CHRF-SAM3: CONSTRAINED HETEROGENEOUS REPRESENTATION FUSION FOR CROSS-DOMAIN FEW-SHOT ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-domain few-shot visual learning aims to adapt models to unseen visual domains with only a small number of labeled samples. Benefiting from large-scale pretraining, SAM3, as a new-generation vision foundation model, exhibits strong zero-shot generalization, yet it struggles to effectively adapt to novel cross-domain visual distributions under extremely limited supervision. To address this issue, we introduce DINOv3, which is heterogeneous from SAM3 and yields multi-scale visual representations learned via a distinct pretraining paradigm. Specifically, we adopt its convolutional variant. Building upon this, we develop Convolutional Residual Fusion (CRF) and Layer-Shared Conditional LoRA (LSC-LoRA) to inject the heterogeneous multi-scale representations into the visual encoder and enable conditional low-rank adaptation. On the decoder side, we employ SoMA to constrain low-rank parameter updates. Extensive experiments demonstrate that our method achieves state-of-the-art performance across multiple benchmarks.

1 INTRODUCTION

Cross-domain few-shot visual learning aims to adapt models to unseen visual domains with only a small number of labeled samples (Xiong, 2023; Wang et al., 2022). In recent years, vision-language models (VLMs), pretrained on large-scale data, have demonstrated strong capabilities in cross-domain few-shot visual tasks (Liu et al., 2026; Feng et al., 2026; Ma et al., 2026). As a new-generation vision foundation model, SAM3 integrates concept recognition, object detection, and instance segmentation into a unified perception framework (Carion et al., 2026).

Powered by the unified perception framework, SAM3 exhibits excellent zero-shot generalization capability (Carion et al., 2026). However, the SAM family still struggles to effectively learn domain-specific visual knowledge with only a small set of target-domain annotations, especially under severe domain shifts (Zhang et al., 2024b; Wallace et al., 2025). Existing few-shot methods for SAM predominantly focus on prompt-side adaptation, such as constructing or learning visual prompts by leveraging support samples, and rarely directly modify the core visual representations of the model (Sun et al., 2024; Xu et al., 2025b; Tsai et al., 2026).

Heterogeneous representation fusion offers a promising direction for addressing the aforementioned problem. By introducing a visual foundation model whose training paradigm and network architecture differs from those of SAM3, we can expand its visual representation space while barely altering its pretrained weights, without requiring large-scale retraining.

Motivated by this idea, we introduce DINOv3 with a self-supervised pretraining paradigm, adopt its convolutional architecture variant as the heterogeneous representation source, and freeze its pretrained weights during training (Siméoni et al., 2025). It delivers complementary heterogeneous visual representations to the SAM3 encoder. Specifically, Convolutional Residual Fusion (CRF) directly injects DINOv3 features into the last three layers of the SAM3 encoder. It supplements and reinforces target-domain visual responses during the final formation phase of high-level representations. Layer-Shared Conditional LoRA (LSC-LoRA) operates on a wider span of the high-level encoding pipeline through Layer-Shared Conditional Mapping (LSCM). It transforms DINOv3 multi-scale representations into image-dependent conditioning features and feeds them into low-rank branches, allowing heterogeneous information to continuously contribute to the construction and refinement of SAM3's high-level visual representations. For the decoder, we adopt SoMA for

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

low-rank adaptation of its linear layers (Yun et al., 2025) , enabling it to accommodate the new visual representations produced by heterogeneous encoder fusion while preserving pretrained knowledge. Together, these modules form a full pipeline from encoder to decoder, enabling layered fusion and constrained adaptation of heterogeneous visual representations within SAM3.

To evaluate the effectiveness of our method, we conduct extensive experiments on two representative visual tasks: object detection and semantic segmentation. Our method achieves state-of-the-art performance on the CD-FSOD (Xiong, 2023), ODinW-13 (Li et al., 2022), and PASCAL-5^i (Shaban et al., 2017) benchmarks. In addition, we establish a cross-domain few-shot segmentation evaluation protocol on eight datasets spanning diverse visual domains, including medical imaging, remote sensing, and underwater imagery, and evaluate the proposed method under multiple SAM3 prompting paradigms.

The main contributions of this work are summarized as follows:

1. We propose a constrained heterogeneous representation fusion framework for adapting SAM3 to cross-domain few-shot visual tasks, which enhances the model's ability to learn visual knowledge from novel domains.
2. To achieve cross-architecture fusion of heterogeneous representations, we integrate a frozen convolutional variant of DINOv3 into the visual encoder and devise the CRF and LSC-LoRA modules. For the decoder, we directly adopt SoMA for fine-tuning.
3. We select two classic visual tasks, detection and segmentation, to validate the effectiveness of our method. The proposed method achieves state-of-the-art results on CD-FSOD, ODinW-13 and PASCAL-5^i. We further establish few-shot segmentation evaluation settings under multiple SAM3 prompting conditions on eight cross-domain datasets.

2 Related Work

Few-Shot Visual Tasks. Few-shot visual learning aims to rapidly learn novel categories or adapt to new scenarios using only a limited number of labeled samples (Wang et al., 2020; Shaban et al., 2017). Early studies mainly explored meta-learning, metric learning, and feature enhancement, while more recent work has increasingly explored open-vocabulary models, vision foundation models, and retrieval-based or generative data augmentation (Kang et al., 2019; Wang et al., 2019; Tian et al., 2020; Han & Lim, 2024; Zhang et al., 2023; Nguyen Vu et al., 2025; Pan et al., 2025).

Cross-domain few-shot visual learning further considers the distribution gap between source and target domains (Xu et al., 2025a). CD-FSOD is a representative benchmark for cross-domain few-shot object detection, providing a unified evaluation setting that covers heterogeneous target domains such as natural images, remote sensing imagery, underwater scenes, and industrial scenarios (Xiong, 2023; Fu et al., 2024). Building on this benchmark, subsequent studies have explored different directions, including feature adaptation, vision foundation model integration, instance retrieval, and generative data augmentation (Liu et al., 2026; Huang et al., 2025; Li et al., 2026).

Vision Foundation Models. Vision foundation models acquire generalizable visual representations through large-scale pretraining and have increasingly become an important basis for few-shot visual adaptation (Radford et al., 2021; Li et al., 2022; Caron et al., 2021; Kirillov et al., 2023). The DINO series learns transferable visual representations through large-scale self-supervised learning, with DINOv3 further enhancing dense visual representations (Caron et al., 2021; Siméoni et al., 2025). The SAM series establishes a general promptable visual perception paradigm, while the latest SAM3 introduces concept prompting and further extends unified visual perception capabilities (Kirillov et al., 2023; Carion et al., 2026). In this work, we adapt SAM3 to cross-domain few-shot visual tasks and introduce DINOv3 as an auxiliary visual encoder to leverage the heterogeneous visual representations provided by the two pretrained models.

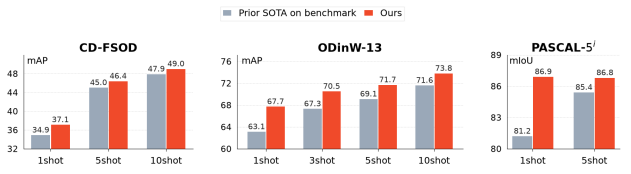

Figure 1: Comparison with prior state-of-the-art methods on CD-FSOD, ODinW-13 and PASCAL-5^i.

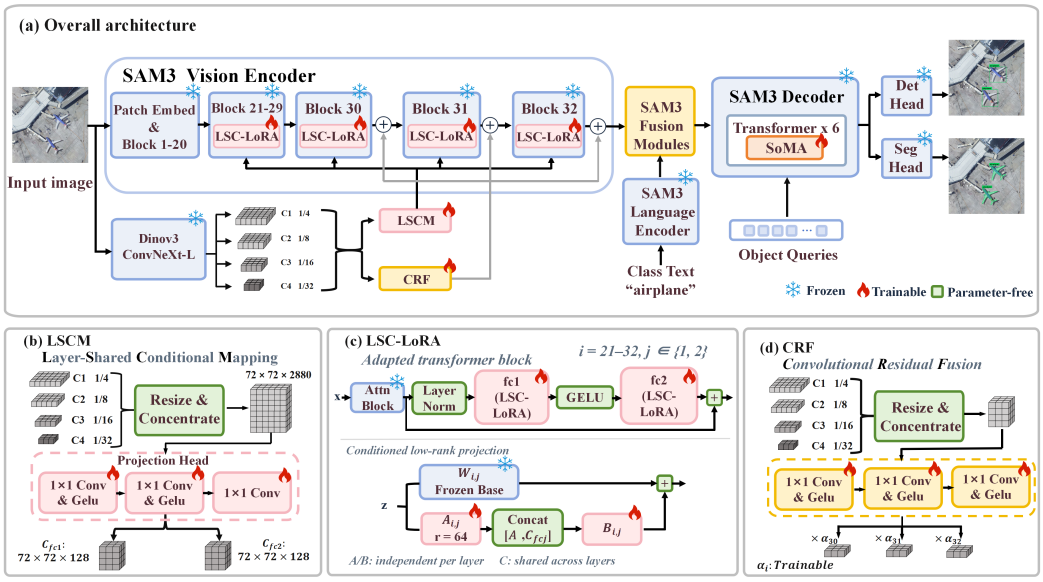

Figure 2: **Overview of CHRF-SAM3.** CHRF-SAM3 improves cross-domain few-shot adaptation through constrained heterogeneous representation fusion. DINOv3 multi-scale convolutional representations are incorporated into the SAM3 vision encoder through CRF and LSCM-guided LSC-LoRA, while SoMA is employed for low-rank adaptation of the SAM3 decoder.

