SAM3 FOR CROSS-DOMAIN FEW-SHOT SEGMENTATION VIA DUAL-BRANCH CANDIDATE CALIBRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-Domain Few-Shot Segmentation (CD-FSS) aims to segment novel categories in unseen domains from a few annotated support examples. Existing methods often rely on source-domain episodic training and remain sensitive to domain shifts. Motivated by advances in vision foundation models, we explore a fully frozen Segment Anything Model 3 (SAM3) for training-free CD-FSS. We observe that class-name text improves visual prompting in some domains but impairs it in others. To address this inconsistency, we propose a dual-branch candidate calibration framework. Support and query images are arranged on a shared canvas, where visual-only and visual-text branches generate complementary candidate masks. We assess candidate reliability using SAM3 confidence, support-query prototype consistency, and cross-branch agreement, then select and merge reliable masks into the final prediction. The framework requires no source-domain training, parameter updates, or additional learnable modules. Extensive experiments on diverse CD-FSS benchmarks demonstrate its effectiveness and superior average performance over the compared training-required and training-free methods.

1 INTRODUCTION

Semantic segmentation Chen et al. (2017); Long et al. (2015) has advanced substantially with the availability of large-scale annotated datasets Benenson et al. (2019). However, dense pixel-level annotation remains costly and labor-intensive, motivating Few-Shot Segmentation (FSS), which aims to segment novel categories from only a few annotated support samples. Existing FSS methods Fan et al. (2022); Li et al. (2021); Lu et al. (2021), however, often suffer substantial performance drops when target-domain appearances differ from those of the source domain Liu et al. (2020); Yue et al. (2021). Cross-Domain Few-Shot Segmentation (CD-FSS) addresses this setting, requiring a model trained on a labeled source domain to segment novel classes in unseen target domains using limited support examples Lei et al. (2022); Wang et al. (2022). Owing to domain shifts in illumination, texture, style, and imaging conditions, CD-FSS is considerably more challenging than standard FSS and often yields unreliable segmentation results.

Existing CD-FSS methods mainly follow two directions. One line inherits the support-query matching paradigm from conventional FSS, using either class prototypes from annotated support regions Dong & Xing (2018); Tian et al. (2020) or dense support-query correspondences at high-resolution feature levels Min et al. (2021); Zhang et al. (2022). The other line improves cross-domain robustness through feature transformation, domain rectification, or frequency-aware adaptation Lei et al. (2022); Su et al. (2024); Tong et al. (2024). Despite their effectiveness, these methods usually require episodic training on source-domain categories and their performance highly depends on the quality of learned feature representations. Foundation segmentation models provide a different route, benefiting from large-scale training on prompt-image-mask triplets, the Segment Anything Model (SAM) Kirillov et al. (2023) exhibits strong segmentation ability across a wide range of objects and mask granularities. Motivated by SAM's strong ability to segment prompt-specified objects using prompt points, as shown in Figure 1(b), recent works He et al. (2024); Nie et al. (2026) have explored SAM-based methods for CD-FSS, They can generate high-quality segmentation masks from prompts such as points, boxes, and masks.

As the latest member of the Segment Anything family, SAM3 Carion et al. (2025) introduces Promptable Concept Segmentation (PCS), which allows segmentation to be guided by textual de-

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.

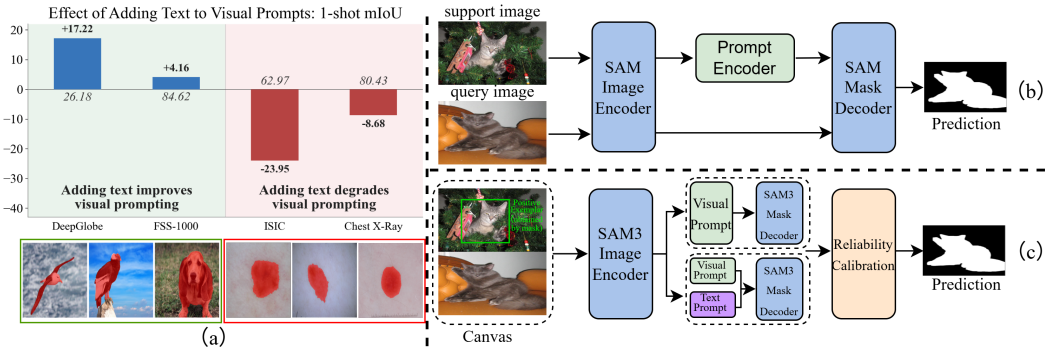

Figure 1: (a) Effect of adding class-name text to visual prompts using only the original fully frozen SAM3. (b) SAM-based method rely on visual-prompting strategy. (c) Our SAM3 with dual-branch candidate calibration.

scriptions, visual exemplars, or their combination. Different from earlier SAM models that mainly generate class-agnostic masks from spatial prompts, SAM3 is built to recognize and segment semantic concepts. This raises a natural question: *Can CD-FSS be achieved by directly leveraging such a pretrained model, without any additional training?*

In this work, we explore a fully frozen SAM3 for CD-FSS. To adapt SAM3 to CD-FSS, we construct a shared spatial canvas containing both the support and query images inspired by Tsai et al. (2026), enabling SAM3 to model cross-image correspondence within a single forward pass (see Figure 1(c)). Because SAM3 can incorporate class-name text as an additional semantic cue, we examine whether textual prompting can provide reliable complementary guidance for CD-FSS. A preliminary analysis in Figure 1(a) shows that adding class-name text can improve performance in some domains, but may degrade it in others, such as medical imaging. We guess that class-name text may alter SAM3's interpretation of visual evidence. When a class name does not precisely characterize the target in the query image, its broad, ambiguous, or weakly discriminative semantics may compete with the visual prompt, yielding unreliable candidate masks. This prompt conflict explains why adding text can yield negative gains and motivates not committing to a single prompting strategy.

Therefore, we propose a dual-branch candidate calibration framework for CD-FSS (Figure 1(c)). The key idea is to avoid committing to a single prompting strategy across all target domains. Specifically, we place the support and query images in a shared spatial canvas. On this canvas, visual-only and visual-text prompts are used as complementary cues to generate candidate masks, allowing our method to select reliable mask predictions from different prompting branches and produce the final query segmentation without updating SAM3. Despite its simplicity, this approach achieves strong performance on multiple cross-domain benchmarks, outperforming many existing methods. In summary, our contributions are:

- We propose a training-free framework that applies a fully frozen SAM3 to CD-FSS without introducing additional learnable modules or updating model parameters.
- We employ visual-only and visual-text prompts as complementary branches, avoiding reliance on a fixed prompting strategy across target domains and generating a diverse candidate pool for query segmentation.
- Experimental results on multiple cross-domain benchmarks under 1-shot and 5-shot settings demonstrate that our method achieves superior performance to existing CD-FSS methods.

