CROSS-DOMAIN FEW-SHOT SEGMENTATION VIA DUAL-LEVEL DISTILLATION ON SAM3

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-Domain Few-Shot Segmentation (CD-FSS) aims to segment target objects in unseen domains using only a few annotated references. Recent Segment Anything Model (SAM)-based methods require less task-specific training but rely on pairwise visual matching, whose reliability decreases when reference and target appearances differ, leading to substantial variation in per-target IoU. SAM3 supports promptable concept segmentation and provides a semantic prompt space for learning a reusable category prototype. This prototype produces stable predictions across targets but cannot fully represent the different appearances. Based on these insights, we propose Dual-level Distillation on SAM3 (D^2-SAM3), a lightweight framework that uses self-distillation at complementary category and image levels. *(i)* Category Prototype Distillation transfers GT-box-prompted predictions into a reusable category prototype; *(ii)* Visual Residual Distillation calibrates reference-target visual correspondence separately for each category and expresses the resulting visual prediction as a residual relative to the prototype-based prediction; and *(iii)* Reliability-aware Fusion controls this residual based on correspondence reliability and uncertainty of the prototype-based prediction. Together, they produce a category-stable yet target-adaptive representation. With SAM3 frozen, D^2-SAM3 optimizes only approximately 8.2K parameters per category. Extensive experiments demonstrate strong CD-FSS performance, robustness to reference selection, and generalization across tasks and semantic granularities.

1 INTRODUCTION

Semantic segmentation has achieved remarkable progress in recent years (Cheng et al., 2022; Kirillov et al., 2023; Wang et al., 2024), but state-of-the-art performance still typically relies on large-scale pixel-level annotations. This reliance is particularly challenging when segmenting objects in previously unseen domains with scarce annotations, such as medical, satellite, and underwater imagery (Codella et al., 2019; Demir et al., 2018; Islam et al., 2020). Cross-Domain Few-Shot Segmentation (CD-FSS) (Lei et al., 2022) aims to address this problem by using only a few annotated references to segment target objects in previously unseen domains. Most existing methods first train a few-shot segmentation model on large-scale source-domain data and then adapt it to each target domain (Su et al., 2024; Nie et al., 2026a; Fan et al., 2025; Tong et al., 2024), incurring heavy training and adaptation costs.

The emergence of vision foundation models offers an alternative that requires less task-specific training (Kirillov et al., 2023; Ravi et al., 2024). Pretrained on large-scale and diverse visual data, the Segment Anything Model (SAM) can segment objects from a wide range of categories (Kirillov et al., 2023). However, SAM operates primarily as a class-agnostic mask generator conditioned on geometric prompts (*e.g.*, points) (Wei et al., 2024; Huang et al., 2024). Existing SAM-based CD-FSS methods therefore use external visual encoders (*e.g.*, DINOv2 (Oquab et al., 2023)) to match features between reference and target images and convert the resulting correspondences into prompt points for segmentation (Liu et al., 2024b; Zhang et al., 2024a; Nie et al., 2026b). Because these points are generated separately for each target image, their reliability depends heavily on the visual similarity between the reference and target. Even with a fixed reference, differences in target appearances cause substantial variations in per-target IoU, as illustrated in Fig. 1 (Left). Sensitivity to different references is further discussed in Sec. 3.4. These observations reveal a fundamental limitation of relying exclusively on pairwise visual matching: although it can identify target regions

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.

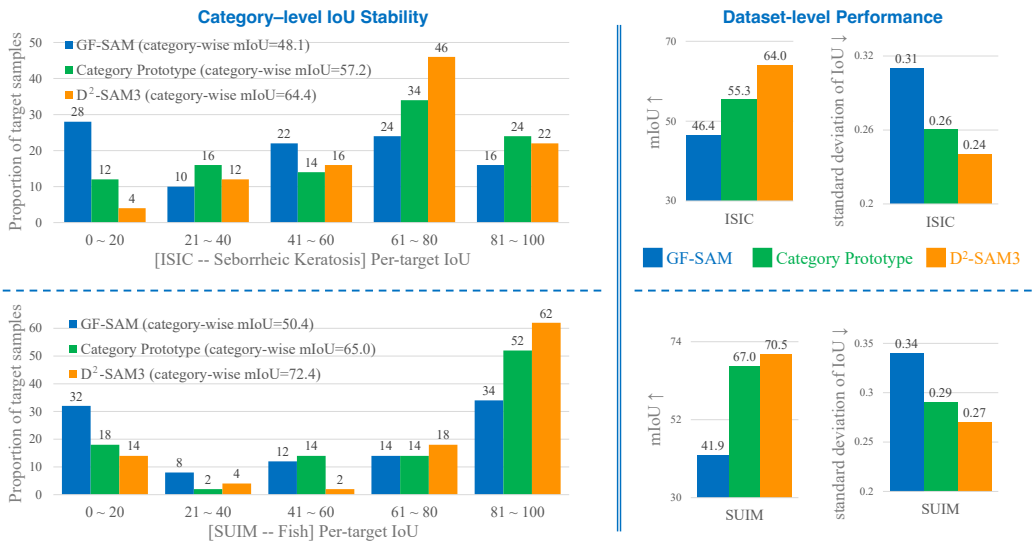


Figure 1: Motivation for Dual-level Distillation on SAM3 (D^2-SAM3). **Left**: For the Seborrheic Keratosis category in ISIC and the Fish category in SUIM, the reference image and target images are fixed for the methods being compared. The per-target IoU distributions are plotted using identical intervals. GF-SAM (Zhang et al., 2024a) exhibits a widely dispersed IoU distribution, whereas the distilled category prototype yields a more concentrated distribution. **Right**: Results on the full datasets show that D^2-SAM3 achieves higher mIoU (↑) and lower standard deviation of per-target IoU (↓) than both GF-SAM and the distilled category prototype.

that visually correspond to the reference, it does not provide a stable category prototype that can be reused across target images.

By supporting promptable concept segmentation, SAM3 (Carion et al., 2026) provides a semantic prompt space in which each target category can be represented by a prototype shared across different target images. The key challenge is how to learn such a prototype from only a few annotated references in a previously unseen domain. The reference mask can be converted into a GT-box prompt, which frozen SAM3 uses to accurately segment the objects in the reference image (refer to Sec. 3.3). This naturally enables self-distillation: inspired by prior work on self-distillation for vision foundation models (Wu et al., 2024b), we use the GT-box-prompted prediction as the teacher for learning the category prototype. Once distilled, this prototype produces more stable segmentation results across target images, as indicated by the reduced variation in per-target IoU in Fig. 1 (Left). However, despite reducing variation in per-target IoU, the category prototype alone yields only a modest mIoU gain over pairwise visual matching, as shown in Fig. 1 (Right). This observation indicates that a stable category prototype alone cannot fully represent the different appearances across target images. As analyzed in Sec. 3.5, pairwise visual correspondence provides complementary information when the reference and target are visually similar, but becomes unreliable when their appearances differ. Moreover, prior work shows that the distributions of visual features differ across categories (Nie et al., 2026b). Since visual correspondence is computed from these features, its distribution may also vary across categories. We therefore calibrate the visual correspondence separately for each category. The key is therefore to use the prototype-based prediction as the primary and refine it with a visual residual derived from reference-target correspondence.

Based on these insights, we propose Dual-level Distillation on SAM3 (D^2-SAM3), which performs self-distillation at two complementary levels using annotated references. At the category level, Category Prototype Distillation (CPD) transfers the GT-box-prompted prediction into a reusable category prototype, which SAM3 uses to produce the primary prediction for each target image. At the image level, Visual Residual Distillation (VRD) learns category-specific calibration of visual correspondence through self-distillation on the annotated reference. During target inference, VRD applies this calibration to the reference-target correspondence and expresses the visual prediction as a residual relative to the prototype-based prediction. Finally, Reliability-aware Fusion (RAF) controls the residual contribution based on the reliability of the reference-target correspondence and the uncertainty of the prototype-based prediction. Together, these components produce a category-stable yet target-adaptive representation. As shown in Fig.1 (Right), D^2-SAM3 achieves higher

mIoU and a lower standard deviation of per-target IoU than both pairwise visual matching and the distilled category prototype. SAM3 remains frozen and only approximately 8.2K parameters per category are optimized during adaptation. Extensive experiments show that D^2-SAM3 outperforms existing CD-FSS methods, is robust to reference selection, and generalizes to related segmentation tasks and different semantic granularities.

Our contributions are summarized as follows. ***First***, we identify a trade-off in CD-FSS: pairwise visual correspondence is reliable when the reference and target are visually similar but becomes unreliable when their appearances differ, whereas a category prototype produces stable predictions but cannot fully represent the different appearances. ***Second***, we propose D^2-SAM3, a lightweight framework that uses self-distillation at the category and image levels to learn a reusable category prototype and category-specific calibration of visual correspondence from annotated references. This produces a category-stable yet target-adaptive representation while optimizing only approximately 8.2K parameters per category. ***Third***, experiments on CD-FSS benchmarks demonstrate strong segmentation performance and robustness to reference selection, while results on related benchmarks show generalization across tasks and semantic granularities.

2 Related Work

Cross-Domain Few-Shot Segmentation. Cross-Domain Few-Shot Segmentation (CD-FSS) aims to segment target objects with a few annotated references under significant domain shifts (Lu et al., 2021; Wang et al., 2022; Tong et al., 2025; Wu et al., 2024a; Chen et al., 2024; Zhu et al., 2025; Tian et al., 2025; Nie et al., 2026a; Cong et al., 2026). Early methods (Lei et al., 2022; Fu et al., 2024) usually follow a two-stage paradigm, first training on large-scale source-domain data and then adapting to each target domain through dynamic refinement (Fan et al., 2023), knowledge transfer (Huang et al., 2023), frequency disentanglement (Tong et al., 2024), or lightweight adaptation (Fan et al., 2025; Su et al., 2024). With the advent of vision foundation models (Kirillov et al., 2023; Oquab et al., 2023; Rombach et al., 2022; Ravi et al., 2024), recent studies (Liu et al., 2024b; Zhang et al., 2024a; Zhu et al., 2024; Sun et al., 2024; Liu et al., 2024a; Cuttano et al., 2025; Peng et al., 2025; Xu et al., 2025; Ma et al., 2026) typically establish reference-target visual correspondences and convert them into geometric prompts for SAM. While pairwise visual correspondence helps localize the target object in each image, it does not provide a stable category prototype, and its reliability varies across reference-target pairs. SAM3 (Carion et al., 2026) introduces promptable concept segmentation, allowing target categories to be represented in its semantic prompt space. We leverage this capability to distill annotated references into a reusable category prototype. To adapt the category-level prediction to different target images, we use reference-target visual correspondence to provide a residual correction.