Parameter-Efficient Fine-Tuning. Parameter-efficient fine-tuning (PEFT) enables efficient adaptation of large pretrained models by restricting the number of trainable parameters (Jia et al., 2022; Chen et al., 2022; Lian et al., 2022; Hu et al., 2021). Among existing approaches, LoRA parameterizes weight updates through low-rank decomposition and has been widely adopted for adapting visual models (Hu et al., 2021). Beyond parameter efficiency, prior studies have shown that constraining parameter updates to low-rank spaces can also reduce disturbance to pretrained knowledge and improve out-of-domain generalization (Hu et al., 2024; Biderman et al., 2024). Recent studies have further exploited the singular spectrum of pretrained weights to construct more targeted low-rank adaptation spaces (Meng et al., 2024; Yun et al., 2025). As a representative approach, SoMA preserves the principal singular components while adapting the minor singular components according to the spectral structure of pretrained weights, thereby reducing disruption to the generalizable knowledge encoded in the pretrained model (Yun et al., 2025). Such methods provide a promising direction for balancing target-domain adaptation and pretrained knowledge preservation under cross-domain few-shot settings.

3 Method

As illustrated in Fig. 2, CHRF-SAM3 builds upon pretrained SAM3 (Carion et al., 2026) and introduces a frozen DINOv3 ConvNeXt-L (Siméoni et al., 2025) to provide heterogeneous representations for cross-domain few-shot adaptation. In the vision encoder, the DINOv3 representations are incorporated into SAM3 through Convolutional Residual Fusion described in Sec. 3.1 and LSCM-guided LSC-LoRA introduced in Sec. 3.2, enabling constrained heterogeneous representation fusion. In the decoder, we employ SoMA (Yun et al., 2025) as described in Sec. 3.3 for low-rank adaptation, while the detection and segmentation heads remain frozen.

3.1 Convolutional Residual Fusion

To introduce heterogeneous visual representations complementary to the features of SAM3, we employ a frozen DINOv3 ConvNeXt-L to extract multi-scale features. Given an input image I, the outputs of its four stages are denoted as

$$D_s = \mathcal{E}^{(s)}_{\text{DINO}}(I), \qquad s \in \{1, 2, 3, 4\}. \tag{1}$$

Here, s denotes the stage index, $\mathcal{E}^{(s)}_{\text{DINO}}$ denotes the feature extraction mapping of the frozen DINOv3 ConvNeXt-L at the s-th stage, and D_s denotes the corresponding feature map.

Since the features from different stages have different spatial resolutions, CRF first aligns them to the spatial grid of the SAM3 visual features and concatenates them along the channel dimension. The fused heterogeneous representation is then projected into convolutional residuals injected into ViT blocks 30–32:

$$\{R_i\}_{i=30}^{32} = \mathcal{P}_C\big(\mathrm{Concat}_{s=1}^{4}\,\rho(D_s)\big)\,. \tag{2}$$

Here, $\rho(\cdot)$ denotes spatial alignment, Concat denotes concatenation along the channel dimension, $\mathcal{P}_C$ is a learnable convolutional projection for residual features, and R_i denotes the residual corresponding to the i-th ViT block.

For the i-th ViT block, we first retain the original SAM3 computation:

$$\widehat{Z}_i = \mathcal{B}_i(Z_{i-1}), \tag{3}$$

where Z_{i-1} denotes the input to the i-th block, $\mathcal{B}_i$ denotes the original mapping of this block, and $\widehat{Z}_i$ is the block output before convolutional residual injection.

The corresponding convolutional residual is then injected into the block output:

$$Z_i = \widehat{Z}_i + \alpha_i R_i, \qquad i \in \{30, 31, 32\}, \tag{4}$$

where α_i is a learnable fusion coefficient initialized to 0.1. CRF performs residual fusion only after the complete ViT block and does not modify the original attention or feed-forward computations inside the block.

3.2 Layer-Shared Conditional LoRA

To further incorporate heterogeneous DINOv3 representations into the low-rank adaptation of the SAM3 vision encoder, we design Layer-Shared Conditional LoRA (LSC-LoRA). Instead of relying solely on the ViT features for adaptation, LSC-LoRA introduces image-dependent convolutional conditions into the low-rank branches of the feed-forward networks.

To construct these conditions, LSC-LoRA employs Layer-Shared Conditional Mapping (LSCM) to transform the four-level DINOv3 features $\{D_s\}_{s=1}^{4}$. Specifically, the multi-scale features are first spatially aligned and concatenated along the channel dimension. The fused heterogeneous representation is then projected by $\mathcal{P}$ into two condition feature maps:

$$(C_{\mathrm{fc1}}, C_{\mathrm{fc2}}) = \mathcal{P}\big(\mathrm{Concat}_{s=1}^{4}\,\rho(D_s)\big)\,. \tag{5}$$

The two condition feature maps correspond to the two linear mappings in the feed-forward network and provide input-dependent convolutional information for their low-rank updates.

Given the condition features generated by LSCM, LSC-LoRA incorporates the corresponding condition into each low-rank branch. For the j-th frozen linear mapping in the i-th ViT block, where $j \in \{1, 2\}$ denotes the two linear mappings in the feed-forward network, the adapted mapping is formulated as

$$y_i^{(j)} = W_i^{(j)} x_i^{(j)} + b_i^{(j)} + B_i^{(j)} \left[A_i^{(j)} x_i^{(j)}; C_{\mathrm{fc}j}\right]. \tag{6}$$

LSC-LoRA is applied to ViT blocks 21–32. For each input image, the condition feature maps C_{fc1} and C_{fc2} are shared across these blocks, whereas the low-rank parameters $A_i^{(j)}$ and $B_i^{(j)}$ are independently learned for each block. We set the rank of LSC-LoRA to $r = 64$. The residual branch is zero-initialized to preserve the original pretrained mapping at the beginning of training.

3.3 SoMA-Based Decoder Adaptation

In the SAM3 decoder, we adopt the tail singular component adaptation strategy from SoMA (Yun et al., 2025) for low-rank fine-tuning on the 6-layer Transformer backbone, while both the detection head and segmentation head remain frozen. For each pretrained linear weight W_0 in the Transformer layers, we select its smallest $r = 32$ singular components to initialize the low-rank factors:

$$A_0 = \sqrt{\Sigma}_{[-r:]}(V^\top)_{[-r:,:]}, \qquad B_0 = U_{[:,-r:]}\sqrt{\Sigma}_{[-r:]}. \tag{7}$$

The corresponding tail singular components are then separated from the original weight, while the remaining pretrained weights are kept frozen. During training, only the low-rank factors A and B are optimized, so the final linear mapping can be written as

$$y = (W_0 - B_0 A_0 + BA)\,x + b. \tag{8}$$

This adaptation strategy is applied to self-attention projections, visual cross-attention projections, text cross-attention projections, and feed-forward network linear mappings in the decoder Transformer backbone, while the rest of the pretrained parameters remain frozen.

4 EXPERIMENTS

4.1 BENCHMARKS AND BASE MODELS

Object Detection. We evaluate our method on CD-FSOD(Fu et al., 2024) and ODinW-13 (Li et al., 2022). CD-FSOD consists of six visually diverse target domains, including ArTaxOr, Clipart1k, DIOR, DeepFish, NEU-DET, and UODD, and is evaluated under the 1/5/10-shot settings. ODinW-13 contains 13 downstream detection datasets spanning diverse visual domains. We evaluate under the 1/3/5/10-shot settings using three official few-shot data splits generated with split seeds 3, 30, and 300.

Semantic Segmentation. For semantic segmentation, we evaluate on PASCAL-5^i and eight cross-domain segmentation datasets. PASCAL-5^i is a widely adopted benchmark for few-shot semantic segmentation on natural images, where the 20 PASCAL VOC categories are partitioned into four disjoint folds with five categories per fold(Shaban et al., 2017). The cross-domain evaluation includes CWFID, CorrosionCS, DeepCrack, DeepFish, KvasirInstrument, PAXRay4, PST900, and Potsdam, covering diverse visual domains such as industrial defects, underwater scenes, medical imaging and remote sensing(Haug & Ostermann, 2014; Bianchi & Hebdon, 2021; Liu et al., 2019; Saleh et al., 2020; Jha et al., 2021; Seibold et al., 2022; Shivakumar et al., 2020; Rottensteiner et al., 2014). We evaluate both 1-shot and 5-shot settings with text and point prompts, where the point-prompt setting considers one, two, and three positive points.

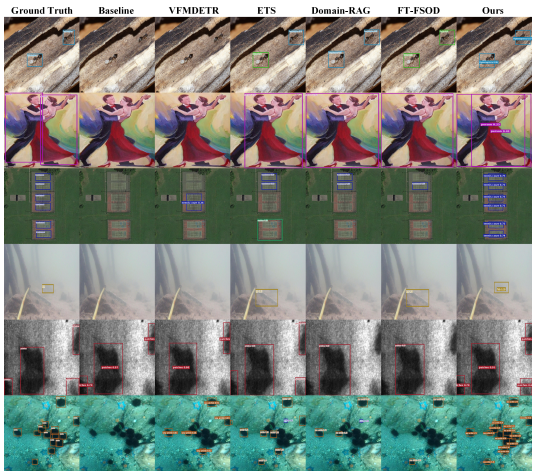

Figure 3: Visualization of detection results on CD-FSOD under the 5-shot setting.

Base Models. All experiments are built upon the official pretrained SAM3 model with PE-L+ as its visual backbone. In addition, we employ DINOv3 ConvNeXt-L pretrained on LVD-1689M as an auxiliary visual encoder. During adaptation, the original SAM3 visual encoder, language encoder, and DINOv3 ConvNeXt-L are kept frozen, while only the proposed adaptation modules and task-specific parameters are optimized.

Evaluation Metrics. For object detection, we adopt the COCO-style AP metric averaged over IoU thresholds from 0.50 to 0.95 with an interval of 0.05. For CD-FSOD, we report the average AP over the six target domains. For ODinW-13, we compute the macro-average over the 13 datasets for each split and report the mean and standard deviation across the three official few-shot data splits. For semantic segmentation, mIoU is used as the primary evaluation metric, while Dice and B-IoU(Cheng et al., 2021) are additionally reported for the cross-domain segmentation experiments.