2 Related Work

2.1 Crooss-Domain Few-Shot Segmentation

Cross-Domain Few-Shot Segmentation (CD-FSS) aims to segment target objects from a few annotated support examples under substantial domain shifts Lu et al. (2021); Boudiaf et al. (2021);

Tong et al. (2025). Existing methods generally learn support-query matching from source-domain episodes and improve robustness through domain adaptation, feature refinement, or frequency-aware modeling Lei et al. (2022); Su et al. (2024); Tong et al. (2024). Despite their effectiveness, these methods typically require training or adaptation on source-domain data and may become unreliable under severe target-domain shifts. Visual foundation models provide an alternative route for CD-FSS Kirillov et al. (2023); Oquab et al. (2023). Recent studies exploit SAM's promptable segmentation capability by transferring support information into prompts for query segmentation He et al. (2024); Zhu et al. (2024); Sun et al. (2024). However, many of these approaches still rely on additional training stages or trainable components Liu et al. (2023); Zhang et al. (2024a); Herzog (2024). In contrast, our work focuses on a training-free solution that no longer relies on domain-specific data for CD-FSS.

2.2 SAM3 AND CONCEPT SEGMENTATION

As a recent member of the SAM series, SAM3 Carion et al. (2025) advances promptable segmentation through Promptable Concept Segmentation (PCS), enabling segmentation based on semantic concepts expressed through text or visual exemplars. Unlike SAM Kirillov et al. (2023) and SAM2 Ravi et al. (2025), which primarily focus on class-agnostic objectness, SAM3 supports concept-level reasoning across different instances. This capability offers a more direct way to formulate CD-FSS, reducing the need for extensive fine-tuning or auxiliary modules previously introduced to align support and query semantics. Motivated by this observation, we revisit CD-FSS from a training-free perspective and explore how pretrained concept segmentation can establish reliable support-query correspondence under domain shifts.

3 PROBLEM SETTING

Cross-Domain Few-Shot Segmentation (CD-FSS) aims to transfer meta-knowledge from a source domain $(\mathcal{X}_s, \mathcal{Y}_s)$ with abundant annotations to a target domain $(\mathcal{X}_t, \mathcal{Y}_t)$ with limited labeled data. The two domains differ in data distribution and label space, i.e., $\mathcal{X}_s \neq \mathcal{X}_t$ and $\mathcal{Y}_s \cap \mathcal{Y}t = \varnothing$. Unlike conventional CD-FSS methods trained on source-domain episodes Fan et al. (2020), our method does not access source-domain training data or update model parameters. At test time, each target-domain episode consists of a 1-way K-shot support set $\mathcal{S} = (I_s^k, M_s^k)k = 1^K$ and a query image I_q. Here, $I_s^k \in \mathbb{R}^{H\times W\times 3}$ and $M_s^k \in \{0,1\}^{H\times W}$ denote the k-th support image and corresponding binary mask. Each episode is also associated with a class name T. The query ground-truth mask $M_q \in \{0,1\}^{H\times W}$ is used only for evaluation and not provided to the model.

4 METHODOLOGY

4.1 OVERALL FRAMEWORK

Our framework performs CD-FSS with a fully frozen SAM3 and introduces no additional learnable modules. Given a support image, its annotation mask, a query image, and a class name, we first derive a tight bounding box from the support mask as a support-conditioned visual prompt. The support and query images are then placed on a shared spatial canvas, allowing SAM3 to model cross-image interaction during inference. As illustrated in Figure 2, the framework consists of two main stages. In the first stage, we construct two complementary prompting branches on the shared canvas. The visual-only branch uses the support-derived geometric prompt, whereas the visual-text branch additionally incorporates the class name as a semantic cue. In the second stage, the frozen SAM3 decoder processes the prompts from both branches to generate query mask candidates. These candidates are merged into a unified candidate pool and assessed for reliability to predict the final target mask. Throughout the entire pipeline, SAM3 remains unchanged and no source-domain training is required.

4.2 DUAL-BRANCH CANDIDATE GENERATION

To enable support-conditioned query segmentation with a fully frozen SAM3, we arrange the support image and query image into a shared spatial canvas. The geometric prompt derived from the

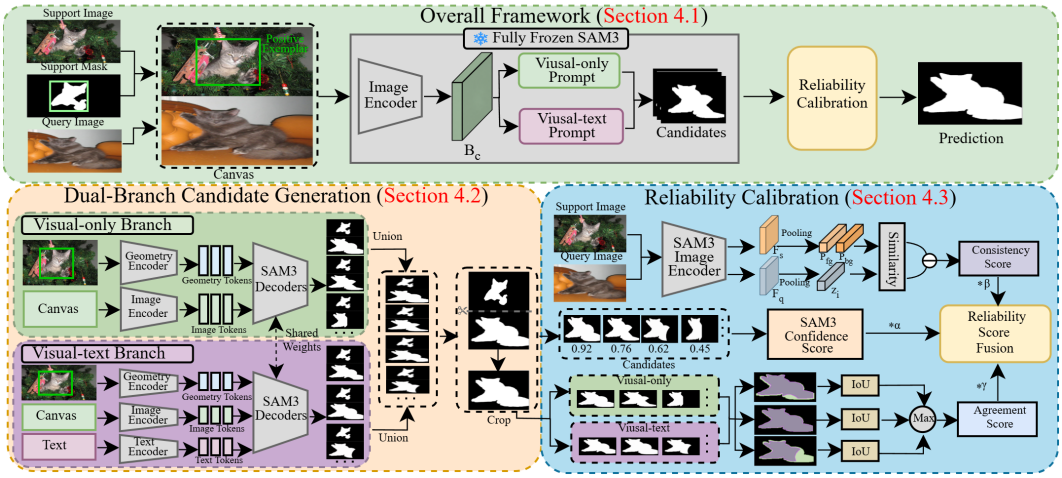

Figure 2: Overview of the proposed framework. Given a support image-mask pair and a query image, we construct a shared canvas and use frozen SAM3 with visual-only and visual-text prompts to generate candidate masks. The candidates are calibrated by SAM3 confidence, prototype consistency, and cross-branch agreement, and reliable masks are merged to obtain the final query prediction.

support annotation is transformed into the corresponding canvas coordinate system. Under this unified representation, visual tokens from the support and query images are processed jointly by SAM3, allowing the model to associate the support target with semantically relevant query regions through its native attention mechanism. This formulation casts support-query correspondence as spatial reasoning within a common visual context, without introducing additional learnable modules.

