VKITTI images and masks using the same click-sampling and evaluation protocol. Compared with the original backbone, DINOv3 substantially improves

SimpleClick on DRAM (+13.42 mIoU), COCO Large (+6.72), and PIDRay (+23.21), while performance decreases on COCO Small (-7.96) and COCO Medium (-2.89). Compared with Prompt2Seg, SimpleClick-DINOv3 performs better on DRAM and COCO Small and is approximately tied on COCO Large, whereas Prompt2Seg remains stronger on COCO Medium and PIDRay. These results confirm that backbone pretraining explains part, but not all, of the original performance gap. On the other hand, SAM 3 is, in general, significantly stronger than SAM and therefore stronger than Prompt2Seg; however, SAM 3 was trained on a significantly larger corpus of task-specific downstream task data than Prompt2Seg.

Given that, we reiterate that our goal is not to claim uniform state-of-the-art performance over all discriminative, interactive segmentation methods, but to demonstrate the possibility of repurposing diffusion models for interactive segmentation with substantially less task-specific dense annotation than SAM. Prompt2Seg introduces a flexible spatial-conditioning mechanism for diffusion-based segmentation models. Beyond sparse point prompts, the same image-like conditioning interface can consume dense confidence maps, as demonstrated by the road-anomaly experiments in Figure 4. Our evaluated SimpleClick configuration uses its standard click-based conditioning, whereas Prompt2Seg demonstrates both interactive point prompting and dense anomaly-map refinement. This highlights the broader utility of spatially conditioning generative segmentation priors. Finally, the controlled Prompt2Seg-Gen2Seg comparison remains the direct evidence for the effectiveness of the proposed conditioning mechanism: both methods use the same diffusion segmentation backbone, while differing in direct learned spatial conditioning versus post-hoc feature-similarity retrieval.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly state the contributions and scope in the Abstract and Introduction.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We describe the limitations in Section 5.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include theoretical results.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have presented the information in Section 4.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code will be publicly available soon. Code release updates will be posted on the project page: `https://prompt2seg.github.io/`.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper and supplementary material describe the datasets, evaluation protocol, training setup, and key hyperparameters needed to interpret the results. Our implementation builds on the released Gen2Seg codebase, and we ablate the main hyperparameters introduced by our method. Additional low-level implementation details will be made available with the code release upon acceptance.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper does not report error bars within the quantitative evaluation.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

Justification: Except for inference latency measurements, we do not report specific information on the computer resources as our experiments were conducted on a cluster with different GPU types (such as A100, H100, and H200).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The conducted research in this paper conforms to the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: We do not discuss the broader impact of the research conducted in this paper.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The released model does not have a high risk for misuse.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite corresponding papers for the asserts we use.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A].

    Justification: This work does not involve the introduction of new assets.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: This work does not involve crowdsourcing nor research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as part of the proposed method, experiments, data generation, annotation, or evaluation. Their use was limited to auxiliary assistance for writing/editing, implementation debugging, and feedback on presentation; all scientific decisions, experiments, and conclusions were made and verified by the authors.