4.2 COMPARISON WITH COMPETING METHODS

CD-FSOD. We first compare our method with representative cross-domain few-shot object detection approaches on CD-FSOD, including VFMDETR (Liu et al., 2026), ETS (Pan et al., 2025), Domain-RAG (Li et al., 2026), LMP (Wang et al., 2026), and FT-FSOD (Yu et al., 2026). As shown in Tab. 1, our method achieves the best average performance under all 1-shot, 5-shot, and 10-shot settings, with advantages across most target domains, demonstrating robust cross-domain adaptation under limited supervision. For visual comparison, Fig. 3 presents representative detection examples from the six target domains. The examples show that our method produces fewer false positives and missed detections, together with more accurate localization.

ODinW-13. We compare our method with GLIPv2-H (Zhang et al., 2022), GLEE-Pro (Wu et al., 2024), MQ-GLIP-L (Xu et al., 2023), Grounding DINO 1.5 Pro (Ren et al., 2024), SAM3 (Carion et al., 2026), and FT-FSOD (Yu et al., 2026) on ODinW-13. As shown in Tab. 2, our method achieves the best performance across all few-shot settings. Notably, although SAM3 itself already

Table 1: Comparison on CD-FSOD (mAP, %). The best result in each column is shown in **bold**.

	Method	Backbone	ArTaxOr	Clipart1k	DIOR	DeepFish	NEU-DET	UODD	Avg
1-shot	VFMDETR	ResNet-50	26.1	20.1	20.6	24.2	9.1	9.0	18.2
	ETS	Swin-B	28.1	57.5	12.7	40.7	11.7	21.2	28.7
	Domain-RAG	Swin-B	57.2	56.1	18.0	38.0	12.1	20.2	33.6
	LMP	Swin-B	**58.5**	58.6	17.2	36.6	13.3	21.6	34.3
	FT-FSOD	Swin-B	49.1	55.6	24.6	42.7	15.5	22.1	34.9
	Ours	PE-L+	39.2	**65.2**	**29.2**	**44.4**	**18.9**	**25.8**	**37.1**
5-shot	VFMDETR	ResNet-50	63.3	45.1	32.1	29.5	19.0	19.6	34.7
	ETS	Swin-B	64.5	60.1	29.3	44.9	23.5	**28.6**	41.8
	Domain-RAG	Swin-B	70.0	59.8	31.5	43.8	24.2	26.8	42.7
	LMP	Swin-B	75.0	60.1	31.3	43.9	25.1	28.3	44.0
	FT-FSOD	Swin-B	76.8	59.4	35.3	45.5	25.2	27.5	45.0
	Ours	PE-L+	**77.5**	**65.4**	**36.3**	**47.7**	**25.6**	25.8	**46.4**
10-shot	VFMDETR	ResNet-50	71.3	49.9	37.8	34.1	23.7	22.1	39.8
	ETS	Swin-B	71.2	61.5	37.5	44.1	26.1	29.8	45.0
	Domain-RAG	Swin-B	73.4	61.1	39.0	41.3	26.3	31.2	45.4
	LMP	Swin-B	75.1	62.6	40.4	44.4	25.7	31.6	46.6
	FT-FSOD	Swin-B	79.2	59.6	41.5	46.3	**28.4**	**32.3**	47.9
	Ours	PE-L+	**81.6**	**67.0**	**42.8**	**46.9**	27.7	27.7	**49.0**

Table 2: Comparison on ODinW-13 (mAP, %). Results are reported as mean ± standard deviation over three official few-shot data splits generated with split seeds 3, 30, and 300. The best result in each column is shown in **bold**.

Method	Backbone	1-shot	3-shot	5-shot	10-shot
GLIPv2-H	Swin-H	61.7±0.5	64.1±0.8	64.4±0.6	65.9±0.3
GLEE-Pro	EVA02-L	59.4±1.5	61.7±0.5	64.3±1.3	65.6±0.4
MQ-GLIP-L[1]	Swin-L	62.4	64.2	65.4	66.6
Grounding DINO 1.5 Pro	EVA02-L	62.4±1.1	66.3±1.0	66.9±0.2	67.9±0.3
SAM3	PE-L+	63.0±2.4	67.3±0.7	69.1±1.1	71.6±0.2
FT-FSOD	Swin-L	63.1±0.6	65.8±0.7	67.6±0.3	68.6±0.6
Ours	PE-L+	$\mathbf{67.7} \pm 0.6$	$\mathbf{70.5} \pm 0.6$	$\mathbf{71.7} \pm 0.5$	$\mathbf{73.8} \pm 0.6$

exhibits strong few-shot adaptation capability, our method consistently improves upon SAM3 with the same PE-L+ base model, with the largest gain observed in the most challenging 1-shot setting. These results demonstrate that the proposed adaptation strategy can effectively exploit extremely limited target-domain supervision and consistently improve the cross-domain adaptation capability of SAM3 across different sample sizes.

Cross-Domain Few-Shot Semantic Segmentation. We evaluate cross-domain adaptation for semantic segmentation on eight target domains with substantial visual distribution differences, comparing against two SAM3 adaptation baselines: LoRA fine-tuning of the decoder and full fine-tuning of the decoder. As shown in Tab. 3, our method consistently outperforms both baselines across different supervision scales and prompt types, with particularly substantial improvements on CWFID, PAXRay4, and PST900, whose visual distributions differ considerably from natural images. The advantage becomes more pronounced as additional target-domain samples are provided, indicating that our method can more effectively exploit limited target-domain supervision. Moreover, consistent improvements are observed under both text and point prompts, suggesting that the learned target-domain representations do not rely on a particular prompting format. Similar trends are also observed for Dice and B-IoU, indicating that the improvements extend to both region prediction and boundary quality.

PASCAL-5^i. We also evaluate our method on PASCAL-5^i, a classical few-shot segmentation benchmark on natural images, to verify whether our method remains effective in natural-image

[1]MQ-GLIP-L reports only single values in the original paper while standard deviations are not providedXu et al. (2023).

Table 3: Comparison on eight cross-domain few-shot semantic segmentation datasets (mIoU, %). In the horizontal header, Fish denotes the DeepFish dataset, while the remaining dataset names retain their original forms; Avg. denotes the average over the eight target datasets. In the vertical method entries, D LoRA and D FFT denote decoder LoRA fine-tuning and decoder full fine-tuning, respectively, while Ours denotes the proposed method. The best result under each setting is shown in **bold**.

	Prompt	Method	CWFID	CorrosionCS	DeepCrack	Fish	Kvasir	PAXRay4	PST900	Potsdam	Avg
1-shot	Text	D LoRA	41.4	30.5	**71.0**	86.5	84.9	41.5	26.3	49.5	53.9
		D FFT	41.3	**31.2**	70.8	86.6	85.8	44.3	27.5	**50.2**	54.7
		Ours	**41.6**	30.5	69.9	**86.7**	**89.7**	**62.8**	**29.6**	49.0	**57.5**
	Point	D LoRA	43.5	33.8	64.3	86.8	61.6	55.6	44.8	53.8	55.5
		D FFT	44.8	34.2	62.9	**86.9**	63.0	55.0	43.0	55.6	55.7
		Ours	**48.6**	**37.0**	**66.6**	86.8	**88.9**	**60.2**	**54.9**	**57.3**	**62.5**
5-shot	Text	D LoRA	41.1	31.6	**71.4**	87.0	85.6	65.4	30.4	63.3	59.5
		D FFT	41.5	32.3	71.2	86.5	86.9	63.1	26.9	62.3	58.8
		Ours	**70.6**	**32.8**	71.2	**89.7**	**91.1**	**84.5**	**41.7**	**71.1**	**69.1**
	Point	D LoRA	44.9	39.3	69.0	87.7	90.1	60.7	55.3	62.1	63.6
		D FFT	44.4	37.0	68.9	87.9	89.5	56.4	50.4	61.7	62.0
		Ours	**70.4**	**42.6**	**72.1**	**90.1**	**93.4**	**75.9**	**59.4**	**65.1**	**71.1**

Table 4: Few-shot semantic segmentation results on PASCAL-5^i (mIoU, %). The best result in each column is shown in **bold**.

	1-shot						5-shot					
Method	5^0	5^1	5^2	5^3	Mean	FB-IoU	5^0	5^1	5^2	5^3	Mean	FB-IoU
Matcher	67.7	70.7	66.9	67.0	68.1	–	71.4	77.5	74.1	72.8	74.0	–
VRP-SAM	73.9	78.3	70.6	65.0	71.9	–	–	–	–	–	–	–
GF-SAM	71.1	75.7	69.2	73.3	72.1	–	81.5	86.3	79.7	82.9	82.6	–
FSSAM	81.6	84.9	81.6	76.0	81.0	89.4	84.1	88.5	83.8	**85.0**	85.4	91.9
FSS-SAM3	76.2	87.5	81.6	79.5	81.2	89.4	76.2	86.7	80.3	81.9	81.3	89.0
Ours	**83.4**	**88.6**	**89.2**	**86.3**	**86.9**	**92.9**	**84.9**	**88.8**	**90.3**	83.3	**86.8**	**93.1**

scenarios. We compare our method with representative few-shot segmentation approaches, including Matcher (Liu et al., 2023), VRP-SAM (Sun et al., 2024), GF-SAM (Zhang et al., 2024a), FSSAM (Xu et al., 2025b), and FSS-SAM3 (Tsai et al., 2026). As shown in Tab. 4, our method achieves new state-of-the-art results under both 1-shot and 5-shot settings and outperforms existing methods on most evaluation entries. These results indicate that the proposed adaptation strategy is effective not only in cross-domain scenarios but also for standard few-shot semantic segmentation on natural images.

4.3 Ablation Studies

To verify the effectiveness of the proposed modules, we conduct ablation experiments across CD-FSOD, ODinW-13, and eight cross-domain semantic segmentation datasets. We start from a baseline that freezes both encoders and adapts only the decoder using plain LoRA with a rank of 32 (Hu et al., 2021). We then progressively introduce CRF and LSC-LoRA, and finally replace the decoder LoRA with SoMA (Yun et al., 2025). CD-FSOD and ODinW-13 are evaluated using average mAP over their respective target domains, while Cross-Domain FSS reports the average mIoU over the eight semantic segmentation datasets. In Tab. 5, C and L denote the addition of CRF and LSC-LoRA, while S denotes replacing LoRA with SoMA in decoder.