Given a support image-mask pair (I_s, M_s) and a query image I_q, where I_s and I_q are the support and query images and M_s is the binary support mask, we derive a tight box prompt from M_s:

$$B_s = \mathrm{Box}(M_s), \tag{1}$$

where B_s is the support box and $\mathrm{Box}(\cdot)$ returns the smallest axis-aligned box enclosing the foreground of M_s.

The support-query pair is composed into a shared canvas:

$$I_c = \mathrm{Canvas}(I_s, I_q), \qquad B_c = \Phi_s(B_s), \tag{2}$$

where I_c is the shared canvas, B_c is the support box in canvas coordinates, and $\Phi_s(\cdot)$ maps support-image coordinates to canvas coordinates.

We apply frozen SAM3, denoted by f_θ with fixed parameters θ, under two prompting configurations. The visual-only branch uses the support-derived box prompt:

$$\mathcal{C}_V = f_\theta(I_c, B_c) = \left\{(M_i^V, s_i^V, V)\right\}_{i=1}^{N_V}. \tag{3}$$

The visual-text branch additionally uses the class name T:

$$\mathcal{C}_{VT} = f_\theta(I_c, B_c, T) = \left\{(M_j^{VT}, s_j^{VT}, VT)\right\}_{j=1}^{N_{VT}}. \tag{4}$$

Here, M_i^V and M_j^{VT} denote the canvas-level candidate masks obtained from the corresponding canvas predictions, while s_i^V and s_j^{VT} are their SAM3 confidence scores. N_V and N_{VT} denote the numbers of candidates produced by the two branches. The last element of each tuple, V or VT, records the prompting source of the candidate. The visual-only branch provides instance-level guidance from the support example, whereas the visual-text branch additionally provides category-level semantic information.

Finally, we merge the candidate sets generated by the visual-only and visual-text branches into a unified candidate pool:

$$\mathcal{C} = \mathcal{C}_V \cup \mathcal{C}_{VT}. \tag{5}$$

where $\mathcal{C}$ contains all canvas-level candidates from both prompting configurations and is used for subsequent reliability assessment.

4.3 CANDIDATE RELIABILITY CALIBRATION

Before calibration, each canvas-level candidate mask in $\mathcal{C}$ is cropped to the query region and resized to the original query resolution, yielding the query-side mask M_i. Although SAM3 provides an intrinsic confidence score for each candidate mask, this score alone may not fully reflect prediction reliability on unseen target domains. Therefore, we further assess each query-side candidate using support-query prototype consistency and cross-branch agreement, together with its original SAM3 confidence score.

To evaluate whether a query candidate is visually consistent with the support target, we compute support-query prototype consistency in the frozen SAM3 feature space. Let $E(\cdot)$ denote the frozen SAM3 image encoder. The support and query features are extracted as

$$F_s = E(I_s), \qquad F_q = E(I_q), \tag{6}$$

where F_s and F_q are the feature maps of the support and query images, respectively. We construct support foreground and background prototypes by masked average pooling:

$$p_{\mathrm{fg}} = \mathrm{Avg}(F_s \mid M_s), \qquad p_{\mathrm{bg}} = \mathrm{Avg}(F_s \mid 1 - M_s). \tag{7}$$

For each query candidate M_i, its regional feature is computed as

$$z_i = \mathrm{Avg}(F_q \mid M_i). \tag{8}$$

The prototype consistency score is defined as

$$r_i^{\mathrm{proto}} = \mathrm{Norm}_{[0,1]}\left[\cos(z_i, p_{\mathrm{fg}}) - \cos(z_i, p_{\mathrm{bg}})\right], \tag{9}$$

where a higher score indicates that the candidate is more similar to the support foreground than to the background.

We further measure whether each candidate is supported by a spatially similar prediction from the other prompting branch:

$$r_i^{\mathrm{agree}} = \max_{M_j \in \mathcal{C}_{\mathrm{other}}} \mathrm{IoU}(M_i, M_j), \tag{10}$$

where $\mathcal{C}_{\mathrm{other}} = \mathcal{C}_{VT}$ for a candidate from $\mathcal{C}_V$, and $\mathcal{C}_{\mathrm{other}} = \mathcal{C}_V$ for a candidate from $\mathcal{C}_{VT}$.

All scores are normalized to $[0, 1]$ before fusion. Let $\bar{s}_i$ denote the normalized SAM3 confidence score of candidate M_i. The final quality score is computed as

$$q_i = \frac{\alpha \bar{s}_i + \beta r_i^{\mathrm{proto}} + \gamma r_i^{\mathrm{agree}}}{\alpha + \beta + \gamma}, \tag{11}$$

where $\alpha = 1.0$,$\beta = 0.2$ and $\gamma = 0.1$ in our experiments. Candidates with quality scores above a threshold δ are retained:

$$\mathcal{R} = \{M_i \in \mathcal{C} \mid q_i > \delta\}. \tag{12}$$

Finally, the query segmentation mask is obtained by taking the union of the retained candidates:

$$M_{\mathrm{final}} = \bigcup_{M_i \in \mathcal{R}} M_i. \tag{13}$$

5 EXPERIMENTS

5.1 DATASETS

We evaluate our method on a standard cross-domain few-shot segmentation (CD-FSS) benchmark spanning natural, remote-sensing, and medical image domains. The benchmark consists of pixel-level annotated images collected from FSS-1000 Li et al. (2020), DeepGlobe Demir et al. (2018), ISIC2018 Tschandl et al. (2018), and Chest X-ray datasets Candemir et al. (2013). These datasets cover diverse visual appearances, imaging modalities, and semantic categories, exhibiting substantial variations in texture, scale, color distribution, and imaging characteristics. This diverse evaluation setting provides a challenging testbed for assessing the ability of CD-FSS methods to transfer to unseen target domains.

5.2 IMPLEMENTATION DETAILS

Table 1: Comparison with training-required CD-FSS methods and SAM-based training-free methods on cross-domain few-shot segmentation benchmarks. Results are reported in mIoU (%). The best and second-best results are highlighted in **bold** and underline, respectively.

