FoRIS: Progressive Foreground Refinement for Training-Free In-Context Segmentation

Anonymous authors
Paper under double-blind review

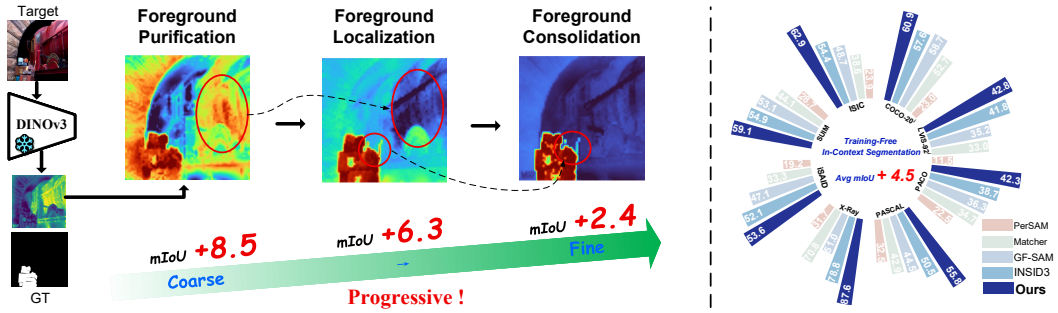


Figure 1: **Left:** Our core coarse-to-fine progressive refinement paradigm, which sequentially performs Foreground Purification, Foreground Localization, and Foreground Consolidation. Ablation results on COCO-20^i show that each stage progressively improves the segmentation mIoU. **Right:** Performance comparison with existing methods across diverse datasets spanning multiple domains, demonstrating the consistent effectiveness of FoRIS and its improvement in average mIoU.

Abstract

In-Context Segmentation (ICS) aims to precisely segment arbitrary semantic concepts, such as objects or parts, given one or a few annotated visual exemplars. In this paper, we revisit ICS from a more classical segmentation perspective, viewing it as a coarse-to-fine progressive refinement process. Rather than directly predicting the final mask through reference-query matching, we progressively refine the segmentation from coarse and ambiguous foreground responses to precise and complete foreground structures. Building upon this perspective, we propose a training-free in-context segmentation framework, termed FoRIS. Specifically, FoRIS consists of three key stages: Foreground Purification, Foreground Localization, and Foreground Consolidation, which progressively suppress background distractions, localize discriminative target regions, and recover complete foreground structures through semantic aggregation. Experimental results demonstrate that FoRIS achieves SOTA performance across semantic and part segmentation tasks, with average improvements of 4.5 and 4.8 mIoU points over existing approaches in the 1-shot and 5-shot settings, respectively. Code: https://anonymous.4open.science/r/FoRIS.

1 Introduction

Segmentation is a fundamental task in computer vision with broad applications across natural, medical, and remote-sensing imagery Hu et al. (2025; 2026b; 2024; 2026a); Ronneberger et al. (2015); Yang et al. (2023; 2025); He et al. (2017); Badrinarayanan et al. (2017); Chen et al. (2017); Arbelaez et al. (2010); Felzenszwalb & Huttenlocher (2004); Achanta et al. (2012); Shi & Malik (2000); Long et al. (2015). Traditional segmentation methods are typically restricted to predefined semantic categories and require substantial task-specific supervision, limiting their flexibility in open-world scenarios Cuttano et al. (2026). With the rapid development of visual foundation models (VFMs) Caron et al. (2021); Oquab et al. (2024); Siméoni et al. (2025); Kirillov et al. (2023); Carion et al. (2025), *in-context segmentation* (ICS) has emerged as a promising paradigm for open-world segmentation Cuttano et al. (2026); Zhang et al. (2024b;a); Liu et al. (2024b). Given one or more annotated reference examples at inference time, ICS aims to segment arbitrary target concepts without category-specific retraining, enabling flexible generalization across diverse concepts, domains, and annotation settings Cuttano et al. (2026).

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.

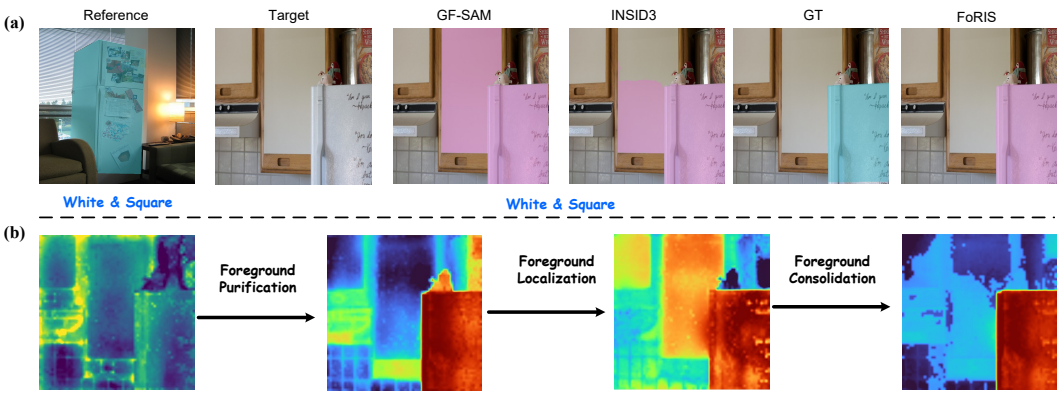


Figure 2: **Motivation for progressive foreground refinement.** (a) Existing training-free ICS methods rely on direct reference-query matching, which may confuse visually similar background regions with the target. (b) We reinterpret ICS as a progressive foreground refinement process consisting of foreground purification, foreground localization, and foreground consolidation.

Existing ICS methods generally follow two paradigms. *Fine-tuned methods* adapt pretrained visual or diffusion architectures with task-specific segmentation supervision, translating pretrained representations into dense pixel-level predictions. Although effective within the training distribution, they require additional optimization and supervision, which can limit open-world generalization. In contrast, *training-free methods* combine complementary pretrained models, typically leveraging visual representations for reference-query correspondence and segmentation priors for mask generation. While avoiding task-specific optimization, these methods still primarily rely on reference-query correspondence to transfer foreground semantics to the target image.

Despite their methodological differences, existing training-free ICS methods largely adopt a *correspondence-centric* formulation, where foreground semantics are transferred from reference exemplars to query images through feature correspondence or similarity propagation Liu et al. (2024b); Zhang et al. (2024a); Cuttano et al. (2026). As illustrated in Fig. 2(a), given a reference image containing a white square-shaped refrigerator, such correspondence-based methods may incorrectly associate visually similar white square-shaped wall regions in the query image with the target, resulting in inaccurate segmentation. This failure exposes a fundamental mismatch between *visual similarity* and *target membership*: a query region may closely resemble the reference foreground while being semantically irrelevant to the target.

More fundamentally, an annotated foreground mask specifies *where* the target lies, but does not guarantee that the corresponding frozen features provide a semantically pure representation of the target. Contextual, positional, and background-correlated information may remain entangled with foreground features. Meanwhile, variations in viewpoint, appearance, and object structure can cause true target regions in the query image to produce only partial or fragmented semantic responses. These factors introduce three coupled sources of uncertainty: *foreground contamination*, *ambiguous target localization*, and *fragmented foreground responses*. Consequently, the central challenge of training-free ICS is not simply to establish more accurate correspondence, but to progressively transform noisy visual evidence into reliable, foreground-centric semantic representations.

This observation motivates us to reconsider the role of correspondence in ICS. Rather than discarding correspondence, we view it as an intermediate source of semantic evidence within a *progressive foreground refinement* process, as illustrated in Fig. 2(b). Specifically, foreground semantics progressively emerge through three stages: *foreground purification*, *foreground localization*, and *foreground consolidation*. Foreground purification suppresses contextual and background interference in reference features, establishing a cleaner semantic basis for correspondence. Foreground localization then exploits the purified foreground cues to identify discriminative target regions in the query image, reducing localization ambiguity. Finally, foreground consolidation aggregates semantically related but spatially scattered responses, progressively recovering complete object structures from fragmented local evidence. In this view, correspondence is not treated as a sufficient endpoint, but as one component of a broader process that progressively refines semantic evidence into coherent object-level predictions.

Based on this perspective, we propose **FoRIS** (**Fo**reground **R**efinement for Training-Free **I**n-context **S**egmentation), a training-free ICS framework that explicitly implements this progressive foreground refinement paradigm. FoRIS sequentially performs foreground purification, foreground localization, and foreground consolidation using a frozen VFM, without additional segmentation supervision or auxiliary segmentation models. By placing correspondence within a progressive refinement pipeline, FoRIS shifts the focus from one-step semantic transfer toward the reliable extraction, localization, and consolidation of foreground evidence.

We further investigate a challenging and underexplored limitation of current VFM-based training-free ICS methods: their limited generalization to *slender and fragmented structures*. Such structures often exhibit weak semantic responses, limited spatial support, and substantial appearance variations, amplifying the ambiguities that correspondence-based methods struggle to resolve. We systematically evaluate this challenge across diverse slender-structure datasets and find that existing training-free ICS methods remain substantially less reliable in such scenarios. These findings reveal a pronounced generalization gap despite the strong representations provided by VFMs, highlighting the need for training-free ICS methods that can accommodate heterogeneous target geometries and fragmented structures. Our contributions are summarized as follows:

- We introduce *progressive foreground refinement* as a new perspective for training-free in-context segmentation, repositioning reference-query correspondence as an intermediate source of semantic evidence within a three-stage refinement process: foreground purification, localization, and consolidation.
- We propose **FoRIS**, a training-free ICS framework that explicitly implements this progressive refinement paradigm and achieves average IoU improvements of over 4.5 and 4.8 percentage points over existing methods under the 1-shot and 5-shot settings, respectively.
- We systematically investigate the generalization limitations of current VFM-based training-free ICS methods on slender and fragmented structures, revealing a challenging and underexplored scenario that remains an open problem for future ICS research.

2 Related Work

In-context segmentation. In-context segmentation (ICS) extends the in-context learning paradigm of LLMs Brown et al. (2020); Ouyang et al. (2023); Chowdhery et al. (2023); Touvron et al. (2023) to visual segmentation. Early works such as SegGPT Wang et al. (2023b) and Painter Wang et al. (2023a) demonstrated that segmentation can be conditioned on contextual examples, while recent approaches leverage visual foundation models (VFMs) for one-shot semantic, part, and personalized segmentation Liu et al. (2024b); Zhang et al. (2024b). Unlike few-shot segmentation Wang et al. (2019); Cuttano et al. (2025); Hong et al. (2022); Lang et al. (2022), which typically targets predefined novel classes, ICS aims to segment arbitrary concepts specified at inference time across diverse semantic granularities.

Existing ICS methods can be broadly categorized into *training-free* and *supervised* approaches. Training-free methods Liu et al. (2024b); Zhang et al. (2024a); Espinosa et al. (2025); Cuttano et al. (2026) typically combine pretrained visual representations with external segmentation priors, while supervised methods Meng et al. (2024); Zhu et al. (2024) adapt pretrained architectures through task-specific training. Despite their different implementations, existing approaches largely rely on reference-target correspondence as the primary mechanism for transferring foreground semantics and inferring target masks. However, correspondence alone may be insufficient when reference features are contaminated by background information or target responses are ambiguous and fragmented. Rather than discarding correspondence, we reinterpret its role within a *progressive foreground refinement* process. Inspired by the coarse-to-fine nature of segmentation, our approach progressively purifies foreground representations, localizes discriminative target regions, and consolidates fragmented semantic responses, allowing correspondence-derived evidence to evolve into coherent object masks.

Progressive Segmentation and Mask Refinement. A central challenge in training-free segmentation is to transform the rich visual representations provided by VFMs into reliable foreground-background separation. This objective is closely related to a long-standing principle in classical

segmentation: accurate object masks often emerge through *progressive refinement* rather than a single prediction step. Classical approaches progressively reduce segmentation uncertainty by updating region assignments, object boundaries, or spatial consistency. Region Growing Adams & Bischof (1994) expands initial seeds according to local similarity, while Graph Cut Boykov & Jolly (2001) and GrabCut Rother et al. (2004) iteratively optimize foreground-background assignments through global region relationships. Active Contour Kass et al. (1988) progressively evolves object boundaries, and multi-scale approaches Adelson et al. (1984) refine object structures from coarse to fine resolutions. Despite their methodological differences, these approaches share a common principle: reliable segmentation emerges by progressively transforming uncertain local evidence into coherent foreground structures.

