In-Context Learning for Remote Sensing Vision: A Semantic-Aware Rotation-Robust Diffusion Framework

Jing Geng[1]* **Chuanlong Zhang**[1]* **Jiayi Su**[2] **Wencheng Han**[2]†
[1]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
[2]School of Control Science and Engineering, Shandong University, Jinan, China

Abstract

Remote sensing imagery supports diverse visual tasks, such as object detection and semantic segmentation, but these tasks vary across scenes, imaging modalities, and target objects. Since many remote sensing scenarios lack sufficient annotations, conventional supervised learning is often difficult to apply, motivating in-context learning as a flexible alternative. We present an in-context learning framework for remote sensing visual tasks, where a new task is specified by a single annotated example and applied to unseen images without retraining at inference time. Built upon diffusion models, our framework establishes target-reference associations through cross-attention and transfers the structural relationship between the reference image and its annotation to produce task-specific predictions. Unlike natural images, remote sensing objects are often small and provide weak semantic cues in early diffusion steps, making semantic alignment between target and reference images difficult. We address this limitation with a semantic-aware cross-attention that combines semantic and fine-grained details for more accurate cross-image matching. Remote sensing objects also appear at arbitrary orientations, causing large appearance variations and unstable in-context transfer. To improve robustness, we introduce a rotation-robust learning strategy that reduces sensitivity to orientation changes. Experiments show that our method achieves strong one-shot performance, establishing a flexible paradigm for adapting remote sensing models to diverse visual tasks.

1 Introduction

Remote sensing imagery plays a fundamental role in real-world applications such as urban planning, environmental monitoring, and disaster response, where object detection and semantic segmentation are often required[1–4]. Unlike natural image benchmarks, remote sensing tasks are highly application-specific and vary across scenes, sensors, and target categories[5–8]. Since collecting dense annotations is expensive, especially for large-scale coverage or rare events, many practical scenarios lack sufficient labeled data for conventional supervised learning[9, 10]. Recent advances in large-scale vision models and diffusion-based architectures have enabled in-context learning (ICL), where a new visual task can be specified by a single annotated example[11–14]. However, directly extending ICL to remote sensing imagery remains challenging: remote sensing objects are often small and provide weak semantic cues, and they appear at arbitrary orientations, both of which disrupt the cross-image correspondence that ICL relies on.

*These authors contributed equally.
†Corresponding author: wencheng256@gmail.com.

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

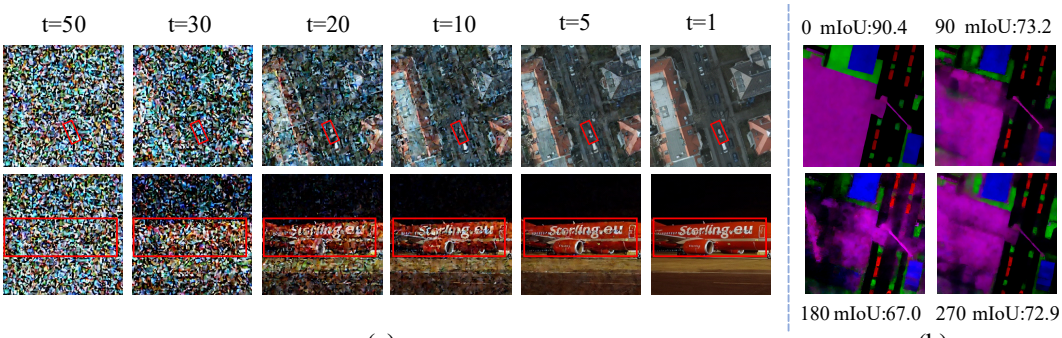


Figure 1: **Denoising-stage semantics and rotation-induced instability.** (a) In remote sensing imagery, semantic features of small targets emerge only at later stages of the diffusion denoising process, leaving early features dominated by low-level structures and making early matching unreliable. In contrast, objects in natural images become semantically recognizable at relatively earlier denoising stages. (b) When the target and reference are the same image, rotating only the reference image by $0°$, $90°$, $180°$, and $270°$ still causes substantial changes in the segmentation output, showing that visual in-context learning is sensitive to arbitrary orientations in remote sensing scenes.

To address these challenges, we propose Semantic-Aware Rotation-Robust Diffusion model (SRDM), a diffusion framework for example-driven remote sensing visual tasks. SRDM uses a reference image-label pair to define the task and transfers the relationship between the reference image and its annotation to a target image by establishing semantic correspondence between their features. It adopts four diffusion streams, including the reference image, target image, reference label, and target label branches, where cross-image feature exchange builds alignment and reference-label conditioning guides target-label generation. This design enables flexible task specification across remote sensing scenes within a unified diffusion framework.

Diffusion-based ICL faces two challenges that are particularly pronounced in remote sensing. First, semantic information in diffusion features emerges progressively over the denoising process[15–18], and the weak visual cues of small remote-sensing objects further delay their semantic distinguishability. As shown in Fig. 1(a), this exacerbated delay causes early-stage cross-attention to mismatch reference and target features, weakening the use of reference-label cues. To address this issue, we propose a semantic-aware cross-attention (SACA) module that fuses reference-label semantics with structural diffusion features for more reliable alignment. Second, arbitrary object orientations further disrupt cross-image correspondence even within the same category. Fig. 1(b) shows that rotating the reference image can induce large variations in the segmentation structure even when the target and reference contents are unchanged. We therefore introduce RotationFormer in the diffusion encoder and train it with a rotation-robust learning strategy that enforces prediction consistency across prompt rotations. Overall, SRDM addresses semantic misalignment and geometric inconsistency in diffusion-based ICL for remote sensing imagery. Experiments on multiple benchmarks show strong one-shot performance on object detection and semantic segmentation, demonstrating flexible adaptation to diverse tasks with minimal annotation effort. In summary, the main contributions of this work are as follows:

- We introduce in-context learning into remote sensing vision and propose SRDM, a semantic-aware and rotation-robust diffusion framework that enables flexible, example-driven task adaptation without task-specific retraining at inference time.
- We identify the delayed emergence of semantic information in diffusion processes for remote sensing imagery and propose a semantic-aware cross-attention module to improve feature alignment and prediction accuracy, particularly for small objects.
- We develop a rotation-robust module, RotationFormer, along with a rotation-robust learning strategy, to enhance robustness to arbitrary object orientations.
- We conduct extensive experiments across multiple remote sensing benchmarks, demonstrating strong performance in few-shot and single-shot settings with superior generalization capability.

2 Related Works

2.1 Remote sensing visual task

For remote sensing semantic segmentation, early CNN-based methods, such as FCN[19] and U-Net[20], established the basic dense prediction framework, while later studies introduced contextual modeling, multi-scale aggregation, and attention mechanisms to improve scene understanding[21–23]. More recently, Transformer-based methods further strengthened long-range dependency modeling and global context extraction, leading to representative models such as SegFormer[24], RSSFormer[25], BSFDT[26], and DESFormer[27]. AFENet[28] continues this line by emphasizing adaptive modeling under diverse land-cover distributions and enhancing the interaction between spatial and frequency information, reflecting the recent trend of combining local details with global contextual representations in remote sensing segmentation.

For remote sensing object detection, the arbitrary orientations and large appearance variations of aerial targets have made oriented detection the dominant research setting, with progress achieved through improved object representations and robust regression losses[3, 29–33]. Recent studies have further explored large-kernel designs to capture long-range context, as exemplified by LSKNet[34] and PKINet[35]. Strip R-CNN[36] builds on this direction and highlights the challenge of elongated objects, showing that direction-aware feature modeling is increasingly important for accurate localization in remote sensing detection. Overall, remote sensing detection research is moving toward stronger geometric sensitivity and more effective contextual modeling.

2.2 Visual in-context learning

In Context Learning (ICL) has shown strong few shot generalization in Large Language Models (LLMs), where a small number of exemplar pairs provided in the prompt can guide models to solve unseen tasks, as demonstrated by GPT 3 [37]. Inspired by this success, the computer vision community has witnessed a surge of research in Visual In-Context Learning (VICL), aiming to guide vision models toward similar task generalization using visual exemplar pairs. Current vision-only VICL approaches generally fall into two dominant paradigms: masked modeling methods based on ViT[11, 38, 39] and generative methods based on diffusion models[12, 40, 13].

Within the masked modeling paradigm, MAE-VQGAN[11] pioneered a grid-based visual prompting strategy. Subsequently, groundbreaking works such as Painter[38] uniformly redefined the output space of dense prediction tasks as RGB images, thereby leveraging standard Masked Image Modeling to perform general vision tasks via image inpainting. A series of subsequent studies have explored strategies to enhance the effectiveness of VICL[41–45]. Zhang et al.[44] revealed the high sensitivity of these models to prompt selection and designed an automatic retrieval framework to match optimal exemplars. Alternatively, numerous methods are built upon diffusion models. For example, SD-VICL [13] directly adapts off-the-shelf Stable Diffusion models for visual in-context learning by restructuring the self-attention computation, bypassing the need for any additional fine-tuning or task-specific data.