SAM Adaptation for Downstream Tasks. Despite their strong promptable segmentation ability, SAM and its variants (Kirillov et al., 2023; Ravi et al., 2024) often require adaptation to specialized domains, including medical (Chen et al., 2023; Ma et al., 2024; Wu et al., 2025) and satellite imagery (Li et al., 2025; Ding et al., 2024; Liu et al., 2025). Existing approaches adapt SAM using large-scale domain-specific annotations (Ma et al., 2024), lightweight trainable modules (Chen et al., 2023; Ke et al., 2023), or few-shot target-domain tuning (Xiao et al., 2024a;b). These methods generally modify the visual encoder or mask decoder for downstream domains. In contrast, we keep SAM3 (Carion et al., 2026) frozen and use distillation to learn a small set of category-specific parameters from annotated references.

Self-Distillation in Vision. Self-distillation has been studied in visual self-supervised learning (Zhang et al., 2019; Caron et al., 2021; Oquab et al., 2023). Unlike conventional distillation (Hinton et al., 2015), where the teacher and student are separate models, self-distillation uses a single model for both roles (Zhang et al., 2019). Teacher and student may correspond to different training steps (Tarvainen & Valpola, 2017; Grill et al., 2020), augmentations (Grill et al., 2020), or cropped views (Caron et al., 2021). This paradigm has proven effective in large-scale label-free pretraining (Caron et al., 2021; Oquab et al., 2023; Siméoni et al., 2025). Recent studies further apply patch-level self-distillation to pretrained vision-language models for open-vocabulary segmentation (Dong et al., 2023; Wu et al., 2024b; Jing et al., 2024). In FSS, DistillFSS (De Marinis et al., 2025) distills reference knowledge into a student model. Leveraging SAM3's geometric prompts (Carion et al., 2026), we distill predictions from annotated references directly into category prototypes.

3 Preliminary Study and Motivation

3.1 Problem Formulation

Cross-Domain Few-Shot Segmentation (CD-FSS) aims to segment objects in a target image under domain shifts, conditioned on a few annotated reference images. For each task, a reference set $\mathcal{R} = \{(I_i^r, M_i^r)\}_{i=1}^K$ is provided, where I_i^r and M_i^r denote the i-th reference image and its corresponding mask, respectively. Given a target image I^t containing objects of the same category, the goal is to predict its segmentation mask $\hat{M}^t$. Our method follows a two-stage process consisting of reference adaptation and target inference (Herzog, 2024; Nie et al., 2026a). During adaptation, we use $\mathcal{R}$ to distill task-specific representations for the target category. During inference, the learned representations and the reference set are jointly used to segment the target image.

3.2 Segment Anything Model 3

SAM3 (Carion et al., 2026) extends the SAM family (Kirillov et al., 2023; Ravi et al., 2024) to promptable concept segmentation (PCS), where textual descriptions, image exemplars, or their combination prompt SAM3 to segment all instances of the specified concept. SAM3 also retains geometric prompts, such as bounding boxes, for localized instance segmentation. For PCS, it encodes the prompt and input image into prompt and visual tokens, respectively. Its DETR-style mask decoder (Carion et al., 2020) allows object queries to interact with prompt tokens and produces candidate masks with corresponding scores. These two prompt types play different roles in reference adaptation: we use the concept prompt space to learn a reusable category prototype and prompt the same frozen SAM3 model with a box derived from the annotated reference mask to obtain a GT-box-prompted prediction for self-distillation. We discuss the potential of using manually specified text as a prompt in Appendix Sec. A2.2.

3.3 Teacher Predictions from GT-Box Prompts

The reference mask provides pixel-level annotation and can be converted into a GT-box prompt, which frozen SAM3 uses to segment the annotated object. Following prior work on knowledge distillation and self-distillation (Hinton et al., 2015; Zhang et al., 2019), we use the resulting GT-box-prompted prediction as a teacher supervision. To assess whether this prediction provides reliable supervision, we evaluate frozen SAM3 using GT-box prompts. As shown in Tab. 1, it achieves 72.6%, 91.0%, 79.5%, and 86.6% mIoU on the four datasets. These scores substantially exceed those of GF-SAM (Zhang et al., 2024a), CPS (Nie et al., 2026b), and INSID3 (Cuttano et al., 2026), confirming that GT-box-prompted predictions provide reliable teacher supervision for subsequent distillation.

Table 1: GT-box-prompted SAM3 performance in mIoU (%).

Method	ISIC	Chest.	DeepG.	SUIM
GF-SAM	48.7	47.6	49.5	41.9
CPS	51.3	70.0	46.0	42.8
INSID3	54.4	78.8	–	54.9
SAM3+GT box	**72.6**	**91.0**	**79.5**	**86.6**

3.4 Sensitivity to Reference Selection with Fixed Target Images

Fig. 1 (Left) fixes the reference image and varies the target images. Here, we reverse the setting by fixing each target image and varying the annotated references. We evaluate three target images of different difficulty levels from the Seborrheic Keratosis category in ISIC and the Fish category in SUIM, as shown in Fig. 2. The IoUs differ substantially across the selected targets, indicating different levels of segmentation difficulty. Nevertheless, for every fixed target image, the IoU of GF-SAM (Zhang et al., 2024a) varies considerably across reference selections. Because each target image remains unchanged, this variation cannot be attributed to target difficulty. The variation therefore results from reference selection, confirming the sensitivity of pairwise visual matching and motivating a stable category prototype.

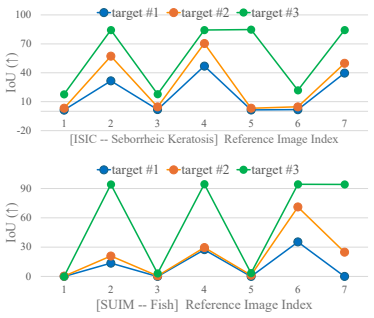


Figure 2: Reference sensitivity of GF-SAM under fixed target images.

3.5 Effect of Reference-Target Similarity

Table 2: Comparison between category prototypes and visual matching across reference-target similarity levels within each category.

Dataset / Category	Low Similarity		High Similarity	
	prototype	visual	prototype	visual
ISIC / Nevus	**66.2**	61.5	73.9	**78.4**
ISIC / Melanoma	**56.6**	50.3	64.9	**70.4**
SUIM / Human	**77.6**	68.6	71.4	**77.1**
SUIM / Reef	**56.9**	38.1	67.2	**69.1**

Although the distilled category prototype reduces variation in IoU across target images, its improvement in mIoU over pairwise visual matching remains modest, as shown in Fig. 1 (Right). To investigate why, we divide reference-target pairs into groups with low and high visual similarity using a DINOv2 correspondence score computed without target masks. The grouping protocol is provided in Appendix Sec. A2.1. For a fair comparison, we replace the original SAM backbone (Kirillov et al., 2023) in GF-SAM (Zhang et al., 2024a) with the same SAM3 backbone (Carion et al., 2026) used to produce the category prototype predictions. The category prototype does not access the reference during inference, but the two groups may contain target images with different difficulty. We therefore compare the two methods within each group. As shown in Tab. 2, the category prototype consistently outperforms visual matching on low-similarity pairs, whereas GF-SAM with SAM3 performs better on high-similarity pairs. This crossover explains the limited improvement in overall mIoU and shows that the two predictions are complementary across similarity levels: the category prototype performs better on pairs with low visual similarity, whereas visual matching performs better on pairs with high visual similarity.

4 Method

Our Dual-level Distillation on SAM3 (D^2-SAM3) constructs a representation that is category-stable yet target-adaptive. As illustrated in Fig. 3, it consists of reference adaptation and target inference, with SAM3 kept frozen. During reference adaptation, Category Prototype Distillation learns a reusable category prototype, while Visual Residual Distillation learns category-specific calibration of visual correspondence. During target inference, the category prototype produces the primary prediction, which is refined by a visual residual derived from calibrated reference-target correspondence. Reliability-aware Fusion controls the contribution of this residual based on the reliability of the correspondence and the uncertainty of the prototype-based prediction.

4.1 Category Prototype Distillation

Given an annotated reference (I^r, M^r) of category c, Category Prototype Distillation (CPD) learns a reusable category prototype from the GT-box-prompted prediction of frozen SAM3. We derive a bounding box $B^r = \mathrm{Box}(M^r)$ and obtain the GT-box-prompted teacher prediction:

$$P_T^r = \mathcal{S}_{\mathrm{box}}(I^r, B^r), \tag{1}$$

where $\mathcal{S}_{\mathrm{box}}$ denotes frozen SAM3 inference with a box prompt. Since the box localizes the annotated object, P_T^r provides reliable teacher supervision for learning the category prototype (refer to Sec. 3.3). We represent category c using learnable prompt embeddings $\widetilde{Z}_c \in \mathbb{R}^{L\times d}$ initialized from its textual embedding. Let $\mathcal{S}_{\mathrm{sem}}$ denote SAM3 inference with a category prototype and without geometric prompts. Starting from $\widetilde{Z}_c$, we first obtain a preliminary prototype $\bar{Z}_c$ using binary cross-entropy and Dice supervision from M^r, with prediction $\bar{P}_S^r = \mathcal{S}_{\mathrm{sem}}(I^r, \bar{Z}_c)$. We then refine it into Z_c, whose prediction is $P_S^r = \mathcal{S}_{\mathrm{sem}}(I^r, Z_c)$.