In contrast, existing training-free ICS methods largely adopt a *correspondence-centric* formulation, where foreground semantics are transferred from annotated reference images to target images through feature correspondence or similarity propagation Liu et al. (2024b); Zhang et al. (2024a); Cuttano et al. (2026). Although some methods further employ clustering, propagation, or mask refinement, these operations are primarily built upon correspondence-derived evidence rather than explicitly modeling foreground refinement as the central inference process. Consequently, they remain vulnerable to background interference, ambiguous semantic correspondence, and fragmented target responses, which may lead to incomplete or disconnected predictions and semantic drift. This observation motivates us to reinterpret training-free ICS as a *progressive foreground refinement* process, where correspondence serves as intermediate semantic evidence that is progressively purified, localized, and consolidated into coherent object-level predictions.

3 In-context Segmentation with FoRIS

Given reference image–mask pairs $\{(I_s, M_s)\}_{s=1}^{S}$ and a target image I_t, FoRIS predicts $\hat{M}_t$ without fine-tuning or auxiliary heads. We extract patch features from the last intermediate layer of the encoder and ℓ_2-normalize them along the channel dimension:

$$\mathbf{F} = \phi\big([I_1, \ldots, I_S, I_t]\big) \in \mathbb{R}^{(S+1)\times C\times H\times W}, \qquad \mathbf{f}(\mathbf{p}) = \mathbf{F}(\mathbf{p})/\|\mathbf{F}(\mathbf{p})\|_2. \tag{1}$$

FoRIS refines a foreground response through three stages—*foreground purification* (FP), *foreground localization* (FL), and *foreground consolidation* (FC). algorithm 1 summarizes the full data flow.

3.1 Foreground Purification

FP removes positional layout bias and background clutter via adaptive positional debiasing (APD) and two-stage foreground refinement (FR).

Adaptive Positional Debiasing. Positional bias can hinder prototype matching when reference and target objects occupy inconsistent locations. However, in domains with strong spatial regularity—e.g., lung CT, where anatomical structures follow consistent layouts—positional cues can serve as informative priors rather than noise. We therefore debias *adaptively* instead of indiscriminately. Following INSID3 Cuttano et al. (2026), we offline estimate a low-rank positional subspace $\mathbf{U}$ via SVD on centered features from a zero-input image. At inference, we measure reference–target semantic alignment:

$$s_{\text{sem}} = \frac{\langle \boldsymbol{\mu}^{\text{fg}}, \bar{\mathbf{f}}_t \rangle}{\|\bar{\mathbf{f}}_t\|_2}, \qquad \boldsymbol{\mu}^{\text{fg}} = \text{norm}\Big(\text{mean}_{\mathbf{p}\in\Omega^{\text{fg}}} \mathbf{f}_r(\mathbf{p})\Big), \tag{2}$$

where $\bar{\mathbf{f}}_t = \text{mean}_{\mathbf{p}} \mathbf{f}_t(\mathbf{p})$. Debiasing is applied to *all* reference and target tokens only when $s_{\text{sem}} < \theta$:

$$\mathbf{P}_\perp = \mathbf{I}_C - \mathbf{U}\mathbf{U}^\top, \qquad \tilde{\mathbf{f}}(\mathbf{p}) = \begin{cases} \text{norm}\big(\mathbf{P}_\perp \mathbf{f}(\mathbf{p})\big), & s_{\text{sem}} < \theta, \\ \mathbf{f}(\mathbf{p}), & \text{otherwise.} \end{cases} \tag{3}$$

We denote the resulting feature maps as $\tilde{\mathbf{F}}$ and use $\tilde{\mathbf{f}}$ for all subsequent stages unless stated otherwise. We set θ=0.80 by default (cf. section 4.2 and fig. 5).

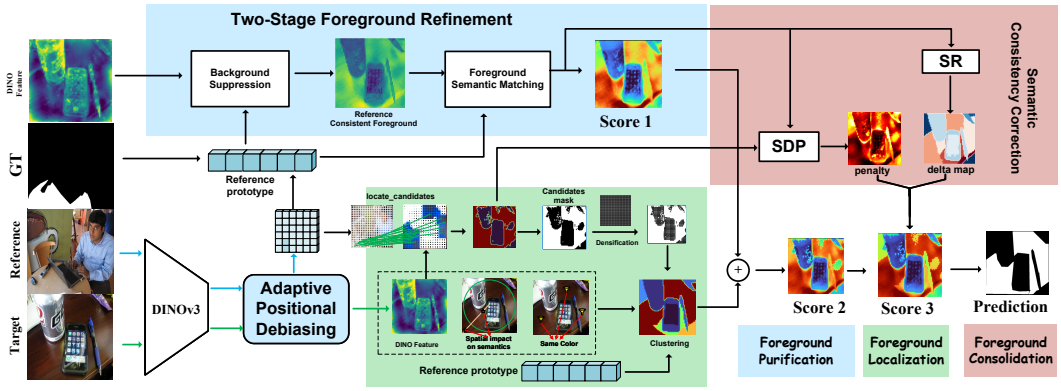


Figure 3: **Overview of FoRIS.** Given reference image(s) with foreground mask(s) and a target image, FoRIS performs through three sequential stages. **(1) Foreground purification (FP):** patch features from a frozen DINOv3 encoder are adaptively debiased and purified via two-stage foreground refinement, including contrastive background suppression and foreground semantic matching, yielding an initial response (*Score 1*). **(2) Foreground localization (FL):** bidirectional feature matching identifies reliable foreground candidates, which are densified and integrated with clustering-based seed priors built from DINO, spatial, and color cues to produce a refined response (*Score 2*). **(3) Foreground consolidation (FC):** semantic disagreement penalization (SDP) and semantic reweighting (SR) suppress conflicting activations and strengthen semantically pure regions, producing the final response (*Score 3*).

Two-stage foreground refinement. FR operates on debiased features $\tilde{\mathbf{f}}$ and produces both an initial response and a gated feature map used later in FC.

Stage 1 (background supression). We first construct foreground and background prototypes from debiased reference tokens. Foreground tokens from all references are concatenated and partitioned by average-linkage agglomerative clustering under cosine distance, yielding J cluster prototypes $\boldsymbol{\mu}_j = \mathrm{norm}\big(\mathrm{mean}_{i:\ell_i=j}\mathbf{x}_i^{\mathrm{fg}}\big)$ for $j = 1, \ldots, J$, where J is determined by the clustering procedure rather than fixed manually and ℓ_i is the cluster label of foreground token i. The hard-negative background prototype targets background patches most confusable with the foreground. We rank all debiased background tokens by their cosine similarity to $\boldsymbol{\mu}^{\mathrm{fg}}$ and retain the top 20%:

$$\mathcal{I}_{\mathrm{hn}} = \mathrm{TopK}_{20\%}\Big(\big\{\langle \mathbf{x}_i^{\mathrm{bg}}, \boldsymbol{\mu}^{\mathrm{fg}}\rangle\big\}_{i=1}^{N_{\mathrm{bg}}}\Big), \tag{4}$$

and define $\boldsymbol{\mu}^{\mathrm{bg}} = \mathrm{norm}\big(\mathrm{mean}_{i\in\mathcal{I}_{\mathrm{hn}}}\mathbf{x}_i^{\mathrm{bg}}\big)$. A contrastive gate is then computed on all reference and target tokens and applied multiplicatively. We aggregate similarities to the J foreground cluster prototypes via temperature-scaled log-sum-exp:

$$g(\mathbf{p}) = \sigma\Big(\beta \log \sum_{j=1}^{J} \exp\Big(\frac{\langle \tilde{\mathbf{f}}(\mathbf{p}), \boldsymbol{\mu}_j\rangle}{\beta}\Big) - \langle \tilde{\mathbf{f}}(\mathbf{p}), \boldsymbol{\mu}^{\mathrm{bg}}\rangle\Big), \qquad \hat{\mathbf{f}}(\mathbf{p}) = g(\mathbf{p}) \cdot \tilde{\mathbf{f}}(\mathbf{p}), \tag{5}$$

where β=0.07 and reference foreground locations receive a gate floor. Stage 1 yields gated features $\hat{\mathbf{f}}$, which are passed to FC for cluster reweighting (section 3.3).

Stage 2 (foreground semantic matching). Score 1 is computed on *debiased but ungated* target features $\tilde{\mathbf{f}}_t$, using prototypes rebuilt from debiased reference tokens. This decouples contrastive gating from prototype matching: gating suppresses background-dominated activations for consolidation, while matching preserves the full debiased semantic geometry for cross-image correspondence. The orthogonalized background direction is

$$\boldsymbol{\mu}_{\perp}^{\mathrm{bg}} = \frac{\boldsymbol{\mu}^{\mathrm{bg}} - \langle \boldsymbol{\mu}^{\mathrm{bg}}, \boldsymbol{\mu}^{\mathrm{fg}}\rangle \boldsymbol{\mu}^{\mathrm{fg}}}{\|\boldsymbol{\mu}^{\mathrm{bg}} - \langle \boldsymbol{\mu}^{\mathrm{bg}}, \boldsymbol{\mu}^{\mathrm{fg}}\rangle \boldsymbol{\mu}^{\mathrm{fg}}\|_2}, \tag{6}$$

and the initial purified response is

$$S^{(1)}(\mathbf{p}) = \beta \log \sum_{j=1}^{J} \exp\Big(\frac{\langle \tilde{\mathbf{f}}_t(\mathbf{p}), \boldsymbol{\mu}_j\rangle}{\beta}\Big) - \langle \tilde{\mathbf{f}}_t(\mathbf{p}), \boldsymbol{\mu}_{\perp}^{\mathrm{bg}}\rangle. \tag{7}$$

3.2 FOREGROUND LOCALIZATION

FL refines $S^{(1)}$ into Score 2 using debiased reference and target features $\tilde{\mathbf{f}}$.

Cross-image candidate voting. Dense patch-to-patch matching assigns a foreground vote to each target location:

$$V(\mathbf{p}) = \frac{1}{S}\sum_{s=1}^{S} \mathbb{1}\Big[M_s\big(\arg\max_{\mathbf{q}} \langle \tilde{\mathbf{f}}_s(\mathbf{q}), \tilde{\mathbf{f}}_t(\mathbf{p})\rangle\big)\Big]. \tag{8}$$

Locations with majority support form $\mathcal{C} = \{\mathbf{p} \mid V(\mathbf{p}) > \frac{1}{2}\}$, which is lightly densified on a coarse spatial grid.

Multi-cue clustering and seed-cluster prior. Target patches are clustered on a joint representation of ℓ_2-normalized DINO features, RGB color, and 2D coordinates:

$$\mathbf{z}_i = \mathrm{norm}\big([\mathbf{f}_i;\, \mathbf{c}_i;\, \mathbf{p}_i]\big). \tag{9}$$

The same agglomerative clustering protocol as above assigns each patch a cluster id $\kappa(\mathbf{p}) \in \{1, \ldots, J'\}$. Let $\mathcal{G}_j = \{\mathbf{p} : \kappa(\mathbf{p}) = j\}$ denote the patch set of cluster j. If no cluster overlaps $\mathcal{C}$, we set $P(\mathbf{p}) = \mathbb{1}[\mathbf{p} \in \mathcal{C}]$; otherwise, the seed cluster $j^\star$ maximizes reference affinity weighted by candidate area:

$$j^\star = \arg\max_{j \in \mathcal{J}_\mathcal{C}} \underbrace{\mathrm{mean}_{\mathbf{p}\in\mathcal{G}_j} \langle \tilde{\mathbf{f}}_t(\mathbf{p}), \boldsymbol{\mu}^{\mathrm{fg}}\rangle}_{\text{cross-image affinity}} \cdot \underbrace{\frac{|\mathcal{G}_j \cap \mathcal{C}|}{|\mathcal{G}_j|}}_{\text{area weight}}, \tag{10}$$

where $\mathcal{J}_\mathcal{C} = \{j : \mathcal{G}_j \cap \mathcal{C} \neq \emptyset\}$. All clusters receive

$$\rho_j = \mathrm{mean}_{\mathbf{p}\in\mathcal{G}_j} \langle \tilde{\mathbf{f}}_t(\mathbf{p}), \boldsymbol{\mu}^{\mathrm{fg}}\rangle \cdot \max\big(\langle \boldsymbol{\mu}_{j^\star}, \boldsymbol{\mu}_j\rangle, 0\big) \cdot \frac{|\mathcal{G}_j \cap \mathcal{C}|}{|\mathcal{G}_j|}, \tag{11}$$

which is min–max normalized across clusters and mapped to patches as

$$P(\mathbf{p}) = \frac{\rho_{\kappa(\mathbf{p})} - \min_i \rho_i}{\max_i \rho_i - \min_i \rho_i + \epsilon} \in [0, 1]. \tag{12}$$

Score 2 is

$$S^{(2)} = S^{(1)} + (V - \tfrac{1}{2}) + (P - \tfrac{1}{2}). \tag{13}$$

3.3 FOREGROUND CONSOLIDATION

FC transforms Score 2 into Score 3 via semantic disagreement penalization (SDP) and semantic reweighting (SR).