Overall, the two backbone-side fusion mechanisms, CRF and LSC-LoRA, account for the majority of performance gains, though their relative contributions differ across benchmarks. CRF yields more pronounced improvements on CD-FSOD, while LSC-LoRA delivers the largest gains on cross-domain semantic segmentation, and the two together produce consistent performance boosts on ODinW-13. These two modules address the core challenge of representation mismatch between the pretrained backbone and target domains from complementary perspectives. CRF augments the

Table 5: Cumulative module ablation on CD-FSOD, ODinW-13, and Cross-Domain FSS. The best result in each setting is shown in **bold**.

Benchmark	Method	1-shot	3-shot	5-shot	10-shot
CD-FSOD Avg. mAP (%)	Baseline	33.6	–	40.1	41.9
	+ C	34.7	–	45.5	48.3
	+ C + L	36.5	–	45.2	48.2
	+ C + L + S	**37.1**	–	**46.4**	**49.0**
ODinW-13 (Seed 30) Avg. mAP (%)	Baseline	65.9	68.0	69.1	71.3
	+ C	66.4	68.9	70.1	72.2
	+ C + L	**67.3**	**70.1**	71.1	72.9
	+ C + L + S	67.1	69.8	**71.2**	**73.0**
Cross-Domain FSS Avg. mIoU (%)	Baseline	53.9	–	59.5	–
	+ C	55.2	–	61.5	–
	+ C + L	57.3	–	69.0	–
	+ C + L+ S	**57.5**	–	**69.1**	–

visual backbone with multi-scale convolutional representations, while LSC-LoRA further closes the remaining domain gap via conditional low-rank adaptation. This finding confirms that backbone-side fusion serves as the primary driver of performance gains.

No single module, however, achieves consistent optimum across all benchmarks and shot settings. The gains of backbone fusion diminish on target domains whose representations are already well covered, LSC-LoRA tends to saturate or slightly fluctuate as supervision increases, and applying SoMA to the decoder also shows limited gains in some settings. This indicates that the benefit of each module is closely tied to target-domain characteristics and the scale of supervision, and no individual design can cover all scenarios.

Nevertheless, applying SoMA to the decoder brings further improvements in many settings, while only minor performance fluctuations are observed in the remaining cases. Overall, its gains are generally smaller than those of the backbone-side modules. This suggests that applying low-rank adaptation along the tail singular directions to the decoder helps it more effectively utilize the new-domain knowledge introduced by the encoder, while avoiding noticeable degradation of the adapted representations, thereby complementing backbone-side fusion.

Table 6: Ablation on the rank of SoMA under the CD-FSOD 5-shot setting.

Rank	Avg. mAP (%)
16	46.1
24	45.9
32	**46.4**
40	46.0
48	46.3

Table 7: Ablation on DINOv3 ConvNeXt scale under the CD-FSOD 10-shot setting.

Scale	Avg. mAP (%)
None (Baseline)	41.9
T	46.0
S	46.7
B	47.5
L (Ours)	**49.0**

The three modules complement each other in their weakest scenarios: backbone fusion resolves the representation gap, conditional adaptation bridges the remaining domain differences, and decoder-side low-rank adaptation improves the utilization of new-domain knowledge. When used jointly, the full model achieves the best or near-best average performance across all benchmarks and shot settings, verifying the effectiveness and complementarity of the proposed modules and achieving competitive cross-domain few-shot adaptation.

Ablation on the rank of SoMA in the decoder. We further investigate the sensitivity of SoMA in the decoder to its rank under the CD-FSOD 5-shot setting, while keeping the remaining model architecture and optimization settings unchanged.

As shown in Tab. 6, the average performance remains relatively stable across ranks from 16 to 48, with only small fluctuations across individual target domains. The difference between the highest and lowest average performance is only 0.5 mAP, indicating that SoMA is not particularly sensitive to the exact rank choice within this range. A moderate rank of 32 provides a balanced configuration and achieves the highest average performance.

Table 8: Effect of different auxiliary CNN backbones on CD-FSOD under the 10-shot setting. Δ denotes the improvement over the Decoder-LoRA baseline.

Auxiliary Backbone	Avg. mAP	Δ
None (Baseline)	41.9	–
ResNet-50	44.9	+3.0
OpenCLIP ConvNeXt-L	44.1	+2.2
PKINet-S	44.5	+2.6
PKINetv2-S	44.0	+2.1
DINOv3 ConvNeXt-L(Ours)	**49.0**	**+7.1**

Table 9: Zero-shot PASCAL-5^i performance after 10-shot target-domain adaptation. Values in parentheses denote changes from the original SAM3.

	Adaptation Domain	
Method	FISH	NEU-DET
SAM3 Original	84.6	
Decoder LoRA	81.0 (-3.6)	83.7 (-0.9)
Decoder Full FT	78.7 (-5.9)	81.2 (-3.4)
Whole-model Full FT	39.5 (-45.1)	84.3 (-0.3)
Ours	**83.1 (-1.5)**	**84.8 (+0.2)**

Ablation on Auxiliary Encoder Scale. We also study the effect of scaling the DINOv3 ConvNeXt auxiliary encoder under the CD-FSOD 10-shot setting. We retain the complete framework with CRF, LSC-LoRA, and SoMA in the decoder, with the SoMA rank fixed to 32, and vary the auxiliary encoder from ConvNeXt-T to ConvNeXt-L. As shown in Tab. 7, the average performance consistently increases as the auxiliary encoder becomes larger, from 46.0 mAP with ConvNeXt-T to 49.0 mAP with ConvNeXt-L. This indicates that scaling the pretrained auxiliary encoder remains effective for cross-domain few-shot adaptation.

4.4 Analysis Studies

Inference Cost Analysis. We compare the inference cost of SAM3 and our method on ArTaxOr. Our method indeed introduces some additional computational overhead, but the overall increase is modest and does not hinder its deployment in practical scenarios.

Pretrained Knowledge Preservation. To examine whether few-shot target-domain adaptation disrupts the pretrained generalization ability of SAM3, we directly evaluate the checkpoints adapted on FISH and NEU-DET, two target domains with visual distributions substantially different from natural images, on PASCAL-5^i without any additional training.

As shown in Tab. 9, different target domains cause markedly different degrees of degradation in the original zero-shot capability of SAM3. Full fine-tuning can severely disrupt the pretrained representation, whereas our method consistently preserves the original generalization ability across both domains. This suggests that the proposed adaptation strategy limits the disruption of pretrained knowledge while learning target-domain information.

Table 10: Comparison of object detection inference costs on ArTaxOr in terms of parameter count and FLOPs.

Method	Params.	FLOPs
SAM3	838.2M	5.7T
Ours	1066.5M	7.4T
Increase	+27.2%	+29.8%

Auxiliary Backbone Analysis. To analyze the impact of different auxiliary CNN backbones on cross-domain few-shot adaptation performance, we replace DINOv3 ConvNeXt-L with ResNet-50 (He et al., 2016), OpenCLIP ConvNeXt-L (Cherti et al., 2023), PKINet-S (Cai et al., 2024) and PKINetv2-S Cai et al. (2026) while keeping the remaining model structure and training settings unchanged.

As shown in Tab. 8, different auxiliary CNN backbones all improve performance over the baseline, but the magnitude of the gains varies considerably. DINOv3 provides the largest improvement among them. Compared with OpenCLIP, which shares the same network architecture, the larger gains of DINOv3 may be attributed to the visual representations learned through self-supervised pretraining on large-scale data. These results suggest that the pretrained representations of the auxiliary backbone play an important role in cross-domain few-shot adaptation performance.

5 Conclusion

In this paper, we propose CHRF-SAM3 for adapting SAM3 to cross-domain few-shot object detection and semantic segmentation under limited supervision. Under constrained parameter adaptation, CHRF-SAM3 enables SAM3 to fuse heterogeneous visual representations. Experimental results demonstrate that our method achieves strong performance across multiple benchmarks while preserving the pretrained capabilities of SAM3.

REFERENCES

Eric Bianchi and Matthew Hebdon. Corrosion condition state semantic segmentation dataset. *University Libraries, Virginia Tech: Blacksburg, VA, USA*, 3(2):12, 2021.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*, 2024.

Xinhao Cai, Qiuxia Lai, Yuwei Wang, Wenguan Wang, Zeren Sun, and Yazhou Yao. Poly kernel inception network for remote sensing detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27706–27716. IEEE, 2024.

Xinhao Cai, Liulei Li, Gensheng Pei, Zeren Sun, Yazhou Yao, and Wenguan Wang. Pkinet-v2: Towards powerful and efficient poly-kernel remote sensing object detection. *arXiv preprint arXiv:2603.16341*, 2026.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

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.

Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15329–15337. IEEE, 2021.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829. IEEE, 2023.

Chen-Bin Feng, Youyang Sha, Longfei Liu, Yongjun Yu, Chi-Man VONG, Xuanlong Yu, and Xi Shen. Fsod-vfm: Few-shot object detection with vision foundation models and graph diffusion. In *International Conference on Learning Representations*, volume 2026, pp. 1644–1661, 2026.

Yuqian Fu, Yu Wang, Yixuan Pan, Lian Huai, Xingyu Qiu, Zeyu Shangguan, Tong Liu, Yanwei Fu, Luc Van Gool, and Xingqun Jiang. Cross-domain few-shot object detection via enhanced open-set object detector. In *European conference on computer vision*, pp. 247–264. Springer, 2024.

Guangxing Han and Ser-Nam Lim. Few-shot object detection with foundation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28608–28618. IEEE, 2024.

Sebastian Haug and Jörn Ostermann. A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks. In *European conference on computer vision*, pp. 105–116. Springer, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Jiajun Hu, Jian Zhang, Lei Qi, Yinghuan Shi, and Yang Gao. Learn to preserve and diversify: Parameter-efficient group with orthogonal regularization for domain generalization. In *European Conference on Computer Vision*, pp. 198–216. Springer, 2024.

Yali Huang, Jie Mei, Yiming Yang, Mi Guo, Mingyuan Jiu, and Mingliang Xu. Instance feature caching for cross-domain few-shot object detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1558–1566. IEEE, 2025.