Let Ω_d contain pixels correctly predicted by P_T^r but incorrectly by $\bar{P}_S^r$, and let Ω_p contain pixels correctly predicted by $\bar{P}_S^r$, with correctness evaluated against M^r. The refinement objective is:

$$\mathcal{L}_{\mathrm{CPD}} = \lambda_d \mathcal{L}_{\mathrm{dist}}(P_S^r, P_T^r; \Omega_d) + \lambda_p \mathcal{L}_{\mathrm{pres}}(P_S^r, \bar{P}_S^r; \Omega_p) + \lambda_r \|Z_c - \bar{Z}_c\|_2^2. \tag{2}$$

Here, $\mathcal{L}_{\mathrm{dist}}$ distills the teacher prediction on pixels misclassified by the preliminary prototype, while $\mathcal{L}_{\mathrm{pres}}$ preserves the preliminary prototype's correct predictions. After reference adaptation, the final prototype Z_c is reused across target images to produce the prototype-based prediction:

$$P_{\mathrm{cat}}^t = \mathcal{S}_{\mathrm{sem}}(I^t, Z_c), \tag{3}$$

where P_{cat}^t is the primary prediction for the target image.

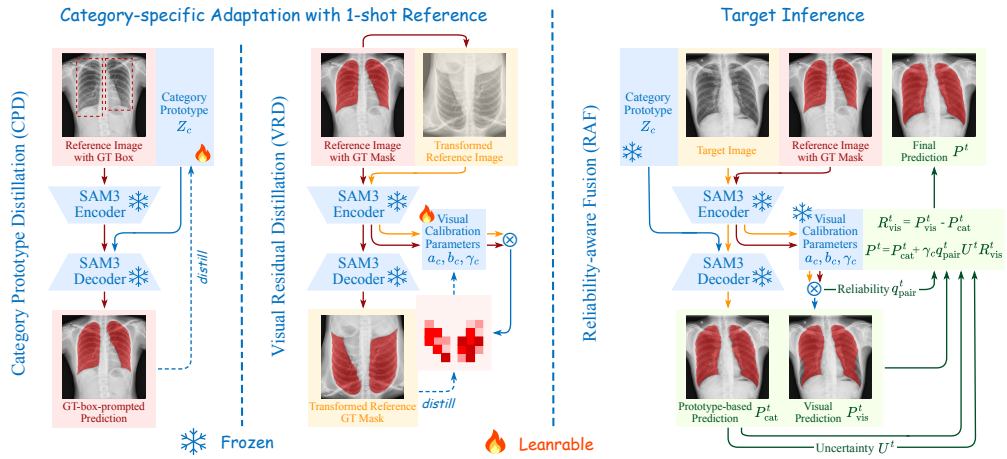


Figure 3: Overview of our proposed D^2-SAM3. **Left**: Category Prototype Distillation (CPD) transfers GT-box-prompted SAM3 predictions into a reusable category prototype. **Middle**: Visual Residual Distillation (VRD) learns category-specific calibration for reference-target visual correspondence. **Right**: Reliability-aware Fusion (RAF) adaptively incorporates this residual according to pairwise reliability and semantic uncertainty.

4.2 Visual Residual Distillation

Although the category prototype produces stable predictions across target images, it does not use the visual correspondence between the annotated reference and the current target and therefore cannot use their visual similarity to refine the target prediction (refer to Sec. 3.5). We therefore retain the annotated reference and extract $F^r = \mathcal{E}(I^r)$ using the frozen SAM3 image encoder. Let M^r denote the reference mask resized to the feature resolution. Using M^r, we average the reference features over the foreground and background regions:

$$v_c^{\mathrm{fg}} = \mathrm{Norm}\left(\frac{\sum_x M^r(x)F^r(x)}{\sum_x M^r(x)}\right), \quad v_c^{\mathrm{bg}} = \mathrm{Norm}\left(\frac{\sum_x (1-M^r(x))F^r(x)}{\sum_x (1-M^r(x))}\right), \tag{4}$$

where $\mathrm{Norm}(\cdot)$ denotes L_2 normalization. Given target features $F^t = \mathcal{E}(I^t)$, their foreground-background correspondence margin is:

$$A^t(x) = \mathrm{sim}\big(F^t(x), v_c^{\mathrm{fg}}\big) - \mathrm{sim}\big(F^t(x), v_c^{\mathrm{bg}}\big), \tag{5}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity. We rank the target locations using A^t according to the foreground proportion of the reference. The resulting threshold τ^t centers the correspondence values, while the foreground-background contrast κ^t normalizes their scale. The definitions of τ^t and κ^t are provided in Appendix Sec. A3.1. Because the distribution of visual correspondence may vary across categories, Visual Residual Distillation (VRD) learns a category-specific scale a_c and shift b_c to calibrate the correspondence:

$$P^t_{\mathrm{vis}}(x) = \sigma\left(a_c \frac{A^t(x) - \tau^t}{T_v \kappa^t} + b_c\right), \tag{6}$$

where T_v is a fixed temperature and $\sigma(\cdot)$ is the sigmoid function. The parameters a_c and b_c are learned through self-distillation using appearance-preserving views of the annotated reference and the corresponding GT-box-prompted predictions. This requires no additional annotated images; the complete objective is provided in Appendix Sec. A3.2. Rather than treating P^t_{vis} as an independent prediction, we express it as a residual relative to the prototype-based prediction:

$$R^t_{\mathrm{vis}} = P^t_{\mathrm{vis}} - P^t_{\mathrm{cat}}. \tag{7}$$

Here, R^t_{vis} is the difference between the visual and prototype-based predictions for the current reference-target pair. The term "residual" refers to this difference in prediction space rather than a feature-level residual connection.

4.3 Reliability-aware Fusion

Even after category-specific calibration, the reliability of the visual residual still depends on the reference-target correspondence. Reliability-aware Fusion (RAF) estimates this reliability using

Table 3: Quantitative comparison between the proposed D^2-SAM3 and existing methods on four CD-FSS benchmarks in mIoU (%). Results are taken from the original papers whenever available. Methods marked by * require reference adaptation and are averaged over ten random reference selections. The best result within each method group is highlighted in **bold**.

Method	ISIC		Chest X-Ray		DeepGlobe		SUIM	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Conventional ResNet-/ViT-based Methods								
ISA ICCV'25	37.2	56.1	83.4	86.3	44.3	52.7	**44.1**	–
DR-Adapter CVPR'24	40.8	48.9	82.4	82.3	41.3	50.1	–	–
PATNet ECCV'22	41.2	53.6	66.6	70.2	37.9	43.0	32.1	40.2
ABCDFSS CVPR'24	45.7	53.3	79.8	81.4	42.6	49.0	35.1	**41.3**
SDRC ICML'25	46.6	55.0	82.9	84.8	43.2	46.8	–	–
DFN ICML'25	50.4	58.5	83.2	**87.1**	39.5	47.7	–	–
LoEC CVPR'25	52.9	62.4	**83.9**	84.1	42.1	51.5	–	–
IFA CVPR'24	66.3	69.8	74.0	74.6	50.6	58.8	–	–
DATO CVPR'25	**68.8**	**70.3**	79.6	81.1	**51.1**	**59.3**	–	–
Vision Foundation Model-based Methods								
PerSAM ICLR'24	23.9	–	–	–	31.4	–	–	–
Matcher ICLR'24	38.6	35.0	66.8	66.3	48.1	50.9	37.9	41.4
APSeg CVPR'24	45.4	54.0	84.1	84.5	35.9	40.0	–	–
TAVP arXiv'24	54.9	–	–	–	46.1	–	–	–
CPS arXiv'26	51.3	56.6	70.0	70.2	46.0	49.8	42.8	48.8
GF-SAM NeurIPS'24	48.7	55.2	47.6	47.4	49.5	**57.7**	41.9	42.9
PerSAM-F* ICLR'24	20.5	–	38.7	–	46.5	–	10.3	–
VLP-SAM* arXiv'25	34.8	45.9	54.7	59.1	48.1	52.4	16.9	26.9
INSID3 CVPR'26	54.4	63.9	78.8	80.1	–	–	54.9	61.7
D^2-SAM3*	**63.0**	**64.5**	**89.1**	**89.6**	**56.5**	57.1	**71.0**	**72.6**

correspondence contrast and cycle consistency. Specifically, we pool the target features over the estimated foreground and background regions and match the pooled features back to the reference features to reconstruct the reference mask. Comparing this reconstruction with the known reference annotation produces a chance-normalized cycle reliability q^t_{cyc}. We additionally compare the target contrast κ^t with the reference self-contrast κ^r:

$$q^t_{\mathrm{ctr}} = \mathrm{clip}\left(\frac{\kappa^t}{\kappa^r}, 0, 1\right), \qquad q^t_{\mathrm{pair}} = q^t_{\mathrm{ctr}}\left(q^t_{\mathrm{cyc}}\right)^{\eta}, \tag{8}$$

where $\mathrm{clip}(z, l, u)$ restricts z to $[l, u]$, and η controls the effect of cycle reliability. The detailed cycle construction is provided in Appendix Sec. A3.3.

RAF weights the visual residual according to the uncertainty of the prototype-based prediction:

$$U^t(x) = \left[4P^t_{\mathrm{cat}}(x)\left(1 - P^t_{\mathrm{cat}}(x)\right)\right]^{\beta}, \tag{9}$$

where β controls how strongly the weight is concentrated in uncertain regions. Let γ_c denote the category-specific residual weight learned during reference adaptation. The final prediction is:

$$P^t(x) = \mathrm{clip}\left(P^t_{\mathrm{cat}}(x) + \gamma_c q^t_{\mathrm{pair}} U^t(x) R^t_{\mathrm{vis}}(x), 0, 1\right). \tag{10}$$

Thus, the prototype-based prediction remains primary, while the visual residual contributes when the reference-target correspondence is reliable and the prototype-based prediction is uncertain.

