FROM PRIOR TO PRECISION: UNIFYING GLOBAL STRUCTURAL PRIORS AND LOCAL GENERATIVE REFINEMENT FOR MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical image segmentation is acutely bottlenecked by the scarcity of pixel-level annotations, especially in the low-data regime. Vision foundation models (VFMs) such as DINO provide rich semantic and structural priors that can alleviate this dependency, yet cross-domain discrepancies erode their direct transfer to the fine-grained representations required by medical segmentation. In this work, we propose **P**rior-t**o**-**P**recision **Net**work (PoPNet), a unified framework that bridges generic foundation-model priors with task-specific segmentation precision through global structural regularization and local generative refinement. To capture global structures, PoPNet uses a spectral clustering regularization to anchor the task-specific latent manifold toward structural priors derived from VFM cross-layer attention affinities, encouraging the learned representation to preserve dominant structural patterns while reducing reliance on domain-specific feature details. To further improve local precision, PoPNet incorporates target-aware region-of-interest masked generation, which performs conditional reconstruction within target regions, compelling the network to capture fine-grained textures and subtle boundary details critical for accurate delineation. Extensive experiments across five supervised benchmarks and semi-supervised segmentation settings demonstrate that PoPNet achieves state-of-the-art performance with strong generalization across datasets and foundation-model priors.

1 INTRODUCTION

Medical image segmentation plays an indispensable role in disease diagnosis (Hu et al., 2021; Addimulam et al., 2020), surgical planning Conze et al. (2023); Kordon et al. (2019); Minnema et al. (2022), and lesion localization (Hesamian et al., 2019; Hatamizadeh et al., 2019). In recent years, deep learning methods (Ronneberger et al., 2015; Ibtehaz & Rahman, 2020; Ibtehaz & Kihara, 2023) have achieved remarkable performance on these tasks, yet their success heavily depends on large-scale pixel-level annotations. Acquiring such annotations is expensive, labor-intensive, and requires substantial domain expertise, making data scarcity a persistent bottleneck in medical segmentation. Under limited supervision, segmentation models are prone to overfitting dataset-specific appearance patterns rather than learning robust structural representations, leading to degraded generalization, particularly for complex pathological structures.

To alleviate this limitation, previous studies have explored more expressive architectures and external priors. Early efforts (Cao et al., 2022; Wang et al., 2022a) introduced transformer-based designs to capture long-range dependencies and global context, while recent architectures (Li et al., 2025a; Ma et al., 2024b) further improve representation capacity. However, increasing architectural complexity alone does not resolve the limitation caused by insufficient supervision. Recently, Vision Foundation Models (VFMs) pretrained on large-scale natural image corpora have emerged as promising sources of transferable visual knowledge. Existing medical segmentation approaches mainly follow two strategies. One direction (Li et al., 2025b; Ma et al., 2024a; 2025; Liu et al., 2025) adapts VFMs to medical domains through large-scale fine-tuning, which requires substantial data and computational resources and may alter the general-purpose representations learned during pretraining. Another direction introduces frozen VFM representations through feature concatenation, fusion, or lightweight adaptation (Yang et al., 2025; Gao et al., 2025). However, directly transferring VFM features may

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.

suffer from domain discrepancies between natural and medical images, while shallow integration strategies do not fully exploit the relational structures encoded by VFMs. Therefore, a key challenge remains in how to effectively transform generic VFM knowledge into task-specific representations that are both robust to domain variation and beneficial for precise medical segmentation.

In this work, we propose the Prior-to-Precision Network (PoPNet), a unified encoder–decoder framework that follows a global-to-local and abstract-to-refinement strategy to bridge generic foundation-model priors with task-specific segmentation representations. PoPNet realizes this strategy through two complementary objectives. First, PoPNet introduces spectral clustering regularization to abstract cross-layer VFM attention affinities into structural priors that regularize the task-specific latent manifold. Rather than directly aligning task-specific features with VFM representations, this objective transfers structural priors through affinity-derived spectral constraints, reducing reliance on domain-specific feature details. Building on this globally regularized representation, PoPNet further introduces target-aware ROI Masked Generation for local refinement. By restricting reconstruction supervision to segmentation-relevant regions, this objective focuses learning on fine-grained target appearance and boundary cues required for precise delineation. Moreover, spectral clustering regularization requires no pixel-level annotations, while ROI Masked Generation uses predicted masks only as coarse spatial anchors for unlabeled images rather than as pixel-level pseudo-label supervision, enabling a natural extension to semi-supervised segmentation. Extensive experiments on five medical segmentation benchmarks demonstrate state-of-the-art performance under fully supervised settings. Further evaluations on semi-supervised segmentation, zero-shot generalization to unseen datasets, and different VFM priors demonstrate the effectiveness and generalizability of PoPNet across annotation regimes and foundation-model priors.

Our main contributions are summarized as follows:

- We propose PoPNet, a unified global-to-local framework that transfers foundation-model structural priors into medical segmentation through spectral abstraction and ROI-guided local refinement.
- We develop two complementary learning objectives: a spectral clustering regularization that preserves global structural patterns from cross-layer VFM affinity graphs, and a target-aware ROI masked generation objective that concentrates reconstruction supervision on fine-grained local appearance and boundary details.
- Extensive experiments across five supervised benchmarks and semi-supervised polyp segmentation settings demonstrate state-of-the-art performance, together with strong generalization across datasets and diverse foundation models.