Method	DeepGlobe		ISIC		Chest X-Ray		FSS-1000		Average	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Training-required CD-FSS Methods										
HSNet Min et al. (2021)	29.65	35.08	31.20	35.10	51.88	54.36	77.53	80.99	47.57	51.38
DMTNet Chen et al. (2024)	34.52	47.62	34.34	40.81	71.14	72.69	70.81	72.89	52.70	58.50
PATNet Lei et al. (2022)	37.89	42.97	41.16	53.58	66.61	70.20	78.59	81.23	56.06	61.99
DR-Adapter Su et al. (2024)	41.29	50.12	40.77	48.87	82.35	82.31	79.05	80.40	60.86	65.42
ABCDFSS Herzog (2024)	42.60	49.00	45.70	53.30	79.80	81.40	74.60	76.20	60.67	64.97
APM Tong et al. (2024)	40.86	44.92	41.71	51.16	78.25	82.81	79.29	81.83	60.03	65.18
LoEC Liu et al. (2025)	44.10	49.67	38.21	47.04	81.02	82.73	78.51	80.60	60.46	65.01
HSL Sun et al. (2026)	46.13	53.80	48.01	55.56	84.57	**85.34**	78.22	80.36	64.23	68.77
SAM-based Training-free Methods										
PerSAM Zhang et al. (2024b)	33.39	37.90	21.27	25.87	31.12	36.34	79.65	80.86	41.35	45.24
Matcher Liu et al. (2023)	48.12	50.92	38.60	35.03	66.85	66.37	87.01	89.65	60.15	60.49
GF-SAM Zhang et al. (2024a)	**49.50**	**57.76**	48.79	55.24	47.65	47.40	88.00	88.98	58.49	62.35
CPS Nie et al. (2026)	46.02	49.85	51.33	56.60	70.06	70.27	81.76	82.89	62.30	64.90
Ours	46.05	50.72	**65.55**	**72.54**	**84.60**	83.99	**90.04**	**91.14**	**71.56**	**74.60**

We conduct all experiments using the official SAM3 checkpoint, which remains fully frozen throughout inference. No source-domain training, fine-tuning, or parameter updates are performed. For each episode, the support mask is converted into a tight bounding box, and the support and query images are arranged on a shared spatial canvas following the native preprocessing pipeline of SAM3. The visual-only branch uses the support-derived box prompt, while the visual-text branch additionally uses the class name as a textual prompt. All implementations are based on PyTorch and executed on a single NVIDIA RTX 3090 GPU. We report mean Intersection over Union (mIoU) under both 1-shot and 5-shot settings.

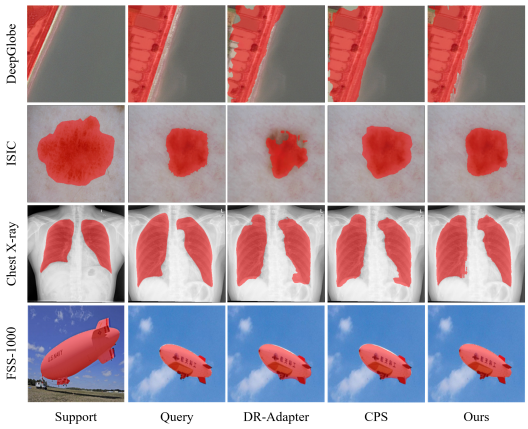

Figure 3: Qualitative comparisons among our method, DR-Adapter, and CPS under the 1-way 1-shot setting on challenging scenarios with significant domain shifts.

5.3 Comparison Experiments

Table 1 compares our method with training-required CD-FSS approaches and SAM-based training-free methods under the mIoU metric. Unlike the training-required methods, our method uses a fully frozen SAM3 without source-domain training or parameter updates. Despite this simple setting, our method achieves the best average performance in both 1-shot and 5-shot settings, reaching 71.56% and 74.60%, respectively.

Our method achieves substantial improvements on the ISIC and FSS-1000 datasets. On ISIC, it obtains 65.55% and 72.54% under the 1-shot and 5-shot settings, surpassing the strongest compared training-free method by 14.22% and 15.94%, respectively. On FSS-1000, our method reaches 90.04% and 91.14%, exceeding the best compared result by 2.04% and 1.49%. Our method also delivers competitive performance on DeepGlobe and Chest X-Ray. On DeepGlobe, it achieves 46.05% and 50.72% under the 1-shot and 5-shot settings, respectively. On Chest X-Ray, it obtains 84.60% and 83.99% under the 1-shot and 5-shot settings, respectively. In particular, it achieves the best 1-shot result on Chest X-Ray. These results indicate that the proposed dual-branch candidate calibration effectively selects reliable predictions from complementary visual and visual-text prompts across diverse target domains. Figure 3 presents qualitative segmentation comparisons on representative samples from DeepGlobe, ISIC, Chest X-Ray, and FSS-1000 under the 1-shot setting. Across

Table 2: Ablation study of shared-canvas layouts under the 1-shot setting. H and V denote horizontal and vertical compositions, respectively.

Layout	DeepGlobe	ISIC	Chest X-Ray	FSS-1000	Avg.
H-Left	45.32	64.01	83.56	88.99	70.47
H-Right	45.16	64.54	83.50	89.12	70.58
V-Top	46.05	**65.55**	**84.60**	**90.04**	**71.56**
V-Bottom	**46.09**	65.37	84.51	89.90	71.47

Table 3: Ablation study of the auxiliary calibration terms and the candidate selection threshold under the 1-shot setting.

(a) Calibration Term Ablation			(b) Threshold Analysis	
Prototype	Cross-Branch	mIoU	δ	mIoU
–	–	70.16	0.3	70.31
✓	–	71.15	0.4	70.98
–	✓	71.02	0.5	**71.56**
✓	✓	**71.56**	0.6	71.24

Table 4: Sensitivity analysis of the score fusion weights under the 1-shot and 5-shot settings.

Score Weight Ablation								
Effect of α			Effect of β			Effect of γ		
α	1-shot	5-shot	β	1-shot	5-shot	γ	1-shot	5-shot
0.5	70.31	72.09	0.1	70.85	73.08	**0.1**	**71.56**	**74.60**
1.0	**71.56**	**74.60**	**0.2**	**71.56**	**74.60**	0.2	70.98	73.41
2.0	70.89	73.39	0.3	71.12	73.85	0.3	70.34	72.39

these diverse target domains, our method produces masks that better align with the target regions and preserve more complete object structures.

5.4 Ablation Studies

Canvas layout.

We compare four shared-canvas layouts by placing the support image to the left, right, top, or bottom of the query image. H and V denote horizontal and vertical compositions, respectively. This examines whether the support-query arrangement affects cross-image interaction on the shared canvas. As shown in Table 2, V-Top achieves the best average performance across four target domains, outperforming both horizontal layouts and slightly exceeding V-Bottom. It also performs best on ISIC, Chest X-Ray, and FSS-1000, indicating that placing the support image above the query provides a more favorable configuration for support-guided segmentation. We therefore adopt V-Top as the default canvas arrangement.