5 Experiments

We evaluate D^2-SAM3 on four Cross-Domain Few-Shot Segmentation (CD-FSS) datasets covering medical imaging, remote sensing, and underwater imagery. **ISIC2018** (Codella et al., 2019; Tschandl et al., 2018) contains images for skin lesion analysis across three lesion categories. **Chest X-Ray** (Candemir et al., 2013; Jaeger et al., 2013) comprises X-ray images collected for tuberculosis screening. **DeepGlobe** (Demir et al., 2018) contains satellite images from six terrain categories: urban, agriculture, rangeland, forest, water, and barren. Following PATNet (Lei et al., 2022), we divide the original images into small patches for evaluation. **SUIM** (Islam et al., 2020) is an underwater image dataset with six object categories: fish, plants, divers, robots, ruins, and rocks.

5.1 Comparison with State-of-the-Art Methods

Comparison with Conventional Methods. As shown in Tab. 3, our proposed D^2-SAM3 outperforms conventional ResNet- and ViT-based approaches in most settings, with particularly strong results on Chest X-Ray and SUIM [1]. It also remains competitive on ISIC and DeepGlobe, showing strong performance across target domains while keeping SAM3 frozen.

Comparison with Vision Foundation Model-based Methods. Our D^2-SAM3 achieves the best performance in seven of the eight settings. Compared with the strongest competing results, it improves ISIC by 8.1% and 0.6%, Chest X-Ray by 5.0% and 5.1%, and SUIM by 16.1% and 10.9% under the 1-shot and 5-shot settings, respectively. It also improves the 1-shot DeepGlobe result by 7.0%, while achieving a competitive 57.1% mIoU in the 5-shot setting. Overall, D^2-SAM3 outperforms methods based on vision foundation models in seven of the eight settings while keeping SAM3 frozen. Fig. 4 presents qualitative predictions of D^2-SAM3 on the four CD-FSS datasets. The backbone-controlled comparisons in Appendix Sec. A4.8 further show that these improvements arise from our designs rather than the SAM3 backbone alone.

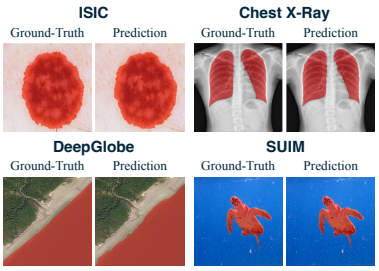


Figure 4: Qualitative results (red mask) of D^2-SAM3 on samples from four datasets. More samples are in Appendix Fig. A2.

Reference Efficiency. Our D^2-SAM3 exhibits strong annotation efficiency in the few-shot setting. Under the 1-shot setting, it achieves 63.0%, 89.1%, 56.5%, and 71.0% mIoU on ISIC, Chest X-Ray, DeepGlobe, and SUIM, respectively. Increasing the number of references to five brings further gains of 1.5%, 0.5%, 0.6%, and 1.6%. On average, the 1-shot result reaches 98.5% of the corresponding 5-shot mIoU. These results show that a single annotated reference is sufficient for D^2-SAM3 to approach its 5-shot performance. The extension to the 5-shot setting is described in Appendix Sec. A3.5.

5.2 Ablation Study

Tab. 4 evaluates the three key designs of our D^2-SAM3. CPD alone substantially outperforms VRD alone across all four datasets, supporting our use of the prototype-based prediction as the primary prediction. Combining CPD with VRD using static fusion improves ISIC, Chest X-Ray, DeepGlobe, and SUIM by 5.4%, 1.1%, 7.9%, and 2.4%, respectively. This shows that reference-target visual correspondence improves the prototype-based prediction. Adding RAF, which controls the residual contribution based on correspondence reliability and prediction uncertainty, further improves static fusion by 3.3%, 1.0%, 2.3%, and 1.1% across the four datasets. These results validate the complementary contributions of the category prototype, visual residual, and adaptive fusion. Detailed ablations of each design are provided in Appendix Sec. A4.3.

Table 4: Ablation studies of Category Prototype Distillation (CPD), Visual Residual Distillation (VRD), and Reliability-aware Fusion (RAF) under the 1-shot setting in mIoU (%). All use the same frozen SAM3.

Incorporated design	ISIC	Chest.	DeepG.	SUIM
VRD alone	32.0	68.8	33.1	29.4
CPD alone	55.3	86.8	47.4	67.0
CPD + VRD	60.7	87.9	55.3	69.4
CPD + VRD + RAF	**64.0**	**88.9**	**57.6**	**70.5**

5.3 Robustness to Reference Image Selection

Because D^2-SAM3 learns category-specific parameters from annotated references, its robustness to reference selection is important. We evaluate ten randomly selected references for each category and compare D^2-SAM3 with PerSAM-F (Zhang et al., 2024b), VLP-SAM (Sakurai et al.,

Table 5: Stability comparison over ten random reference selections, reported as mean mIoU (%) ± standard deviation.

Method	ISIC	Chest X-Ray	DeepGlobe	SUIM
PerSAM-F ICLR'24	20.5±7.3	38.7±8.6	46.5±11.7	10.3±6.9
VLP-SAM arXiv'25	34.8±6.0	54.7±13.4	48.1±17.8	16.9±7.2
CAT-SAM ECCV'24	52.3±5.6	75.0±3.6	29.7±20.3	18.6±6.3
CPD alone	55.3±6.4	88.3±1.2	50.2±3.4	63.2±4.1
D^2-SAM3	**63.0±2.5**	**89.1±0.6**	**56.5±1.4**	**71.0±1.4**

[1] Unless otherwise specified, all comparisons use a single randomly selected reference shared across methods, without averaging over multiple reference selections. Consequently, D^2-SAM3 results reported in other tables may differ slightly from those in Tab. 3, which are averaged over multiple reference selections.

2025), and CAT-SAM (Xiao et al., 2024a). As shown in Tab. 5, D^2-SAM3 achieves the highest mean mIoU and the lowest standard deviation across reference selections on all four datasets. In particular, the standard deviations are only 2.5, 0.6, 1.4, and 1.4 mIoU on ISIC, Chest X-Ray, DeepGlobe, and SUIM, respectively, demonstrating substantially lower sensitivity to the selected references. Compared with CPD alone, the full method further reduces the standard deviation, showing that incorporating the visual residual through RAF improves robustness to reference selection. Hyperparameter sensitivity is discussed in Appendix Sec. A4.4.

5.4 Efficiency Analysis

Tab. 6 compares representative CD-FSS methods using the same hardware and evaluation protocol. Conventional methods such as PATNet (Lei et al., 2022) and IFA (Nie et al., 2024) require more than 500 minutes of source training and target adaptation, resulting in total runtimes exceeding 540 minutes. In contrast, D^2-SAM3 completes reference adaptation in only 2 minutes by keeping SAM3 frozen and optimizing approximately 8.2K parameters per category (refer to Appendix Sec. A4.9). It subsequently processes the evaluation set in 5 minutes, reducing the total runtime to 7 minutes. Compared with Matcher and GF-SAM, D^2-SAM3 also requires less inference time and fewer total model parameters (840M), while requiring comparable FLOPs. These results demonstrate efficient reference adaptation and target inference with few trainable parameters.

Table 6: Efficiency comparison under the 1-shot CD-FSS setting. Preparation refers to the optimization performed before target inference, including source training, target adaptation, or self-distillation, depending on the method.

ResNet-based Methods					
Method	Runtime (min)				
	Preparation		Inference	Total	
PATNet	538		5	543	
IFA	541		5	546	
SAM-based Methods					
Method	Runtime (min)			Model size	FLOPs
	Prep.	Inf.	Total	(M)	(T)
Matcher	–	41	41	945	**9**
GF-SAM	–	9	9	945	**9**
D^2-SAM3	2	**5**	**7**	**840**	11

5.5 Generalization across Tasks

Beyond CD-FSS, we evaluate the generalization of D^2-SAM3 on three In-Domain Few-Shot Segmentation benchmarks and two One-Shot Part Segmentation benchmarks, with further details provided in Appendix Sec. A4.5. As shown in Tab. 7, D^2-SAM3 achieves the best performance on four of the five benchmarks. Compared with the strongest competing results, it improves PASCAL-5i (Shaban et al., 2017) and COCO-20i (Nguyen & Todorovic, 2019) by 14.3% and 16.6%, respectively. It also surpasses GF-SAM by 17.7% on PASCAL-Part (Chen et al., 2014) and 8.8% on PACO-Part (Ramanathan et al., 2023), while remaining competitive on FSS-1000 (Li et al., 2020). These results demonstrate strong generalization across object- and part-level segmentation tasks. Qualitative results are provided in Appendix Fig. A3.

Table 7: Generalization under the 1-shot setting in mIoU (%). The best results are highlighted in **bold**.

Method	In-Domain FSS			Part Segmentation	
	Pascal.	COCO.	FSS.	PASCAL.	PACO.
PerSAM	43.1	23.0	71.2	32.5	22.5
PerSAM-F	48.5	23.5	75.6	32.9	22.7
Matcher	68.1	52.7	87.0	42.9	34.7
VRP-SAM	71.9	53.9	–	–	–
GF-SAM	72.1	58.7	**88.0**	44.5	36.3
CPS	–	49.2	81.7	–	–
INSID3	–	57.6	–	–	–
D^2-SAM3	**86.4**	**75.3**	86.8	**62.2**	**45.1**

6 Conclusion

In this work, we revisit pairwise visual matching in CD-FSS and identify a stability-adaptability trade-off between category prototypes and reference-target visual correspondence. We propose D^2-SAM3, which combines a reusable category prototype with visual residuals derived from reference-target correspondence. Category Prototype Distillation transfers GT-box-prompted predictions into a stable category prototype, Visual Residual Distillation learns category-specific calibration of visual correspondence, and Reliability-aware Fusion controls the residual contribution based on correspondence reliability and uncertainty of the prototype-based prediction. With SAM3 frozen, D^2-SAM3 optimizes only approximately 8.2K parameters per category. Extensive experiments demonstrate strong segmentation performance, robustness to reference selection, and generalization across tasks and semantic granularities.