Debesh Jha, Sharib Ali, Krister Emanuelsen, Steven A Hicks, Vajira Thambawita, Enrique Garcia-Ceja, Michael A Riegler, Thomas De Lange, Peter T Schmidt, Håvard D Johansen, et al. Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. In *International Conference on Multimedia Modeling*, pp. 218–229. Springer, 2021.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pp. 709–727. Springer, 2022.

Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8419–8428. IEEE, 2019.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10955–10965. IEEE, 2022.

Yu Li, Xingyu Qiu, Yuqian Fu, Jie Chen, Tianwen Qian, Xu Zheng, Danda Pani Paudel, Yanwei Fu, Xuanjing Huang, Luc V Gool, et al. Domain-rag: Retrieval-guided compositional image generation for cross-domain few-shot object detection. *Advances in Neural Information Processing Systems*, 38:147594–147614, 2026.

Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. *Advances in Neural Information Processing Systems*, 35: 109–123, 2022.

Changhan Liu, Zixuan Duan, Wenbin Li, Qi Fan, Yang Gao, et al. Don't need retraining: A mixture of detr and vision foundation models for cross-domain few-shot object detection. *Advances in Neural Information Processing Systems*, 38:146347–146371, 2026.

Yahui Liu, Jian Yao, Xiaohu Lu, Renping Xie, and Li Li. Deepcrack: A deep hierarchical feature learning architecture for crack segmentation. *Neurocomputing*, 338:139–153, 2019.

Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. *arXiv preprint arXiv:2305.13310*, 2023.

Junyuan Ma, Xunzhi Xiang, Wenbin Li, Qi Fan, and Yang Gao. Selective, regularized, and calibrated: Harnessing vision foundation models for cross-domain few-shot semantic segmentation. *arXiv preprint arXiv:2605.19340*, 2026.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37:121038–121072, 2024.

Anh-Khoa Nguyen Vu, Quoc Truong Truong, Vinh-Tiep Nguyen, Thanh Ngo, Thanh-Toan Do, and Tam Nguyen. Multi-perspective data augmentation for few-shot object detection. In *International Conference on Learning Representations*, volume 2025, pp. 65620–65638, 2025.

Jiancheng Pan, Yanxing Liu, Xiao He, Long Peng, Jiahao Li, Yuze Sun, and Xiaomeng Huang. Enhance then search: An augmentation-search strategy with foundation models for cross-domain few-shot object detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1539–1547. IEEE, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, et al. Grounding dino 1.5: Advance the" edge" of open-set object detection. *arXiv preprint arXiv:2405.10300*, 2024.

Franz Rottensteiner, Gunho Sohn, Markus Gerke, Jan Dirk Wegner, Uwe Breitkopf, and Jaewook Jung. Results of the isprs benchmark on urban object detection and 3d building reconstruction. *ISPRS journal of photogrammetry and remote sensing*, 93:256–271, 2014.

Alzayat Saleh, Issam H Laradji, Dmitry A Konovalov, Michael Bradley, David Vazquez, and Marcus Sheaves. A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Scientific reports*, 10(1):14671, 2020.

Constantin Seibold, Simon Reiß, Saquib Sarfraz, Matthias A Fink, Victoria Mayer, Jan Sellner, Moon Sung Kim, Klaus H Maier-Hein, Jens Kleesiek, and Rainer Stiefelhagen. Detailed annotations of chest x-rays via ct projection for report understanding. *arXiv preprint arXiv:2210.03416*, 2022.

Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017.

Shreyas S Shivakumar, Neil Rodrigues, Alex Zhou, Ian D Miller, Vijay Kumar, and Camillo J Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network. In *2020 IEEE international conference on robotics and automation (ICRA)*, pp. 9441–9447. IEEE, 2020.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Yanpeng Sun, Jiahui Chen, Shan Zhang, Xinyu Zhang, Qiang Chen, Gang Zhang, Errui Ding, Jingdong Wang, and Zechao Li. Vrp-sam: Sam with visual reference prompt. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23565–23574. IEEE, 2024.

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.

Yi-Jen Tsai, Yen-Yu Lin, and Chien-Yao Wang. Few-shot semantic segmentation meets sam3. *arXiv preprint arXiv:2604.05433*, 2026.

Steven Wallace, Aiden Durrant, William David Harcourt, Richard Hann, and Georgios Leontidis. Exploring segment anything foundation models for out of domain crevasse drone image segmentation. In *Northern Lights Deep Learning Conference 2025*, 2025.

Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9196–9205. IEEE, 2019.

Wanqi Wang, Jingcai Guo, Yuxiang Cai, and Zhi Chen. Learning multi-modal prototypes for cross-domain few-shot object detection. *arXiv preprint arXiv:2602.18811*, 2026.

Wenjian Wang, Lijuan Duan, Yuxi Wang, Qing En, Junsong Fan, and Zhaoxiang Zhang. Remember the difference: Cross-domain few-shot semantic segmentation via meta-memory transfer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7055–7064. IEEE, 2022.

Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. *arXiv preprint arXiv:2003.06957*, 2020.

Junfeng Wu, Yi Jiang, Qihao Liu, Zehuan Yuan, Xiang Bai, and Song Bai. General object foundation model for images and videos at scale. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3783–3795. IEEE, 2024.

Wuti Xiong. Cd-fsod: A benchmark for cross-domain few-shot object detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Huali Xu, Shuaifeng Zhi, Shuzhou Sun, Vishal Patel, and Li Liu. Deep learning for cross-domain few-shot visual recognition: A survey. *ACM Computing Surveys*, 57(8):1–37, 2025a.

Qianxiong Xu, Lanyun Zhu, Xuanyi Liu, Guosheng Lin, Cheng Long, Ziyue Li, and Rui Zhao. Unlocking the power of sam 2 for few-shot segmentation. *arXiv preprint arXiv:2505.14100*, 2025b.

Yifan Xu, Mengdan Zhang, Chaoyou Fu, Peixian Chen, Xiaoshan Yang, Ke Li, and Changsheng Xu. Multi-modal queried object detection in the wild. *Advances in Neural Information Processing Systems*, 36:4452–4469, 2023.

Xuanlong Yu, Youyang Sha, Longfei Liu, Xi Shen, and Di Yang. A closer look at cross-domain few-shot object detection: Fine-tuning matters and parallel decoder helps. *arXiv preprint arXiv:2603.28182*, 2026.

Seokju Yun, Seunghye Chae, Dongheon Lee, and Youngmin Ro. Soma: Singular value decomposed minor components adaptation for domain generalizable representation learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25602–25612. IEEE, 2025.

Anqi Zhang, Guangyu Gao, Jianbo Jiao, Chi H Liu, and Yunchao Wei. Bridge the points: Graph-based few-shot segment anything semantically. *Advances in neural information processing systems*, 37:33232–33261, 2024a.

Haojie Zhang, Yongyi Su, Xun Xu, and Kui Jia. Improving the generalization of segmentation foundation model under distribution shift via weakly supervised adaptation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23385–23395. IEEE, 2024b.

Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *Advances in neural information processing systems*, 35:36067–36080, 2022.

Xinyu Zhang, Yuhan Liu, Yuting Wang, and Abdeslam Boularias. Detect everything with few examples. *arXiv preprint arXiv:2309.12969*, 2023.

Appendix

A Implementation and Reproducibility Details **15**

A.1 AI Use Statement 15

A.2 Model Configuration and Parameter Settings 15

A.3 Input Preprocessing and Data Augmentation 15

A.4 Optimization Settings and Computational Resources 16

B Object Detection Experiments and Module Ablations **16**

B.1 ODinW-13 16

B.2 CD-FSOD 16

C Cross-Domain Few-Shot Semantic Segmentation **16**

D Model Configuration Ablations **17**

A Implementation and Reproducibility Details

A.1 AI Use Statement

In this work, we used generative AI tools to assist with several aspects of research execution, including method implementation, dataset cleaning and reformatting, interpretation of experimental results, and translation. We also used generative AI tools to assist with drafting and language polishing of parts of the manuscript, as well as literature retrieval, discovery of potentially relevant related work, and information search.

We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, propose or refine research hypotheses, design research methodologies or experiments, or provide feedback on research methodology or experimental design. We also did not use generative AI tools to formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in writing proofs, or conduct qualitative or thematic data analysis.

All AI-assisted code, data-processing outputs, interpretations, text, and suggested references were reviewed and verified by the authors and revised where necessary. The final implementation of the proposed method, experimental settings, experimental results, and their final interpretation were confirmed by the authors. We take full responsibility for the final content of this work, including its text, research results, claims, citations, and other artifacts produced or refined with the aid of generative AI.

A.2 Model Configuration and Parameter Settings

All experiments are built upon the official pretrained SAM3 (Carion et al., 2026) model with PE-L+ as its visual encoder. We employ DINOv3 ConvNeXt-L (Siméoni et al., 2025) pretrained on LVD-1689M as the auxiliary visual encoder. During adaptation, the SAM3 visual encoder, language encoder, DINOv3 ConvNeXt-L, and the original pretrained weights in the decoder are kept frozen.

For the auxiliary visual encoder, we extract features from the four stages of DINOv3 ConvNeXt-L. CRF spatially aligns the four-level features with the SAM3 visual features and generates three residual branches through convolutional projection. These residuals are injected into ViT Blocks 30, 31, and 32 of the SAM3 visual encoder, respectively. The fusion coefficient of each residual branch is initialized to 0.1 and optimized during training.

LSCM maps the spatially aligned and channel-wise concatenated multi-scale DINOv3 features into two condition features corresponding to the two linear layers in the feed-forward network. LSC-LoRA is applied to the feed-forward networks in ViT Blocks 21–32 of the SAM3 visual encoder, with its rank set to $r = 64$. Different ViT blocks use independent low-rank parameters while sharing the condition features generated from the current input image. The low-rank residual branches are zero-initialized so that the original pretrained mappings are preserved at the beginning of training.