Semantic disagreement penalization. We penalize inter-map disagreement and foreground–background coupling:

$$\mathcal{D}(\mathbf{p}) = |S_{\mathrm{fg}}(\mathbf{p}) - V(\mathbf{p})| + |S_{\mathrm{fg}}(\mathbf{p}) - P(\mathbf{p})|, \qquad \mathcal{B}(\mathbf{p}) = \min\big(S_{\mathrm{fg}}(\mathbf{p}), S_{\mathrm{bg}}(\mathbf{p})\big). \tag{14}$$

An uncertainty-modulated penalty concentrates corrections on ambiguous patches:

$$\Pi(\mathbf{p}) = \big(\mathcal{D}(\mathbf{p}) + \mathcal{B}(\mathbf{p})\big) \cdot \mathcal{U}(\mathbf{p}), \qquad \mathcal{U}(\mathbf{p}) = 1 - 2\big|S_{\mathrm{fg}}(\mathbf{p}) - \tfrac{1}{2}\big|. \tag{15}$$

Semantic reweighting. Target patches are re-clustered on gated features $\hat{\mathbf{f}}_t$, yielding a new cluster assignment $\kappa'(\mathbf{p})$. Each cluster j receives a reweighting factor based on foreground–background separation:

$$\Delta_j = \big(\overline{S}^{(j)}_{\mathrm{fg}} - \overline{S}^{(j)}_{\mathrm{bg}}\big)_+ - \min\big(\overline{S}^{(j)}_{\mathrm{fg}}, \overline{S}^{(j)}_{\mathrm{bg}}\big), \tag{16}$$

where $\overline{S}^{(j)}_{\cdot}$ is the mean over $\{\mathbf{p} : \kappa'(\mathbf{p}) = j\}$. The patch-level map is $\Delta(\mathbf{p}) = \Delta_{\kappa'(\mathbf{p})}$, and the consolidated response is

$$S^{(3)} = S^{(2)} - \Pi + \Delta. \tag{17}$$

Table 1: **Comparison of FoRIS (mIoU in %, ↑) on one-shot semantic, part, and personalized segmentation.** State-of-the-art methods are grouped into task-specific fine-tuning and training-free approaches. Previous training-free methods rely on SAM, pre-trained with mask-level supervision, whereas FoRIS uses only frozen self-supervised DINOv3 features. Gray indicates the model was trained on the corresponding train split of the dataset; best results **bold**, 2nd best underlined.

Method	Encoder	#Param	Semantic						Part		Avg
			LVIS-92	COCO-20	ISIC	SUIM	iSAID	X-Ray	PASCAL	PACO	
Task-specific fine-tuning: *Semantic + mask supervision*											
Painter Wang et al. (2023a)	ViT	354 M	10.5	33.1	–	–	–	–	30.4	14.1	–
SegGPT Wang et al. (2023b)	ViT	354 M	18.6	56.1	37.5	34.9	30.9	87.5	35.8	13.5	39.4
SINE Liu et al. (2024a)	DINOv2	373 M	31.2	64.5	25.8	50.7	38.3	39.8	36.2	23.3	38.7
DiffewS Zhu et al. (2024)	Stable Diffusion	890 M	31.4	71.3	27.8	48.9	47.5	41.6	34.0	22.8	40.7
SegIC Meng et al. (2024)	DINOv2	310 M	44.6	76.1	25.3	52.5	46.1	34.5	39.9	25.9	43.1
SegIC (COCO) Meng et al. (2024)	DINOv2	310 M	35.7	75.6	22.5	52.9	40.8	30.8	38.6	25.1	40.3
Training free: *Mask-supervised pre-training*											
PerSAM Zhang et al. (2024b)	SAM	640 M	11.5	23.0	23.9	28.7	19.2	31.7	32.5	22.5	24.1
Matcher Liu et al. (2024b)	DINOv2 + SAM	945 M	33.0	52.7	38.6	44.1	33.3	70.8	42.9	34.7	43.7
GF-SAM Zhang et al. (2024a)	DINOv2 + SAM	945 M	35.2	58.7	48.7	53.1	47.1	51.0	44.5	36.3	46.8
GF-SAM Zhang et al. (2024a)	DINOv3+SAM	945M	31.8	54.8	50.9	50.5	46.7	56.1	44.9	34.4	46.3
Training-free: *Unsupervised pre-training*											
INSID3	DINOv3	304M	41.8	57.6	54.4	54.9	52.1	78.8	50.5	38.7	53.6
FoRIS (ours)	DINOv3	**304M**	**42.8**	**60.9**	**62.9**	**59.1**	**53.6**	**87.6**	**55.8**	**42.3**	**58.1**

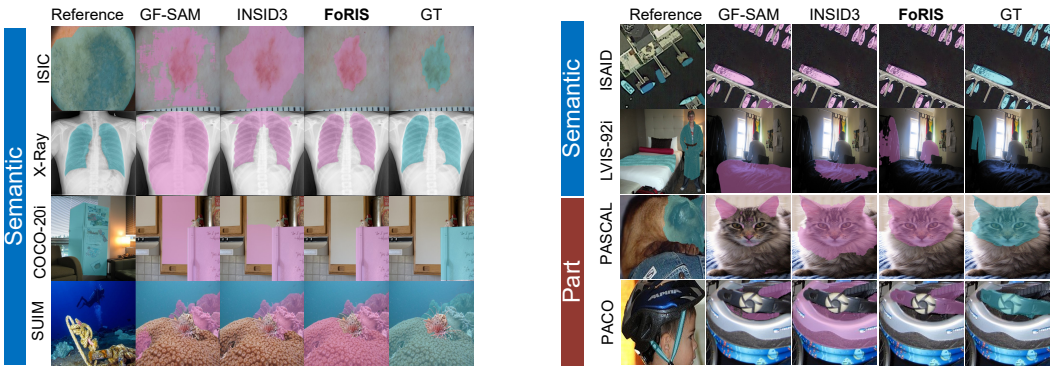


Figure 4: Comparison of **FoRIS** with GF-SAM Zhang et al. (2024a) and INSID3 Cuttano et al. (2026) on one-shot semantic and part segmentation.

4 Experiments

We evaluate FoRIS on one-shot *semantic* and *part* segmentation. Following INSID3 Cuttano et al. (2026), we adopt the same datasets, baselines, and evaluation protocols for fair comparison. All results are reported as mIoU; implementation details are provided in Appendix A.

4.1 Main Results

One-shot semantic segmentation. FoRIS achieves the best performance on all semantic benchmarks (cf. Tab. 1). Compared with the GF-SAM, FoRIS improves mIoU by 7.6 %, 2.2 %, 14.2 %, 6.0 %, 6.5 %, and 36.6 % pts. on LVIS-92^i, COCO-20^i, ISIC, SUIM, iSAID, and Chest X-ray, with the largest gains on out-of-domain datasets. Under the same frozen DINOv3 backbone and parameter budget, FoRIS consistently surpasses INSID3 by 1.0 %–8.8 % pts. across all six benchmarks. FoRIS also generalizes better than task-specific fine-tuned methods without any segmentation training. Qualitative results in Fig. 4 show that FoRIS produces clean, semantically coherent masks directly from frozen features.

One-shot part segmentation. FoRIS also sets a new SOTA on part-level benchmarks. It outperforms GF-SAM by 11.3 % and 6.0 % pts. on PASCAL-Part and PACO-Part, and INSID3 by 5.3 % and 3.6 % pts., respectively. Compared with fine-tuned SegIC and DiffewS, FoRIS achieves up to 21.8 % pts. higher mIoU while remaining fully training-free. These results indicate that progressive foreground refinement benefits fine-grained structural understanding beyond object-level segmentation. Qualitative comparisons in Fig. 4 confirm improved boundary quality and part integrity.

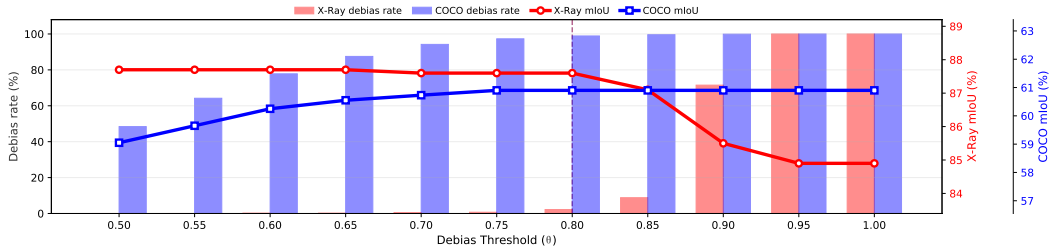


Figure 5: Effect of the adaptive debiasing threshold θ on COCO and Chest X-ray.

4.2 Ablation Study

We conduct ablation experiments to validate each component of FoRIS (cf. Tab. 2). Starting from the baseline, progressively adding FP, FL, and FC consistently improves performance across all benchmarks, confirming the complementarity of the three-stage design.

Table 2: Ablation study of different modules.

Base	FP		FL	FC	mIoU			
	APD	FR			COCO	ISIC	SUIM	PASCAL
✓					43.7	46.8	44.4	38.7
✓	✓				46.9	41.2	43.9	39.3
✓		✓			48.3	57.1	50.2	44.4
✓			✓		51.9	53.1	53.2	51.8
✓	✓	✓			52.2	58.7	52.9	48.3
✓	✓		✓		52.9	49.3	55.6	50.8
✓		✓	✓		56.2	58.0	55.8	53.8
✓	✓	✓	✓		58.5	60.2	58.9	55.1
✓		✓	✓	✓	58.0	59.2	54.8	54.2
✓	✓	✓	✓	✓	**60.9**	**62.9**	**59.1**	**55.8**

(1)Foreground Purification. We examine the dual role of positional bias: it can interfere with correspondence when reference and target layouts differ, yet provide useful structural priors in domains with stable spatial organization, such as Chest X-ray. APD therefore adaptively applies debiasing only when the semantic alignment score $s_{\text{sem}} < \theta$, projecting features onto the orthogonal complement of the positional subspace. We sweep $\theta \in \{0.50, 0.55, \ldots, 1.00\}$ on COCO and Chest X-ray (Fig. 5). On COCO, mIoU increases from 59.1 % to 60.9 % and saturates at $\theta \geq 0.75$, indicating that suppressing positional bias benefits cross-layout matching. In contrast, Chest X-ray remains stable at 87.6 %–87.7 % for $\theta \leq 0.80$ but drops to 84.9 % at θ=1.0, showing that excessive debiasing can remove useful anatomical priors.

Tab. 2 further confirms the dataset-dependent role of APD: enabling APD alone improves COCO by 3.2 % pts. (43.7 $\rightarrow$ 46.9) but reduces ISIC by 5.6 % pts. (46.8 $\rightarrow$ 41.2). With all modules enabled, FoRIS achieves 60.9 %/62.9 %/59.1 %/55.8 % mIoU on COCO/ISIC/SUIM/PASCAL-Part, with up to 4.3 % pts. improvement over the best variant without APD. These results support *adaptive*, rather than always-on, positional debiasing for cross-domain generalization.

(2)Foreground Localization. *(a)Multi-cue clustering.* We study whether augmenting frozen DINOv3 features with positional coordinates and RGB appearance during clustering improves FL (cf. Tab. 3c). Positional cues benefit Chest X-ray (+1.1 % pts.), where anatomical structures follow strong spatial regularity, while RGB cues help SUIM (+0.2 % pts.), where appearance variation is more pronounced. Combining both cues maintains strong performance across domains (87.6 % on Chest X-ray, 59.1 % on SUIM), and we adopt this joint representation by default. *(b)Dense candidate aggregation.* We further evaluate coarse-grid densification of the candidate set (cf. Tab. 3b). Dense aggregation consistently improves mIoU on COCO (60.1 % $\rightarrow$ 60.9 %), Chest X-ray (87.4 % $\rightarrow$ 87.6 %), and PACO-Part (41.8 % $\rightarrow$ 42.3 %). The gains are largest on fine-grained PACO-Part, where richer local context helps disambiguate subtle part-level differences.

(3)Foreground Consolidation. Tab. 3d evaluates the two FC components: semantic disagreement penalization (SDP) and semantic reweighting (SR). SDP consistently improves the baseline by suppressing unreliable activations caused by inter-map disagreement and foreground–background coupling, with the largest gain on PACO-Part (+1.9 % pts.). SR provides stronger improvements on ISIC and COCO by correcting region-level semantic bias through cluster-wise reweighting. Combining SDP and SR achieves the best performance on all benchmarks, confirming their complementarity.

Table 3: Ablation studies of different components in FoRIS.

(a) Comparison between original and adaptive positional debiasing on X-Ray, COCO, and PACO.