AI USE STATEMENT

Generative AI tools were used to assist with code implementation, feedback on experimental design and analysis, and the polishing of manuscript text. The research problem and core method were developed by the authors. All AI-assisted code was reviewed and tested, and all reported results, claims, citations, and figures were manually verified by the authors. Generative AI was not used to generate experimental datasets or determine the final scientific conclusions. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses existing datasets made publicly available for research and does not involve new data collection or direct interaction with human participants. We followed the applicable licenses and evaluation protocols of these datasets. The proposed method is intended for research purposes; its performance may be affected by biases or limitations in the training and evaluation data and should not be directly used in safety-critical or clinical settings without appropriate validation and oversight. We are not aware of any additional ethical concerns or conflicts of interest associated with this work.

REPRODUCIBILITY STATEMENT

We provide the methodological and experimental details necessary for reproducing our results in the main paper and appendix. These include descriptions of the proposed method, experimental settings, datasets, evaluation protocols, and relevant implementation details.

REFERENCES

Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P Musco, Rahul K Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *TMI*, 2013.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *ICLR*, 2026.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

Hao Chen, Yonghan Dong, Zheming Lu, Yunlong Yu, and Jungong Han. Pixel matching network for cross-domain few-shot segmentation. In *WACV*, 2024.

Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, medical image segmentation, and more. *arXiv preprint arXiv:2304.09148*, 2023.

Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, 2014.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

Runmin Cong, Anpeng Wang, Bin Wan, Cong Zhang, Xiaofei Zhou, and Wei Zhang. Divide-and-conquer decoupled network for cross-domain few-shot segmentation. In *AAAI*, 2026.

Claudia Cuttano, Gabriele Trivigno, Giuseppe Averta, and Carlo Masone. Sansa: Unleashing the hidden semantics in SAM2 for few-shot segmentation. In *NeurIPS*, 2025.

Claudia Cuttano, Gabriele Trivigno, Christoph Reich, Daniel Cremers, Carlo Masone, and Stefan Roth. Insid3: Training-free in-context segmentation with dinov3. In *CVPR*, 2026.

Pasquale De Marinis, Pieter M Blok, Uzay Kaymak, Rogier Brussee, Gennaro Vessio, and Giovanna Castellano. Distillfss: Synthesizing few-shot knowledge into a lightweight segmentation model. *arXiv preprint arXiv:2512.05613*, 2025.

Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *CVPRW*, 2018.

Lei Ding, Kun Zhu, Daifeng Peng, Hao Tang, Kuiwu Yang, and Lorenzo Bruzzone. Adapting segment anything model for change detection in vhr remote sensing images. *TGRS*, 2024.

Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *CVPR*, 2023.

Haoran Fan, Qi Fan, Maurice Pagnucco, and Yang Song. Darnet: Bridging domain gaps in cross-domain few-shot segmentation with dynamic adaptation. *arXiv preprint arXiv:2312.04813*, 2023.

Qi Fan, Kaiqi Liu, Nian Liu, Hisham Cholakkal, Rao Muhammad Anwer, Wenbin Li, and Yang Gao. Adapting in-domain few-shot segmentation to new domains without retraining. In *ICCV*, 2025.

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 *ECCV*, 2024.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *NeurIPS*, 2020.

Jonas Herzog. Adapt before comparison: A new perspective on cross-domain few-shot segmentation. In *CVPR*, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Xiaoke Huang, Jianfeng Wang, Yansong Tang, Zheng Zhang, Han Hu, Jiwen Lu, Lijuan Wang, and Zicheng Liu. Segment and caption anything. In *CVPR*, 2024.

Xinyang Huang, Chuang Zhu, and Wenkai Chen. Restnet: Boosting cross-domain few-shot segmentation with residual transformation network. In *BMVC*, 2023.

Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic segmentation of underwater imagery: Dataset and benchmark. In *IROS*, 2020.

Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les Folio, Jenifer Siegelman, Fiona Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K Singh, Sameer Antani, et al. Automatic tuberculosis screening using chest radiographs. *TMI*, 2013.

Dong Jing, Xiaolong He, Yutian Luo, Nanyi Fei, Guoxing Yang, Wei Wei, Huiwen Zhao, and Zhiwu Lu. FineCLIP: Self-distilled region-based CLIP for better fine-grained understanding. In *NeurIPS*, 2024.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. In *NeurIPS*, 2023.

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 *ICCV*, 2023.

Shuo Lei, Xuchao Zhang, Jianfeng He, Fanglan Chen, Bowen Du, and Chang-Tien Lu. Cross-domain few-shot semantic segmentation. In *ECCV*, 2022.

Kaiyu Li, Ruixun Liu, Xiangyong Cao, Xueru Bai, Feng Zhou, Deyu Meng, and Zhi Wang. Segearth-ov: Towards training-free open-vocabulary segmentation for remote sensing images. In *CVPR*, 2025.

Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *CVPR*, 2020.

Nanqing Liu, Xun Xu, Yongyi Su, Haojie Zhang, and Heng-Chao Li. Pointsam: Pointly-supervised segment anything model for remote sensing images. *TGRS*, 2025.

Yang Liu, Chenchen Jing, Hengtao Li, Muzhi Zhu, Hao Chen, Xinlong Wang, and Chunhua Shen. A simple image segmentation framework via in-context examples. *NeurIPS*, 2024a.

Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. In *ICLR*, 2024b.

Zhihe Lu, Sen He, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Simpler is better: Few-shot semantic segmentation with classifier weight transformer. In *ICCV*, 2021.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 2024.

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. In *CVPR*, 2026.

Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *ICCV*, 2019.

Jiahao Nie, Yun Xing, Gongjie Zhang, Pei Yan, Aoran Xiao, Yap-Peng Tan, Alex C Kot, and Shijian Lu. Cross-domain few-shot segmentation via iterative support-query correspondence mining. In *CVPR*, 2024.

Jiahao Nie, Guanqiao Fu, Wenbin An, Yap-Peng Tan, Alex C Kot, and Shijian Lu. Cross-domain few-shot segmentation via multi-view progressive adaptation. In *CVPR*, 2026a.

Jiahao Nie, Yun Xing, Wenbin An, Qingsong Zhao, Jiawei Shao, Yap-Peng Tan, Alex C Kot, Shijian Lu, and Xuelong Li. Boosting sam for cross-domain few-shot segmentation via conditional point sparsification. *arXiv preprint arXiv:2602.05218*, 2026b.

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.

Shi-Feng Peng, Guolei Sun, Yong Li, Hongsong Wang, and Guo-Sen Xie. Sam-aware graph prompt reasoning network for cross-domain few-shot segmentation. In *AAAI*, 2025.

Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *CVPR*, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Kosuke Sakurai, Ryotaro Shimizu, and Masayuki Goto. Vision and language reference prompt into sam for few-shot segmentation. *arXiv preprint arXiv:2502.00719*, 2025.

Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In *BMVC*, 2017.

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.

Jiapeng Su, Qi Fan, Wenjie Pei, Guangming Lu, and Fanglin Chen. Domain-rectifying adapter for cross-domain few-shot segmentation. In *CVPR*, 2024.

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 *CVPR*, 2024.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017.

Pinzhuo Tian, Xiaojie Zhou, Hang Yu, and Shaorong Xie. High specificity guided cross-domain few-shot segmentation. *TMM*, 2025.

Jintao Tong, Yixiong Zou, Yuhua Li, and Ruixuan Li. Lightweight frequency masker for cross-domain few-shot semantic segmentation. In *NeurIPS*, 2024.

Jintao Tong, Yixiong Zou, Guangyao Chen, Yuhua Li, and Ruixuan Li. Self-disentanglement and re-composition for cross-domain few-shot segmentation. In *ICML*, 2025.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 2018.

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 *CVPR*, 2022.

Yuan Wang, Rui Sun, Naisong Luo, Yuwen Pan, and Tianzhu Zhang. Image-to-image matching via foundation models: A new perspective for open-vocabulary semantic segmentation. In *CVPR*, 2024.

Zhaoyang Wei, Pengfei Chen, Xuehui Yu, Guorong Li, Jianbin Jiao, and Zhenjun Han. Semantic-aware sam for point-prompted instance segmentation. In *CVPR*, 2024.

Jiamin Wu, Xin Liu, Xiaotian Yin, Tianzhu Zhang, and Yongdong Zhang. Task-adaptive prompted transformer for cross-domain few-shot learning. In *AAAI*, 2024a.

Junde Wu, Ziyue Wang, Mingxuan Hong, Wei Ji, Huazhu Fu, Yanwu Xu, Min Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *Medical image analysis*, 2025.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. In *ICLR*, 2024b.

Aoran Xiao, Weihao Xuan, Heli Qi, Yun Xing, Ruijie Ren, Xiaoqin Zhang, Ling Shao, and Shijian Lu. Cat-sam: Conditional tuning for few-shot adaptation of segment anything model. In *ECCV*, 2024a.

Aoran Xiao, Weihao Xuan, Heli Qi, Yun Xing, Naoto Yokoya, and Shijian Lu. Segment anything with multiple modalities. *arXiv preprint arXiv:2408.09085*, 2024b.

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. In *ICML*, 2025.

Anqi Zhang, Guangyu Gao, Jianbo Jiao, Chi Liu, and Yunchao Wei. Bridge the points: Graph-based few-shot segment anything semantically. In *NeurIPS*, 2024a.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *ICCV*, 2019.

Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Xianzheng Ma, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. In *ICLR*, 2024b.

Muzhi Zhu, Yang Liu, Zekai Luo, Chenchen Jing, Hao Chen, Guangkai Xu, Xinlong Wang, and Chunhua Shen. Unleashing the potential of the diffusion model in few-shot semantic segmentation. In *NeurIPS*, 2024.

Yazhou Zhu, Shidong Wang, Tao Zhou, Zechao Li, Haofeng Zhang, and Ling Shao. Cross-domain few-shot medical image segmentation via dynamic semantic matching. *TIP*, 2025.