3 Method

3.1 Overview of SRDM Framework

We study VICL for remote sensing vision, where a target task is specified by a reference image-label pair and applied to a target image by transferring the task definition encoded in the reference example. To achieve this, we propose **Semantic-Aware Rotation-Robust Diffusion model (SRDM)**, a semantic-aware and rotation-robust diffusion framework for example-driven task inference. SRDM maps all inputs into a latent space using an off-the-shelf inversion model[46], and then generates the target output through conditional reverse denoising.

As illustrated in Fig. 2, SRDM adopts a four-branch design, including a reference image branch, a reference label branch, a target image branch, and a target label branch. The reference image and target image provide structural and appearance information from their respective scenes, while the reference label encodes the task definition. The target label branch serves as the generation pathway, where the final prediction is progressively constructed.

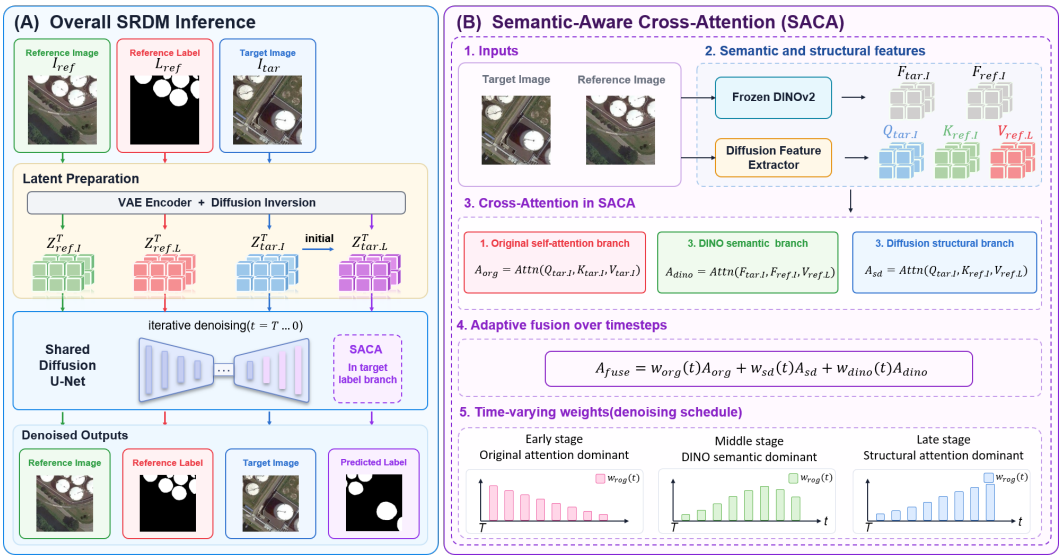


Figure 2: **Pipeline of SRDM.** The overall framework contains four branches, including the target image, target label, reference image, and reference label branches. The right part illustrates the Semantic-Aware Cross-Attention module, which combines semantic and structural cues to align the target and reference features.

A key design of SRDM lies in cross-image interaction. During generation, the model establishes correspondence between the reference and target images by aligning their feature representations. The target image guides where to predict, while the reference example provides what to predict. By integrating structural cues from the target image and semantic information from the reference label, SRDM incrementally constructs the target output and transfers annotation patterns across images.

Despite its flexibility, applying diffusion-based VICL to remote sensing imagery introduces two key challenges. First, small objects lead to delayed semantic emergence during generation, making early-stage alignment unreliable. Second, arbitrary object orientations introduce substantial geometric variations, which further disrupt correspondence. To address these issues, SRDM incorporates a semantic-aware attention mechanism and a rotation-robust learning (RRL) strategy, which are detailed in Sec. 3.3 and Sec. 3.4.

3.2 In-Context Diffusion Formulation

We formulate in-context learning in remote sensing as a conditional generation problem within a diffusion framework. Given a reference example and a target image, the objective is to generate the corresponding target output conditioned on the provided context. This can be expressed as modeling the conditional distribution:

$$p(L_{tar} \mid I_{tar}, I_{ref}, L_{ref}). \tag{1}$$

SRDM encodes the reference image, reference label, and target image into the latent space with the VAE encoder[47], and then applies an inversion operator to obtain their noisy latent states at timestep T:

$$\mathbf{Z}_{ref.I}^{(T)} = \mathcal{I}_T\big(\mathcal{E}_{\text{vae}}(I_{ref})\big), \quad \mathbf{Z}_{ref.L}^{(T)} = \mathcal{I}_T\big(\mathcal{E}_{\text{vae}}(L_{ref})\big), \quad \mathbf{Z}_{tar.I}^{(T)} = \mathcal{I}_T\big(\mathcal{E}_{\text{vae}}(I_{tar})\big). \tag{2}$$

The target label branch is initialized with $\mathbf{Z}_{tar.I}^{(T)}$. The generation process starts from an initialized latent representation of the target output, and proceeds through iterative denoising steps:

$$\mathbf{Z}_{tar.L}^{(t-1)} = \mathcal{D}\big(\mathbf{Z}_{tar.L}^{(t)}, \mathbf{Z}_{tar.I}^{(t)}, \mathbf{Z}_{ref.I}^{(t)}, \mathbf{Z}_{ref.L}^{(t)}\big) \tag{3}$$

where $\mathcal{D}(\cdot)$ denotes the conditional denoising function and t indexes the diffusion steps.

A key component in this formulation is cross-attention, which enables information transfer from the reference example to the target. At each denoising step, the interaction is defined as:

$$\mathbf{A}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right)=\mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d}}\right)\mathbf{V}, \tag{4}$$

where

$$\mathbf{Q}=\phi_q(\mathbf{F}_{tar.I}),\quad \mathbf{K}=\phi_k(\mathbf{F}_{ref.I}),\quad \mathbf{V}=\phi_v(\mathbf{F}_{ref.L}). \tag{5}$$

Here, $\mathbf{F}$ denotes the feature tensor produced by the corresponding U-Net branch at the current attention block and denoising step. Specifically, $\mathbf{F}_{tar.I}$, $\mathbf{F}_{ref.I}$, and $\mathbf{F}_{ref.L}$ denote the target-image, reference-image, and reference-label features, respectively. The functions ϕ_q, ϕ_k, and ϕ_v are learned linear projections that map these tensors to query, key, and value embeddings. Accordingly, the query is derived from the target representation, the key encodes structural information from the reference image, and the value carries semantic information from the reference label. This design allows the model to retrieve task-relevant semantics from the reference example while preserving structural consistency with the target image. The resulting attention output is incorporated into the denoising process to progressively refine the generated representation.

This formulation enables SRDM to perform example-driven task generalization without parameter updates. However, for remote sensing imagery, two limitations arise. First, semantic information may not be reliably captured in early denoising stages, leading to weak supervision. Second, geometric variations such as arbitrary rotations degrade feature alignment. These issues motivate the semantic-aware attention design in Sec. 3.3 and the rotation-robust learning strategy in Sec. 3.4.

3.3 Semantic-Aware Cross-Attention

Remote sensing images often contain dense small targets, while early diffusion features mainly encode structures and contain weak semantic cues. Since reliable semantic information appears only in later denoising steps, early cross-attention based solely on diffusion features may fail to align target and reference regions correctly.

Existing diffusion-based visual ICL methods usually assume that the intermediate features used for attention are already semantically sufficient. This assumption is less problematic for natural images, where objects are often large and visually salient. In remote sensing scenes, however, small objects and background structures can have similar local appearances, causing diffusion-only matching to retrieve label information from the wrong semantic category.

To mitigate this semantic mismatch, we decouple semantic matching from structural matching. As illustrated in the Fig. 2, SACA combines a DINOv2-based semantic branch for early-stage correspondence with a diffusion-based structural branch for generation-compatible transfer. We introduce DINOv2[48] features to provide explicit semantic cues for early-stage matching, and use diffusion features for structure-aware matching when the denoising process becomes more reliable. Following Eq. 4, the diffusion branch uses the target image feature as query, the reference image feature as key, and the reference label feature as value.

$$\mathbf{A}_{\mathrm{sd}}^{(t)}=\mathrm{Attn}\left(\phi_q(\mathbf{F}_{tar.I}^{(t)}),\phi_k(\mathbf{F}_{ref.I}^{(t)}),\phi_v(\mathbf{F}_{ref.L}^{(t)})\right). \tag{6}$$

This branch preserves generation-compatible structural transfer. In parallel, we compute a semantic branch using DINOv2 features from the clean target and reference images:

$$\mathbf{A}_{\mathrm{dino}}^{(t)}=\mathrm{Attn}\left(\pi(\psi(I_{tar})),\pi(\psi(I_{ref})),\phi_v(\mathbf{F}_{ref.L}^{(t)})\right), \tag{7}$$

where $\psi(\cdot)$ denotes the frozen DINOv2 encoder and $\pi(\cdot)$ aligns its features to the U-Net attention resolution. Thus, DINOv2 provides explicit semantic guidance for where to match, while the reference label branch provides the task-specific information to transfer.