2 Related Work

Vision Foundation Models. Pretrained on large-scale visual data, Vision Foundation Models (VFMs) provide transferable representations for diverse downstream tasks. Among self-supervised VFMs, the DINO series (Caron et al., 2021; Oquab et al., 2023; Siméoni et al., 2025) employs teacher–student self-distillation with ViT backbones to learn semantically rich representations without manual annotations. DINOv2 has shown strong transferability to medical imaging (Ayzenberg et al., 2024), while DINOv3 further improves dense feature quality through large-scale training and regularization such as Gram anchoring. Other representative paradigms include CLIP (Radford et al., 2021) for vision–language contrastive learning, I-JEPA (Assran et al., 2023) for latent predictive learning, and Stable Diffusion 3 (Esser et al., 2024), which models visual distributions with Multimodal Diffusion Transformers and flow-based objectives (Lipman et al., 2022). Despite their strong general-purpose priors, transferring these representations to medical imaging remains challenging due to the domain gap between natural and medical imagery.

Supervised Medical Image Segmentation. Existing supervised methods can be broadly grouped into discriminative, generative, and foundation-model-based paradigms. Discriminative approaches include CNN-based models (Ronneberger et al., 2015; Ibtehaz & Rahman, 2020; Ibtehaz & Kihara, 2023; Isensee et al., 2021; Ruan et al., 2023), which provide strong local inductive biases, and ViT-based architectures (Cao et al., 2022; Wang et al., 2022b;a), which better capture long-range dependencies but typically require more data and computation. Generative methods based on GANs, VAEs, and diffusion models (Goodfellow et al., 2020; Kingma & Welling, 2013; Ho et al., 2020) are used either for data synthesis (Huo et al., 2022; Zhang et al., 2025) or for direct segmentation through

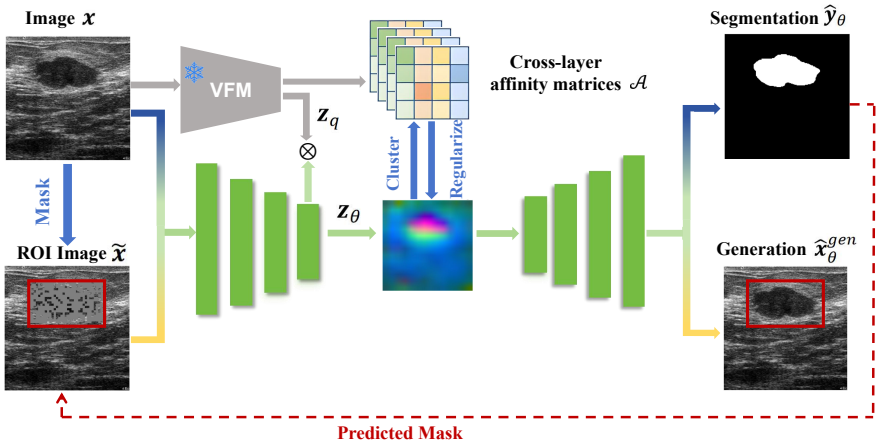

Figure 1: **Overview of the proposed PoPNet.** The unified encoder-decoder architecture anchors global representations to foundation-model priors via spectral clustering regularization, while concurrently enhancing fine-grained local precision through region-of-interest masked generation.

conditional generation or latent translation (Wu et al., 2024; Lin et al., 2024; Chen et al., 2023; Huo et al., 2025; Ahmed et al., 2026), but may introduce substantial inference cost or strong dependence on pretrained decoders. More recently, foundation-model-based methods either integrate frozen VFM features through fusion or adaptation (Yang et al., 2025; Gao et al., 2025), or further specialize VFMs on medical corpora (Li et al., 2025b; Ma et al., 2024a; 2025; Liu et al., 2025). These approaches improve transferability, yet effectively exploiting general-purpose priors while mitigating domain mismatch and computational overhead remains challenging.

Semi-Supervised Medical Image Segmentation. Existing semi-supervised methods can be broadly grouped into pseudo-labeling, consistency regularization, and masked self-supervision. Teacher–student and pseudo-labeling approaches (Xu et al., 2023; Bai et al., 2023; Wang et al., 2021; Thompson et al., 2022; Qiu et al., 2023), sometimes assisted by foundation models such as SAM (Kirillov et al., 2023; Zhang et al., 2024b), exploit unlabeled data through model-generated supervision but remain susceptible to confirmation bias from inaccurate predictions. Consistency-based methods (Chen et al., 2022; Zeng et al., 2023; Assefa et al., 2025; Luo et al., 2022; Yang et al., 2023; Chi et al., 2024) enforce prediction or representation invariance under perturbations, improving robustness but still relying on discriminative supervision. More recently, masked image modeling (Zhu et al., 2025; Zeng et al., 2025; Hondru et al., 2025) provides denser self-supervision by reconstructing masked content; however, global masking may overemphasize background regions, whereas aggressive target-aware masking can remove useful structural context, limiting the learning of fine-grained local features.

3 Methodology

In this section, we introduce PoPNet, a unified framework that bridges foundation-model priors with task-specific segmentation through global structural regularization and local generative refinement. As shown in Figure 1, PoPNet is built on two complementary objectives: spectral clustering regularization, which anchors the task-specific latent manifold toward structural priors derived from VFM affinity graphs, and ROI Masked Generation, which concentrates reconstruction supervision on segmentation-relevant regions to enhance fine-grained target appearance and boundary cues.