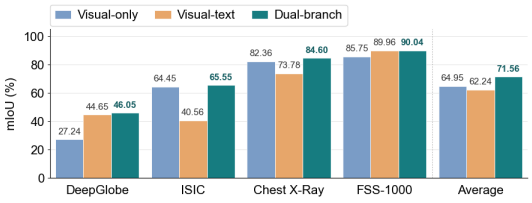

Figure 4: Quantitative comparison of visual-only, visual-text, and dual-branch prompting under the 1-shot setting.

Hyperparameter analysis. Table 3 evaluates the auxiliary calibration terms and the candidate selection threshold. Building upon the original SAM3 confidence, prototype consistency and cross-branch agreement individually improve mIoU to 71.15% and 71.02%, respectively, while their combination achieves the best result of 71.56%. This confirms that the two cues provide complementary information for candidate selection. We further vary the threshold δ and observe the best performance at $\delta = 0.5$; lower values retain more unreliable candidates, whereas higher values may discard useful predictions.

Beyond the calibration terms and selection threshold, we further analyze the score fusion weights α, β, and γ. As shown in Table 4, when varying one weight, the other two are held constant.

The combination $\alpha = 1.0$, $\beta = 0.2$, and $\gamma = 0.1$ achieves the best performance, with mIoU values of 71.56% and 74.60% under the 1-shot and 5-shot settings, respectively. Both overly small and overly large values of α lead to performance degradation, indicating that the SAM3 confidence should serve as the primary cue while being properly balanced with the two auxiliary signals. The prototype consistency term performs best when $\beta = 0.2$, whereas assigning it a larger weight slightly reduces

Table 5: Effect of the maximum number of negative prompts under the 1-shot setting. Results are reported in mIoU (%).

Max Negative Prompts	Target Domain				
	DeepGlobe	ISIC	Chest X-Ray	FSS-1000	Average
$N_{\text{neg}} = 0$ (Positive only)	46.05	65.55	84.60	90.04	71.56
$N_{\text{neg}} \leq 1$	41.15	62.83	79.31	83.22	66.63
$N_{\text{neg}} \leq 2$	39.22	56.34	75.37	81.64	63.14
$N_{\text{neg}} \leq 3$	36.39	54.82	72.41	79.96	60.90

performance. Similarly, increasing the cross-branch agreement weight beyond $\gamma = 0.1$ consistently degrades performance. Overall, prototype consistency and cross-branch agreement provide effective complementary calibration cues, while excessive emphasis on either auxiliary term may impair the final candidate ranking.

Effect of Dual-Branch Prompting. Figure 4 compares visual-only, visual-text, and dual-branch prompting under the 1-shot setting. The two single branches show distinct domain preferences: visual-text prompting performs better on DeepGlobe and FSS-1000, whereas visual-only prompting is more effective on ISIC and Chest X-Ray. This result indicates that the usefulness of textual guidance varies across target domains. By combining candidates from both prompting branches, the dual-branch setting consistently achieves the best performance on all datasets. It improves the average mIoU to 71.56%, exceeding visual-only and visual-text prompting by 6.61% and 9.32%, respectively.

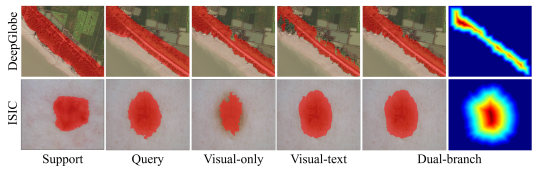

Figure 5: Qualitative comparison of visual-only, visual-text, and dual-branch prompting on DeepGlobe and ISIC. The dual-branch setting produces more complete and coherent segmentation masks.

Figure 5 further presents qualitative examples from DeepGlobe and ISIC. The visual-only and visual-text branches may produce incomplete masks or introduce noisy regions under different domain conditions. In contrast, the dual-branch prediction better preserves the target structure and yields a more coherent segmentation result, demonstrating the benefit of using complementary prompting cues.

Effect of negative prompting. Positive prompts directly guide target localization, but may be insufficient when query images contain visually similar backgrounds or cluttered structures. Negative prompts may complement this guidance by indicating regions to exclude, thereby reducing foreground-background ambiguity. To examine this, we extend the positive-only setting by providing support-derived negative prompts to SAM3. Specifically, we apply SAM3 to the support image with a positive box prompt and convert low-confidence candidate masks into negative box prompts.

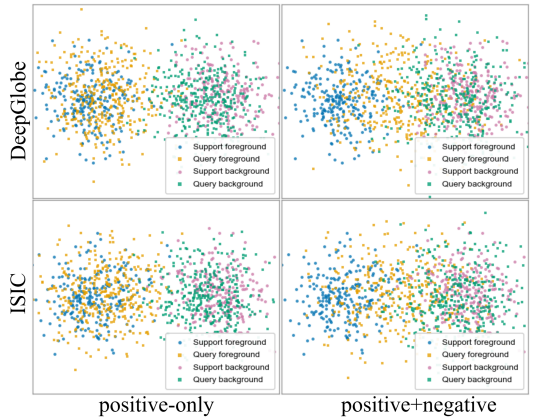

Figure 6: t-SNE visualization of support and query prototypes under positive-only prompting (left) and positive-negative prompting (right) on DeepGlobe and ISIC.

Contrary to the above intuition, Table 5 shows that introducing negative prompts leads to a clear performance degradation compared with positive-only prompting. We further investigate the influence of the number of negative prompts by varying the maximum number of negative boxes provided to the model. The results show that performance continues to decrease as more negative prompts are introduced, indicating that SAM3 may be sensitive to negative guidance.

Table 6: Effect of domain-aware text prompts for the visual-text branch. Results are reported in mIoU (%).

Text Prompt	DeepGlobe	ISIC	Chest X-Ray	FSS-1000	Average
Class-only	46.05	65.55	84.60	90.04	**71.56**
Class + domain	44.72	66.38	83.21	**90.36**	71.17

Figure 6 visualizes the support and query prototypes on DeepGlobe and ISIC. With positive-only prompting, query foreground prototypes are closely aligned with support foreground prototypes, whereas query background prototypes remain near the support background prototypes. The foreground and background prototype groups are therefore well separated. After introducing negative prompts, query foreground and background prototypes become more mixed, and the query foreground prototypes shift away from their support foreground counterparts. This reduces support-query prototype consistency. We therefore retain positive-only prompting in the final framework.

Effect of domain-aware text prompts. We investigate whether adding domain characteristics to class-name prompts improves visual-text prompting. For each target domain, we compare the original class-name prompt with one augmented by concise domain-specific attributes. The support box, shared canvas, frozen SAM3, and candidate-selection settings remain identical. As shown in Table 6, domain-aware prompts do not consistently improve performance across target datasets and provide limited average benefit. This suggests that domain-related words may not improve textual concept specification and can introduce redundant constraints. We therefore use concise class-name prompts in the final framework.