The target label branch fuses the original self-attention, the DINOv2-guided semantic cross-attention, and the diffusion-based structural cross-attention:

$$\mathbf{A}_{\mathrm{fuse}}^{(t)}=w_{\mathrm{ori}}(t)\mathbf{A}_{\mathrm{ori}}^{(t)}+w_{\mathrm{dino}}(t)\mathbf{A}_{\mathrm{dino}}^{(t)}+w_{\mathrm{sd}}(t)\mathbf{A}_{\mathrm{sd}}^{(t)}. \tag{8}$$

Here, $\mathbf{A}_{\mathrm{ori}}^{(t)}$ denotes the original self-attention output in the target image branch, $\mathbf{A}_{\mathrm{dino}}^{(t)}$ denotes the DINOv2-guided semantic cross-attention output, and $\mathbf{A}_{\mathrm{sd}}^{(t)}$ denotes the diffusion-based structural

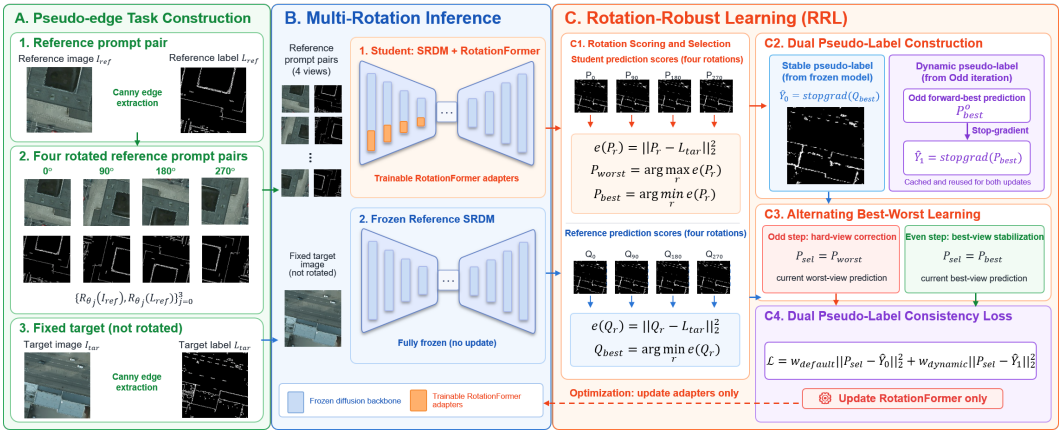


Figure 3: **Rotation-Robust Learning.** The training pipeline runs multi-rotation inference and selects the best- and worst-rotation predictions using Canny edge cues. A frozen reference model and the dynamic best prediction provide pseudo labels to supervise RotationFormer.

cross-attention output. To follow the semantic evolution of the reverse diffusion process, we adopt a three-stage progressive weighting schedule. Let the denoising progress be $\rho_t = 1 - t/T$, where t decreases from T to 0. Two sigmoid gates, $g_1(t) = \sigma(k_1(\rho_t - m_1))$ and $g_2(t) = \sigma(k_2(\rho_t - m_2))$, are used to control the transition among the three attention sources, where $m_1 < m_2$ specify the transition positions and k_1, k_2 control the transition sharpness. The attention weights are defined as:

$$w_{\text{ori}}(t) = 1 - g_1(t), \quad w_{\text{dino}}(t) = g_1(t)\big(1 - g_2(t)\big), \quad w_{\text{sd}}(t) = g_1(t)g_2(t). \tag{9}$$

This schedule makes the original attention dominate in the early denoising stage, semantic guidance dominate in the middle stage, and diffusion-based structural attention dominate in the late stage. As a result, the model first preserves stable target-branch generation, then improves semantic correspondence using DINOv2 features, and finally relies on diffusion-compatible structural transfer when the generated representation becomes more reliable.

3.4 Rotation-Robust Learning

Rotation-robust learning addresses the feature mismatch caused by arbitrary object orientations in remote sensing images. Unlike many natural images, remote sensing images are captured overhead and do not follow a strong gravity constraint. As a result, objects from the same category may appear at different orientations, producing large appearance variations and unstable feature responses.

This instability directly affects visual in-context learning, as shown in Fig. 1(b): even when the target image and reference content are unchanged, rotating the reference image can induce large variations in the predicted segmentation structure. SRDM transfers task information by measuring cross-image feature similarity between the target image and the reference example. When the reference object and target object share the same semantic category but appear at different rotations, their features may become poorly aligned. This rotation-induced mismatch weakens semantic transfer and makes the prediction sensitive to the orientation of the visual prompt.

We improve rotation robustness from both the representation and training perspectives. At the representation level, we insert a lightweight *RotationFormer* into the encoder of the diffusion U-Net to adapt the feature representation to orientation variations. At the training level, we introduce a rotation-aware consistency learning strategy that explicitly aligns predictions produced under different prompt rotations, without using manually annotated labels.

RotationFormer. RotationFormer follows an adapter-style design similar to AdaptFormer[49] and is inserted into the encoder blocks of the U-Net. Given an encoder feature $\mathbf{F}^\ell$ at layer ℓ, RotationFormer applies a bottleneck transformation and adds the adapted feature back to the original representation:

$$\widetilde{\mathbf{F}}^\ell = \mathbf{F}^\ell + \alpha \mathbf{W}_{\text{up}} \sigma\left(\mathbf{W}_{\text{down}} \text{LN}(\mathbf{F}^\ell)\right), \tag{10}$$

Table 1: **Remote Sensing Semantic Segmentation.** Per-class IoU and mIoU are reported on the Potsdam dataset for general VICL methods and SRDM.

Model	Surf	Building	Low Veg	Tree	Car	Clutter	mIoU
Visual Prompting[11]	33.32	5.55	20.05	8.87	0.96	3.67	12.07
SD-VICL[13]	54.78	59.01	**47.38**	38.90	51.26	**12.07**	43.90
Ours	**57.67**	**68.64**	46.44	**39.89**	**60.59**	11.57	**47.47**

where $\mathbf{W}_{\text{down}}$ and $\mathbf{W}_{\text{up}}$ are down- and up-projection layers, $\sigma(\cdot)$ denotes activation function, and α controls the adapter strength. This design preserves the pre-trained diffusion representation while introducing a small number of trainable parameters to model orientation-aware feature adaptation.

Rotation-aware consistency learning. As illustrated in Fig. 3, we train RotationFormer without real task labels by constructing a pseudo edge prediction task using Canny[50] edge maps. Let $\mathcal{R}$ denote a set of rotation angles. For each training image, we run inference under different rotations in $\mathcal{R}$ and obtain a set of predictions. At each odd step, we select the best and worst rotations according to the pseudo edge target, denoted as r_b and r_w, and keep the best prediction as a dynamic pseudo label $\hat{\mathbf{Y}}_1$. In parallel, a frozen reference model performs the same multi-rotation inference and provides a stable pseudo label $\hat{\mathbf{Y}}_0$, which regularizes the training and prevents the model from drifting too far from the pre-trained diffusion prior.

Training alternates between the two rotations: each odd step optimizes r_w, whereas the subsequent even step reuses $\hat{\mathbf{Y}}_1$ without reselection and optimizes r_b. For a selected rotation $r \in \{r_w, r_b\}$, the loss is defined as:

$$\mathcal{L}_{\text{rot}}(r) = w_{\text{default}} \left\| \mathbf{P}_r - \hat{\mathbf{Y}}_0 \right\|_2^2 + w_{\text{dynamic}} \left\| \mathbf{P}_r - \hat{\mathbf{Y}}_1 \right\|_2^2, \tag{11}$$

where $\mathbf{P}_r$ is the trainable model prediction under rotation r. The worst rotation is pulled toward the best pseudo prediction to reduce rotation sensitivity, while the best rotation is constrained to maintain prediction quality. The reference-model pseudo label further stabilizes optimization. As a result, object features under different rotations become more consistent, improving rotation robustness in remote sensing VICL.

4 Experiments

4.1 Experiments Setting

4.1.1 Datasets

Our method has been tested on a variety of tasks, including semantic segmentation, single-object detection, and foreground segmentation, each evaluated on distinct remote sensing datasets. Below, we provide a detailed description of the datasets used for each task.

Semantic segmentation was tested on the Potsdam dataset, a high-resolution remote sensing dataset that contains urban scenes. It consists of 38 very high-resolution images, each with a size of 6000 $\times$ 6000 pixels and a spatial resolution of 5 cm per pixel. The dataset includes six labeled classes: impervious surfaces, buildings, low vegetation, trees, cars, and clutter/background. We followed the preprocessing scheme described in the [38], using 24 images for training and the remaining 14 for testing. We report the mIoU metric.

Single-object detection was tested on the HRSC2016 dataset, a high-resolution remote sensing dataset focused on ship detection. The dataset contains 1,061 images with 2,976 instances of ships. Following the preprocessing method in [11], we cleaned and organized the dataset to extract instances for single-object detection task. We report the mIoU and mAP results.

Foreground segmentation was tested on the iSAID-5^i dataset, which is derived from the iSAID dataset and has been widely used for few-shot segmentation in remote sensing[4]. The iSAID-5^i dataset consists of 15 classes, which are evenly divided into three folds. All images have a size of 256 $\times$ 256 pixels and are cropped from larger images. We report the mIoU results.