3.1 Preliminary

Spectral clustering (Von Luxburg, 2007) formulates data grouping as a graph partitioning problem. Given an affinity matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ and its degree matrix $\boldsymbol{D}$, the relaxed Normalized Cut (NCut) problem leads to the generalized eigenvalue problem $(\boldsymbol{D}-\boldsymbol{A})\boldsymbol{Z} = \lambda \boldsymbol{D}\boldsymbol{Z}$, where the columns of $\boldsymbol{Z} \in \mathbb{R}^{N\times K}$ correspond to eigenvectors associated with the smallest eigenvalues and characterize the intrinsic grouping structure of the graph.

Classical spectral clustering relies on explicit eigendecomposition and is therefore difficult to integrate directly into scalable representation learning. Zhang *et al.* (Zhang et al., 2024a) recast spectral

clustering from a random-walk perspective and further extend it to a layer-distributed setting, seeking a shared spectral embedding that preserves affinity structures across multiple representation graphs. Specifically, by exploiting the bounded spectrum of the transition operator $\boldsymbol{D}_A^{-1}\boldsymbol{A}$, they optimize the spectral embedding through an affinity-preserving surrogate rather than explicit eigendecomposition (see Appendix A). However, the spectral embedding is optimized with respect to fixed affinity graphs rather than being parameterized by task-specific network representations, and thus does not directly regularize feature learning through end-to-end backpropagation.

Building on this formulation, we parameterize the spectral embedding with task-specific latent features and optimize it through differentiable layer-distributed spectral regularization, enabling frozen foundation-model affinities to guide end-to-end representation learning.

3.2 Spectral Clustering Regularization for Global Structural Abstraction

VFMs provide rich semantic and structural priors, but they are typically pretrained on natural images and may exhibit a domain discrepancy when transferred to medical segmentation. Rather than directly transferring VFM features, PoPNet uses spectral clustering regularization to abstract cross-layer VFM attention affinities into global structural priors that regularize the task-specific latent manifold.

VFM Feature Adaptation. Given an input image $\boldsymbol{x} \in \mathbb{R}^{C\times H\times W}$, where C denotes the number of channels and $H \times W$ the spatial resolution, the task-specific encoder produces the bottleneck feature $\boldsymbol{z}_\theta^e \in \mathbb{R}^{C_e\times H_e\times W_e}$. Meanwhile, the frozen VFM extracts a query representation $\boldsymbol{z}_q \in \mathbb{R}^{C_q\times H_q\times W_q}$ from a selected deep transformer layer. After spatial and channel alignment, a lightweight gated fusion adaptively integrates the VFM representation with the task-specific bottleneck:

$$\boldsymbol{w}_\theta = \sigma\big(f_\theta^{gate}\left[\phi_e(\boldsymbol{z}_\theta^e), \phi_q(\boldsymbol{z}_q)\right]\big) \qquad \boldsymbol{z}_\theta^g = \phi_e(\boldsymbol{z}_\theta^e) \odot \boldsymbol{w}_\theta + \phi_q(\boldsymbol{z}_q) \odot (1 - \boldsymbol{w}_\theta) \tag{1}$$

where $\phi_e(\cdot)$ and $\phi_q(\cdot)$ denote channel projections, $[\cdot,\cdot]$ denotes channel-wise concatenation, and $\odot$ denotes element-wise multiplication. During training, prior dropout sets the VFM feature to zero with probability p, preventing excessive dependence on the frozen prior.

Spectral Regularization. Classical spectral clustering characterizes intrinsic structural patterns by computing eigenvectors of an affinity-derived graph operator. Accordingly, we introduce a spectral head to predict a learnable spectral embedding from the task-specific feature:

$$\boldsymbol{z}_\theta = f_\theta^{sp}(\boldsymbol{z}_\theta^g) \in \mathbb{R}^{N_e\times K}, \qquad N_e = H_e \times W_e. \tag{2}$$

where K denotes the dimension of the learned spectral embedding. Each column of $\boldsymbol{z}_\theta$ defines a spectral vector over the N_e bottleneck locations and is regularized by the cross-layer affinity matrices introduced below. Together, these vectors form a compact structural abstraction of the task-specific representation, with K controlling its capacity independently of the number of segmentation classes.

To construct the affinity graphs required by spectral clustering, we extract self-attention matrices $\mathcal{A} = \{\boldsymbol{A}_l\}_{l=1}^L$ from L selected transformer layers of the frozen VFM. Each $\boldsymbol{A}_l \in \mathbb{R}^{N_a\times N_a}$ encodes pairwise attention relations among N_a spatial tokens. Since self-attention is row-normalized by Softmax, $\boldsymbol{A}_l$ is row-stochastic and can be interpreted as a transition operator under the random-walk spectral formulation. When the bottleneck grid and VFM token grid differ in spatial resolution, a resampling function $g(\cdot)$ aligns the spectral embedding $\boldsymbol{z}_\theta$ as $g(\boldsymbol{z}_\theta) \in \mathbb{R}^{N_a\times K}$.

In the layer-distributed spectral formulation of (Zhang et al., 2024a), the spectral embedding is optimized as an independent variable over affinity matrices. In contrast, PoPNet predicts the spectral embedding from task-specific features through Eq. 2, and reformulates the spectral clustering objective as an end-to-end regularization term:

$$\min_{\boldsymbol{z}_\theta} \mathbb{E}_{\boldsymbol{A}_l\in\mathcal{A}} \left\|\mathrm{diag}\left(g(\boldsymbol{z}_\theta)^\top \boldsymbol{A}_l g(\boldsymbol{z}_\theta)\right) - \boldsymbol{1}_K\right\|_1 + \left\|g(\boldsymbol{z}_\theta)^\top g(\boldsymbol{z}_\theta) - \boldsymbol{I}_K\right\|_F^2 \tag{3}$$

where $\boldsymbol{1}_K$ is the all-one vector and $\boldsymbol{I}_K$ is the identity matrix. After spatial resampling, each column of $g(\boldsymbol{z}_\theta)$, corresponding to one spectral vector, is L_2-normalized along the spatial dimension; hereafter, $g(\boldsymbol{z}_\theta)$ denotes the normalized spectral embedding. The first term encourages the learned spectral modes to preserve the structural priors encoded by the VFM affinity matrices, while the second term maintains the orthogonality constraint required by spectral clustering. Accordingly, we instantiate the spectral clustering regularization with an eigenvalue approximation loss and an orthogonality loss.