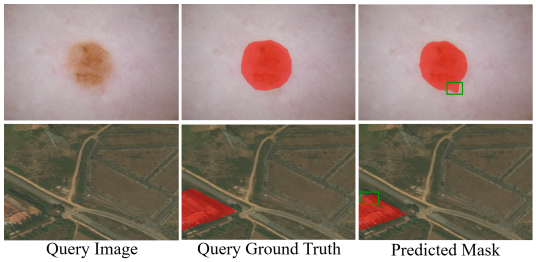

Figure 7: Representative failure cases of the proposed method. Ambiguous boundaries and irregular target structures can lead to incomplete target masks.

5.5 Limitation

Despite the strong performance of our training-free framework, several limitations remain. First, support-derived negative prompts may degrade performance by providing ambiguous spatial guidance and suppressing target-relevant regions, suggesting that negative guidance should be used more cautiously in CD-FSS. Second, although SAM3 supports textual concept prompts, adding domain descriptions to class-name prompts does not yield consistent improvements, leaving more effective ways of incorporating domain information to future work. Finally, our method relies on candidate masks generated by fully frozen SAM3. As shown in Figure 7, ambiguous boundaries, irregular structures, or visually similar backgrounds may lead to under- or over-segmentation. When the initial candidates fail to cover the target adequately, reliability assessment cannot recover the missing semantics. Future work may explore more robust candidate generation or lightweight adaptation while preserving the training-free setting.

6 Conclusion

In this paper, we propose a training-free dual-branch candidate calibration framework for CD-FSS. Using a fully frozen SAM3, our method places support and query images on a shared canvas and generates complementary candidates through visual-only and visual-text prompting. Candidate reliability is assessed using SAM3 confidence, support-query prototype consistency, and cross-branch agreement. Extensive experiments across diverse target domains demonstrate the effectiveness of the framework under both 1-shot and 5-shot settings. These results suggest that complementary prompting and candidate-level reliability assessment provide an effective route for applying frozen foundation segmentation models to CD-FSS.

7 AI USE STATEMENT

We used AI tools to assist with translation and language editing to improve the clarity and readability of the manuscript. The authors take full responsibility for the final content.

REFERENCES

Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11700–11709, 2019.

Malik Boudiaf, Hoel Kervadec, Ziko Imtiaz Masud, Pablo Piantanida, Ismail Ben Ayed, and Jose Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13979–13988, 2021.

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. *IEEE transactions on medical imaging*, 33(2):577–590, 2013.

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. *arXiv preprint arXiv:2511.16719*, 2025.

Jiayi Chen, Rong Quan, and Jie Qin. Cross-domain few-shot semantic segmentation via doubly matching transformation. *arXiv preprint arXiv:2405.15265*, 2024.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848, 2017.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 172–181, 2018.

Nanqing Dong and Eric P Xing. Few-shot semantic segmentation with prototype learning. In *Bmvc*, volume 3, pp. 4, 2018.

Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4013–4022, 2020.

Qi Fan, Wenjie Pei, Yu-Wing Tai, and Chi-Keung Tang. Self-support few-shot semantic segmentation. In *European conference on computer vision*, pp. 701–719. Springer, 2022.

Weizhao He, Yang Zhang, Wei Zhuo, Linlin Shen, Jiaqi Yang, Songhe Deng, and Liang Sun. Apseg: Auto-prompt network for cross-domain few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23762–23772, 2024.

Jonas Herzog. Adapt before comparison: A new perspective on cross-domain few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23605–23615, 2024.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Shuo Lei, Xuchao Zhang, Jianfeng He, Fanglan Chen, Bowen Du, and Chang-Tien Lu. Cross-domain few-shot semantic segmentation. In *European conference on computer vision*, pp. 73–90. Springer, 2022.

Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8334–8343, 2021.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2869–2878, 2020.

Bingyu Liu, Zhen Zhao, Zhenpeng Li, Jianan Jiang, Yuhong Guo, and Jieping Ye. Feature transformation ensemble model with batch spectral regularization for cross-domain few-shot classification. *arXiv preprint arXiv:2005.08463*, 2020.

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.

Yuhan Liu, Yixiong Zou, Yuhua Li, and Ruixuan Li. The devil is in low-level features for cross-domain few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4618–4627, 2025.

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

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8741–8750, 2021.

Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6941–6952, 2021.

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*, 2026.

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.

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. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Jiapeng Su, Qi Fan, Wenjie Pei, Guangming Lu, and Fanglin Chen. Domain-rectifying adapter for cross-domain few-shot segmentation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 24036–24045, 2024.

Sujun Sun, Haowen Gu, Cheng Xie, Yanxu Ren, Mingwu Ren, and Haofeng Zhang. Bridging granularity gaps: Hierarchical semantic learning for cross-domain few-shot segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9215–9223, 2026.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23565–23574, 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.

Jintao Tong, Yixiong Zou, Yuhua Li, and Ruixuan Li. Lightweight frequency masker for cross-domain few-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 37:96728–96749, 2024.

Jintao Tong, Yixiong Zou, Guangyao Chen, Yuhua Li, and Ruixuan Li. Self-disentanglement and re-composition for cross-domain few-shot segmentation. *arXiv preprint arXiv:2506.02677*, 2025.

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

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*, 5(1): 180161, 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7065–7074, 2022.

Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, and Alberto Sangiovanni Vincentelli. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13834–13844, 2021.

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.

Jian-Wei Zhang, Yifan Sun, Yi Yang, and Wei Chen. Feature-proxy transformer for few-shot segmentation. *Advances in neural information processing systems*, 35:6575–6588, 2022.

Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Yu Qiao, Gao Peng, and Hongsheng Li. Personalize segment anything model with one shot. In *International Conference on Learning Representations*, volume 2024, pp. 18250–18279, 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. *Advances in Neural Information Processing Systems*, 37:42672–42695, 2024.

A Appendix

A.1 Overview

This supplementary material presents extended analyses, detailed implementation descriptions, and additional experimental results that further support the main paper. For better readability, the contents are structured as follows:

1. an overview of the datasets employed in our experiments;
2. an analysis of computational efficiency;
3. additional qualitative results under the 5-shot evaluation setting;
4. a summary of the experimental settings and hyperparameters;
5. the full pseudocode of the proposed method for ensuring reproducibility in experiments.

A.2 Dataset Overview

Extending the concise dataset descriptions in the main paper, this supplementary section presents a more detailed examination of the datasets used in our experiments (Figure 8), focusing on the specific challenges they introduce for cross-domain few-shot segmentation (CD-FSS). DeepGlobe consists of high-resolution satellite imagery that differs substantially from natural images in terms of large-scale spatial organization, irregular object boundaries, and complex land-cover patterns, which hinders few-shot models—typically trained on natural-image statistics—from learning stable and transferable representations.