Method	X-Ray ↑	COCO ↑	PACO ↑
ORD	84.9	60.9	42.3
APD	**87.6**	60.9	42.3

(b) Ablation study of densification in the FL module on COCO, X-Ray, and PACO.

Method	COCO ↑	X-Ray ↑	PACO ↑
w/o Dense	60.1	87.4	41.8
w/ Dense	**60.9**	**87.6**	**42.3**

(c) Ablation study of positional and RGB cues in the FL module on X-Ray and SUIM.

Method	X-Ray ↑	SUIM ↑
None	86.7	59.0
Pos	**87.8**	58.8
RGB	86.8	**59.2**
Pos + RGB	87.6	59.1

(d) Ablation study of consolidation components in the FC module on ISIC, COCO, and PACO.

Method	ISIC ↑	COCO ↑	PACO ↑
None	60.2	58.5	39.8
SDP	61.1	59.9	41.7
SR	62.6	60.2	40.9
SDP + SR	**62.9**	**60.9**	**42.3**

(4) Additional Ablation and Analysis. We additionally investigate the effect of positional debiasing on semantic correspondence using SPair-71k. The results are provided in Appendix C. The computational cost is analyzed in Appendix D, the effect of CRF refinement is evaluated in Appendix E, the 5-shot results are reported in Appendix F, and the additional results on FSS-1000 dataset Li et al. (2020) are presented in Appendix G.

5 Challenges and the future of VFM-based Training-Free ICS

When extending ICS to the slender-structure scenarios illustrated in Fig. 6, Tab. 4 shows that existing SAM- and DINO-based ICS methods achieve consistently low IoU on Fundus Jin et al. (2022) and DeepGlobe-18 Demir et al. (2018) , revealing their limitations in slender-structure segmentation. This raises two questions: why are slender structures particularly challenging for VFM-based training-free ICS, and how can segmentation performance be improved in such scenarios? Further analysis and experiments are provided in Appendix H.



Figure 6: Visualization of the Fundus and DeepGlobe-18 datasets.

Table 4: Comparison of results between SAM-based and DINO-based methods.

Method	Fundus		DeepGlobe 18	
	1-shot	5-shot	1-shot	5-shot
GF-SAM	8.6	8.4	8.5	9.3
INSID3	12.3	11.7	7.2	7.3
FoRIS	**19.7**	**21.5**	**13.3**	**13.5**

6 Conclusion

In this work, inspired by the coarse-to-fine paradigm in classical image processing, we rethink the fundamental formulation of ICS and propose a foreground progressive ICS framework, **FoRIS**. Rather than treating ICS as a direct matching problem between reference and target images, we reformulate it as a *progressive foreground refinement* process, which progressively suppresses background interference, localizes target regions, and recovers coherent object structures through three stages: foreground purification, foreground localization, and foreground consolidation. Extensive experiments demonstrate that FoRIS achieves competitive performance across semantic segmentation, part segmentation, and cross-domain benchmarks, while exhibiting strong out-of-distribution generalization. Meanwhile, we reveal the limitations of existing training-free ICS methods on *slender and fragmented structures*: weak semantic responses, limited spatial support, and substantial appearance variations make direct feature matching inadequate for preserving structural continuity. How to fully exploit VFM representations to handle such weakly salient, fine-grained, and topology-sensitive targets remains an important open problem for training-free ICS. We hope this observation will encourage future ICS research to move beyond general object segmentation toward more robust and structure-aware open-world segmentation.

AI USE STATEMENT

In this work, we used generative AI tools solely for language polishing, including improving grammar, phrasing, clarity, and readability. We did not use generative AI tools for research ideation, methodological or experimental design, data analysis, code generation, or the generation or interpretation of scientific claims or results. All AI-assisted text was carefully reviewed and revised by the authors. We take responsibility for the final content of this work, including all text, claims, and results.

REFERENCES

Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

Rolf Adams and Leanne Bischof. Seeded region growing. *IEEE Transactions on pattern analysis and machine intelligence*, 16(6):641–647, 1994.

Edward H Adelson, Charles H Anderson, James R Bergen, Peter J Burt, and Joan M Ogden. Pyramid methods in image processing. *RCA engineer*, 29(6):33–41, 1984.

Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010.

Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

Yuri Y Boykov and M-P Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, volume 1, pp. 105–112. IEEE, 2001.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, pp. 1877–1901, 2020.

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 Trans. Med. 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.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pp. 9650–9660, 2021.

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 Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2017.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, et al. PaLM: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24 (240):1–113, 2023.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv:1902.03368 [cs.CV]*, 2019.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21638–21648, 2026.

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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 172–17209. IEEE, 2018.

Miguel Espinosa, Chenhongyi Yang, Linus Ericsson, Steven McDonagh, and Elliot J Crowley. No time to train! training-free reference-based instance segmentation. *arXiv preprint arXiv:2507.02798*, 2025.

Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004.

Oliver Hahn, Christoph Reich, Nikita Araslanov, Daniel Cremers, Christian Rupprecht, and Stefan Roth. Scene-centric unsupervised panoptic segmentation. In *CVPR*, pp. 24485–24495, 2025.

Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T. Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In *ICLR*, 2022.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pp. 2961–2969, 2017.

Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4D convolutional Swin transformer for few-shot segmentation. In *ECCV*, pp. 108–126, 2022.

Ming Hu, Jianfu Yin, Jing Wang, Yuqi Wang, Bingliang Hu, and Quan Wang. Specslice-convlstm: Medical hyperspectral image segmentation using spectral slicing and convlstm. In *International Conference on Pattern Recognition*, pp. 211–225. Springer, 2024.

Ming Hu, Jianfu Yin, Zhuangzhuang Ma, Jianheng Ma, Feiyu Zhu, Bingbing Wu, Ya Wen, Meng Wu, Cong Hu, Bingliang Hu, et al. beta-fft: Nonlinear interpolation and differentiated training strategies for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30839–30849, 2025.

Ming Hu, Mingyu Dou, Jianfu Yin, Miaomiao Zhang, Cong Hu, Yao Wang, Bingliang Hu, and Quan Wang. Whereedit: Mask-aware local latent editing for one-step image editing. *arXiv preprint arXiv:2607.20883*, 2026a.

Ming Hu, Yongsheng Huo, Mingyu Dou, Jianfu Yin, Peng Zhao, Yao Wang, Cong Hu, Bingliang Hu, and Quan Wang. Fb-clip: Fine-grained zero-shot anomaly detection with foreground-background disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35659–35669, 2026b.

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*, pp. 1769–1776, 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. *IEEE Trans. Med. Imaging*, 33(2):233–245, 2013.

Kai Jin, Xingru Huang, Jingxing Zhou, Yunxiang Li, Yan Yan, Yibao Sun, Qianni Zhang, Yaqi Wang, and Juan Ye. Fives: A fundus image dataset for artificial intelligence based vessel segmentation. *Scientific data*, 9(1):475, 2022.

Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment Anything. In *ICCV*, pp. 4015–4026, 2023.

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *NIPS*, pp. 109–117, 2011.

Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. Learning what not to segment: A new perspective on few-shot segmentation. In *CVPR*, pp. 8057–8067, 2022.

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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2866–2875. IEEE, 2020.

Yang Liu, Chenchen Jing, Hengtao Li, Muzhi Zhu, Hao Chen, Xinlong Wang, and Chunhua Shen. A simple image segmentation framework via in-context examples. In *NeurIPS*, pp. 25095–25119, 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.

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.

Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *CVPR*, pp. 8364–8375, 2022.

Lingchen Meng, Shiyi Lan, Hengduo Li, Jose M. Alvarez, Zuxuan Wu, and Yu-Gang Jiang. SegIC: Unleashing the emergent correspondence for in-context segmentation. In *ECCV*, volume 38, pp. 203–220, 2024.

Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *ICCV*, pp. 622–631, 2019.

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. *Trans. Mach. Learn. Res.*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, pp. 27730–27744, 2023.

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

Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

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:2508.10104 [cs.CV]*, 2025.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv:2302.13971 [cs.CV]*, 2023.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data*, 5(1):1–9, 2018.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *ICCV*, pp. 10052–10062, 2021.

Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. PaNet: Few-shot image semantic segmentation with prototype alignment. In *ICCV*, pp. 9197–9206, 2019.

Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *CVPR*, pp. 6830–6839, 2023a.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. SegGPT: Towards segmenting everything in context. In *ICCV*, pp. 1130–1140, 2023b.

Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7236–7246, 2023.

Lihe Yang, Zhen Zhao, and Hengshuang Zhao. Unimatch v2: Pushing the limit of semi-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4): 3031–3048, 2025.

Xiwen Yao, Qinglong Cao, Xiaoxu Feng, Gong Cheng, and Junwei Han. Scale-aware detailed matching for few-shot aerial image semantic segmentation. *IEEE Trans. Geosci. Remote. Sens.*, 60:1–11, 2021.

Anqi Zhang, Guangyu Gao, Jianbo Jiao, Chi Harold Liu, and Yunchao Wei. Bridge the points: Graph-based few-shot segment anything semantically. In *NeurIPS*, pp. 33232–33261, 2024a.

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. In *NeurIPS*, pp. 42672–42695, 2024.

APPENDIX

A Experiments setting **15**

B Algorithm Summary **15**

C Semantic Correspondence Analysis of Positional Debiasing **15**

D Computational Cost of FoRIS **16**

E Effect of CRF Refinement **17**

F n-shot segmentation **17**

G Additional Results on FSS-1000 **18**

H Limitations of VFM-based Training-Free ICS for Slender Structure Segmentation **19**

H.1 Experimental Setup 19

H.2 Quantitative Comparison on Slender Structures 20

H.3 Layer-wise and Resolution Sensitivity 20

H.4 Feature Representation Analysis 20

A EXPERIMENTS SETTING

We evaluate FoRIS on one-shot *semantic* and *part* segmentation. In each setting, a single annotated reference mask is provided, and the model is tasked with segmenting the corresponding concept in the target image: *(1) semantic* – all instances of a given class (e.g., "dog"); *(2) part* – same object part (e.g., "dog ear").

For **one-shot *semantic* segmentation**, we use six datasets across a range of imaging scenarios: COCO-20^i Nguyen & Todorovic (2019) with 80 object categories; LVIS-92^i Liu et al. (2024b) with 920 categories and a strong long-tail distribution; ISIC2018 Codella et al. (2019); Tschandl et al. (2018) for skin lesion segmentation; Chest X-Ray Candemir et al. (2013); Jaeger et al. (2013), an X-ray dataset of lung screening; iSAID-5^i Yao et al. (2021), a remote sensing dataset with 15 categories; and SUIM Islam et al. (2020) with underwater imagery and 8 categories. For **one-shot *part* segmentation**, we use PASCAL-Part Liu et al. (2024b), providing 56 object parts across 15 categories, and PACO-Part Liu et al. (2024b) with 303 object parts from 75 categories.

Implementation details. We adopt the Large version of the DINOv3 Siméoni et al. (2025) encoder. Input images are resized to 1024×1024, following SAM-based approaches Liu et al. (2024b); Zhang et al. (2024b;a). The final segmentation masks are predicted at patch resolution: we bilinearly interpolate them to original resolution, and apply mask refinement with a CRF Krähenbühl & Koltun (2011), following Hamilton et al. (2022); Hahn et al. (2025); Van Gansbeke et al. (2021); Melas-Kyriazi et al. (2022).