The SAM3 decoder consists of six Transformer layers. We apply SoMA (Yun et al., 2025) to the linear layers in self-attention, visual cross-attention, text cross-attention, and the feed-forward networks, with a default rank of 32. The low-rank factors are initialized using the smallest r singular components of each pretrained weight matrix, and only the corresponding low-rank parameters are updated during training.

A.3 Input Preprocessing and Data Augmentation

During training, each input image is first randomly resized such that its shorter-side length lies between 480 and 1008 pixels, and is then padded to 1008×1008. Training augmentations include horizontal flipping with a probability of 0.5, color jitter with a strength of 0.4 for brightness, contrast, and saturation and 0.1 for hue, and grayscale conversion with a probability of 0.2. During testing, deterministic resizing, padding, and normalization are applied without random data augmentation.

A.4 Optimization Settings and Computational Resources

Experiments are conducted on a server equipped with eight NVIDIA GeForce RTX 4090 GPUs. Each training run uses six GPUs with a per-GPU batch size of 1, resulting in a global batch size of 6. No gradient accumulation is used. All models are trained with BF16 mixed precision.

We use AdamW as the optimizer. The initial learning rate is set to 1×10^{-4} for CRF, LSCM, and LSC-LoRA, and 1×10^{-5} for all other trainable parameters. An inverse-square-root learning-rate schedule is used with a timescale of 50, one warm-up step, and a cooldown of 50. The weight decay is set to 0.1, excluding bias and LayerNorm parameters. Gradients are clipped to a maximum ℓ_2 norm of 0.1.

All experiments are trained for 50 epochs by default, with two exceptions: the 10-shot Clipart1k experiment in CD-FSOD runs for 10 epochs, and all NEU-DET configurations use 200 training epochs. The optimization process uses random seed 123 for CD-FSOD and ODinW-13. The three few-shot splits of ODinW-13 are generated using split seeds 3, 30, and 300. Cross-domain few-shot segmentation uses random seed 3407, while PASCAL-5^i uses random seed 656.

B Object Detection Experiments and Module Ablations

B.1 ODinW-13

ODinW-13 consists of 13 object detection datasets spanning diverse scenes, categories, and visual distributions. Following the official protocol, we construct 1-shot, 3-shot, 5-shot, and 10-shot training sets. Three few-shot splits are generated using split seeds 3, 30, and 300. All three splits share the same model architecture, training strategy, and optimization seed.

Table 11 reports the AP of each dataset under the three splits, together with their arithmetic mean. Each dataset is assigned equal weight in the final macro-average, regardless of its number of images or categories. The means and standard deviations reported in Table 2 of the main paper are computed from the results of these three splits.

Tables 12 and 13 use split seed 30 and conduct cumulative module ablations. The baseline freezes both visual encoders and adapts the SAM3 decoder using only plain LoRA (Hu et al., 2021) with rank 32. '+C' indicates the addition of CRF, '+L' indicates the integration of LSC-LoRA into the encoder, and '+S' indicates replacing the vanilla decoder LoRA with SoMA. Except for the modules under comparison, all configurations use the same model architecture, training data, and optimization settings.

B.2 CD-FSOD

CD-FSOD consists of six target domains: ArTaxOr, Clipart1k, DIOR, DeepFish, NEU-DET, and UODD. These domains cover natural images, artistic imagery, remote sensing, industrial defects, and underwater scenes. Following the benchmark protocol, we perform adaptation under the 1-shot, 5-shot, and 10-shot settings and use COCO-style AP as the evaluation metric. Results are reported separately for each target domain, and the overall result is computed as the arithmetic mean of AP over the six domains.

Table 14 follows the same cumulative ablation setting as ODinW-13. The baseline uses plain LoRA to adapt the decoder. CRF and LSC-LoRA are then progressively introduced, after which the plain decoder LoRA is replaced with SoMA applied to the decoder . All configurations use the same few-shot splits, input processing, number of training epochs, and optimization settings.

C Cross-Domain Few-Shot Semantic Segmentation

We construct a cross-domain few-shot segmentation evaluation on eight datasets: CWFID Haug & Ostermann (2014), CorrosionCS Bianchi & Hebdon (2021), DeepCrack Liu et al. (2019), DeepFish Saleh et al. (2020), KvasirInstrument Jha et al. (2021), PAXRay4 Seibold et al. (2022), PST900 Shivakumar et al. (2020), and Potsdam Rottensteiner et al. (2014). Each image–foreground-class pair is treated as an independent binary segmentation unit. When an image contains multiple

foreground classes, it is divided into multiple evaluation units. Only non-empty units containing the corresponding foreground class are retained, while the background and clutter classes are not evaluated separately.

CWFID uses 21 images as the test set and the remaining images to construct the training set. PAXRay4 follows its official training, validation, and test splits, where the four potentially overlapping anatomical structures—lungs, mediastinum, bones, and diaphragm—are treated as independent binary segmentation tasks. Potsdam uses the split containing 14 test tiles, from which non-overlapping 1024×1024 crops are extracted from the IRRG images; the clutter class is not evaluated.

For each foreground class, the 1-shot and 5-shot support sets contain one and five non-empty segmentation units, respectively. Using random seed 3407, support samples are selected through a deterministic within-class ordering based on SHA-256. Each 1-shot support set is always nested within its corresponding 5-shot support set. Under the same shot setting, all methods share the same training samples, test units, and prompts.

Under text prompting, each image–class unit uses the corresponding English class name without point coordinates. For example, CWFID uses "crop" and "weed", while PAXRay4 uses "lungs", "mediastinum", "bones", and "diaphragm".

Under point prompting, the category text is left blank, and only the image and normalized point coordinates are provided to the model. A fixed nested sequence of three foreground points is generated in advance for each test unit. The first point is located at the maximum of the Euclidean distance transform within the foreground region. The second and third points are then selected sequentially within the foreground using a weighted farthest-point criterion that jointly considers boundary distance and spatial coverage. The one-, two-, and three-point settings use the corresponding prefixes of this sequence. If a foreground region contains fewer than three pixels, the last valid point is repeated.

For each dataset, IoU, Dice, and B-IoU are first computed independently for every non-empty image–class evaluation unit and are then macro-averaged over all evaluation units in that dataset. Under point prompting, the aggregate value of each metric is computed as the arithmetic mean of its values obtained with one, two, and three points. Tables 15 and 16 report the complete results under the 1-shot and 5-shot settings, respectively. Table 17 follows the cumulative ablation setting described in Section B.1. Tables 18 and 19 report detailed results under different point counts for the 1-shot and 5-shot settings, respectively.

D Model Configuration Ablations

The main text reports the average performance of different ranks of SoMA. To provide the complete results of this ablation, Table 20 further presents the AP on all six target datasets of CD-FSOD. The experiment is conducted under the 5-shot setting by varying only the rank of SoMA in each adapted linear layer of the decoder over $r \in \{16, 24, 32, 40, 48\}$, while keeping the remaining model architecture and optimization settings unchanged.

As shown in Table 20, most target domains are relatively insensitive to the choice of rank. The results on Clipart1k vary only from 64.7 to 65.5, while those on DIOR range from 35.5 to 36.3. DeepFish is even more stable, with all settings producing results between 47.7 and 47.8, and UODD also exhibits only limited variation. ArTaxOr and NEU-DET show somewhat larger fluctuations, but neither exhibits a monotonic trend as the rank increases. Moreover, the best rank differs across target domains: NEU-DET performs best at $r = 24$, Clipart1k at $r = 40$, and DeepFish at $r = 48$, whereas $r = 32$ achieves the best results on ArTaxOr, DIOR, and UODD. Overall, these results indicate that SoMA in the decoder maintains relatively stable performance over a broad range of ranks and is not highly sensitive to this hyperparameter.

The main text also reports the average performance of DINOv3 ConvNeXt auxiliary encoders at different scales. To provide the complete results of this ablation, Table 21 presents the AP of ConvNeXt-T, ConvNeXt-S, ConvNeXt-B, and ConvNeXt-L (Siméoni et al., 2025) on all six CD-FSOD target datasets, together with the baseline without an auxiliary encoder. The experiment is conducted under the 10-shot setting using the complete framework with CRF, LSC-LoRA, and

SoMA in the decoder, where the rank of SoMA is fixed to 32. The baseline freezes the SAM3 visual encoder and adapts only the decoder using plain LoRA with rank 32, without introducing the auxiliary DINOv3 encoder.

The effect of auxiliary-encoder scale varies substantially across target domains. ArTaxOr is the most sensitive to model scale, with AP increasing from 66.6 using ConvNeXt-T to 81.6 using ConvNeXt-L, showing a clear benefit from larger convolutional representations. In contrast, Clipart1k is largely insensitive to scale, with all four auxiliary encoders producing results around 66–67 AP. DIOR benefits substantially from introducing an auxiliary encoder compared with the baseline, but the differences among ConvNeXt variants are relatively small, with ConvNeXt-S obtaining the highest result. DeepFish shows a similar pattern, with all four variants remaining close to 47 AP, indicating that further increasing the auxiliary-encoder scale has only a limited effect on this domain. NEU-DET exhibits moderate fluctuations across different scales, with ConvNeXt-B performing best, while some configurations remain close to or slightly below the baseline. For UODD, the smaller variants provide limited changes, whereas ConvNeXt-L yields a more noticeable improvement. These results show that the influence of auxiliary-encoder scale is strongly domain-dependent: ArTaxOr benefits substantially from larger models, while domains such as Clipart1k and DeepFish are comparatively insensitive to scale.

Table 11: Per-dataset AP on ODinW-13 for the three dataset-split seeds s_3, s_{30}, and s_{300}, and their arithmetic mean μ. Each dataset block lists the four shot settings; the final block is the macro average over the 13 datasets.