4.1.2 Implementation Details

For training, we utilized the AdamW optimizer with a cosine learning rate scheduler. The batch size was set to 1, and the model was trained for 200 iterations. The learning rate was set to 1×10^{-4}, with dynamic weight $w_{dynamic} = 0.2$ and default weight $w_{default} = 0.8$. The feature fusion hyperparameters were set as follows: $m_1 = 0.1$, $m_2 = 0.9$, $k_1 = 10$, and $k_2 = 30$. Denoising steps for the diffusion model were set to $T = 50$. The input resolution for the diffusion model VAE was 512×512 pixels, while the input resolution for DINOv2 was 896×896 pixels. For reference selection, we encode candidate images with RemoteCLIP [5] and choose the reference image with the highest image-level similarity to the target image. All experiments were conducted on an NVIDIA A40 GPU.

4.2 Experimental Results

Semantic segmentation. Table 1 compares semantic segmentation performance on Potsdam using per-class IoU and mIoU. Our method achieves the best overall result with 47.47 mIoU, outperforming SD-VICL and Visual Prompting by 3.57 and 35.40 points, respectively. Compared with SD-VICL, our method shows clear gains on structured categories such as Building (+9.63 IoU) and Car (+9.33 IoU), and also improves Surf and Tree by 2.89 and 0.99 points. These results indicate that the proposed semantic-aware feature matching is effective for structured urban objects and fine-grained remote sensing targets. The qualitative results in Figure 4 show the same trend: Visual Prompting produces unreliable masks, while our method better preserves object completeness and category boundaries, especially for Building and Car.

Single-object detection. Table 2 reports the single-object detection results on HRSC. Our method achieves the best performance across all metrics, reaching 68.10 mIoU, 74.01 AP_{50}, 47.79 AP_{75}, and 46.18 $mAP_{50:95}$, outperforming SD-VICL by 18.49, 31.29, 27.49, and 23.82 points, respectively. The large gains on AP_{75} and $mAP_{50:95}$ indicate more accurate localization under stricter IoU thresholds. The qualitative results in Figure 4 show the same trend: compared with Visual Prompting and SD-VICL, our method reduces false responses on harbor structures and produces detections that better match the ship extent.

Table 2: **Single-Object Detection on HRSC.** mIoU, AP_{50}, AP_{75}, and $mAP_{50:95}$ are reported for general VICL methods and SRDM.

Model	mIoU	AP_{50}	AP_{75}	$mAP_{50:95}$
Visual Prompting[11]	25.87	20.57	0.27	5.04
SD-VICL[13]	49.61	42.72	20.30	22.36
Ours	**68.10**	**74.01**	**47.79**	**46.18**

Foreground segmentation. Table 3 compares our method with representative few-shot remote sensing segmentation methods under the one-shot setting on the iSAID-5^i benchmark. Our method achieves the best performance on all three folds and obtains a mean mIoU of 64.00, outperforming the strongest baseline ViRefSAM by 10.87 points. It also surpasses ICLIP by 12.60 points in mean mIoU and exceeds MGANet and PI-CLIP by clear margins, indicating that the proposed feature matching and fusion strategy provides stronger query-prompt correspondence for remote sensing objects. The qualitative results in Fig. 4 further show that our method preserves clearer object boundaries and reduces fragmented responses for dense small targets such as vehicles, demonstrating effective and robust one-shot foreground segmentation in remote sensing scenarios.

Table 3: **One-Shot Foreground Segmentation on iSAID-5^i.** Fold-wise and mean mIoU are reported for few-shot remote sensing segmentation methods and SRDM.

Model	fold0	fold1	fold2	mean
PANet[51]	27.56	17.23	24.60	23.13
PFENet[52]	35.84	23.35	27.20	28.80
BAM[53]	39.43	21.69	28.64	29.92
R^2Net[54]	41.22	21.64	35.28	32.71
DMNet[55]	40.72	26.47	40.37	35.85
PI-CLIP[56]	35.54	26.13	37.63	33.10
MGANet[10]	39.90	30.72	44.18	38.27
ICLIP[9]	58.63	38.76	56.82	51.40
HRL[57]	41.50	27.84	34.51	34.62
INSID3[58]	–	–	–	52.10
ViRefSAM[59]	58.13	45.26	56.01	53.13
Ours	**68.89**	**58.58**	**64.53**	**64.00**

4.3 Ablation Studies

Ablation on Feature Fusion and Training Strategy. Table 4 evaluates the feature-fusion design and training strategy on Potsdam. Using only diffusion features obtains 40.74 mIoU, while only DINOv2 features improve the result to 44.63, showing stronger semantic correspondence. Fusing the two features further increases mIoU to 46.69, outperforming the Only SD and Only DINOv2 settings by 5.95 and 2.06 points, respectively. This confirms their complementarity: diffusion features provide structural cues, while DINOv2 features provide semantic matching. With the proposed training strategy, the performance reaches 47.47 mIoU.

Table 4: **Feature Fusion and Training Strategy Ablation.** Per-class IoU and mIoU on Potsdam are reported for variants using diffusion features, DINOv2 features, SACA, and SACA+RRL.

Model	Surf	Building	Low Veg	Tree	Car	Clutter	mIoU
Only SD	52.52	58.94	44.38	33.54	43.56	11.49	40.74
Only Dinov2	52.96	67.13	40.09	37.13	59.56	10.88	44.63
SACA	56.39	67.38	45.42	39.08	60.40	11.44	46.69
SACA+RRL	**57.67**	**68.64**	**46.44**	**39.89**	**60.59**	**11.57**	**47.47**

Table 5: **Temporal Attention Schedule Ablation on HRSC2016.** We compare the proposed temporal schedule with average fusion across denoising steps.

Strategy	mIoU ↑
Average fusion	63.45
Temporal schedule (ours)	**68.10**

Effect of the Temporal Attention Schedule. We evaluate the temporal attention schedule on HRSC2016 against an average-fusion baseline that assigns equal weights to the three attention sources at every denoising step. As shown in Table 5, the proposed temporal schedule improves the mIoU from 63.45 to 68.10, a gain of 4.65 points, demonstrating the benefit of stage-dependent attention weighting during denoising. Average fusion treats the three sources as equally reliable throughout denoising, although their usefulness changes as semantic content emerges and the generated structure stabilizes. The gain therefore validates the intended coarse-to-fine transition from target-branch self-attention to DINOv2 semantic guidance and finally diffusion-based structural transfer.

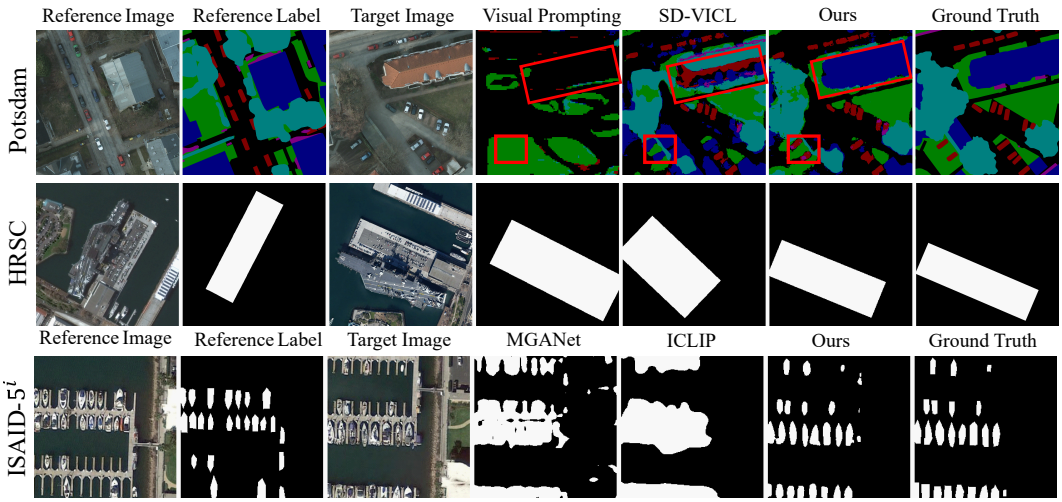


Figure 4: **Qualitative Comparison.** We compare SRDM with Visual Prompting and SD-VICL on Potsdam and HRSC, and with MGANet and ICLIP under the one-shot setting on iSAID-5^i. In the Potsdam examples, red boxes highlight Building and Car regions, where SRDM better preserves object completeness and category boundaries.

Effect of Training on Rotation Robustness. To evaluate rotation robustness, we introduce the *Best-Worst Gap*. For each query, we rotate the prompt image and label at multiple angles, select the best and worst prediction among the rotated prompts, and then compute mIoU over the best-prediction set and the worst-prediction set on the whole dataset. Their difference is the Best-Worst Gap, where a smaller value indicates lower sensitivity to prompt rotation. As shown in Table 6, the proposed training strategy reduces the mIoU Gap from 6.81 to 5.54, with notable reductions for Building (8.66 to 5.80) and Surf (8.29 to 6.69). These results show that the training strategy improves rotation robustness for remote sensing tasks.

Table 6: **Rotation Robustness Ablation.** Per-class IoU and mean Best-Worst mIoU gaps are reported on Potsdam for SACA and SACA+RRL.