Table 7: Inference speed of different prompting configurations on four target datasets. Results are reported in FPS.

Dataset	Visual-only	Visual-text	Dual-branch
DeepGlobe	23.28	18.96	9.82
ISIC	19.45	16.09	9.40
Chest X-Ray	10.20	9.25	7.52
FSS-1000	22.98	18.34	9.04

While FSS-1000 remains within the natural-image domain, it exhibits significant intra-class diversity and fine-grained object characteristics; under few-shot supervision, such variability can impair prototype consistency and increase susceptibility to noisy support–query associations. ISIC corresponds to a distinct medical imaging domain, where dermoscopic images diverge sharply from natural images in illumination, color distribution, and morphological structure, thereby intensifying domain sensitivity and often resulting in unstable cross-instance correspondence when annotated data are scarce. Chest X-Ray data further magnify the domain discrepancy, as grayscale radiographs are dominated by subtle anatomical structures and limited texture cues, making support–query alignment particularly fragile and challenging a model's ability to generalize beyond RGB natural-image priors. Overall, these datasets collectively constitute a highly heterogeneous evaluation benchmark that simultaneously tests domain transferability and sample efficiency, offering a more demanding assessment of few-shot segmentation methods than conventional in-domain benchmarks.

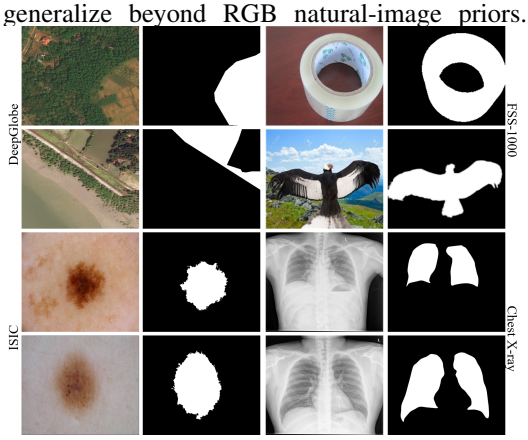

Figure 8: The four segmentation datasets used in our experiments are DeepGlobe, FSS-1000, ISIC, and Chest X-Ray, covering natural scenes, object-centric images, and medical imagery.

A.3 Computational Efficiency

We evaluate the inference speed of visual-only, visual-text, and dual-branch prompting on four target datasets, as summarized in Table 7. All three configurations are evaluated on a single NVIDIA RTX 3090 GPU with a shared-canvas resolution of 1008×1008. Inference speed is measured in frames per second (FPS), where higher values indicate faster processing.

Visual-only prompting achieves the highest inference speed across all four datasets, ranging from 10.20 to 23.28 FPS, while visual-text prompting runs at 9.25–18.96 FPS. The full dual-branch framework achieves 7.52–9.82 FPS. Its lower inference speed reflects the additional computation involved in generating candidates from both prompting routes and performing candidate reliability calibration. Although the framework requires no additional training or learnable modules, these inference operations introduce extra computational cost. The results therefore characterize the inference overhead associated with the segmentation improvements reported in the main paper.

A.4 Qualitative Results under the 5-Shot Setting

For the 5-shot setting, we compared four shared-canvas layouts: a uniform 2×3 grid, a horizontal strip, a vertical strip, and an inverse L-shaped arrangement. The inverse L-shaped layout achieved the highest average mIoU, so we place the five support images along the top and left sides of the canvas, leaving the larger bottom-right region for the query image. To complement the quantitative evaluation in the main paper, we present additional qualitative comparisons under the 5-shot setting in Figure 9. The examples cover DeepGlobe, ISIC, Chest X-Ray, and FSS-1000, representing di-

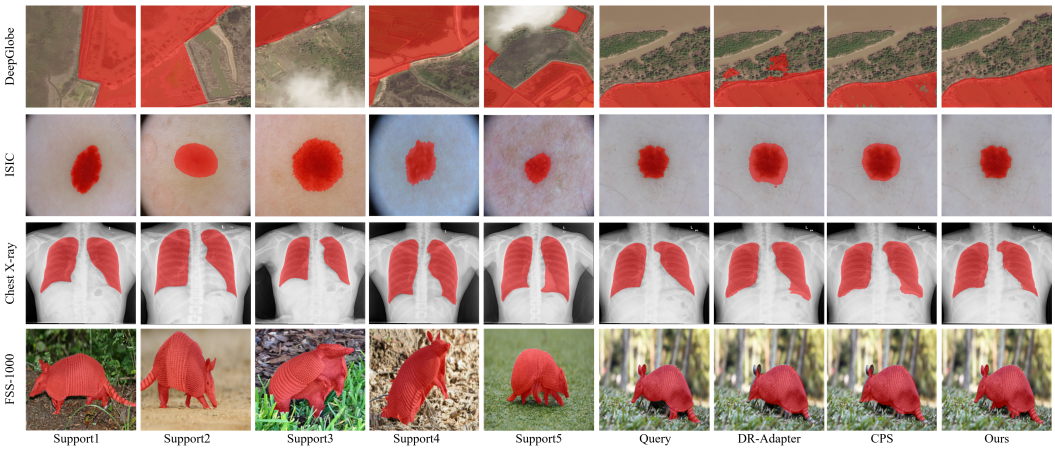

Figure 9: Qualitative comparisons under the 5-shot setting on DeepGlobe, ISIC, Chest X-Ray, and FSS-1000. From left to right: five support images with ground-truth masks, the query image with its ground-truth mask, and predictions from DR-Adapter, CPS, and our method.

Table 8: Summary of experimental settings and hyperparameters used in our experiments.

Item	**Setting**
Model and prompting	
Model	Fully frozen SAM3
Canvas resolution	1008×1008
Support shots	1 or 5
1-shot layout	Vertical: support above query
5-shot layout	Inverse L-shape (query bottom-right)
Prompting branches	Visual-only and visual-text
Visual prompt	Bounding box derived from the support mask
Text prompt	Class name, used only in the visual-text branch
Candidate calibration	
Candidate pool	All non-empty masks from both branches
SAM3 confidence weight α	1.0
Prototype consistency weight β	0.2
Cross-branch agreement weight γ	0.1
Selection threshold δ	0.5
Accepted-mask aggregation	Pixel-wise union
No accepted candidates	Empty mask
Evaluation	
Target datasets	DeepGlobe, ISIC, Chest X-Ray, FSS-1000
Metric	mIoU
Inference hardware	NVIDIA RTX 3090

verse target appearances and domain shifts. Each row shows five annotated support images, a query image with its ground-truth mask, and predictions from DR-Adapter, CPS, and our method.