Eigenvalue Approximation. For each normalized spectral vector, we compute its quadratic response under the VFM affinity graph and define:

$$\ell_\theta^{\text{eig}} = \frac{1}{LK} \sum_{l=1}^{L} \sum_{k=1}^{K} \left| g(\boldsymbol{z}_\theta)_k^\top \boldsymbol{A}_l g(\boldsymbol{z}_\theta)_k - 1 \right|, \tag{4}$$

where $g(\boldsymbol{z}_\theta)_k \in \mathbb{R}^{N_a}$ denotes the k-th normalized spectral vector. Following the layer-distributed spectral formulation, we use 1, corresponding to the dominant eigenvalue of the row-normalized affinity operator, as a surrogate target for the quadratic response. This regularization encourages the learned spectral vectors to capture dominant structural patterns encoded by the VFM affinity matrices.

Orthogonality Regularization. The eigenvalue approximation loss alone may cause different spectral vectors to converge to similar solutions. To preserve distinct structural components, we impose an orthogonality constraint on different eigenvectors:

$$\ell_\theta^{\text{ortho}} = \frac{1}{K(K-1)} \sum_{i \neq j} \left(\left[g(\boldsymbol{z}_\theta)^\top g(\boldsymbol{z}_\theta) \right]_{i,j} \right)^2, \tag{5}$$

where i and j index different spectral vectors, and the summation considers all off-diagonal entries of the Gram matrix. After L_2 normalization of each spectral vector, the diagonal entries of the Gram matrix are fixed to one, so $\ell_\theta^{\text{ortho}}$ only penalizes correlations between different modes and encourages complementary structural components.

The final spectral clustering regularization is formulated as: $\ell_\theta^{\text{sc}} = \lambda_{\text{eig}} \ell_\theta^{\text{eig}} + \lambda_{\text{ortho}} \ell_\theta^{\text{ortho}}$.

Structural Feedback. The learned spectral embedding is fed back to the task-specific representation through a lightweight spatial modulator f_θ^{sm}:

$$\boldsymbol{z}_\theta^m = \boldsymbol{z}_\theta^g + f_\theta^{sm}(\boldsymbol{z}_\theta) \odot \boldsymbol{z}_\theta^g, \tag{6}$$

where f_θ^{sm} transforms the spectral embedding into a spatial modulation map for refining the bottleneck representation. Since $\boldsymbol{z}_\theta$ is predicted from the task-specific bottleneck, gradients from ℓ_θ^{sc} propagate through the spectral head to the encoder, encouraging the learned representation to capture structural patterns encoded by the cross-layer VFM affinity matrices.

Segmentation Prediction. The modulated bottleneck feature $\boldsymbol{z}_\theta^m$ is progressively decoded by f_θ^{dec} to obtain the dense representation $\boldsymbol{z}_\theta^{\text{dec}}$. The segmentation head f_θ^{seg} then produces $\hat{\boldsymbol{y}}_\theta$, which is optimized by

$$\ell_\theta^{\text{seg}} = \lambda_{\text{ce}} \ell_\theta^{\text{ce}}(\hat{\boldsymbol{y}}_\theta, \boldsymbol{y}) + \lambda_{\text{dice}} \ell_\theta^{\text{dice}}(\hat{\boldsymbol{y}}_\theta, \boldsymbol{y}), \tag{7}$$

where $\boldsymbol{y} \in \mathbb{R}^{C_y \times H \times W}$ denotes the ground-truth segmentation mask and C_y is the number of segmentation classes.

3.3 Region-of-Interest Masked Generation for Local Refinement

While spectral clustering regularization captures global structural priors from VFM, accurate segmentation also requires fine-grained local appearance and boundary cues. Therefore, we introduce region-of-interest masked generation as an auxiliary generative objective for local refinement.

ROI Construction and Masking. Conventional masked generation methods, such as MAE (He et al., 2022), randomly mask patches over the entire image. However, global masking may allocate reconstruction supervision to regions irrelevant to the segmentation target. PoPNet instead restricts masking to segmentation-relevant ROIs, concentrating the generative objective on local structures and boundaries associated with the target region.

Specifically, given a segmentation mask, we identify connected target components and construct bounding boxes with a small margin around each component. The union of these boxes forms the ROI mask $\boldsymbol{M}_{\text{box}}$. Following the masking ratio adopted in MAE (He et al., 2022), we randomly mask 75% of patches within $\boldsymbol{M}_{\text{box}}$ to obtain the pixel-level binary mask $\boldsymbol{M}_{\text{pixel}}$. Applying $\boldsymbol{M}_{\text{pixel}}$ to the original image $\boldsymbol{x}$ produces the masked input $\tilde{\boldsymbol{x}}$.