Model	Surf	Building	Low Veg	Tree	Car	Clutter	mIoU Gap↓
SACA	8.29	8.66	7.13	7.71	5.95	3.10	6.81
SACA+RRL	6.69	5.80	6.52	6.37	5.18	2.68	5.54

Table 7: **Comparison with Alternative Rotation Strategies on Potsdam.** We compare RRL with rotation augmentation during training and rotation ensembling during inference.

Method	mIoU ↑
SACA	46.69
SACA + rotation augmentation	44.22
SACA + rotation ensemble	41.49
SACA + RRL (ours)	**47.47**

Comparison with Alternative Rotation Strategies. We compare RRL with two controlled rotation-handling baselines. Rotation augmentation follows the same training pipeline as SACA but rotates the reference image and label with 30% probability, whereas rotation ensembling aggregates predictions from rotated inputs at inference time. As shown in Table 7, both strategies underperform the SACA baseline, while RRL achieves the highest mIoU of 47.47 without requiring rotation ensembling at inference time, improving the SACA baseline by 0.78 points. The lower results of conventional augmentation and ensembling suggest that simply exposing the model to rotated inputs or aggregating rotated predictions does not resolve cross-image feature mismatch. In contrast, RRL explicitly aligns the best- and worst-rotation predictions during training, improving orientation consistency while preserving generation quality.

5 Conclusion

In this paper, we proposed SRDM, a semantic-aware and rotation-robust diffusion framework for example-driven remote sensing task adaptation. SRDM formulates in-context learning as conditional generation and transfers task definitions through cross-image interactions. To address delayed semantic emergence for small objects and geometric inconsistency under arbitrary rotations, it combines semantic-aware cross-attention for reliable alignment with rotation-robust learning for geometric consistency. Experiments on multiple remote sensing benchmarks demonstrate strong one-shot performance against general-domain VICL and representative few-shot methods, highlighting the potential of in-context learning for remote sensing vision. Current limitations include the computational overhead of Stable Diffusion denoising with DINOv2 feature extraction and modality coverage centered on RGB with only a preliminary extension to NIR. Improving efficiency and generalization to multispectral and SAR imagery remains future work.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 42571343. The authors declare no competing interests.

References

[1] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine*, 5(4):8–36, 2017.

[2] Qiangqiang Yuan, Huanfeng Shen, Tongwen Li, Zhiwei Li, Shuwen Li, Yun Jiang, Hongzhang Xu, Weiwei Tan, Qianqian Yang, Jiwen Wang, et al. Deep learning in environmental remote sensing: Achievements and challenges. *Remote sensing of Environment*, 241:111716, 2020.

[3] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018.

[4] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 28–37, 2019.

[5] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

[6] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35: 197–211, 2022.

[7] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27672–27683, 2024.

[8] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021.

[9] Zhiyu Jiang, Ye Yuan, Dandan Ma, Qi Wang, and Yuan Yuan. Implicit clip prior decoupling for few-shot remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[10] Shi-Feng Peng, Guo-Sen Xie, Fang Zhao, Xiangbo Shu, and Qingshan Liu. Multi-granularity aggregation network for remote sensing few-shot segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–14, 2024.

[11] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *Advances in neural information processing systems*, 35: 25005–25017, 2022.

[12] Zhendong Wang, Yifan Jiang, Yadong Lu, Pengcheng He, Weizhu Chen, Zhangyang Wang, Mingyuan Zhou, et al. In-context learning unlocked for diffusion models. *Advances in Neural Information Processing Systems*, 36:8542–8562, 2023.

[13] Trevine Oorloff, Vishwanath Sindagi, Wele Gedara Chaminda Bandara, Ali Shafahi, Amin Ghiasi, Charan Prakash, and Reza Ardekani. Stable diffusion models are secretly good at visual in-context learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23604–23613, 2025.

[14] Jiarui Xu, Yossi Gandelsman, Amir Bar, Jianwei Yang, Jianfeng Gao, Trevor Darrell, and Xiaolong Wang. Improv: Inpainting-based multimodal prompting for computer vision tasks. *arXiv preprint arXiv:2312.01771*, 2023.

[15] Tianyi Yan, Wencheng Han, Xueyang Zhang, Kun Zhan, Cheng-Zhong Xu, Jianbing Shen, et al. Rlgf: Reinforcement learning with geometric feedback for autonomous driving video generation. *Advances in Neural Information Processing Systems*, 38:128659–128684, 2026.

[16] Lei Wang, Senmao Li, Fei Yang, Jianye Wang, Ziheng Zhang, Yuhan Liu, Yaxing Wang, and Jian Yang. Not all parameters matter: Masking diffusion models for enhancing generation ability. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12880–12890. IEEE, 2025.

[17] Xiefan Guo, Xinzhu Ma, Haiyu Zhang, and Di Huang. Ctcal: Rethinking text-to-image diffusion models via cross-timestep self-calibration. *arXiv preprint arXiv:2603.20741*, 2026.

[18] Jun Hao Liew, Hanshu Yan, Daquan Zhou, and Jiashi Feng. Magicmix: Semantic mixing with diffusion models. *arXiv preprint arXiv:2210.16056*, 2022.

[19] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[21] Rui Li, Shunyi Zheng, Ce Zhang, Chenxi Duan, Libo Wang, and Peter M Atkinson. Abcnet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery. *ISPRS journal of photogrammetry and remote sensing*, 181:84–98, 2021.

[22] Bo Yu, Lu Yang, and Fang Chen. Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(9):3252–3261, 2018.

[23] Qi Zhao, Jiahui Liu, Yuewen Li, and Hong Zhang. Semantic segmentation with attention mechanism for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2021.

[24] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[25] Rongtao Xu, Changwei Wang, Jiguang Zhang, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang. Rssformer: Foreground saliency enhancement for remote sensing land-cover segmentation. *IEEE Transactions on Image Processing*, 32:1052–1064, 2023.

[26] Jie Zhang, Mingwen Shao, Yecong Wan, Lingzhuang Meng, Xiangyong Cao, and Shuigen Wang. Boundary-aware spatial and frequency dual-domain transformer for remote sensing urban images segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–18, 2024.

[27] Wenshu Liu, Nan Cui, Luo Guo, Shihong Du, and Weiyin Wang. Desformer: A dual-branch encoding strategy for semantic segmentation of very-high-resolution remote sensing images based on feature interaction and multiscale context fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024.

[28] Feng Gao, Miao Fu, Jingchao Cao, Junyu Dong, and Qian Du. Adaptive frequency enhancement network for remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[29] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented r-cnn for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3520–3529, 2021.

[30] Zikai Xiao, Guoye Yang, Xue Yang, Taijiang Mu, Junchi Yan, and Shimin Hu. Theoretically achieving continuous representation of oriented bounding boxes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16912–16922, 2024.

[31] Xue Yang and Junchi Yan. Arbitrary-oriented object detection with circular smooth label. In *European conference on computer vision*, pages 677–694. Springer, 2020.

[32] Xue Yang, Junchi Yan, Qi Ming, Wentao Wang, Xiaopeng Zhang, and Qi Tian. Rethinking rotated object detection with gaussian wasserstein distance loss. In *International conference on machine learning*, pages 11830–11841. PMLR, 2021.

[33] Xue Yang, Xiaojiang Yang, Jirui Yang, Qi Ming, Wentao Wang, Qi Tian, and Junchi Yan. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. *Advances in Neural Information Processing Systems*, 34:18381–18394, 2021.

[34] Yuxuan Li, Qibin Hou, Zhaohui Zheng, Ming-Ming Cheng, Jian Yang, and Xiang Li. Large selective kernel network for remote sensing object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16794–16805, 2023.

[35] Xinhao Cai, Qiuxia Lai, Yuwei Wang, Wenguan Wang, Zeren Sun, and Yazhou Yao. Poly kernel inception network for remote sensing detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 27706–27716, 2024.

[36] Xinbin Yuan, Zhaohui Zheng, Yuxuan Li, Xialei Liu, Li Liu, Xiang Li, Qibin Hou, and Ming-Ming Cheng. Strip r-cnn: large strip convolution for remote sensing object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 12259–12267, 2026.

[37] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[38] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2023.

[39] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023.

[40] Zhong Ji, Weilong Cao, Yan Zhang, Yanwei Pang, Jungong Han, and Xuelong Li. Underlying semantic diffusion for effective and efficient in-context learning. *arXiv preprint arXiv:2503.04050*, 2025.

[41] Yanpeng Sun, Qiang Chen, Jian Wang, Jingdong Wang, and Zechao Li. Exploring effective factors for improving visual in-context learning. *IEEE Transactions on Image Processing*, 2025.

[42] Wei Suo, Lanqing Lai, Mengyang Sun, Hanwang Zhang, Peng Wang, and Yanning Zhang. Rethinking and improving visual prompt selection for in-context learning segmentation. In *European Conference on Computer Vision*, pages 18–35. Springer, 2024.

[43] Chengming Xu, Chen Liu, Yikai Wang, Yuan Yao, and Yanwei Fu. Towards global optimal visual in-context learning prompt selection. *Advances in Neural Information Processing Systems*, 37:74945–74965, 2024.