In the illustrated DeepGlobe example, our method effectively reduces spurious foreground regions and produces a more coherent and complete target mask. For ISIC, the prediction follows the lesion contour more closely, while showing less over-segmentation into the surrounding skin regions. The Chest X-Ray and FSS-1000 examples further illustrate the ability of our method to capture the overall target shape and preserve the main object structure, although some boundary discrepancies still remain. Overall, these qualitative examples provide additional support for the 5-shot results reported in the main paper and further illustrate the applicability of the proposed framework across different target domains with diverse visual appearances and domain shifts.

A.5 Experimental Settings and Hyperparameters

Table 8 summarizes the implementation and evaluation settings of our framework. We use a fully frozen SAM3 without source-domain training or parameter updates. Experiments are conducted under both 1-shot and 5-shot settings using a 1008×1008 shared canvas. In the 1-shot setting, the support image is placed above the query image. In the 5-shot setting, five support images are arranged along the top and left sides of the canvas, leaving the larger bottom-right region for the query image. A tight bounding box is derived from each support mask and mapped to the canvas as a visual prompt for both branches. The visual-text branch additionally uses the class name as a textual prompt, allowing the two branches to generate candidate masks from different cues.

We collect all non-empty masks from both branches into a candidate pool. Each canvas-level mask is cropped to the query region and restored to the original query resolution before reliability assessment. For each candidate, we combine three scores: the confidence predicted by SAM3, its feature consistency with the support foreground relative to the support background, and its spatial agreement with candidates from the other prompting branch. The corresponding fusion weights are $\alpha = 1.0$, $\beta = 0.2$, and $\gamma = 0.1$. Candidates with a fused reliability score above $\delta = 0.5$ are retained and merged by pixel-wise union to obtain the final query prediction. If no candidate passes the threshold, the method returns an empty mask. We report mIoU on DeepGlobe, ISIC, Chest X-Ray, and FSS-1000. All inference experiments are performed on an NVIDIA RTX 3090 GPU.

A.6 Algorithm Overview

Algorithm 1 details the inference pipeline of our dual-branch candidate calibration framework under the 1-shot cross-domain few-shot segmentation setting. Given a support image and its ground-truth mask, a query image, and a class name, the framework predicts the query mask using a fully frozen SAM3. The process requires neither source-domain training nor parameter updates and consists of three stages:

(1) Shared-Canvas Construction. The support image is placed above the query image on a shared spatial canvas. A tight bounding box is derived from the support mask and mapped to canvas coordinates. This box provides a visual prompt that identifies the target in the support region while allowing SAM3 to process both images within the same canvas;

(2) Dual-Branch Candidate Generation. The visual-only branch uses the support-derived box prompt, whereas the visual-text branch additionally uses the class name. Both prompting configurations produce candidate masks and associated SAM3 confidence scores. We pool the candidates while retaining their branch labels, discard empty masks, and crop the remaining masks to the query region. The cropped masks are then resized to the original query resolution;

(3) Candidate Reliability Calibration. For each query-side candidate, we assess its consistency with the support foreground relative to the support background. We also measure its spatial agreement with candidates generated by the other prompting branch. These two scores are combined with the SAM3 confidence score to obtain a final quality score. Candidates exceeding the selection threshold are retained and merged by pixel-wise union to form the query prediction. If no candidate passes the threshold, the framework returns an empty mask.

Algorithm 1 Pipeline of the Proposed Dual-Branch Candidate Calibration Framework

Input: Support image I_s, support mask M_s, query image I_q, class name T
Parameters: Frozen SAM3 f_θ with image encoder E, fusion weights α, β, γ, selection threshold δ
Output: Final query prediction M_{final}
/* Shared-Canvas Construction */
Derive the support box: $B_s \leftarrow \text{Box}(M_s)$
Construct the shared canvas: $I_c \leftarrow \text{Canvas}(I_s, I_q)$
Map the support box to canvas coordinates: $B_c \leftarrow \Phi_s(B_s)$
/* Dual-Branch Candidate Generation */
Generate visual-only candidates: $\mathcal{C}_V \leftarrow f_\theta(I_c, B_c)$
Generate visual-text candidates: $\mathcal{C}_{VT} \leftarrow f_\theta(I_c, B_c, T)$
Pool candidates: $\mathcal{C} \leftarrow \mathcal{C}_V \cup \mathcal{C}_{VT}$, preserving their SAM3 scores and branch labels
Crop each mask to the query region and restore its original resolution
Discard empty masks to obtain $\mathcal{Q} = \{(M_i, s_i, b_i)\}$
/* Support Prototype Construction */
Extract image features: $F_s \leftarrow E(I_s)$, $F_q \leftarrow E(I_q)$
Compute support foreground and background prototypes:

$$p_{\text{fg}} = \text{Avg}(F_s \mid M_s), \qquad p_{\text{bg}} = \text{Avg}(F_s \mid 1 - M_s)$$

/* Candidate Reliability Calibration */
Initialize $M_{\text{final}} \leftarrow \mathbf{0}$
for each $(M_i, s_i, b_i) \in \mathcal{Q}$ **do**
  Compute the candidate representation: $z_i \leftarrow \text{Avg}(F_q \mid M_i)$
  Compute prototype consistency:

$$r_i^{\text{proto}} = \text{clip}\left(\frac{\cos(z_i, p_{\text{fg}}) - \cos(z_i, p_{\text{bg}}) + 2}{4}, 0, 1\right)$$

  Identify candidates from the other branch: $\mathcal{Q}_{\text{other}} \leftarrow \{(M_j, s_j, b_j) \in \mathcal{Q} \mid b_j \neq b_i\}$
  Initialize $r_i^{\text{agree}} \leftarrow 0$
  if $\mathcal{Q}_{\text{other}} \neq \emptyset$ **then**
    Compute cross-branch agreement:

$$r_i^{\text{agree}} = \max_{(M_j, s_j, b_j) \in \mathcal{Q}_{\text{other}}} \text{IoU}(M_i, M_j)$$

  end if
  Obtain the SAM3 confidence score $\bar{s}_i$
  Fuse the three reliability scores:

$$q_i = \frac{\alpha \bar{s}_i + \beta r_i^{\text{proto}} + \gamma r_i^{\text{agree}}}{\alpha + \beta + \gamma}$$

  if $q_i > \delta$ **then**
    Aggregate the retained mask: $M_{\text{final}} \leftarrow M_{\text{final}} \vee M_i$
  end if
end for
return M_{final}