ROI Masked Generation. The masked image $\tilde{\boldsymbol{x}}$ is processed by the shared encoder–decoder to obtain $\tilde{\boldsymbol{z}}_\theta^{\text{dec}}$, from which the generation head f_θ^{gen} reconstructs the image as $\hat{\boldsymbol{x}}_\theta^{\text{gen}}$. The reconstruction

loss is computed only on masked pixels within the ROI:

$$\ell_\theta^{\text{gen}} = \frac{\sum \left(|\hat{\boldsymbol{x}}_\theta^{\text{gen}} - \boldsymbol{x}|_1 \odot \boldsymbol{M}_{\text{pixel}} \right)}{\sum \boldsymbol{M}_{\text{pixel}}} \tag{8}$$

By restricting reconstruction supervision to segmentation-relevant masked regions, ℓ_θ^{gen} encourages the shared representation to preserve local appearance and boundary details, complementing the global structural constraints imposed by spectral regularization.

Dual-Stream Training. PoPNet adopts two forward passes through the shared encoder–decoder. The segmentation pass processes the original image $\boldsymbol{x}$ and jointly optimizes ℓ_θ^{seg} and ℓ_θ^{sc}, whereas the generation pass processes the ROI-masked image $\tilde{\boldsymbol{x}}$ and optimizes ℓ_θ^{gen}. During the generation pass, stop-gradient is applied to the spectral embedding, preventing reconstruction gradients from directly backpropagating through the spectral head while retaining gradient flow to the shared encoder–decoder through the task-feature pathway. The generation objective is activated after an initial warm-up stage. Detailed training procedures are provided in Appendix B.

To this end, the overall training objective is formulated as:

$$\ell_\theta^{\text{total}} = \ell_\theta^{\text{seg}} + \ell_\theta^{\text{sc}} + \lambda_{\text{gen}} \ell_\theta^{\text{gen}}. \tag{9}$$

3.4 Extension to Semi-Supervised Segmentation

PoPNet naturally extends to semi-supervised segmentation by applying its spectral and generative objectives to unlabeled images without pixel-level segmentation annotations.

Label-Free Spectral Regularization. For unlabeled images, the spectral clustering regularization ℓ_θ^{sc} is directly applicable because it only relies on the cross-layer affinity matrices extracted from the frozen VFM. It therefore provides label-free structural constraints that encourage the task-specific representation to capture global priors encoded by the VFM affinity graphs.

Prediction-Guided ROI Generation. ROI Masked Generation does not require ground-truth masks as reconstruction targets. For unlabeled images, the segmentation prediction is used only as a coarse spatial anchor to construct the ROI, while the original image remains the reconstruction target. Thus, the prediction provides localization guidance rather than pixel-level pseudo-label supervision.

Semi-Supervised Objective. Formally, let $\mathcal{B}_{\text{L}}$ and $\mathcal{B}_{\text{U}}$ denote labeled and unlabeled mini-batches, respectively, and $\mathcal{B} = \mathcal{B}_{\text{L}} \cup \mathcal{B}_{\text{U}}$. The semi-supervised objective is:

$$\ell_\theta^{\text{total}} = \frac{1}{|\mathcal{B}_{\text{L}}|} \sum_{(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{B}_{\text{L}}} \ell_\theta^{\text{seg}} + \frac{1}{|\mathcal{B}|} \sum_{\boldsymbol{x} \in \mathcal{B}} \left(\ell_\theta^{\text{sc}} + \lambda_{\text{gen}} \ell_\theta^{\text{gen}} \right), \tag{10}$$

where ℓ_θ^{seg} is applied only to labeled data, while ℓ_θ^{sc} and ℓ_θ^{gen} are optimized on both labeled and unlabeled data.

4 Experiments

We conduct comprehensive experiments to evaluate PoPNet under fully supervised and semi-supervised settings. Extensive ablations and evaluations with diverse vision foundation models further validate the effectiveness and adaptability of the proposed framework.

4.1 Experimental Setup

We adopt a standard U-Net as the segmentation backbone and apply PoPNet at its encoder bottleneck, where the proposed spectral regularization and structural modulation operate on the bottleneck representation. Unless otherwise specified, all main experiments are conducted under the same training configuration. We use the pretrained DINOv3-ViT-B/16 as the frozen Vision Foundation Model, which consists of 12 transformer blocks. The query representation $\boldsymbol{z}_{\text{q}}$ is extracted from the 9th block, while attention affinity matrices from the 7th–11th blocks are used to construct the cross-layer affinity set. The prior dropping probability p is set to 0.3, and the number of spectral modes K is set to 32. We evaluate segmentation performance using three standard metrics: Dice Similarity Coefficient (DSC), Intersection over Union (IoU), and the 95th-percentile Hausdorff Distance (HD95). Additional implementation details and training configurations are provided in Appendix C.

4.2 Comparison with State-of-the-Art Methods

Table 1: Quantitative comparison with state-of-the-art supervised segmentation methods. Baselines are grouped into discriminative and generative paradigms, while PoPNet follows a hybrid paradigm integrating discriminative segmentation and generative reconstruction. Results of PoPNet are highlighted in green, and the best previous results among all baselines are highlighted in red.