Dataset	Shot	s_3	s_{30}	s_{300}	μ
AerialMaritimeDrone_large	1-shot	36.7	34.4	31.6	34.2
	3-shot	37.0	37.4	36.0	36.8
	5-shot	37.2	36.8	37.6	37.2
	10-shot	37.4	37.4	37.3	37.4
Aquarium	1-shot	48.9	48.5	48.2	48.5
	3-shot	52.0	53.4	55.1	53.5
	5-shot	54.5	56.5	55.9	55.6
	10-shot	57.9	58.2	60.3	58.8
CottontailRabbits	1-shot	83.6	83.7	84.6	84.0
	3-shot	85.7	86.0	85.9	85.9
	5-shot	86.0	85.9	85.5	85.8
	10-shot	88.5	87.1	86.0	87.2
EgoHands_generic	1-shot	73.4	73.9	74.4	73.9
	3-shot	74.0	73.8	73.3	73.7
	5-shot	75.3	73.7	73.7	74.2
	10-shot	76.1	74.8	73.9	74.9
NorthAmericaMushrooms	1-shot	96.8	99.3	96.8	97.6
	3-shot	97.7	98.7	100.0	98.8
	5-shot	96.7	99.3	99.3	98.4
	10-shot	100.0	100.0	100.0	100.0
Packages	1-shot	86.0	86.7	78.8	83.8
	3-shot	88.1	90.7	90.5	89.8
	5-shot	88.8	90.1	88.5	89.1
	10-shot	94.5	94.1	95.1	94.6
PascalVOC	1-shot	69.6	67.9	69.1	68.9
	3-shot	69.7	69.4	70.0	69.7
	5-shot	70.2	69.7	70.6	70.2
	10-shot	70.7	70.5	70.5	70.6
Raccoon	1-shot	77.6	67.1	74.4	73.1
	3-shot	84.0	70.6	80.4	78.3
	5-shot	82.3	76.7	85.7	81.6
	10-shot	85.5	85.0	86.2	85.6
ShellfishOpenImages	1-shot	65.1	63.4	63.4	63.9
	3-shot	62.3	67.4	64.1	64.6
	5-shot	63.6	68.8	67.4	66.6
	10-shot	66.4	67.5	71.9	68.6
VehiclesOpenImages	1-shot	68.7	69.1	69.6	69.1
	3-shot	69.4	68.9	67.4	68.5
	5-shot	69.2	68.2	67.3	68.2
	10-shot	71.0	67.4	69.7	69.3
pistols	1-shot	73.3	71.2	73.3	72.6
	3-shot	73.5	69.7	74.0	72.4
	5-shot	73.5	71.9	76.5	74.0
	10-shot	74.3	72.5	76.1	74.3
pothole	1-shot	39.1	37.4	43.4	40.0
	3-shot	48.1	40.7	44.1	44.3
	5-shot	47.2	42.1	45.7	45.0
	10-shot	49.4	46.6	51.6	49.2
thermalDogsAndPeople	1-shot	72.9	69.1	69.9	70.6
	3-shot	77.0	80.4	84.5	80.6
	5-shot	85.9	85.8	87.1	86.3
	10-shot	88.6	88.3	89.5	88.8
Macro avg.	1-shot	68.6	67.1	67.5	67.7
	3-shot	70.6	69.8	71.2	70.5
	5-shot	71.6	71.2	72.4	71.7
	10-shot	73.9	73.0	74.5	73.8

Table 12: Per-dataset ODinW-13 module ablation for 1-shot and 3-shot adaptation (AP, %). C: CRF; L: LSC-LoRA; S: SoMA applied to the decoder with rank 32.

	1-shot					3-shot			
Dataset	Base	+C	+C+L	+C+L+S	Dataset	Base	+C	+C+L	+C+L+S
AerialMaritimeDrone	30.2	31.7	33.8	**34.4**	AerialMaritimeDrone	35.2	36.3	**37.9**	37.4
Aquarium	46.1	47.0	48.4	**48.5**	Aquarium	50.1	**53.4**	52.6	53.4
CottontailRabbits	83.7	84.6	**84.6**	83.7	CottontailRabbits	85.8	85.8	85.7	**86.0**
EgoHands	74.1	74.2	**74.4**	73.9	EgoHands	**74.3**	74.2	74.1	73.8
Mushrooms	94.3	94.7	98.7	**99.3**	Mushrooms	94.2	94.7	**98.7**	**98.7**
Packages	86.0	**87.1**	**87.1**	86.7	Packages	87.1	90.1	89.1	**90.7**
PascalVOC	67.2	67.1	67.5	**67.9**	PascalVOC	69.1	69.1	**69.9**	69.4
Raccoon	**67.6**	66.6	67.0	67.1	Raccoon	69.3	69.5	**73.1**	70.6
Shellfish	62.4	63.3	**64.7**	63.4	Shellfish	66.9	**67.8**	67.5	67.4
Vehicles	68.8	69.2	**69.9**	69.1	Vehicles	68.7	69.3	**69.4**	68.9
pistols	70.8	**71.2**	70.8	71.2	pistols	70.6	**71.2**	70.3	69.7
pothole	35.7	36.6	**37.8**	37.4	pothole	38.1	38.3	**41.3**	40.7
thermalDogsAndPeople	70.4	69.8	**70.9**	69.1	thermalDogsAndPeople	74.8	75.6	**81.6**	80.4
Macro average	65.9	66.4	**67.3**	67.1	Macro average	68.0	68.9	**70.1**	69.8

Table 13: Per-dataset ODinW-13 module ablation for 5-shot and 10-shot adaptation (AP, %). C: CRF; L: LSC-LoRA; S: SoMA applied to the decoder with rank 32.

	5-shot					10-shot			
Dataset	Base	+C	+C+L	+C+L+S	Dataset	Base	+C	+C+L	+C+L+S
AerialMaritimeDrone	33.2	35.5	35.6	**36.8**	AerialMaritimeDrone	36.8	**39.5**	39.1	37.4
Aquarium	51.4	56.0	55.9	**56.5**	Aquarium	52.7	57.2	57.1	**58.2**
CottontailRabbits	85.9	85.8	84.8	**85.9**	CottontailRabbits	85.8	85.6	86.9	**87.1**
EgoHands	**73.8**	73.7	73.6	73.7	EgoHands	74.6	74.7	**75.2**	74.8
Mushrooms	96.8	**100.0**	99.3	99.3	Mushrooms	99.3	**100.0**	**100.0**	**100.0**
Packages	87.1	87.1	89.1	**90.1**	Packages	92.0	92.4	93.1	**94.1**
PascalVOC	69.5	69.1	69.6	**69.7**	PascalVOC	70.2	70.2	70.1	**70.5**
Raccoon	72.6	74.1	**77.8**	76.7	Raccoon	82.7	84.1	83.4	**85.0**
Shellfish	68.5	68.9	**70.5**	68.8	Shellfish	68.2	**68.5**	67.7	67.5
Vehicles	**68.7**	68.6	68.1	68.2	Vehicles	69.4	**69.5**	68.7	67.4
pistols	71.5	71.7	71.8	**71.9**	pistols	70.9	71.7	71.2	**72.5**
pothole	38.4	38.9	**43.6**	42.1	pothole	42.6	42.3	**46.7**	46.6
thermalDogsAndPeople	80.4	82.3	84.3	**85.8**	thermalDogsAndPeople	82.1	83.6	88.0	**88.3**
Macro average	69.1	70.1	71.1	**71.2**	Macro average	71.3	72.2	72.9	**73.0**

Table 14: Per-domain module ablation on CD-FSOD (mAP, %). C: CRF; L: LSC-LoRA; S: SoMA applied to the decoder with rank 32. Bold indicates the best unrounded result within each shot and target domain.

Shot	Configuration	ArTaxOr	Clipart1k	DIOR	DeepFish	NEU-DET	UODD	Avg.
1	Baseline	31.1	65.1	24.3	42.3	12.7	26.3	33.6
	+ C	35.9	**65.6**	26.5	43.2	10.4	**26.3**	34.7
	+ C + L	38.2	65.5	29.0	44.1	16.5	25.5	36.5
	+ C + L + S (Ours)	**39.2**	65.2	**29.2**	**44.4**	**18.9**	25.8	**37.1**
5	Baseline	46.5	65.6	33.1	46.5	23.4	25.2	40.1
	+ C	70.3	65.8	**37.8**	**47.7**	25.4	**26.0**	45.5
	+ C + L	71.0	**65.9**	36.7	47.0	**26.8**	23.7	45.2
	+ C + L + S (Ours)	**77.5**	65.4	36.3	**47.7**	25.6	25.8	**46.4**
10	Baseline	50.7	66.8	35.0	45.6	27.2	25.9	41.9
	+ C	80.7	**67.4**	41.7	47.2	27.0	25.9	48.3
	+ C + L	81.4	64.3	41.4	**47.5**	**28.2**	26.3	48.2
	+ C + L + S (Ours)	**81.6**	67.0	**42.8**	46.9	27.7	**27.7**	**49.0**

Table 15: Cross-domain 1-shot segmentation results (%). The eight dataset columns report mIoU, Dice, and B-IoU for each prompt type and method; the Avg column is the macro-average of the corresponding metric over the eight datasets. Each entry in the point-prompting block is averaged over the 1-, 2-, and 3-point settings. The best average within each prompt type and metric is bold.

Prompt	Metric	Method	CWFID	CorrosionCS	DeepCrack	DeepFish	Kvasir	PAXRay4	PST900	Potsdam	Avg
Text	mIoU	Decoder LoRA	41.4	30.5	71.0	86.5	84.9	41.5	26.3	49.5	53.9
		Decoder full FT	41.3	31.2	70.8	86.6	85.8	44.3	27.5	50.2	54.7
		Ours	41.6	30.5	69.9	86.7	89.7	62.8	29.6	49.0	**57.5**
	Dice	Decoder LoRA	54.5	42.3	82.3	92.0	89.1	53.2	34.3	59.0	63.3
		Decoder full FT	54.5	43.2	82.1	92.1	90.0	55.8	36.4	59.8	64.2
		Ours	54.5	42.5	81.6	92.2	93.8	76.1	38.2	58.4	**67.2**
	B-IoU	Decoder LoRA	41.3	23.8	69.8	80.6	72.6	15.6	20.8	37.7	45.3
		Decoder full FT	41.3	24.3	69.5	80.6	73.6	17.0	22.0	38.4	45.8
		Ours	41.6	23.8	68.8	80.8	76.0	22.3	23.8	37.4	**46.8**
Point	mIoU	Decoder LoRA	43.5	33.8	64.3	86.8	61.6	55.6	44.8	53.8	55.5
		Decoder full FT	44.8	34.2	62.9	86.9	63.0	55.0	43.0	55.6	55.7
		Ours	48.6	37.0	66.6	86.8	88.9	60.2	54.9	57.3	**62.5**
	Dice	Decoder LoRA	57.7	46.5	76.4	92.3	68.4	69.9	56.2	65.6	66.6
		Decoder full FT	59.1	47.0	75.1	92.4	69.6	69.5	54.1	67.5	66.8
		Ours	62.8	50.1	78.8	92.5	92.6	73.6	67.7	68.6	**73.3**
	B-IoU	Decoder LoRA	43.5	23.2	63.0	80.3	53.3	16.5	38.4	38.6	44.6
		Decoder full FT	44.8	23.4	61.7	80.4	54.6	15.3	36.8	40.1	44.6
		Ours	48.6	26.8	65.1	80.1	76.3	21.2	48.1	42.7	**51.1**

Table 16: Cross-domain 5-shot segmentation results (%). Layout and averaging protocol follow Table 15. The best average within each prompt type and metric is bold.