[44] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. What makes good examples for visual in-context learning? *Advances in Neural Information Processing Systems*, 36:17773–17794, 2023.

[45] Jiahao Zhang, Bowen Wang, Liangzhi Li, Yuta Nakashima, and Hajime Nagahara. Instruct me more! random prompting for visual in-context learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2597–2606, 2024.

[46] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12469–12478, 2024.

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[48] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[49] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in neural information processing systems*, 35:16664–16678, 2022.

[50] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986.

[51] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 9197–9206, 2019.

[52] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 44(2):1050–1065, 2020.

[53] Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. Learning what not to segment: A new perspective on few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8057–8067, 2022.

[54] Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. Global rectification and decoupled registration for few-shot segmentation in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–11, 2023.

[55] Hanbo Bi, Yingchao Feng, Zhiyuan Yan, Yongqiang Mao, Wenhui Diao, Hongqi Wang, and Xian Sun. Not just learning from others but relying on yourself: A new perspective on few-shot segmentation in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–21, 2023.

[56] Jin Wang, Bingfeng Zhang, Jian Pang, Honglong Chen, and Weifeng Liu. Rethinking prior information generation with clip for few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3941–3951, 2024.

[57] Xin He, Yun Liu, Yong Zhou, Henghui Ding, Jiaqi Zhao, Bing Liu, and Xudong Jiang. Hierarchical relation learning for few-shot semantic segmentation in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025.

[58] Claudia Cuttano, Gabriele Trivigno, Christoph Reich, Daniel Cremers, Carlo Masone, and Stefan Roth. Insid3: Training-free in-context segmentation with dinov3. In *Third Workshop on Visual Concepts*, 2026.

[59] Hanbo Bi, Yulong Xu, Ya Li, Yongqiang Mao, Boyuan Tong, Chongyang Li, Chunbo Lang, Wenhui Diao, Hongqi Wang, Yingchao Feng, et al. Virefsam: Visual reference-guided segment anything model for remote sensing segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2026.

A Additional Implementation Details

All training and evaluation were conducted on a single NVIDIA A40 GPU with 48GB GPU memory and 64GB system memory. During training, we enabled gradient checkpointing to reduce GPU memory consumption. To train the rotation-robust components, we randomly sampled 200 images from the Potsdam training set and optimized the model for 200 iterations. Only images from the training split were used for model training and pseudo-label generation; no test images or test labels were involved in training, thereby preventing test-set leakage. The weights used for iSAID-5^i were trained on Potsdam and directly evaluated on iSAID-5^i without dataset-specific retraining.

Continuous-rotation construction. RRL is not restricted to the four predefined angles used for hard-view selection. We first sample an angle θ from a continuous distribution, rotate the image, and inward-crop the valid image region to avoid black-border artifacts that could reveal the rotation angle. From the same cropped image $\widetilde{I} = \mathrm{Crop}(R_\theta(I))$, we construct four views $\{R_{k\pi/2}(\widetilde{I})\}_{k=0}^{3}$ and select the highest- and lowest-scoring views for the RRL loss. These views contain identical visual content and differ only in orientation, allowing the loss to measure rotation sensitivity without confounding changes in visible content. Since their orientations relative to the original image are $\theta + k\pi/2$, with θ sampled continuously, the resulting training pairs cover arbitrary continuous rotations rather than only a fixed set of angles.

B Additional Experiments

B.1 Natural Scene

To further examine whether the proposed feature-fusion design is specific to remote sensing data or can generalize to conventional natural images, we conduct an additional foreground segmentation experiment on the PASCAL-5^i. PASCAL-5^i is built from PASCAL VOC classes and splits the 20 semantic categories into four folds.

For this experiment, we use the feature-fusion-only variant of our method, which applies the proposed feature-fusion improvement without the additional RRL training strategy. As shown in Table 8, our method achieves the best result on all four splits and obtains a mean mIoU of 68.72. Compared with SD-VICL and Visual Prompting, the proposed feature fusion improves the mean mIoU by 16.04 and 12.57 points, respectively, with consistent gains across all splits. These results indicate that the benefit of combining diffusion-model features with DINOv2 features is not limited to remote sensing scenes; it also provides robust semantic correspondence for natural-scene foreground segmentation.

Table 8: **Natural-Scene Foreground Segmentation.** Fold-wise and mean mIoU are reported on PASCAL-5^i for Visual Prompting, SD-VICL, and the feature-fusion-only variant of SRDM.

Model	splite0	splite1	splite2	splite3	Mean
Visual Prompting	60.54	56.23	57.47	50.34	56.15
SD-VICL	54.33	52.53	51.78	52.09	52.68
SACA	**70.38**	**69.99**	**69.11**	**65.39**	**68.72**

B.2 Dinov2 Input Image Dim

We isolate the effect of the DINOv2 input resolution under the *Only DINOv2* setting, where only DINOv2 features are used as the query and key representations for attention matching. In our implementation, the self-attention modules in the upsampling path of the Stable Diffusion U-Net are replaced at three spatial resolutions, corresponding to feature maps of 16×16, 32×32, and 64×64. Therefore, the DINOv2 feature resolution controls how well the external semantic features can be aligned with the internal attention maps of the diffusion model.

As shown in Table 9, increasing the input image size from 224×224 to 896×896 consistently improves the mIoU from 35.74 to 44.63. This gain suggests that higher-resolution DINOv2 features provide more accurate spatial correspondences, especially for small or structured categories such as Car, where the IoU increases substantially from 19.13 to 59.56. The best performance is achieved at 896×896, whose 64×64 feature map matches the finest modified attention resolution in the U-Net.

However, further increasing the input size to 1792×1792 produces a 128×128 DINOv2 feature map and degrades the mIoU to 39.55. We attribute this drop to the resolution mismatch between the overly dense DINOv2 features and the U-Net attention maps, which requires additional aggregation or interpolation and may introduce noisy, over-localized correspondences. Based on this trade-off, we use 896×896 as the default DINOv2 input resolution.

Table 9: **DINOv2 Input Resolution Ablation.** Feature-map resolution, per-class IoU, and mIoU are reported on Potsdam under the *Only DINOv2* setting.

Image Dim	Feature Dim	Surf	Building	Low Veg	Tree	Car	Clutter	mIoU
224×224	16×16	48.52	68.85	36.16	30.68	19.13	11.11	35.74
448×448	32×32	51.80	**69.55**	39.68	36.75	47.29	10.94	42.67
896×896	64×64	**52.96**	67.13	**40.09**	**37.13**	**59.56**	**10.88**	**44.63**
1792×1792	128×128	48.58	62.99	35.62	27.64	52.23	10.22	39.55

B.3 Reference-Selection Sensitivity

We evaluate reference-selection sensitivity on HRSC2016 using RemoteCLIP-ranked, random, and bottom-30% reference policies with 1, 3, and 5 references. RemoteCLIP ranks candidate references by the cosine similarity between their image embeddings and the target-image embedding, while the bottom-30% policy randomly samples from the lowest 30% of this ranking. All settings use the same checkpoint and evaluation protocol, and multiple references are jointly used in a single inference pass.

Table 10: **Reference-Selection Sensitivity on HRSC2016.** Results are reported for different reference-selection policies and numbers of references.

Reference Policy	Reference Number	mIoU	AP_{50}	AP_{75}	$mAP_{50:95}$
RemoteCLIP	1	**68.10**	**74.01**	**47.79**	**46.18**
RemoteCLIP	3	72.91	89.47	60.53	54.47
RemoteCLIP	5	73.94	90.13	63.16	55.79
Random reference	1	61.76	68.42	43.42	41.12
Random reference	3	69.77	83.55	51.32	49.01
Random reference	5	71.87	87.50	57.89	52.89
Bottom-30% reference	1	57.41	63.16	32.24	34.41
Bottom-30% reference	3	67.02	76.32	53.29	47.17
Bottom-30% reference	5	72.16	88.12	57.24	53.36

As shown in Table 10, RemoteCLIP yields the strongest result when only one reference is available, confirming the benefit of similarity-based reference selection. The performance gap narrows as more references are provided: with five references, random and bottom-30% sampling reach 71.87 and 72.16 mIoU, respectively. This trend indicates that multi-reference prompting reduces sensitivity to the quality of any individual reference.

B.4 Modality Coverage and Limitations

Table 11: **RGB and NIR Modality Evaluation on Potsdam.** We report mIoU for different methods using RGB or near-infrared (NIR) inputs.

Method	mIoU ↑
Visual Prompting-RGB	12.07
SD-VICL-RGB	43.90
SD-VICL-NIR	36.67
Ours-NIR	44.19
Ours-RGB	**47.47**

We further evaluate modality coverage on the NIR imagery in Potsdam using the same reference-conditioned protocol as in the RGB setting. As shown in Table 11, SRDM achieves 44.19 mIoU with NIR inputs, outperforming SD-VICL-NIR by 7.52 points, while remaining 3.28 points below SRDM-RGB. These results demonstrate that SRDM can transfer to NIR imagery, although the gap relative to RGB indicates room for further improvement in multispectral adaptation.