APPENDIX

A1 LIMITATION AND FUTURE WORK

Although D^2-SAM3 optimizes only a small number of category-specific parameters, it still requires a short reference adaptation process and is therefore not strictly training-free. In addition, retaining reference-derived visual evidence introduces modest computational and memory overhead compared with purely point-conditioned inference, despite requiring no target-time gradient updates. The overall efficiency also remains constrained by the frozen SAM3 backbone. Future work may explore optimization-free prototype generation, more compact visual representations, and efficient correspondence computation to further reduce adaptation and inference costs.

A2 PRELIMINARY STUDY AND MOTIVATION

A2.1 EFFECT OF REFERENCE-TARGET SIMILARITY

Controlled Comparison. To compare the category prototype with pairwise visual matching using the same SAM3 architecture, we construct a controlled GF-SAM variant (Zhang et al., 2024a). Its DINOv2 matcher (Oquab et al., 2023) and mask selection procedure are retained, while the original SAM image encoder, prompt encoder, and mask decoder (Kirillov et al., 2023) are replaced by the corresponding SAM3 modules (Carion et al., 2026). The category prototype baseline uses the distilled prototype learned from the annotated reference mask.

Reference-Target Similarity Score. We compute the similarity score from the foreground correspondences produced by the DINOv2 matcher. Let $D^r(y)$ and $D^t(x)$ denote the L_2-normalized DINOv2 features at reference location y and target location x, respectively, and let M^r denote the reference mask resized to the feature resolution. For each target location, its similarity to the annotated reference foreground is defined as:

$$C^t(x) = \max_y \left[M^r(y) \, \mathrm{sim}\big(D^t(x), D^r(y)\big) \right], \tag{A1}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity. Thus, $C^t(x)$ is the highest similarity between target location x and the reference foreground. To prevent unrelated target regions from dominating the score, we aggregate only the top-ranked target locations. Their number is determined by the foreground proportion of the annotated reference:

$$\rho^r = \frac{\sum_y \mathbb{I}\big[M^r(y) > 0.5\big]}{|M^r|}. \tag{A2}$$

We select the $n^t = \max\{1, \mathrm{round}(\rho^r |\Omega^t|)\}$ target locations with the highest correspondence values, where Ω^t denotes the target feature grid. Denoting these locations by Ω^t_{top}, the reference-target similarity score is:

$$s(I^r, I^t) = \frac{1}{|\Omega^t_{\mathrm{top}}|} \sum_{x \in \Omega^t_{\mathrm{top}}} C^t(x). \tag{A3}$$

This score uses the annotated reference mask but does not access the target mask.

Grouping Protocol. We compute $s(I^r, I^t)$ for every reference–target pair and rank the pairs separately within each category. This prevents differences in score scale across categories from affecting the grouping. Within each category, the 40 pairs with the lowest scores form the low-similarity group, while the 40 pairs with the highest scores form the high-similarity group. The scores and group assignments are used only for this analysis and do not alter either method's predictions. GF-SAM with SAM3 and the distilled category prototype are evaluated on exactly the same selected pairs. Because the grouping changes the composition of target images, the two groups may differ in segmentation difficulty. We therefore compare the two methods within each similarity group.

A2.2 Using Manual Text Prompts as Category Prototype

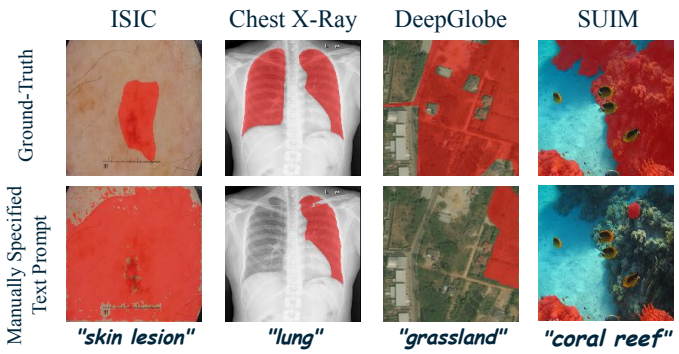


Figure A1: Qualitative results of manually specified text prompts.

Fig. A1 provides qualitative results using manually specified text prompts. Although category names describe the corresponding categories, they produce inaccurate or incomplete predictions across different target domains. This suggests that descriptions in natural language are insufficient to characterize target-category appearances under severe domain shifts. We evaluate five semantically valid text prompts for each category while keeping the model and target images fixed. As shown in Tab. A1, manual prompts consistently yield poor performance and vary substantially with different prompts. For example, on Chest X-Ray, "lung", "pulmonary region", and "lungs in a chest X-ray" obtain 20.2%, 18.8%, and 22.2% mIoU, respectively. For the forest category of DeepGlobe, "forest", "woodland", and "tree-covered land" achieve 21.5%, 20.2%, and 14.6% mIoU. Similar variations are observed for prompts such as "nevus", "skin mole", and "melanocytic nevus" on ISIC. These results show that directly prompting SAM3 with category descriptions does not provide reliable generalization under severe domain shifts.

A3 Additional Method Details

A3.1 Pairwise Visual Normalization

Let $\mathcal{X}^r$ and $\mathcal{X}^t$ denote the spatial grids of the reference and target feature maps, respectively. The foreground proportion derived from the downsampled reference mask is:

$$\rho^r = \frac{1}{|\mathcal{X}^r|} \sum_{x \in \mathcal{X}^r} M^r(x). \quad \text{(A4)}$$

Table A1: Sensitivity to manually specified text prompts under the 1-shot setting in mIoU (%). The ranges report the minimum and maximum performance over five semantically valid prompts evaluated on the same target images.

Prompt setting	ISIC	Chest X-Ray	DeepGlobe	SUIM
category names	10.2	24.4	21.5	16.1
Five prompts (range)	4.8-12.6	18.0-27.6	10.2-22.0	7.7-16.1

We select the top-ranked target locations covering approximately ρ^r of $\mathcal{X}^t$ and denote this set by Ω^t_{fg}. The rank threshold and pair-specific contrast are:

$$\tau^t = \min_{x \in \Omega^t_{\text{fg}}} A^t(x), \quad \kappa^t = \text{mean}_{x \in \Omega^t_{\text{fg}}} A^t(x) - \text{mean}_{x \notin \Omega^t_{\text{fg}}} A^t(x). \quad \text{(A5)}$$

Consequently, $A^t(x) - \tau^t$ determines whether a location lies above the pair-specific rank boundary, while κ^t normalizes the response according to its foreground-background separability. The reference self-contrast κ^r is computed analogously by treating the reference as both the source and target.

A3.2 Visual Residual Distillation Objective

Given the annotated reference (I^r, M^r), we construct V appearance-preserving views $\{(I^r_q, M^r_q)\}_{q=1}^V$ as pseudo target instances. The transformed mask of each view is converted into a box prompt, producing:

$$P^r_{T,q} = \mathcal{S}_{\text{box}}\big(I^r_q, \text{Box}(M^r_q)\big). \quad \text{(A6)}$$

For each view, CPD and VRD produce $P^r_{\text{cat},q}$ and $P^r_{\text{vis},q}$, respectively. Their residual-enhanced composition is:

$$P^r_{\text{res},q} = \text{clip}\left(P^r_{\text{cat},q} + \gamma_c r_q U_q\big(P^r_{\text{vis},q} - P^r_{\text{cat},q}\big), 0, 1\right), \quad \text{(A7)}$$

where r_q is the contrast-based reliability of the q-th view and U_q is computed from $P^r_{\mathrm{cat},q}$ following Eq. 9. We parameterize the category-specific residual weight as $\gamma_c = \gamma_{\max}\sigma(\xi_c)$, where ξ_c is learned during reference adaptation and $\gamma_{\max}$ is a fixed upper bound. Let $\Omega_{T,q}$ denote pixels correctly predicted by the teacher with respect to M^r_q. The calibration parameters a_c and b_c and the residual-weight parameter ξ_c are jointly optimized using:

$$\begin{aligned}\mathcal{L}_{\mathrm{VRD}} = \frac{1}{V}\sum_{q=1}^{V}\big[&\lambda_v \mathcal{L}_{\mathrm{dist}}(P^r_{\mathrm{vis},q}, P^r_{T,q}; \Omega_{T,q}) \\ &+ \lambda_f \mathcal{L}_{\mathrm{dist}}(P^r_{\mathrm{res},q}, P^r_{T,q}; \Omega_{T,q}) \\ &+ \lambda_g \mathcal{L}_{\mathrm{seg}}(P^r_{\mathrm{res},q}, M^r_q)\big],\end{aligned} \tag{A8}$$

where $\mathcal{L}_{\mathrm{dist}}$ is masked binary cross-entropy and $\mathcal{L}_{\mathrm{seg}}$ combines binary cross-entropy and Dice losses. The first term distills the teacher into the calibrated visual prediction, while the remaining terms supervise its residual composition with the category prediction. We use six deterministic views: the original reference, horizontal flipping, brightness factors of 0.75 and 1.25, and contrast factors of 0.75 and 1.25. We set $T_v = 0.25$, $\lambda_v = 0.5$, and $\lambda_f = \lambda_g = 1$. The parameters a_c, b_c, and ξ_c are optimized for 40 iterations.

A3.3 Cycle-consistency Reliability

We pool target foreground and background anchors from the highest and lowest correspondence margins, respectively. These target-derived anchors are matched back to the reference features, producing a reconstructed reference mask $\widehat{M}^r_{\mathrm{cyc}}$ with the same foreground proportion as M^r. Its raw overlap can be biased by foreground size. Under random overlap, the expected IoU between two masks with foreground proportion ρ^r is $q^r_0 = \rho^r/(2-\rho^r)$. We therefore define the chance-normalized cycle reliability as:

$$q^t_{\mathrm{cyc}} = \mathrm{clip}\left(\frac{\mathrm{IoU}(\widehat{M}^r_{\mathrm{cyc}}, M^r) - q^r_0}{1 - q^r_0}, 0, 1\right). \tag{A9}$$

Accordingly, random-level reconstruction yields zero reliability, whereas perfect reconstruction yields one. This verification uses only the annotated reference and does not require the target mask.