Prompt	Metric	Method	CWFID	CorrosionCS	DeepCrack	DeepFish	Kvasir	PAXRay4	PST900	Potsdam	Avg
Text	mIoU	Decoder LoRA	41.1	31.6	71.4	87.0	85.6	65.4	30.4	63.3	59.5
		Decoder full FT	41.5	32.3	71.2	86.5	86.9	63.1	26.9	62.3	58.8
		Ours	70.6	32.8	71.2	89.7	91.1	84.5	41.7	71.1	**69.1**
	Dice	Decoder LoRA	54.1	43.9	82.6	92.2	90.6	78.0	39.7	74.4	69.4
		Decoder full FT	54.1	44.8	82.4	91.8	92.1	76.2	35.8	73.4	68.8
		Ours	81.5	45.3	82.1	94.3	94.9	91.3	52.2	81.1	**77.8**
	B-IoU	Decoder LoRA	41.0	24.1	70.3	81.5	71.8	27.3	25.3	48.3	48.7
		Decoder full FT	41.4	24.0	70.2	81.1	73.1	26.9	22.0	47.5	48.3
		Ours	70.6	26.0	70.8	84.6	78.6	47.5	36.0	55.9	**58.8**
Point	mIoU	Decoder LoRA	44.9	39.3	69.0	87.7	90.1	60.7	55.3	62.1	63.6
		Decoder full FT	44.4	37.0	68.9	87.9	89.5	56.4	50.4	61.7	62.0
		Ours	70.4	42.6	72.1	90.1	93.4	75.9	59.4	65.1	**71.1**
	Dice	Decoder LoRA	58.5	52.7	80.5	92.8	93.7	73.7	68.0	73.4	74.2
		Decoder full FT	57.8	50.2	80.5	93.0	93.2	70.2	62.6	73.1	72.6
		Ours	80.6	56.9	82.8	94.6	96.5	84.9	72.4	75.5	**80.5**
	B-IoU	Decoder LoRA	44.9	28.5	67.6	81.9	77.4	21.4	48.6	46.9	52.2
		Decoder full FT	44.4	25.7	67.6	82.2	77.0	16.3	44.1	46.5	50.5
		Ours	70.4	32.6	71.0	85.0	81.4	35.8	52.8	50.9	**60.0**

Table 17: Per-dataset module ablation for cross-domain few-shot segmentation under text prompting (mIoU, %). C: CRF; L: LSC-LoRA; S: SoMA applied to the decoder with rank 32.

Shot	Configuration	CWFID	CorrosionCS	DeepCrack	DeepFish	Kvasir	PAXRay4	PST900	Potsdam	Avg.
1	Baseline	41.4	**30.5**	**71.0**	86.5	84.9	41.5	26.3	49.5	53.9
	+ C	41.3	30.1	70.8	86.7	87.1	48.7	27.6	**49.5**	55.2
	+ C + L	**43.1**	30.0	69.8	**86.9**	87.2	62.2	**30.5**	48.9	57.3
	+ C + L + S (Ours)	41.6	**30.5**	69.9	86.7	**89.7**	**62.8**	29.6	49.0	**57.5**
5	Baseline	41.1	31.6	**71.4**	87.0	85.6	65.4	30.4	63.3	59.5
	+ C	41.2	31.9	71.2	88.1	87.3	71.0	33.4	68.3	61.5
	+ C + L	68.1	**33.7**	71.2	89.6	**91.4**	**84.6**	**41.8**	**71.3**	69.0
	+ C + L + S (Ours)	**70.6**	32.8	71.2	**89.7**	91.1	84.5	41.7	71.1	**69.1**

Table 18: Detailed 1-shot point-prompting results (mIoU, %). Each dataset occupies three rows, one per method, so that the 1-, 2-, and 3-point columns can be read without slanted or squeezed cells.

Dataset	Method	1 point	2 points	3 points	Avg.
CWFID	Decoder LoRA	37.9	45.3	47.4	43.5
	Decoder full FT	43.0	44.6	46.8	44.8
	Ours	45.7	47.8	52.3	**48.6**
CorrosionCS	Decoder LoRA	33.3	34.7	33.3	33.8
	Decoder full FT	33.6	34.9	34.2	34.2
	Ours	32.5	38.2	40.2	**37.0**
DeepCrack	Decoder LoRA	60.3	65.6	66.9	64.3
	Decoder full FT	59.0	64.1	65.7	62.9
	Ours	63.7	67.6	68.3	**66.6**
DeepFish	Decoder LoRA	86.7	87.0	86.6	86.8
	Decoder full FT	86.7	87.5	86.6	**86.9**
	Ours	86.2	87.7	86.6	86.8
Kvasir	Decoder LoRA	60.2	66.1	58.7	61.6
	Decoder full FT	61.0	66.5	61.6	63.0
	Ours	88.5	89.6	88.6	**88.9**
PAXRay4	Decoder LoRA	56.6	55.7	54.4	55.6
	Decoder full FT	55.5	55.3	54.3	55.0
	Ours	60.3	60.8	59.5	**60.2**
PST900	Decoder LoRA	42.6	45.9	46.0	44.8
	Decoder full FT	41.3	43.9	43.9	43.0
	Ours	51.5	56.0	57.1	**54.9**
Potsdam	Decoder LoRA	44.8	56.3	60.4	53.8
	Decoder full FT	48.1	57.7	60.9	55.6
	Ours	52.2	58.5	61.3	**57.3**
Macro avg.	Decoder LoRA	52.8	57.1	56.7	55.5
	Decoder full FT	53.5	56.8	56.7	55.7
	Ours	**60.1**	**63.3**	**64.2**	**62.5**

Table 19: Detailed 5-shot point-prompting results (mIoU, %). Each dataset occupies three rows, one per method.

Dataset	Method	1 point	2 points	3 points	Avg.
CWFID	Decoder LoRA	44.1	45.0	45.8	44.9
	Decoder full FT	43.1	44.4	45.7	44.4
	Ours	69.7	70.2	71.4	**70.4**
CorrosionCS	Decoder LoRA	35.0	41.4	41.5	39.3
	Decoder full FT	34.3	39.0	37.9	37.0
	Ours	38.1	44.3	45.4	**42.6**
DeepCrack	Decoder LoRA	66.0	70.3	70.7	69.0
	Decoder full FT	65.9	70.1	70.8	68.9
	Ours	69.6	73.0	73.7	**72.1**
DeepFish	Decoder LoRA	87.7	87.8	87.6	87.7
	Decoder full FT	88.0	87.8	87.8	87.9
	Ours	89.8	90.3	90.2	**90.1**
Kvasir	Decoder LoRA	89.5	90.6	90.1	90.1
	Decoder full FT	89.5	89.7	89.3	89.5
	Ours	93.0	93.7	93.6	**93.4**
PAXRay4	Decoder LoRA	58.5	62.1	61.6	60.7
	Decoder full FT	55.5	56.9	56.7	56.4
	Ours	75.5	76.5	75.7	**75.9**
PST900	Decoder LoRA	50.7	56.9	58.3	55.3
	Decoder full FT	45.6	52.3	53.4	50.4
	Ours	56.1	60.1	61.9	**59.4**
Potsdam	Decoder LoRA	58.6	62.9	64.8	62.1
	Decoder full FT	58.1	62.7	64.4	61.7
	Ours	60.7	66.3	68.5	**65.1**
Macro avg.	Decoder LoRA	61.3	64.6	65.0	63.6
	Decoder full FT	60.0	62.9	63.2	62.0
	Ours	**69.1**	**71.8**	**72.6**	**71.1**

Table 20: Ablation study of the rank of SoMA on CD-FSOD under the 5-shot setting (AP, %).

Rank r	ArTaxOr	Clipart1k	DIOR	DeepFish	NEU-DET	UODD	Avg.
16	76.9	65.1	35.5	47.7	26.3	25.1	46.1
24	75.4	64.7	35.8	47.7	**26.4**	25.2	45.9
32	**77.5**	65.4	**36.3**	47.7	25.6	**25.8**	**46.4**
40	76.1	**65.5**	36.0	47.8	25.2	25.1	46.0
48	77.4	65.1	**36.3**	**47.8**	26.0	25.1	46.3

Table 21: Ablation study of different DINOv3 ConvNeXt scales on CD-FSOD under the 10-shot setting (AP, %).

Auxiliary encoder	ArTaxOr	Clipart1k	DIOR	DeepFish	NEU-DET	UODD	Avg.
None (baseline)	50.7	66.8	35.0	45.6	27.2	25.9	41.9
DINOv3 ConvNeXt-T	66.6	**67.1**	42.0	47.1	27.3	25.8	46.0
DINOv3 ConvNeXt-S	69.3	67.1	**43.1**	**47.4**	26.4	26.6	46.7
DINOv3 ConvNeXt-B	75.8	66.3	41.8	47.0	**28.0**	26.1	47.5
DINOv3 ConvNeXt-L	**81.6**	67.0	42.8	46.9	27.7	**27.7**	**49.0**