B.5 Runtime and Large-Scale Practicality

We evaluate inference cost on a single NVIDIA A40 by separately varying the number of denoising steps, the DINOv2 input resolution, and the number of references. Unless otherwise specified, the default configuration uses 50 denoising steps, a DINOv2 input resolution of 896×896, and one reference.

Table 12: **Denoising-Step Runtime Analysis.** The DINOv2 input resolution is fixed at 896×896, and one reference is used.

Denoising steps T	Time/image (s)	Peak VRAM (GB)
10	6.27	14.92
20	11.26	14.92
30	16.14	14.92
40	21.04	14.92
50 (default)	25.94	14.92

Table 13: **DINOv2-Resolution Runtime Analysis.** The denoising process uses 50 steps, and one reference is used.

DINOv2 input	Time/image (s)	Peak VRAM (GB)
224×224	25.11	14.92
448×448	25.22	14.92
896×896 (default)	25.94	14.92
1792×1792	34.38	14.92

Table 14: **Reference-Count Runtime Analysis.** The denoising process uses 50 steps, and the DINOv2 input resolution is fixed at 896×896.

Reference number	Time/image (s)	Peak VRAM (GB)
1 (default)	25.94	14.92
2	41.79	21.79
3	58.20	30.69
4	75.80	37.57
5	93.07	44.47

As shown in Tables 12–14, the default configuration processes one image in 25.94 seconds and uses 14.92 GB of peak reserved GPU memory. Runtime scales approximately linearly with the number of denoising steps, whereas memory remains constant from 10 to 50 steps because the same model and denoising buffers are reused. Increasing the DINOv2 resolution has a limited effect up to 896×896, while 1792×1792 mainly increases latency. In contrast, using additional references increases both runtime and memory, reaching 93.07 seconds and 44.47 GB with five references. These results quantify the efficiency–memory trade-off and show that reducing the number of denoising steps provides a direct option for latency-sensitive large-scale deployment.

B.6 Image-level Statistical Analysis

To quantify variation across test images, we report image-level statistics on the HRSC and Potsdam datasets. For HRSC, the image-level score is foreground IoU. For Potsdam, the image-level score is mIoU computed for each image by averaging class IoUs over classes whose union is nonzero in

that image. We report the mean, standard deviation, standard error of the mean (SEM), variance, and bootstrap 95% confidence interval of the image-level mean. These statistics complement the pixel-aggregated metrics in the main paper by measuring image-sampling uncertainty rather than training-seed variability.

Table 15: **Image-level statistical analysis on HRSC and Potsdam.**Mean, standard deviation, SEM, and confidence intervals are reported in percentage points. Variance is computed on the original $[0, 1]$ metric scale. The 95% confidence intervals are estimated by bootstrap resampling over test images.

Dataset	Model	Mean $\pm$ Std	SEM	Var.	95% CI
HRSC	Visual Prompting	25.87 ± 20.55	1.67	0.0422	[22.58, 29.16]
	SD-VICL	49.61 ± 31.99	2.59	0.1024	[44.48, 54.66]
	Ours	$\mathbf{68.10 \pm 23.27}$	**1.89**	0.0542	**[64.29, 71.72]**
Potsdam	Visual Prompting	8.79 ± 4.51	0.20	0.0020	[8.41, 9.19]
	SD-VICL	35.98 ± 10.90	0.49	0.0119	[35.00, 36.92]
	Ours	$\mathbf{39.36 \pm 11.35}$	0.51	0.0129	**[38.35, 40.34]**

As shown in Table 15, SRDM maintains clear improvements under image-level averaging. On HRSC, SRDM improves the mean foreground IoU by 18.49 points over SD-VICL and 42.23 points over Visual Prompting. On Potsdam, SRDM improves the image-level mIoU by 3.38 points over SD-VICL and 30.57 points over Visual Prompting. The bootstrap confidence intervals of SRDM do not overlap with those of the two baselines on either dataset, providing image-level evidence that the gains are not an artifact of pixel aggregation. We further perform a paired per-image comparison between SRDM and the strongest baseline SD-VICL using the same image-level metric. SRDM obtains a higher score on **76.32%** of HRSC test images and **72.02%** of Potsdam test images, showing that the improvement is not concentrated in a small number of samples. The image-level mIoU values on Potsdam are lower than the pixel-aggregated mIoU values in Table 1 because they follow a different averaging protocol: each image contributes equally, and each image-level mIoU averages only classes with nonzero union. Nevertheless, the relative ranking remains consistent with the main results.

B.7 Additional Qualitative Examples

We provide additional qualitative examples on the three evaluated datasets, including Potsdam for semantic segmentation, HRSC for single-object detection, and ISAID-5^i for foreground segmentation.