A3.4 Multi-query Mask Selection and Aggregation

SAM3 predicts multiple object queries for each image. Let $\mathbf{A}_j$, s_j, and s_p denote the mask logit, classification logit, and global presence logit of the j-th query, respectively. Since different queries may capture complementary parts or instances of the target category, our distillation and inference operate on a set of queries rather than a single query.

Teacher-query aggregation. During category-level distillation, the teacher is prompted using the annotated reference region. We construct $\mathcal{Q}^T_r$ by first selecting the query with the highest IoU against the reference mask, with query masks binarized at 0.5. We then greedily add the query that maximizes the union IoU, stopping when the improvement is less than 0.005 or four queries have been selected. The resulting teacher probability is defined as the pixel-wise union of the selected queries:

$$P^r_T = \max_{j \in \mathcal{Q}^T_r} \sigma(\mathbf{A}^T_j). \tag{A10}$$

The selection is performed only on the labeled reference image and introduces no supervision from target images.

Differentiable student aggregation. For each student query, we compute its presence-aware confidence as:

$$c_j = \sigma(s_j)\sigma(s_p). \tag{A11}$$

To retain differentiability during distillation, the hard query selection is replaced by a soft confidence gate:

$$g_j = \sigma\left(\frac{c_j - \tau_c}{T_q}\right), \tag{A12}$$

where $\tau_c = 0.5$ is the query-confidence threshold and $T_q = 0.05$ is the soft-gating temperature. The aggregated student probability is:

$$P_S^r = \max_j \left[g_j \, \sigma(\mathbf{A}_j^S) \right]. \tag{A13}$$

Within CPD, teacher knowledge is selectively distilled on pixels correctly predicted by the teacher but misclassified by the preliminary category prototype, while the preservation objective retains the prototype's already correct predictions. Specifically, $\mathcal{L}_{\text{pres}}$ is a masked squared error, averaged separately over the correctly predicted foreground and background:

$$\mathcal{L}_{\text{pres}} = \frac{1}{2} \sum_{b \in \{0,1\}} \frac{\sum_{x \in \Omega_p,\, M^r(x)=b} \big(P_S^r(x) - \bar{P}_S^r(x)\big)^2}{\max\big(1, |\{x \in \Omega_p : M^r(x) = b\}|\big)}. \tag{A14}$$

The preliminary prediction $\bar{P}_S^r$ is kept fixed.

Query aggregation during inference. At target inference, we use the corresponding hard query selection:

$$\mathcal{Q}_t = \{ j \mid c_j \geq \tau_c \}. \tag{A15}$$

The prototype-based prediction is obtained by:

$$P_{\text{cat}}^t = \max_{j \in \mathcal{Q}_t} \sigma(\mathbf{A}_j). \tag{A16}$$

It is then combined with the visual residual through RAF according to Eq. 10, yielding the final prediction P^t. The final mask is:

$$\hat{M}^t = \text{Resize}\left(\mathbb{I}\left[P^t \geq \tau_m \right] \right). \tag{A17}$$

The soft aggregation used during distillation can be regarded as the continuous counterpart of the confidence-filtered query union used during inference. This formulation allows multiple complementary queries to contribute to the prototype-based prediction.

A3.5 Extension to the K-Shot Setting

For $K > 1$, all annotated references jointly supervise a shared category prototype and category-specific calibration of visual correspondence, with the objectives averaged across references. For each reference i, we evaluate its visual prediction on the other $K-1$ annotated references. Its leave-one-out utility is the maximum mean IoU improvement over the category-only prediction, evaluated over six uniformly spaced fusion weights in $[0, \gamma_c]$ using the uncertainty-weighted residual fusion. Self-matching is excluded. We retain the two references with the highest utilities in the 5-shot setting. During target inference, their visual predictions are aggregated according to the leave-one-out utility and the reliability of the current reference-target correspondence, and then combined with the prototype-based prediction through RAF. This procedure requires neither target annotations nor gradient updates during inference.

A4 Experiments

A4.1 Computational Environment

All experiments are implemented in Python 3.12.13 using PyTorch 2.10.0 and CUDA 12.8. Each run uses a single NVIDIA H100 80GB GPU. We use the official SAM3 implementation and pretrained checkpoint. Reference adaptation and target inference are performed in FP32, with mixed precision and TF32 disabled for consistent numerical evaluation.

A4.2 Optimization Details

Pretrained SAM3 remains frozen throughout reference adaptation. We first learn a preliminary category prototype from the reference mask using binary cross-entropy and Dice losses for 20 iterations with AdamW and a learning rate of 1×10^{-3}. Starting from this prototype, CPD selectively distills the GT-box-prompted predictions for another 20 iterations using AdamW with a learning rate of

Table A3: Sensitivity to the main hyperparameters under the 1-shot setting in mIoU (%). Default settings are highlighted in **bold**.

Parameter	Value	ISIC	DeepG.	SUIM	Parameter	Value	ISIC	DeepG.	SUIM
	0.025	63.4	57.5	70.4		0.1	63.5	57.3	70.4
	0.05	63.6	57.5	70.5		0.2	63.8	57.5	70.4
λ_d	**0.1**	**64.0**	**57.6**	**70.5**	γ_{max}	**0.25**	**64.0**	**57.6**	**70.5**
	0.2	64.2	57.3	70.5		0.3	63.7	57.4	70.3
	0.4	63.9	57.4	70.3		0.4	63.8	57.4	70.2
						0.5	63.8	57.5	70.5
	0.25	63.8	57.3	70.3	η	**1.0**	**64.0**	**57.6**	**70.5**
	0.5	64.0	57.4	70.4		2.0	63.7	57.4	70.3
λ_p	**1.0**	**64.0**	**57.6**	**70.5**		0.1	63.7	57.3	70.1
	2.0	63.6	57.6	70.4		0.2	63.8	57.5	70.3
	4.0	63.1	57.5	70.3	T_v	**0.25**	**64.0**	**57.6**	**70.5**
						0.3	63.9	57.6	70.3
	0.0025	63.7	57.4	70.2		0.4	63.7	57.5	70.2
	0.005	64.0	57.6	70.5		0.5	63.7	57.3	69.9
λ_r	**0.01**	**64.0**	**57.6**	**70.5**		0.75	63.9	57.4	70.2
	0.02	63.9	57.4	70.4	β	**1.0**	**64.0**	**57.6**	**70.5**
	0.04	63.8	57.3	70.4		1.5	63.8	57.6	70.3
						2.0	63.7	57.5	70.4

5×10^{-3}. After fixing the resulting category prototype, we jointly optimize the VRD calibration parameters a_c and b_c and the RAF residual-weight parameter ξ_c for 40 iterations using Adam with a learning rate of 5×10^{-2}. All stages use only the annotated references. After reference adaptation, all learned parameters remain fixed, and no gradient update is required during target inference.

A4.3 Fine-grained Ablation Studies

Tab. A2 further examines the internal designs of CPD, VRD, and RAF. For CPD, all variants use the same total budget of 40 adaptation iterations. Using reference-mask supervision for all 40 iterations yields 58.1% mIoU on ISIC, while naive teacher distillation increases it to 60.9%. Selective distillation and prediction preservation further increase mIoU to 64.0%. The same progressive improvement is observed on DeepGlobe. For VRD, pairwise normalization improves mIoU from 59.7% to 61.2% on ISIC and from 55.6% to 56.5% on DeepGlobe. Adding category-specific calibration and reliability control further increases mIoU to 64.0% and 57.6%, respectively. Finally, Pair-adaptive Fusion substantially improves Static Fusion, showing that the contribution of the visual residual should depend on the reliability of the reference-target correspondence. Incorporating the uncertainty of the prototype-based prediction further improves DeepGlobe from 56.8% to 57.6% while leaving the ISIC result nearly unchanged. These results show that correspondence reliability and prediction uncertainty provide complementary control over the visual residual.

Table A2: Fine-grained ablations of Category Prototype Distillation (CPD), Visual Residual Distillation (VRD), and Reliability-aware Fusion (RAF) under the 1-shot setting in mIoU (%). Preliminary Prototype uses only reference-mask supervision, while Naive Teacher Distillation transfers all teacher-correct predictions. The VRD variants progressively introduce pairwise normalization and category-specific calibration with reliability control. Pair-adaptive Fusion uses pairwise reliability, while Full RAF additionally incorporates semantic uncertainty. For each block, the other components are kept in their full configurations. Full configurations are highlighted in **bold**.

Incorporated design	ISIC	DeepGlobe
Preliminary Prototype	58.1	55.7
+ Naive Teacher Distillation	60.9	56.2
+ **Full CPD**	**64.0**	**57.6**
Unnormalized Visual Response	59.7	55.6
+ Pair-normalized Response	61.2	56.5
+ **Full VRD**	**64.0**	**57.6**
Static Fusion	60.7	55.3
+ Pair-adaptive Fusion	63.9	56.8
+ **Full RAF**	**64.0**	**57.6**

Table A4: Generalization to In-Domain Few-Shot Segmentation and One-Shot Part Segmentation benchmarks. The best result in each column is highlighted in **bold**.