B ALGORITHM SUMMARY

Algorithm 1 FoRIS inference (single target episode)

Require: Reference images $\{I_s, M_s\}_{s=1}^{S}$, target image I_t, frozen encoder ϕ
Ensure: Binary mask $\hat{M}_t$

$\mathbf{F} \leftarrow \phi([I_1, \ldots, I_S, I_t])$; $\mathbf{f}(\mathbf{p}) \leftarrow \mathbf{F}(\mathbf{p})/\|\mathbf{F}(\mathbf{p})\|_2$
// FP: adaptive positional debiasing
Compute s_{sem}; if $s_{\text{sem}} < \theta$, replace $\mathbf{f}$ with $\tilde{\mathbf{f}}$ via $\mathbf{P}_\perp$ (eq. (3))
// FP: Stage-1 gating
Cluster reference foreground tokens (agglomerative) $\rightarrow \{\boldsymbol{\mu}_j\}_{j=1}^{J}$; build $\boldsymbol{\mu}^{\text{fg}}$, $\boldsymbol{\mu}^{\text{bg}}$ (TopK 20% hard negatives)
Compute gate $g(\mathbf{p})$ and gated features $\hat{\mathbf{f}}(\mathbf{p})$ on all tokens (eq. (5))
// FP: Stage-2 matching (Score 1)
Orthogonalize $\boldsymbol{\mu}_\perp^{\text{bg}}$ w.r.t. $\boldsymbol{\mu}^{\text{fg}}$ (eq. (6))
$S^{(1)} \leftarrow \beta \log \sum_j \exp(\langle \tilde{\mathbf{f}}_t, \boldsymbol{\mu}_j \rangle / \beta) - \langle \tilde{\mathbf{f}}_t, \boldsymbol{\mu}_\perp^{\text{bg}} \rangle$ on *ungated* $\tilde{\mathbf{f}}_t$ (eq. (7))
$S_{\text{fg}}, S_{\text{bg}} \leftarrow$ per-episode min–max normalization
// FL (Score 2)
$V \leftarrow$ cross-image candidate votes; densify $\mathcal{C}$
Cluster target patches on $[\mathbf{f}; \mathbf{c}; \mathbf{p}]$ (eq. (9)); build seed prior P from $j^\star$ and $\{\rho_j\}$
$S^{(2)} \leftarrow S^{(1)} + (V - \frac{1}{2}) + (P - \frac{1}{2})$ (eq. (13))
// FC (Score 3)
$\Pi \leftarrow$ SDP from $S_{\text{fg}}, S_{\text{bg}}, V, P$ (eq. (15))
Re-cluster gated $\hat{\mathbf{f}}_t$; compute $\Delta(\mathbf{p})$ from cluster means
$S^{(3)} \leftarrow S^{(2)} - \Pi + \Delta$ (eq. (17))
$\hat{M}_t \leftarrow$ binarize, upsample, CRF-refine $S^{(3)}$
return $\hat{M}_t$

C SEMANTIC CORRESPONDENCE ANALYSIS OF POSITIONAL DEBIASING

To further investigate the effect of positional debiasing on semantic correspondence, we evaluate the original, fixed-debiasing, and adaptive-debiasing variants on SPair-71k. The results are reported in table 5.

Table 5: **Semantic correspondence on SPair-71k** (PCK@T in %, ↑). Comparison across DINOv3 backbones with original, debiased, and adaptive debias variants.

	Small			**Base**			**Large**		
T	*original*	*debias*	***adaptive***	*original*	*debias*	***adaptive***	*original*	*debias*	***adaptive***
0.05	27.93	28.18	**28.37**	30.07	33.68	**33.69**	33.55	34.57	**34.61**
0.10	44.92	45.65	**45.94**	46.80	52.60	**52.60**	52.04	54.11	**54.12**
0.15	54.31	55.75	**56.01**	55.60	62.49	**62.50**	61.63	64.24	**64.27**
0.20	60.55	62.78	**62.97**	61.22	68.72	**68.73**	67.80	70.76	**70.76**

Applying positional debiasing consistently improves semantic correspondence across different DINOv3 model scales and evaluation thresholds. For example, with the DINOv3-Base backbone, PCK@0.10 improves from 46.80 to 52.60, while PCK@0.20 increases from 61.22 to 68.73. Similar improvements are observed across the Small, Base, and Large backbones, indicating that removing position-dominated components can improve the reliability of dense semantic correspondence.

Adaptive debiasing further achieves the strongest performance across the evaluated settings. Although its improvement over fixed debiasing is relatively modest, the adaptive variant consistently avoids the performance degradation observed when positional information is suppressed indiscriminately. This suggests that positional information in frozen DINOv3 features is not purely harmful: useful spatial regularities can facilitate correspondence when reference and target images are already well aligned.

This observation is also consistent with the segmentation ablation in table 2. Applying debiasing unconditionally does not always improve downstream segmentation and can even degrade performance on some datasets, such as ISIC (46.8→41.2). In contrast, adaptive debiasing dynamically regulates the debiasing strength according to the reliability of semantic correspondence, suppressing position-dominated activations only when necessary.

Overall, the SPair-71k results provide complementary evidence for the design of APD: effective positional debiasing should selectively suppress harmful positional bias while retaining useful geometric priors encoded in frozen visual features.

D Computational Cost of FoRIS

We further analyze the computational cost of FoRIS on COCO-20^i from both the component-wise and resolution-wise perspectives. As shown in Table 6, the total inference time of FoRIS is 1 620.77 ms for a single in-context example at a resolution of 1 024 px on a single RTX 3090. The main computational costs come from the encoder forward pass and the subsequent foreground localization and consolidation stages, while foreground purification introduces only a relatively small overhead. Although FoRIS is slower than INSID3 Cuttano et al. (2026) at the same 1 024 px resolution, directly comparing runtime at a fixed resolution does not fully characterize the accuracy–efficiency trade-off between the two methods.

To further investigate this trade-off, we evaluate FoRIS under different input resolutions on COCO-20^i, as reported in Table 7. Interestingly, FoRIS achieves 59.5 % mIoU even at a resolution of only 512 px, already surpassing the 57.6 % mIoU achieved by INSID3 at 1 024 px. More importantly, the 512 px setting requires only 258 ms per in-context example, compared with 812 ms for INSID3 at 1 024 px. This corresponds to a 68.2 % reduction in inference time while simultaneously improving mIoU by 1.9 percentagepoints. Thus, although FoRIS incurs a higher computational cost when operating at the same resolution, it can achieve stronger segmentation performance with substantially lower-resolution inputs, yielding a more favorable practical accuracy–efficiency trade-off on COCO-20^i.

The resolution study further reveals a clear diminishing-return effect. Increasing the input resolution of FoRIS from 512 px to 1 024 px improves mIoU from 59.5 % to 60.9 %, yielding only a 1.4 percentage−point gain, whereas the inference time increases from 258 ms to 1 621 ms, corresponding to more than a 6× increase in computational cost. Further increasing the resolution provides little additional benefit: mIoU remains around 61 % even at 1 760 px, while the inference time

increases substantially to 11 235 ms. These results indicate that, on COCO-20^i, the segmentation performance of FoRIS is relatively insensitive to increasing input resolution, and that low-resolution inference can already preserve most of its segmentation capability.

Overall, the results on COCO-20^i demonstrate that the computational efficiency of FoRIS should be considered jointly with its achievable segmentation accuracy at different resolutions, rather than solely through runtime comparisons at a fixed resolution. In particular, the 512 px setting provides a compelling operating point, achieving higher mIoU than INSID3 at 1 024 px while requiring only a fraction of its inference time. This favorable accuracy–efficiency trade-off suggests that FoRIS can maintain strong in-context segmentation performance without relying on computationally expensive high-resolution inputs.

Table 6: **Detailed computational analysis of FoRIS on COCO-20^i.** We report the inference time of each component in milliseconds (ms, ↓) for a single in-context example. Runtime is measured at a resolution of 1 024 px on a single RTX 3090.

Component	**Runtime ↓**
Encoder forward	702.27 ms
Foreground purification	38.74 ms
Foreground localization	346.44 ms
Foreground consolidation	342.50 ms
CRF refinement	190.82 ms
Total inference time	**1 620.77 ms**
INSID3 Cuttano et al. (2026)	**812 ms**

Table 7: **Resolution-wise computational analysis of FoRIS on COCO-20^i.** We report the inference time and mIoU of FoRIS under different input resolutions. Runtime is measured for a single in-context example on a single RTX 3090. The 1 024 px setting corresponds to the resolution used for the component-wise analysis in Table 6.

	Low resolution ← **Resolution** → *High resolution*						
	512 px	**720 px**	**896 px**	**1 024 px**	**1 440 px**	**1 600 px**	**1 760 px**
Runtime	258 ms	593 ms	1 108 ms	1 621 ms	5 468 ms	9 018 ms	11 235 ms
mIoU (%)	59.5	60.3	60.7	60.9	60.8	60.1	61.0

E Effect of CRF Refinement

After clustering in the DINOv3 feature space, FoRIS produces a coarse binary mask that is upsampled to the original resolution. As an optional post-processing step, we apply a dense CRF to refine mask boundaries by encouraging label consistency among spatially and appearance-wise similar pixels. This refinement is lightweight and does not require any additional training.

Tab. 8 compares FoRIS with and without CRF on eight one-shot benchmarks. Enabling CRF improves the average mIoU from **57.8%** to **58.1%** (+0.3%), confirming that the core segmentation capability already resides in the frozen backbone and our debiasing module. The gain is consistent across benchmarks: CRF helps most on structurally complex scenes such as LVIS (+0.9%) and SUIM (+0.6%), where object boundaries are irregular, while the improvement is marginal on datasets with already sharp foreground regions, e.g., chest X-ray (+0.1%) and iSAID (unchanged). On PASCAL-Part (+0.4%) and PACO-Part (+0.3%), CRF slightly sharpens part boundaries without altering the overall ranking. Overall, CRF serves as a complementary refinement rather than a critical component: FoRIS w/o CRF already achieves strong performance, and the full pipeline benefits from cleaner contours at a negligible computational cost.

F N-shot segmentation

We further evaluate FoRIS in the 5-shot setting by providing five reference image–mask pairs for each target image. As shown in Table 9, FoRIS achieves the highest average mIoU of **64.7%**, out-

Table 8: **Effect of CRF refinement on FoRIS (mIoU in %, ↑).** We report one-shot results on eight benchmarks with and without the optional CRF post-processing step.

Method	Encoder	#Param	Semantic						Part		Avg
			LVIS-92	COCO-20	ISIC	SUIM	iSAID	X-Ray	PASCAL	PACO	
FoRIS (ours)	DINOv3	**304M**	**42.8**	**60.9**	**62.9**	**59.1**	**53.6**	**87.6**	**55.8**	**42.3**	**58.1**
FoRIS (w/o CRF)	DINOv3	304M	41.9	60.4	62.7	58.5	53.6	87.5	55.4	42.0	57.8

performing the strongest training-free baseline, INSID3, by **4.8 percentage points**. FoRIS ranks first on **7 out of 8 benchmarks** and consistently outperforms INSID3 across all datasets, with particularly notable gains on X-Ray (**+7.9 points**), PASCAL (**+6.2 points**), and PACO (**+5.4 points**). These improvements span diverse domains, including natural images, medical images, underwater scenes, and aerial imagery, demonstrating the robustness of FoRIS under substantial domain shifts. Notably, FoRIS uses the same DINOv3 encoder and parameter count as INSID3, indicating that the performance gains arise from more effective foreground purification, localization, and consolidation rather than increased model capacity. Moreover, all hyperparameters are directly reused from the 1-shot setting without any additional tuning, demonstrating that FoRIS scales effectively to multiple contextual examples.

Table 9: **Comparison of FoRIS (mIoU in %, ↑) on 5-shot semantic and part segmentation.** Models are provided with 5 contextual examples and tasked with segmenting the annotated concept in the target image. FoRIS scales effectively to multiple references, achieving robust performance across domains. All hyperparameters are reused from the 1-shot setting without any tuning, highlighting the versatility of our approach. Gray indicates the model was trained on the corresponding train split of the dataset; best results **bold**, 2nd best underlined.

Method	Encoder	Params	Semantic						Part		Avg
			LVIS-92$_i$	COCO-20$_i$	ISIC	SUIM	iSAID	X-Ray	PASCAL	PACO	
Task-specific fine-tuning: semantic + mask supervision											
SegGPT	ViT	354M	25.4	67.9	45.2	33.7	35.9	**89.1**	42.8	14.1	44.3
SINE	DINOv2	373M	35.5	66.1	28.6	54.8	40.5	40.6	36.4	25.4	41.0
DiffewS	Stable Diffusion	890M	35.4	72.2	32.7	49.8	48.0	45.1	39.7	26.1	43.6
Training-free: mask-supervised pre-training											
Matcher	DINOv2+SAM	945M	40.0	60.7	35.0	50.6	34.3	71.2	45.8	33.6	46.4
GF-SAM	DINOv2+SAM	945M	44.2	66.8	55.2	58.1	52.4	52.9	51.2	41.9	52.8
GF-SAM†	DINOv3+SAM	945M	42.8	64.4	56.7	58.6	53.2	54.7	50.3	39.3	52.5
GF-SAM† + debias	DINOv3+SAM	945M	43.6	64.6	58.2	59.2	54.1	59.1	51.4	40.2	53.8
Training-free: unsupervised pre-training											
INSID3	DINOv3	**304M**	47.2	65.1	63.9	61.7	56.9	80.1	57.1	46.8	59.9
FoRIS	DINOv3	**304M**	**51.0**	**68.4**	**66.5**	**66.9**	**61.7**	88.0	**63.3**	**52.2**	**64.7**

G Additional Results on FSS-1000

To further evaluate the generalization of FoRIS, we conduct additional experiments on the FSS-1000 dataset under both 1-shot and 5-shot settings. The results are summarized in Tab. 10.

FoRIS achieves an mIoU of 87.2% in the 1-shot setting and 89.2% in the 5-shot setting. Compared with INSID3, FoRIS improves the performance by 1.5 and 2.0 percentage points under the 1-shot and 5-shot settings, respectively. Moreover, FoRIS achieves a higher 5-shot performance than GF-

Table 10: Comparison on the FSS-1000 dataset under 1-shot and 5-shot settings. We report the mean IoU (%) following the same evaluation protocol. FoRIS consistently outperforms INSID3 and achieves competitive performance with GF-SAM.