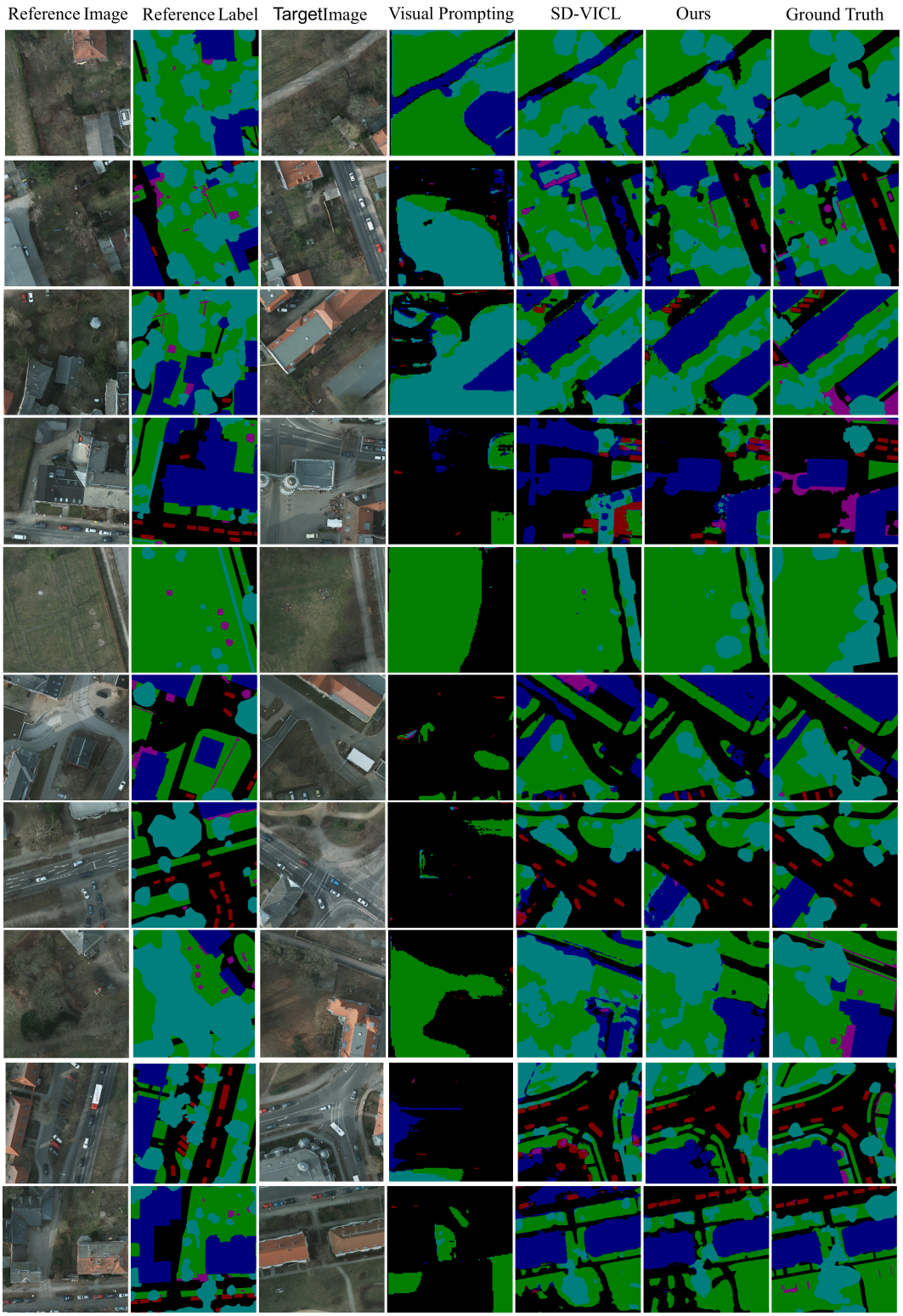


Figure 5: Qualitative comparison on the Potsdam dataset

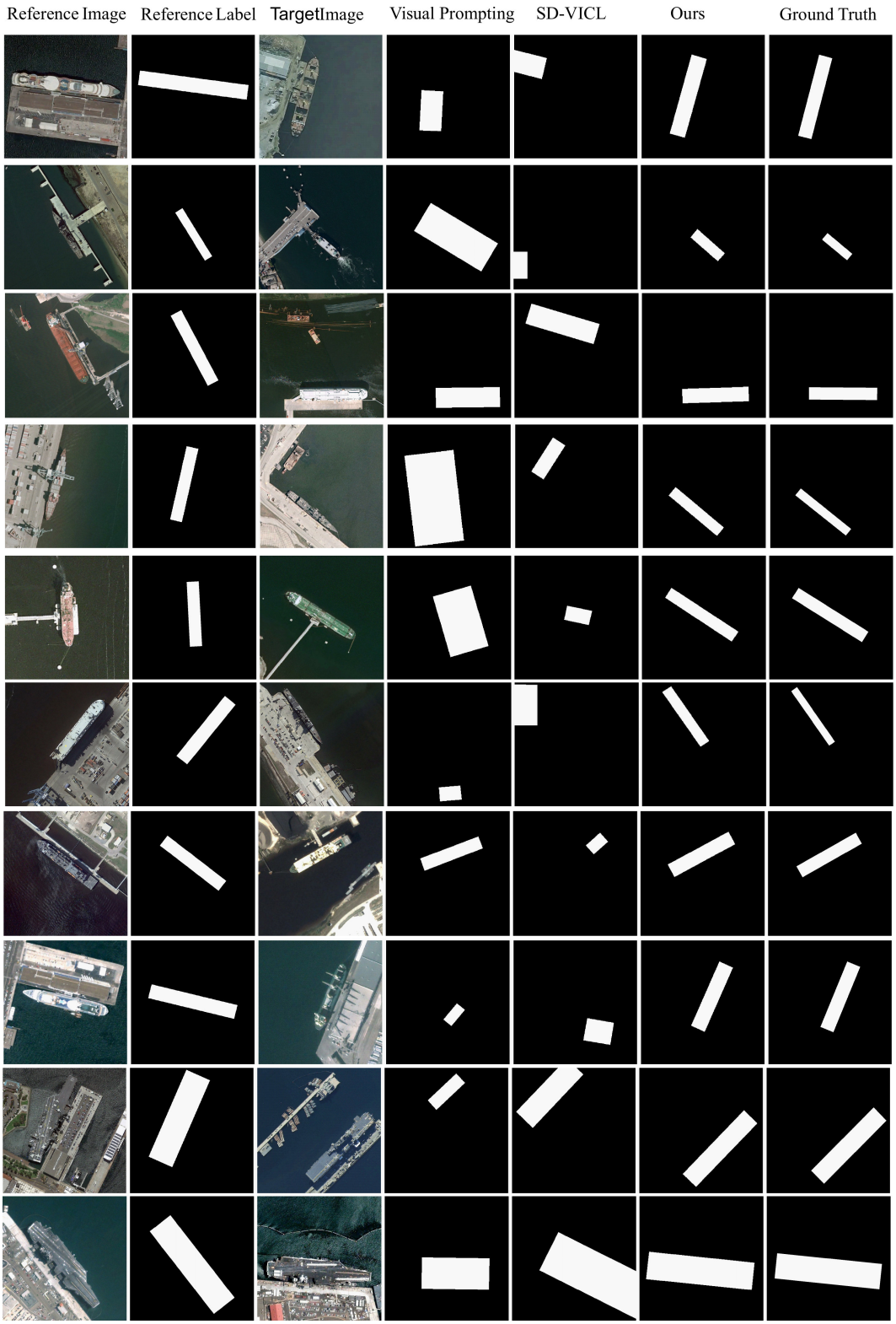


Figure 6: Qualitative comparison on the HRSC dataset

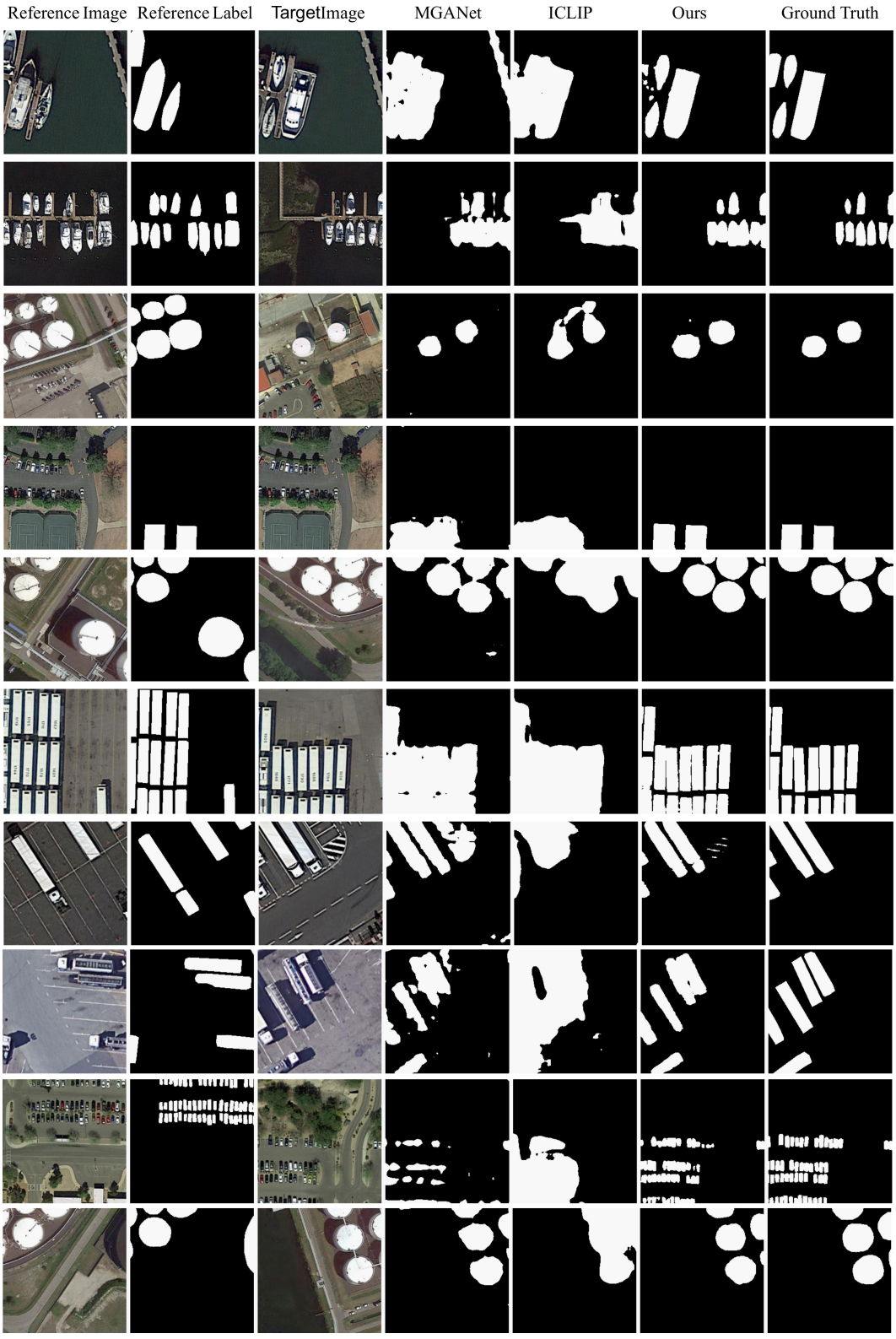


Figure 7: Qualitative comparison on the ISAID-5^i dataset

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: The abstract and introduction state the proposed SRDM framework, semantic-aware cross-attention, and rotation-robust learning strategy.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss limitations in the Conclusion.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

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 present theoretical theorems or formal proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

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 provided experimental details in the experimental section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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: [No]

Justification: All experiments use publicly available datasets (Potsdam, HRSC2016, iSAID-5^i, and PASCAL-5^i). The code is currently being organized and documented, and will be publicly released once it is ready.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

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: Sec. 4.1 reports datasets, data splits or preprocessing protocols, metrics, optimizer, learning rate schedule, batch size, training iterations, diffusion steps, input resolutions, reference selection, and hardware.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

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: [Yes]

Justification: We report image-level statistical analysis in Appendix B.6. For HRSC and Potsdam, we provide mean, standard deviation, SEM, variance, and bootstrap 95% confidence intervals over test images. We also include a paired per-image comparison with the strongest baseline SD-VICL, where SRDM achieves higher image-level scores on 76.32% of HRSC test images and 72.02% of Potsdam test images. These statistics quantify image-sampling uncertainty under the fixed evaluation setting.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

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: [Yes]

   Justification: We have reported relevant information in the experimental details of the main text and supplementary materials.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

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 study uses publicly available remote sensing benchmarks and does not involve human subjects or private personal data; we follow the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [No]

    Justification: Although our method can have positive impacts such as reducing annotation costs, assisting environmental monitoring, and disaster response, we did not discuss it.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

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 paper does not release scraped datasets or high-risk general-purpose generative models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

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: [No]

    Justification: All existing assets used in this work are properly credited by citing their original papers, and they are used in accordance with their academic-use terms. However, the paper does not explicitly list the license name of each asset.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [No]

    Justification: The paper does not introduce new datasets. The code and trained weights are being organized and documented, and will be released together with usage instructions once they are ready.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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: The paper does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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: The paper does not involve crowdsourcing or human subjects research, so IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

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: The core method does not use LLMs as an important, original, or non-standard component; any language-model assistance, if used, was limited to writing or formatting and did not affect the methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.