Method	In-Domain Few-Shot Segmentation						One-Shot Part Segmentation	
	PASCAL-5i (Mean)		COCO-20i (Mean)		FSS-1000		PASCAL-Part	PACO-Part
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot		
PerSAM	43.1	–	23.0	–	71.2	–	32.5	22.5
PerSAM-F	48.5	–	23.5	–	75.6	–	32.9	22.7
Matcher	68.1	74.0	52.7	60.7	87.0	**89.6**	42.9	34.7
VRP-SAM	71.9	–	53.9	–	–	–	–	–
GF-SAM	72.1	82.6	58.7	66.8	88.0	88.9	44.5	36.3
CPS	–	–	49.2	56.9	81.7	82.8	–	–
INSID3	–	–	57.6	65.1	–	–	–	–
D^2-SAM3	**86.4**	**88.4**	**75.3**	**77.4**	**86.8**	87.6	**62.2**	**45.1**

A4.4 Hyperparameter Sensitivity

Tab. A3 evaluates the sensitivity of D^2-SAM3 to its main hyperparameters. The weights λ_d, λ_p, and λ_r control teacher distillation, prediction preservation, and prototype regularization in CPD, respectively. The hyperparameter $\gamma_{\max}$ sets the upper bound on the learned category-specific residual weight γ_c, while η controls the effect of cycle reliability. The temperature T_v is used in visual correspondence normalization, and β controls the concentration of the uncertainty weight. Performance remains stable over the tested ranges, indicating low sensitivity to these choices. We set $\lambda_d = 0.1$, $\lambda_p = 1.0$, $\lambda_r = 0.01$, $\gamma_{\max} = 0.25$, $\eta = 1.0$, $T_v = 0.25$, and $\beta = 1.0$ by default.

A4.5 Additional Generalization Benchmarks

In-Domain Few-Shot Segmentation. We evaluate our proposed D^2-SAM3 on three In-Domain Few-Shot Segmentation benchmarks: PASCAL-5i, COCO-20i, and FSS-1000. PASCAL-5i (Shaban et al., 2017) is constructed from PASCAL VOC with augmented segmentation annotations. Its 20 object categories are partitioned into four folds, each containing five evaluation categories. COCO-20i (Nguyen & Todorovic, 2019) follows a similar protocol on MS COCO, dividing its 80 categories into four folds of 20 categories. For both datasets, we report performance averaged over all four folds. FSS-1000 (Li et al., 2020) contains 1,000 categories covering diverse object appearances. Following the standard split, we evaluate on its 240 test categories.

One-Shot Part Segmentation. We further examine whether D^2-SAM3 generalizes to concepts at a finer granularity using PASCAL-Part and PACO-Part. PASCAL-Part (Chen et al., 2014) extends PASCAL VOC with part-level annotations. Following Matcher (Liu et al., 2024b), we evaluate 56 object-part categories spanning animals, indoor objects, persons, and vehicles. PACO-Part (Ramanathan et al., 2023) provides detailed part masks for common objects. Following Matcher (Liu et al., 2024b), we filter rare and low-quality annotations, resulting in a one-shot benchmark with 303 object-part categories divided into four folds. We report mIoU averaged over all categories and folds.

A4.6 Generalization across Tasks

Tab. A4 presents the complete generalization results, including 1-shot and 5-shot evaluations on three in-domain FSS benchmarks and one-shot evaluations on two part segmentation benchmarks. Our proposed D^2-SAM3 achieves the best performance in most settings and remains competitive in the others, demonstrating strong generalization across tasks and semantic granularities.

A4.7 Additional Qualitative Results

Fig. A2 presents additional predictions on the four CD-FSS datasets under substantial domain shifts. Fig. A3 presents predictions for in-domain objects and fine-grained object parts. Together, these figures illustrate the results of D^2-SAM3 across different domains and semantic granularities.

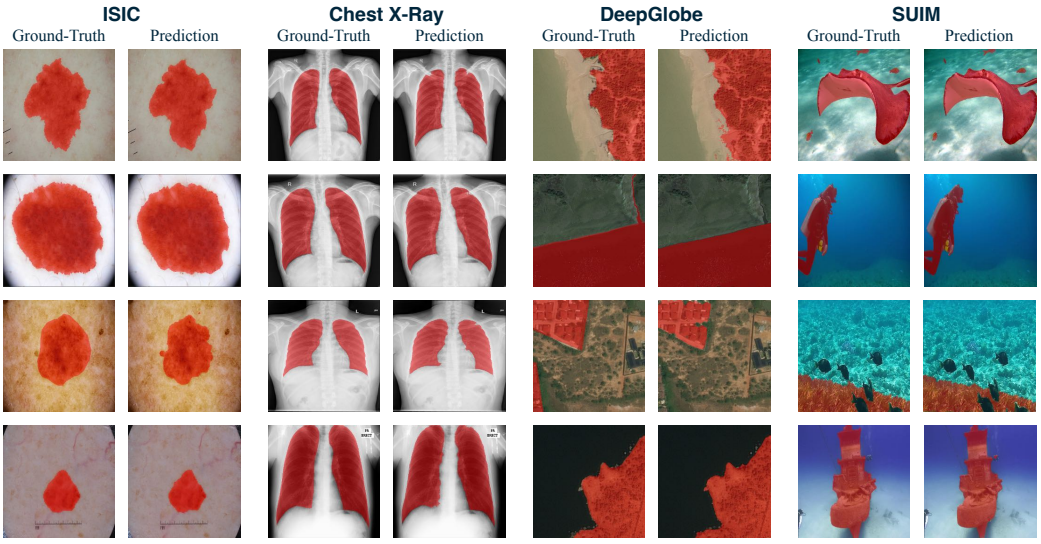


Figure A2: More qualitative segmentation results (red mask) of D^2-SAM3.

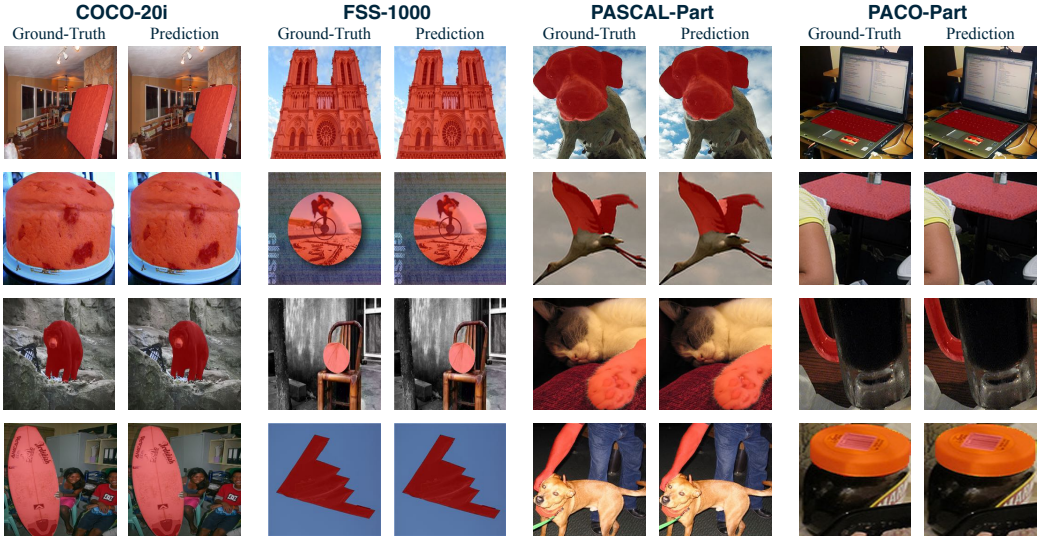


Figure A3: Qualitative segmentation results on In-Domain Few-Shot Segmentation datasets and One-Shot Part Segmentation datasets (red mask) of D^2-SAM3.

A4.8 Backbone-Controlled Comparison

The performance gains of D^2-SAM3 may arise from either the proposed dual-level distillation or the use of SAM3. We therefore construct two controlled baselines. First, we replace the original SAM in GF-SAM with SAM3 while retaining its reference-target visual matching pipeline. Second, we concatenate each reference-target pair and use the GT reference box as a geometric prompt for SAM3. Both baselines use the same seg-

Table A5: Backbone-controlled comparison on four CD-FSS benchmarks under the 1-shot setting. GF-SAM (w/ SAM3) retains its reference-target visual matching pipeline while replacing the original SAM with SAM3. SAM3 (vis. prompt) concatenates each reference-target pair and uses the ground-truth reference bounding box as a visual prompt.

Method	ISIC	Chest X-Ray	DeepGlobe	SUIM
GF-SAM	48.7	47.6	49.5	41.9
GF-SAM (w/ SAM3)	53.7	38.6	43.4	44.2
SAM3 (vis. prompt)	58.6	88.7	34.7	56.2
D^2-SAM3	**64.0**	**88.9**	**57.6**	**70.5**

mentation backbone as D^2-SAM3 while retaining reference-dependent target inference. As shown in Tab. A5, replacing SAM with SAM3 does not consistently improve GF-SAM: it increases mIoU by 5.0 and 2.3 points on ISIC and SUIM but decreases it by 9.0 and 6.1 points on Chest X-Ray and DeepGlobe, respectively. This indicates that the SAM3 backbone alone does not resolve the limitations of visual matching. Reference-target concatenation with the GT reference box provides a stronger geometric-prompt baseline, particularly on Chest X-Ray. Nevertheless, D^2-SAM3 surpasses this baseline by 5.4, 0.2, 22.9, and 14.3 mIoU points on ISIC, Chest X-Ray, DeepGlobe, and SUIM, respectively. These results indicate that the performance gains mainly arise from the proposed designs rather than the SAM3 backbone alone.

A4.9 Number of Learnable Parameters

D^2-SAM3 keeps the SAM3 image encoder, text encoder, mask decoder, classification head, and presence head frozen. For each category, it optimizes only a post-encoder category prototype and three scalar parameters used by VRD and RAF. Let L denote the fixed prompt sequence length and d denote the prototype token dimension. The learned category prototype is $Z_c \in \mathbb{R}^{L\times d}$. Reference adaptation additionally learns the category-specific scale a_c, shift b_c, and residual-weight parameter ξ_c. The maximum residual weight is then obtained as $\gamma_c = \gamma_{\max}\sigma(\xi_c)$. The total number of learnable parameters per category is therefore:

$$N_{\text{learn}} = Ld + 3. \tag{A18}$$

For SAM3, $L = 32$ and $d = 256$, giving:

$$N_{\text{learn}} = 32 \times 256 + 3 = 8{,}195, \tag{A19}$$

which corresponds to approximately 8.2K learnable parameters per category.