Method	1-shot	5-shot
INSID3	85.7	87.2
GF-SAM	**88.0**	88.9
FoRIS	87.2	**89.2**

SAM, with an improvement of 0.3 percentage points, while remaining competitive in the 1-shot setting.

These results further demonstrate that FoRIS can effectively leverage a small number of in-context examples without task-specific training, and maintains consistent performance as the number of support examples increases.

H Limitations of VFM-based Training-Free ICS for Slender Structure Segmentation

Vision foundation models (VFMs) have emerged as powerful visual representation learners and have substantially advanced training-free in-context segmentation (ICS). By leveraging large-scale pre-trained knowledge, existing VFM-based ICS methods perform segmentation without task-specific optimization, typically by measuring feature similarity between support and query images, constructing semantic prototypes, or exploiting correspondence encoded in foundation-model representations. These approaches achieve strong results on objects with distinct semantics and relatively compact spatial extent.

However, the suitability of such training-free paradigms for slender-structure segmentation remains uncertain. Unlike common objects with large, coherent regions, slender structures—such as blood vessels, nerves, and tubular anatomical patterns—exhibit properties that conflict with the assumptions underlying current VFM-based ICS methods. Segmenting them requires not only semantic recognition but also precise localization, boundary preservation, and structural continuity, none of which are explicitly optimized in standard VFM representations.

Patch-level foreground–background mixing. Most foundation models, especially transformer-based architectures, represent images with spatial tokens derived from local patches. For compact objects, individual tokens usually contain sufficient foreground signal to capture object semantics. For slender structures, however, the foreground may occupy only a small fraction of a patch, causing the resulting representation to be dominated by surrounding background. The extracted feature therefore mixes foreground and background information rather than encoding a pure object descriptor. This weakens the separability between slender structures and their context and makes reliable training-free matching difficult.

Dispersed intra-class feature distributions. Training-free ICS methods often assume that instances of the same category form a consistent and relatively compact cluster in feature space. Prototype-based approaches, for instance, summarize foreground regions into representative embeddings and label query regions according to their similarity to these prototypes. Slender structures violate this assumption: scale, orientation, curvature, illumination, and local appearance can vary substantially along a single structure, while different segments may exhibit considerably different visual characteristics. Consequently, the foreground distribution in feature space becomes dispersed rather than compact, and a single prototype may fail to cover the full appearance and topology of an elongated target.

H.1 Experimental Setup

To systematically examine this limitation, we conduct additional experiments on two representative slender-structure segmentation datasets, Fundus and DeepGlobe-18. Following the standard training-free ICS setting, all methods use frozen DINOv3-L features without task-specific training. Unless otherwise specified, the input resolution is set to 1024×1024, and the final transformer layer (Layer 23) is used as the default feature representation. We evaluate both 1-shot and 5-shot settings using mean Intersection-over-Union (mIoU).

We compare FoRIS with representative training-free ICS baselines, including correspondence-based and prototype-based methods. For a fair comparison, all methods operate on the same frozen DINOv3-L backbone and use the same support-query setting.

Table 11: Comparison of training-free ICS methods on slender-structure datasets (mIoU, %). All methods use frozen DINOv3-L at 1024×1024 with default last-layer features.

Method	Fundus		DeepGlobe-18	
	1-shot	5-shot	1-shot	5-shot
GF-SAM	8.6	8.4	8.5	9.3
INSID3	12.3	11.7	7.2	7.3
FoRIS	19.7	21.5	13.3	13.5

H.2 Quantitative Comparison on Slender Structures

The quantitative results are presented in Tab. 11. All methods exhibit a substantial performance degradation on slender structures compared with their performance on conventional object-centric benchmarks. In particular, FoRIS achieves only 19.7% and 21.5% mIoU on Fundus in the 1-shot and 5-shot settings, respectively, and 13.3% and 13.5% mIoU on DeepGlobe-18. These results indicate that the effectiveness of current training-free VFM-based ICS methods does not readily transfer to targets with extremely thin and elongated structures.

Although FoRIS consistently outperforms the compared training-free baselines, the absolute performance remains considerably lower than that observed on datasets containing larger and more compact foreground objects. This gap suggests that the primary bottleneck is not simply the absence of a stronger matching mechanism, but the suitability of the frozen feature representation itself for thin structures.

H.3 Layer-wise and Resolution Sensitivity

To investigate whether the poor performance mainly originates from the choice of feature layer and spatial resolution, we conduct a comprehensive layer-wise analysis. We evaluate DINOv3 features from different transformer layers under multiple input resolutions. The detailed numerical results are reported in Tabs. 12 and 13.

The results reveal a strong dependence on both feature depth and spatial resolution. For Fundus, the default Layer 23 at 1024×1024 produces only 19.7% mIoU in the 1-shot setting. However, changing the feature layer and increasing the input resolution substantially improves the performance. The best 1-shot result reaches 41.7% mIoU using Layer 20 at 2048×2048, while the best 5-shot result reaches 44.9% mIoU using Layer 13 at the same resolution. Similarly, on DeepGlobe-18, the best performance is obtained from an intermediate layer rather than the final layer. The best 1-shot and 5-shot results reach 23.9% and 26.3% mIoU, respectively, using Layer 19 with 2048×2048 inputs.

These observations indicate that the final-layer representation is not necessarily optimal for slender structures. Intermediate features can preserve finer spatial and local appearance information that becomes less explicit in deeper semantic representations. At the same time, increasing the input resolution provides more spatial tokens across thin structures, reducing the degree of foreground–background mixing within individual patches.

The layer-wise curves in Figs. 7 and 8 further show that the performance does not monotonically increase with feature depth. Instead, intermediate layers often provide a better balance between semantic abstraction and spatial detail. This behavior is particularly important for slender structures, where local continuity and boundary information are essential for accurate segmentation.

H.4 Feature Representation Analysis

We further analyze the spatial feature representations across different DINOv3 layers. For each layer group, we visualize the corresponding feature activations on DeepGlobe-18 and Fundus. The results are shown in Figs. 9–11 and Figs. 12–14, respectively.

The visualization provides a qualitative explanation for the quantitative layer-wise results. Early and intermediate layers tend to preserve more localized spatial variations and fine-grained appearance cues. As the network becomes deeper, the representations become increasingly semantic and invari-

Table 12: Layer-wise mIoU (%) on Fundus with FoRIS under 1-shot and 5-shot settings and input resolutions of 512, 1024, and 2048. Bold indicates the best result per column.

Layer	1-shot			5-shot		
	512	**1024**	**2048**	**512**	**1024**	**2048**
0	9.4	9.3	8.9	9.2	9.4	9.5
1	9.9	9.6	9.5	9.0	9.0	9.2
2	9.5	9.6	9.4	9.8	9.4	9.1
3	9.6	9.8	9.5	10.2	9.6	9.3
4	10.2	10.1	9.8	11.1	10.5	9.6
5	10.5	10.1	9.7	11.2	10.5	9.8
6	11.1	12.1	10.8	12.7	14.1	12.6
7	10.1	10.7	10.5	12.8	15.3	13.7
8	12.4	13.2	13.1	14.3	17.6	18.5
9	14.0	15.8	15.3	17.0	22.0	23.8
10	16.3	21.9	22.0	19.8	28.2	33.3
11	18.0	25.4	26.5	21.5	32.2	38.1
12	19.0	28.9	31.9	21.7	33.3	42.0
13	20.6	31.9	37.0	23.1	**35.3**	**44.9**
14	19.9	30.2	38.0	21.7	33.4	43.4
15	20.7	30.4	39.0	21.2	32.3	42.8
16	21.5	31.3	40.0	22.0	32.7	42.0
17	20.9	30.6	38.0	21.7	32.9	41.3
18	21.5	31.7	39.3	22.8	33.6	39.6
19	**21.9**	**33.4**	40.9	23.8	34.6	40.6
20	**21.9**	32.5	**41.7**	24.5	34.8	43.8
21	**21.9**	31.4	40.1	**24.6**	33.4	42.3
22	19.8	29.4	35.7	22.8	30.8	38.6
23	15.1	19.7	20.5	14.6	21.5	23.9

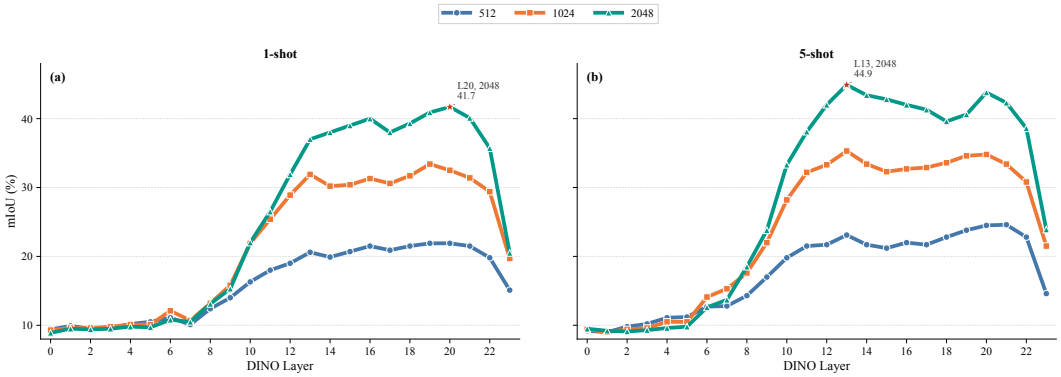


Figure 7: **Layer-wise mIoU on Fundus.** Curves summarize table 12. Mid-to-deep layers (8–21) benefit strongly from higher input resolution, whereas the last layer (23) degrades sharply, confirming that the default feature depth is misaligned with slender-structure matching.

Table 13: Layer-wise mIoU (%) on DeepGlobe-18 with FoRIS under 1-shot and 5-shot settings and input resolutions of 512, 1024, and 2048. Bold indicates the best result per column.

Layer	1-shot			5-shot		
	512	1024	2048	512	1024	2048
0	6.8	7.5	9.1	6.7	7.7	9.6
1	6.6	7.0	8.1	6.4	6.9	8.4
2	6.7	7.4	8.6	6.7	7.5	9.7
3	6.6	7.3	8.3	6.6	7.7	9.9
4	6.4	6.8	7.8	6.4	7.0	8.1
5	6.2	7.0	9.0	6.2	7.4	10.5
6	6.2	6.9	8.5	6.3	7.4	10.1
7	6.1	7.2	8.9	6.4	7.6	10.2
8	6.5	7.6	9.4	6.9	8.0	11.0
9	7.0	8.6	10.9	7.4	9.3	12.6
10	7.5	9.6	12.5	7.9	10.4	14.8
11	7.9	10.9	14.9	8.5	11.8	17.7
12	8.2	11.4	16.2	8.7	12.4	18.8
13	8.4	12.5	17.6	9.0	13.1	19.2
14	8.6	12.7	18.9	9.1	13.6	21.6
15	8.7	13.3	20.7	9.2	14.5	23.5
16	8.7	13.2	20.4	9.3	14.4	22.6
17	8.7	13.1	19.8	9.2	13.9	22.2
18	9.5	15.3	23.2	10.3	16.8	26.1
19	**10.0**	**16.7**	**23.9**	**11.3**	**18.2**	**26.3**
20	9.6	15.5	22.5	10.5	17.2	25.2
21	9.0	14.7	20.8	10.4	16.5	23.8
22	9.0	15.0	21.6	10.2	17.0	24.5
23	8.9	13.3	18.8	7.9	13.5	20.4

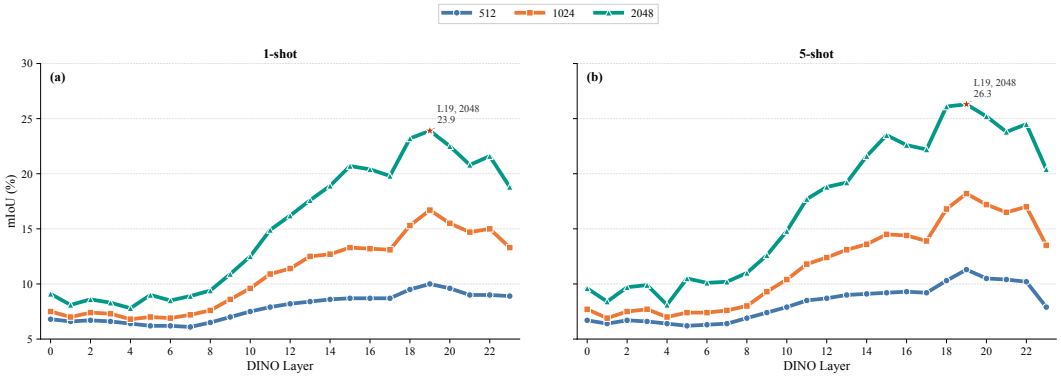


Figure 8: **Layer-wise mIoU on DeepGlobe-18.** Curves summarize table 13. The pattern mirrors Fundus: mid-to-deep layers gain from higher resolution, while the last layer (23) underperforms, indicating that default last-layer features are poorly suited to elongated targets.

ant, which is beneficial for recognizing compact semantic objects but can suppress the subtle local differences required to distinguish thin structures from their surrounding background.