Type	Method	BUS			BUSI			GlaS			HAM10000			Kvasir-Instrument		
		DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓
Discriminative	U-Net	81.50	70.77	17.68	72.27	63.00	35.42	87.99	80.01	18.45	92.24	86.93	13.74	93.82	89.23	8.71
	MultiResUNet	80.41	70.33	19.22	72.43	62.59	34.19	88.34	80.34	17.42	92.74	87.60	13.02	92.31	87.03	9.49
	ACC-UNet	83.40	73.51	16.49	77.19	68.51	25.49	88.60	80.84	17.14	93.20	88.44	10.83	93.91	89.73	8.74
	nnUNet	85.71	78.68	11.43	79.45	70.99	22.13	87.25	78.24	20.07	93.83	89.32	9.43	93.95	90.20	8.51
	EGE-UNet	72.79	61.96	27.73	75.17	60.23	29.51	83.25	71.31	28.79	93.90	88.50	10.01	92.65	86.30	9.04
	SwinUNet	80.37	69.75	20.49	76.06	66.10	28.69	86.44	76.89	19.63	93.51	88.68	10.46	92.02	85.83	9.15
	SME-SwinUNet	78.87	67.13	22.19	73.93	62.70	30.45	83.72	72.77	26.23	92.71	87.21	12.53	93.32	88.27	8.91
	UCTransNet	83.44	73.74	16.33	76.55	67.50	25.46	87.17	78.80	20.79	93.45	88.73	10.91	93.27	88.48	8.84
	Dino-UNet	88.08	79.99	9.41	76.09	67.74	37.89	91.01	84.17	15.26	93.71	88.85	8.87	94.95	91.04	7.81
Generative	MedSegDiff-V2	83.23	74.36	17.02	71.32	62.73	38.47	86.82	77.05	19.96	92.28	87.02	13.02	92.29	87.21	9.06
	SDSeg	82.47	73.45	20.53	72.76	63.52	36.79	86.76	76.23	21.41	92.54	87.53	12.29	91.23	86.54	9.38
	GSS	84.86	77.58	22.42	79.56	71.22	28.20	87.41	79.17	19.81	92.92	87.98	11.29	93.66	89.15	7.25
	GMS	88.42	80.56	6.79	81.43	72.58	19.50	88.98	81.16	16.32	94.11	89.68	9.32	94.24	90.02	7.03
	WAVE-GMS	90.14	82.62	5.36	82.31	73.42	18.46	-	-	-	93.93	89.37	9.25	94.00	89.40	9.24
Hybrid	**PoPNet(Ours)**	**92.48**	**86.36**	**4.84**	**83.85**	**75.68**	**17.64**	**92.54**	**86.74**	**12.49**	**94.99**	**91.04**	**8.04**	**96.71**	**93.86**	**4.12**

Table 2: Semi-supervised segmentation and zero-shot generalization results. The 10% labeled training setting evaluates semi-supervised segmentation on in-domain datasets, while the 0% labeled setting evaluates zero-shot generalization on unseen datasets excluded from training and validation.

Method	**10% labeled: Semi-supervised segmentation**												**0% labeled: Zero-shot generalization**					
	Kvasir			CVC-ClinicDB			CVC-ColonDB			ETIS			CVC-300			BKAI		
	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓
URPC	46.31	36.92	53.37	39.64	31.02	37.54	15.37	10.80	36.43	10.39	7.22	41.91	11.26	8.30	14.45	25.29	19.97	38.54
AC-MT	41.79	31.84	66.05	39.99	30.38	50.66	19.62	14.24	59.73	13.62	9.49	68.54	21.63	16.42	46.06	26.66	20.11	61.93
CauSSL	42.58	32.43	60.50	32.18	23.50	50.60	19.73	14.64	47.58	13.03	9.50	58.46	24.47	19.40	38.66	24.87	18.91	50.19
BCPNet	46.32	35.22	61.30	37.67	30.07	67.58	24.73	19.20	54.54	13.18	9.84	83.02	34.95	27.75	41.02	29.01	21.80	63.05
UniMatch	77.13	67.86	40.03	76.75	69.20	26.78	63.03	54.66	35.96	38.98	33.88	33.13	84.07	76.42	16.50	62.73	54.17	35.23
ABD	41.79	29.07	75.46	36.75	26.03	74.91	24.12	15.83	81.98	22.51	15.84	95.60	25.01	15.84	79.69	29.62	19.74	77.68
DyCON	68.55	57.77	50.31	67.93	59.46	39.74	47.05	39.16	44.77	24.42	19.72	63.42	67.32	59.67	18.31	52.79	43.22	57.41
CVBM	62.60	51.37	54.97	48.26	39.91	47.67	29.39	22.83	45.94	17.00	12.61	51.95	37.41	30.80	22.22	42.37	34.01	50.50
PICK	66.36	56.09	46.43	66.44	57.28	42.30	48.96	39.56	57.72	36.58	29.02	72.66	72.05	62.44	53.10	51.11	42.79	44.02
FPGM	82.24	73.21	31.74	79.83	72.36	24.59	68.03	59.55	30.16	44.84	38.46	39.52	89.55	81.87	12.30	68.68	59.91	28.11
PoPNet	**87.14**	**80.29**	**20.93**	**88.29**	**82.02**	**22.44**	**72.11**	**64.92**	**22.44**	**61.54**	**55.06**	**26.67**	**89.86**	**83.37**	**7.55**	**77.89**	**70.02**	**24.44**

Supervised Segmentation. We evaluate PoPNet on five medical image segmentation datasets, including BUS (Yap et al., 2017), BUSI (Al-Dhabyani et al., 2020), GlaS (Sirinukunwattana et al., 2017), HAM10000 (Tschandl et al., 2018), and Kvasir-Instrument (Jha et al., 2021). We compare against 9 representative discriminative methods and 5 generative segmentation methods; detailed dataset settings and baseline descriptions are provided in Appendix C. As shown in Table 1, although existing fully supervised methods already achieve strong performance on these benchmarks, PoPNet consistently attains the best results across all five datasets, showing that global spectral regularization and local generative refinement provide further gains even under strong pixel-level supervision.