For vascular structures in Fundus, for example, a large proportion of the target occupies only a small number of pixels within each local region. Consequently, even small deviations in feature localization can cause the predicted mask to become disconnected or substantially shrink. Similar behavior can be observed on DeepGlobe-18, where elongated road structures are fragmented when the feature representation does not sufficiently preserve local spatial continuity.

Overall, these experiments suggest that the limitation of training-free VFM-based ICS on slender structures is closely related to the mismatch between the spatial granularity of frozen representations and the geometric characteristics of the target. Although higher resolution and intermediate feature layers can partially alleviate this issue, they do not fundamentally resolve the problem.

Importantly, this observation should not be interpreted as evidence that training-free ICS is inherently unsuitable for slender structures. Rather, it highlights a limitation of current frozen VFM representations and the corresponding feature-matching paradigm. Addressing this challenge may require representations that explicitly preserve thin-structure topology, connectivity, and boundary information, potentially through structure-aware features, topology-preserving refinement, or graph-based reasoning. We leave such extensions for future work.

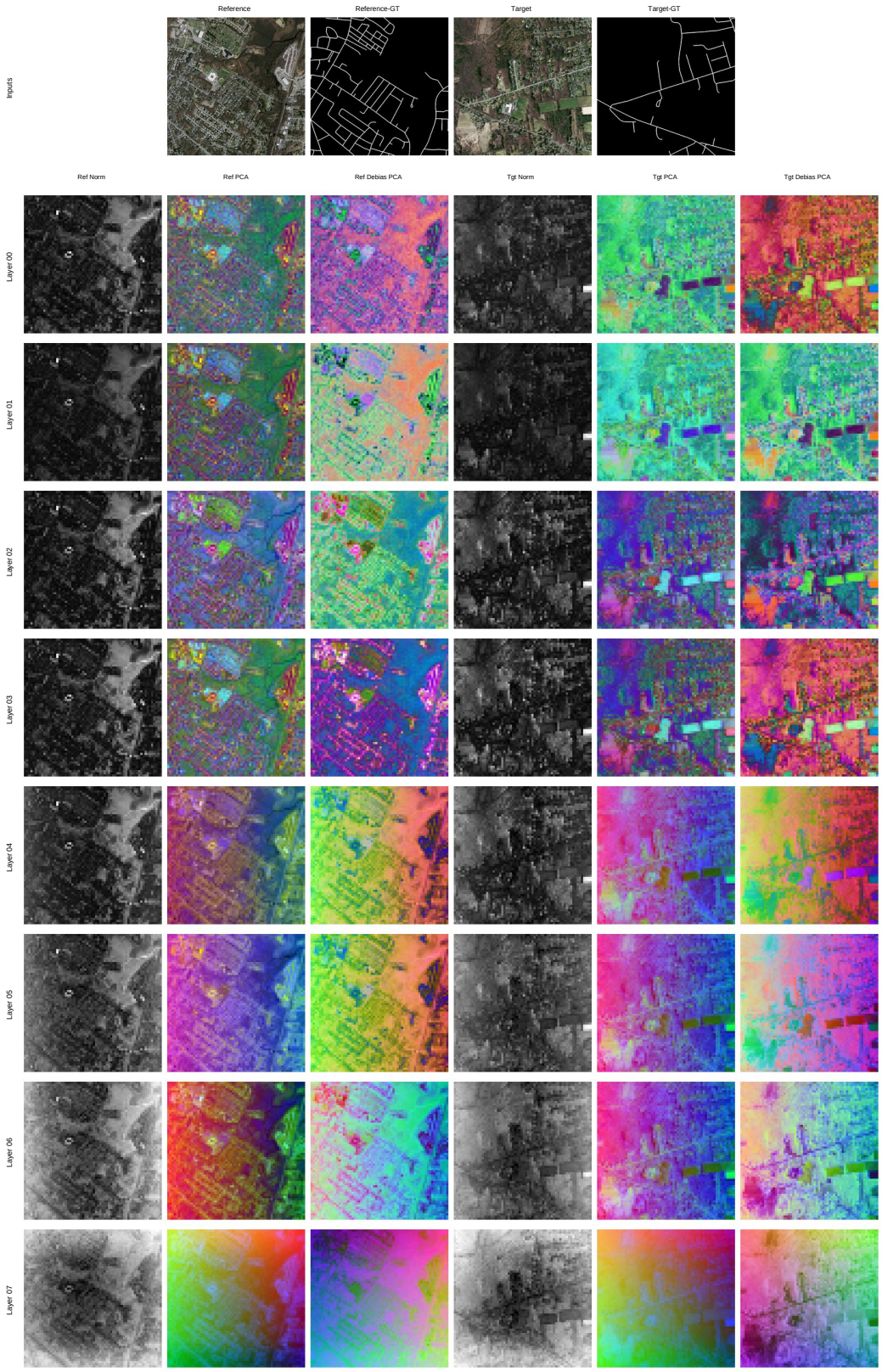


Figure 9: **Layer-wise DINOv3 features on DeepGlobe-18 (layers 0–7).** Top row: reference image, reference mask, target image, and ground truth. Each subsequent row shows, for reference *(left)* and target *(right)*, the feature norm (**Norm**), PCA colorization (**PCA**), and debiased PCA (**Debias PCA**).

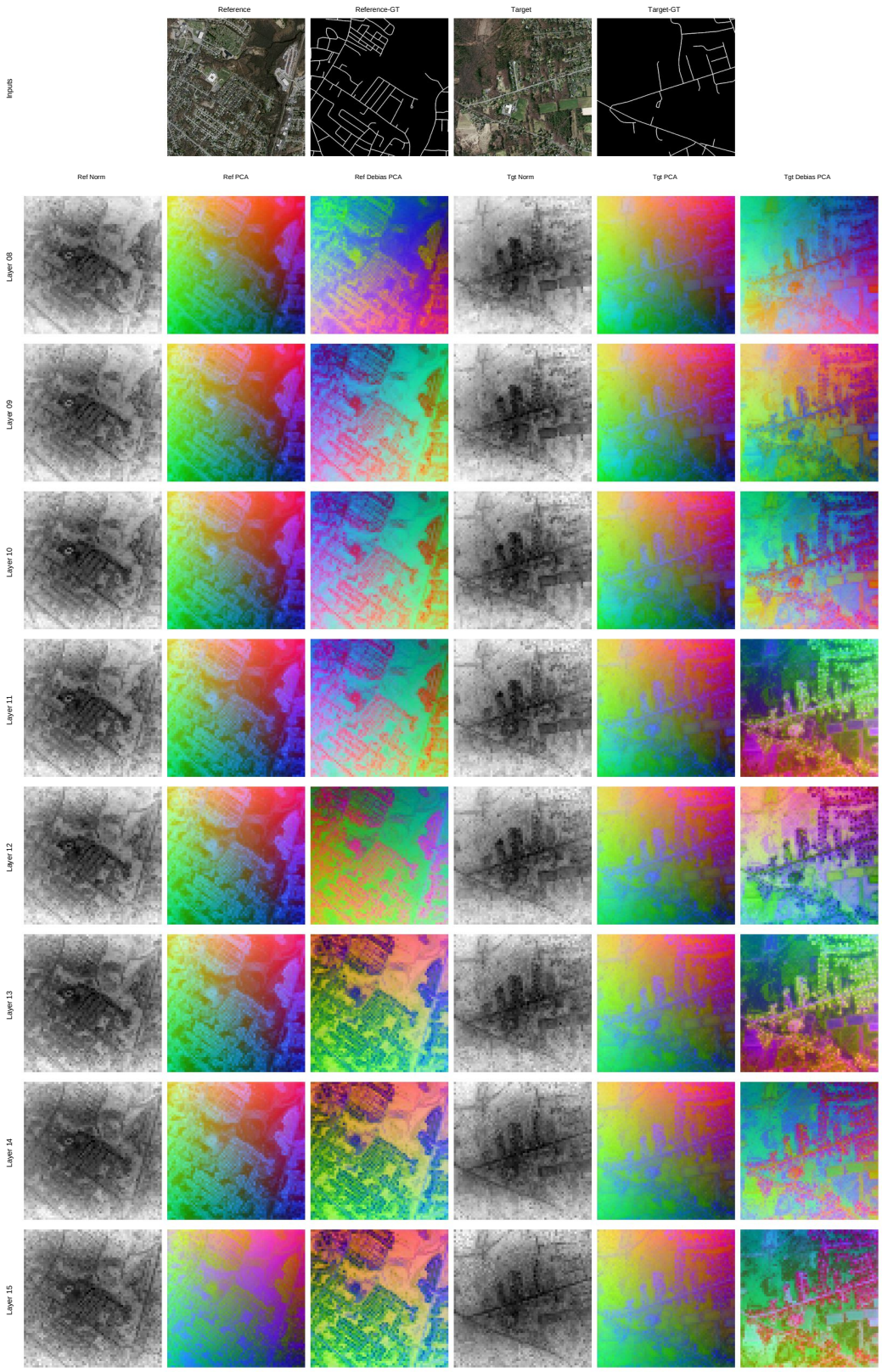


Figure 10: **Layer-wise DINOv3 features on DeepGlobe-18 (layers 8–15).** Same layout as fig. 9.

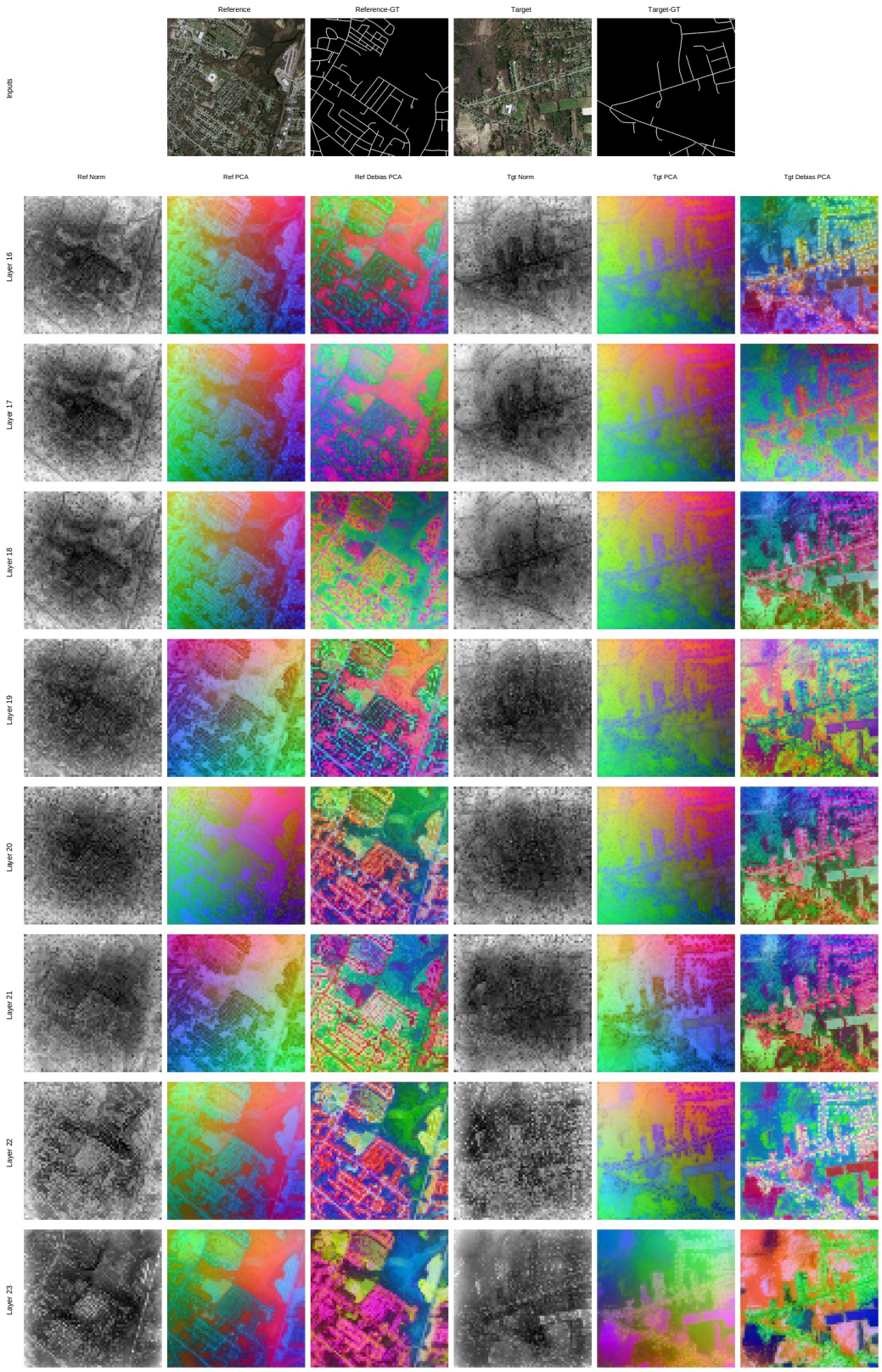


Figure 11: **Layer-wise DINOv3 features on DeepGlobe-18 (layers 16–23).** Same layout as fig. 9.

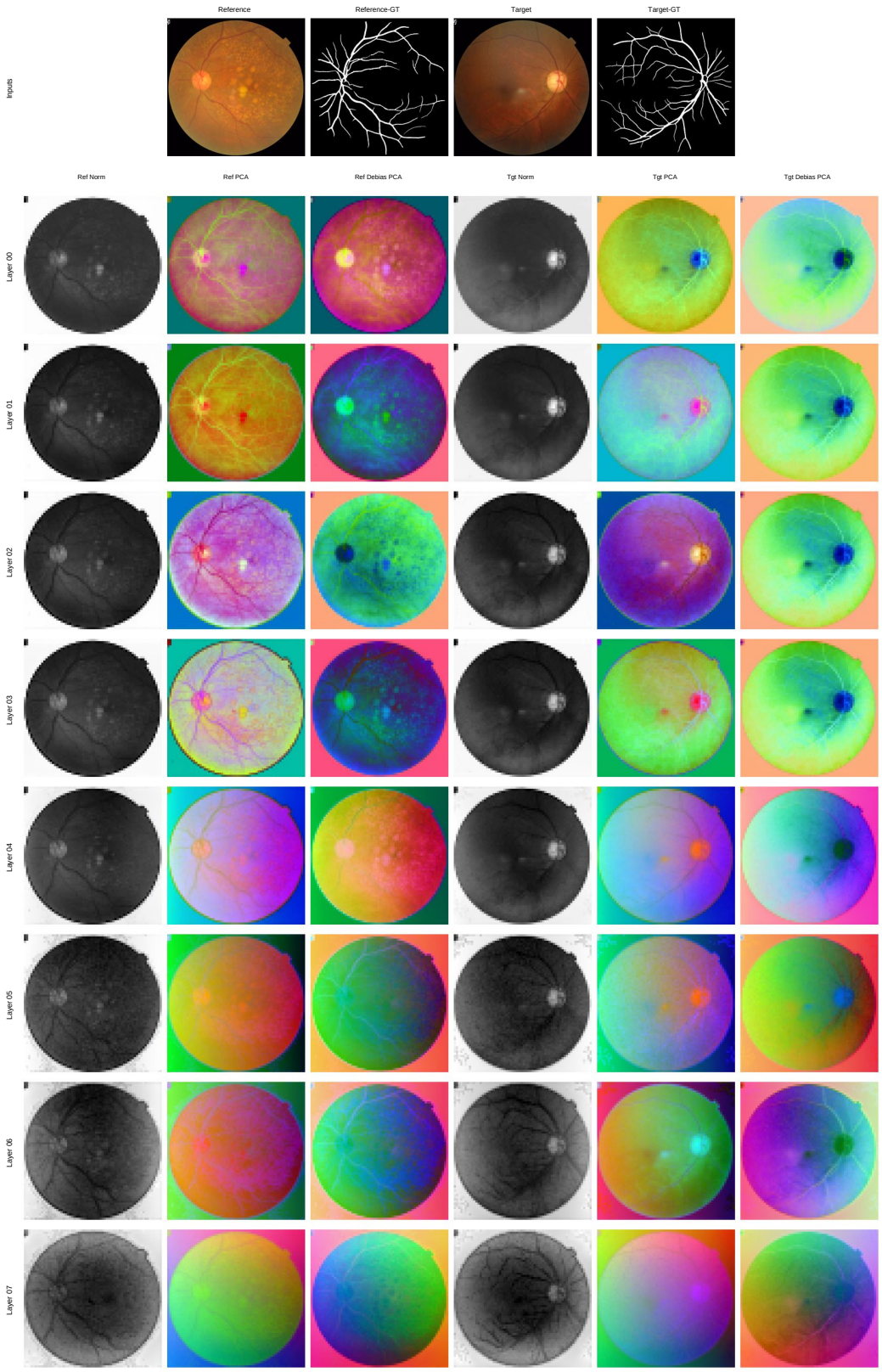


Figure 12: **Layer-wise DINOv3 features on Fundus (layers 0–7).** Same layout as fig. 9.

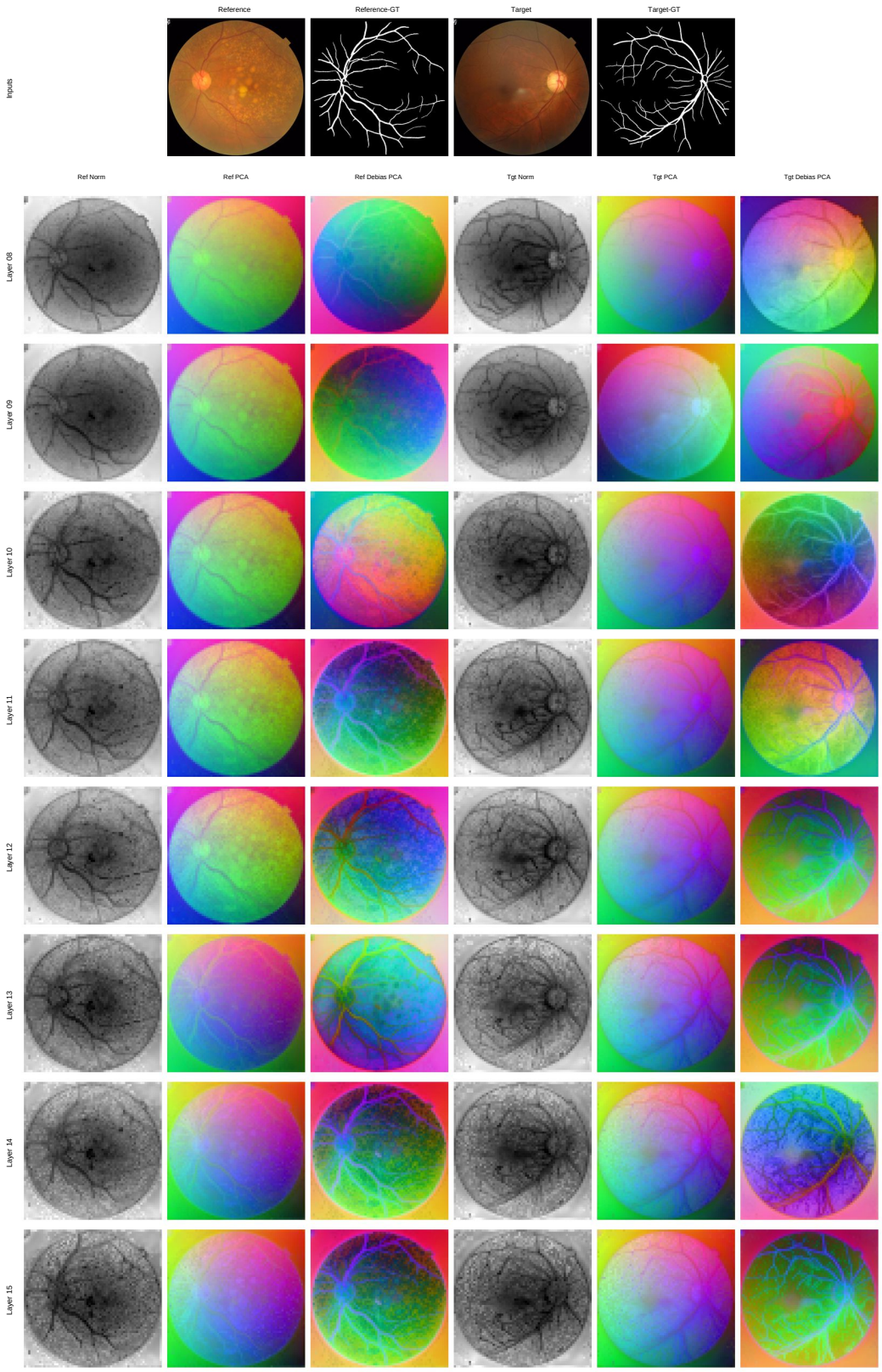


Figure 13: **Layer-wise DINOv3 features on Fundus (layers 8–15).** Same layout as fig. 9.

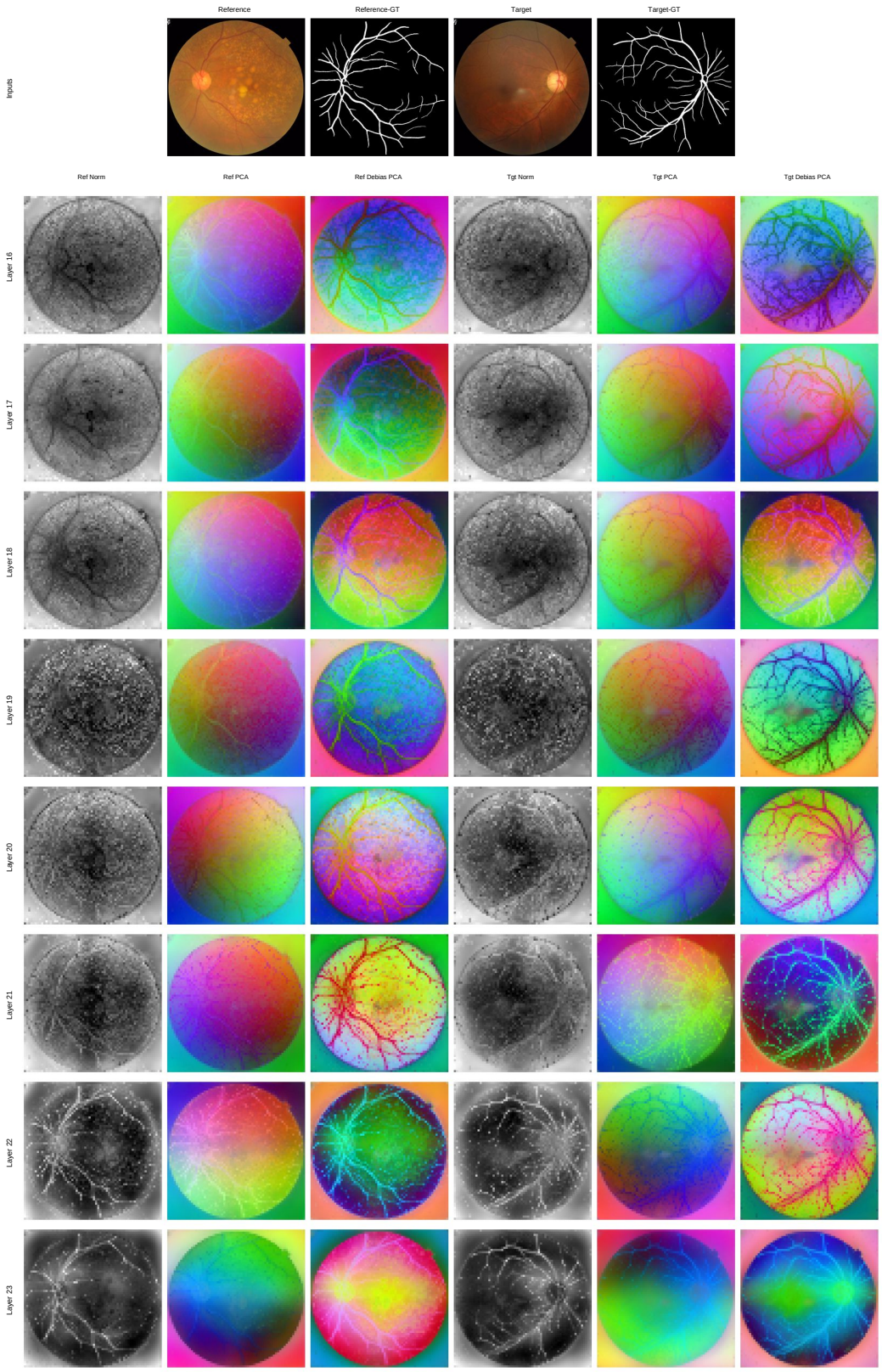


Figure 14: **Layer-wise DINOv3 features on Fundus (layers 16–23).** Same layout as fig. 9.