Semi-Supervised Segmentation and Zero-Shot Generalization. Following the protocol of Xi et al. (2025), we evaluate semi-supervised polyp segmentation on Kvasir (Jha et al., 2019), CVC-ClinicDB (Bernal et al., 2015), CVC-ColonDB (Jain et al., 2023), and ETIS (Silva et al., 2014), and zero-shot generalization on the held-out CVC-300 (Bernal et al., 2012) and BKAI (Tomar et al., 2023) datasets. CVC-300 and BKAI are excluded from both training and validation. Further experimental details are provided in Appendix C.

As shown in Table 2, PoPNet consistently achieves the best performance across all four in-domain datasets under semi-supervised training. Notably, the gains become more pronounced in this more challenging low-annotation setting, showing that our method can effectively exploit unlabeled data when pixel-level supervision is limited. This stronger advantage under scarce annotations highlights the practical value of PoPNet for medical segmentation, where dense labels are costly to obtain. PoPNet also achieves the best results under zero-shot evaluation. These results indicate that the learned representations remain transferable beyond the training domains, consistent with the use of VFM-derived structural priors rather than relying solely on source-domain appearance. Results with 5% and 20% labeled source data are provided in Appendix D.

4.3 Ablation Study

Quantitative Ablation Analysis. Table 3 compares different strategies for exploiting VFM priors and evaluates the contribution of each PoPNet component. Directly decoding frozen DINO features yields limited and inconsistent gains, reflecting the transfer gap between generic VFM representations and medical segmentation. Feature concatenation provides more stable improvements, but remains clearly inferior to our method across all datasets, indicating the limitation of this shallow fusion. In comparison, adding spectral clustering regularization to U-Net substantially improves all three metrics; for example, on BUSI, DSC increases from 74.81 to 82.55, and HD95 decreases from 30.81 to 18.26 compared with direct DINO concatenation. Further incorporating ROI Masked Generation consistently improves the performance across all datasets, especially in HD95, including reducing it from 5.88 to 4.84 on BUS and from 14.48 to 12.49 on GlaS. These results demonstrate the complementary roles of spectral clustering regularization and region-of-interest masked generation in capturing global structural priors and refining local target-aware details, respectively.

Table 3: Component-wise ablation study under the same fully supervised setting with a frozen DINOv3-ViT-B/16 backbone. U-Net serves as the segmentation baseline. DINO+MLP directly decodes frozen DINO features using a lightweight probing head, while U-Net+DINO Concat fuses DINO features with the U-Net bottleneck. For PoPNet variants, SC denotes spectral clustering regularization and RG denotes ROI Masked Generation: U-Net+SC applies only spectral clustering regularization, whereas U-Net+SC+RG represents the full PoPNet framework.

Method	BUS			BUSI			GlaS			Kvasir-Instrument		
	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓
DINO-based Baselines												
U-Net (Baseline)	81.50	70.77	17.68	72.27	63.00	35.42	87.99	80.01	18.45	93.82	89.23	8.71
DINO + MLP	82.90	72.55	10.73	71.54	61.18	32.65	86.47	76.77	19.13	90.64	85.19	12.13
U-Net + DINO Concat	86.83	77.78	10.72	74.81	65.72	30.81	89.91	82.45	17.91	93.95	90.77	7.58
PoPNet Ablation												
U-Net + SC	91.09	83.97	5.88	82.55	74.50	18.26	91.87	85.67	14.48	95.61	91.98	5.74
U-Net + SC + RG (PoPNet)	**92.48**	**86.36**	**4.84**	**83.85**	**75.68**	**17.64**	**92.54**	**86.74**	**12.49**	**96.71**	**93.86**	**4.12**

Qualitative Analysis of Structural Abstraction and Segmentation. As shown in Fig. 2(a), cross-layer VFM affinity maps assign strong responses to regions related to the red anchor point in the input image, indicating that the pretrained VFM encodes meaningful structural priors. Under these affinity constraints, the learned spectral vectors exhibit responses aligned with segmentation-relevant structures, showing how the priors are abstracted into the task-specific spectral embedding. The feature maps further illustrate the evolution from the encoder representation $\boldsymbol{z}_\theta^e$ and VFM query $\boldsymbol{z}_q$ to the fused representation $\boldsymbol{z}_\theta^g$ and spectrally modulated representation $\boldsymbol{z}_\theta^m$, illustrating how VFM priors and spectral feedback progressively modify the task-specific representation. As shown in Fig. 2(b), PoPNet produces segmentation masks that more closely match the ground truth than DINO+MLP and U-Net+SC across the representative datasets.

Analysis of ROI Masked Generation. Table 4 analyzes the effects of masked generation, ROI-based masking, and masking ratio. Compared with the model without generation, global masked generation provides modest overall improvements, showing that the generative objective can benefit representation learning. Restricting the masked region to the ROI yields better performance than global masking consistently, demonstrating that concentrating generation supervision on segmentation-relevant regions provides more effective local refinement. Furthermore, increasing the ROI masking ratio from 50% to 75% consistently improves all metrics, suggesting that a more challenging task encourages the model to learn more informative local representations and boundary cues.

Table 4: Ablation study on ROI Masked Generation. The first row removes the generation objective. The second row applies global masked generation with a 75% masking ratio, while the third and fourth rows apply ROI-based masked generation with 50% and 75% masking ratios, respectively.

ROI	Mask Ratio	BUS			BUSI			GlaS			Kvasir-Instrument		
		DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓
-	-	91.09	83.97	5.88	82.55	74.50	18.26	91.87	85.67	14.48	95.61	91.98	5.74
-	75%	91.26	84.07	5.85	82.86	74.94	18.05	92.15	86.08	13.87	96.26	93.10	5.67
✓	50%	91.34	84.73	5.26	83.28	75.31	17.96	92.51	86.65	13.13	96.50	93.48	4.98
✓	75%	**92.48**	**86.36**	**4.84**	**83.85**	**75.68**	**17.64**	**92.54**	**86.74**	**12.49**	**96.71**	**93.86**	**4.12**

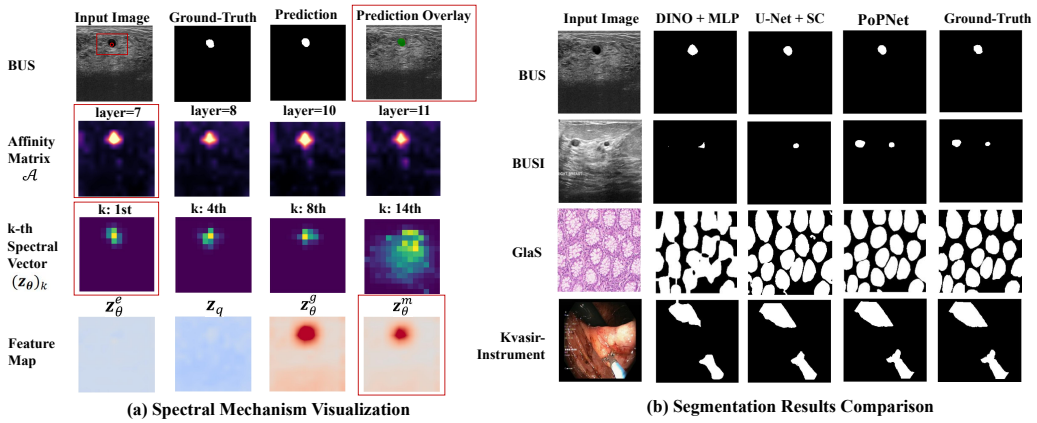

Figure 2: Qualitative analysis of spectral clustering and segmentation. (a) Spectral mechanism visualization on a BUS sample. The first row shows the input image, ground truth, prediction, and prediction overlay. The second row visualizes VFM affinity maps queried from the red anchor point in the input image, where 'layer' denotes the corresponding DINO transformer layer. The third row shows representative spectral vectors from the learned spectral embedding z_θ, where k indexes individual spectral vectors. The fourth row visualizes the encoder feature z_θ^e, VFM query feature z_q, gated-fused feature z_θ^g, and spectrally modulated feature z_θ^m, respectively. (b) Segmentation comparison among DINO+MLP, U-Net+SC, and PoPNet across four datasets.

4.4 Generalizability Across Vision Foundation Models.

To examine whether PoPNet depends on a specific foundation model, we evaluate four VFMs with distinct pre-training paradigms, including I-JEPA Assran et al. (2023), Stable Diffusion 3 (SD3) Esser et al. (2024), DINOv3 Siméoni et al. (2025), and CLIP Radford et al. (2021), while keeping all training settings unchanged. As shown in Table 5, incorporating any of these VFM priors consistently improves the U-Net baseline across all four datasets. Moreover, no single VFM dominates across datasets: DINOv3 achieves the best performance on BUS and Kvasir-Instrument, whereas CLIP performs best on BUSI and GlaS. These consistent gains demonstrate that PoPNet generalizes across different VFM priors and does not rely on a particular foundation model.

Table 5: Generality of PoPNet across different vision foundation models. All VFM variants are evaluated under identical training settings. The best results are highlighted in **bold**.

VFM	BUS			BUSI			GlaS			Kvasir-Instrument		
	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓
Baseline	81.50	70.77	17.68	72.27	63.00	35.42	87.99	80.01	18.45	93.82	89.23	8.71
I-JEPA	91.62	84.92	5.39	83.66	75.52	18.26	92.54	86.62	12.95	96.08	92.81	5.67
SD3	90.56	83.32	5.52	83.45	75.64	16.85	91.93	85.80	13.48	95.73	92.31	6.93
DINOv3	**92.48**	**86.36**	**4.84**	83.85	75.68	17.64	92.54	86.74	12.49	**96.71**	**93.86**	**4.12**
CLIP	91.69	84.99	5.20	**83.99**	**76.09**	**15.61**	**92.70**	**86.89**	**12.25**	96.05	92.88	5.51

5 Conclusion

We present PoPNet, a unified framework that bridges foundation-model priors with task-specific segmentation precision through complementary global structural regularization and local generative refinement. PoPNet introduces spectral clustering regularization to anchor the task-specific latent manifold toward structural priors derived from cross-layer VFM attention affinities, while ROI Masked Generation concentrates reconstruction supervision on fine-grained target appearance and boundary details. These objectives further enable PoPNet to naturally extend to semi-supervised segmentation, where unlabeled images contribute through spectral regularization and prediction-guided ROI construction without pixel-level pseudo-label supervision. Extensive experiments demonstrate state-of-the-art performance under both fully supervised and semi-supervised settings, together with strong zero-shot generalization on unseen datasets and consistent improvements across different vision foundation models.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the writing and checking of mathematical proofs, translation between Chinese and English, implementation of the proposed methods, and thematic analysis of research materials. For proof-related assistance, generative AI was used to help organize, express, and check derivations formulated by the authors; it was not used to formulate the mathematical claims themselves or to provide critical ingredients underlying the proofs. For method implementation, generative AI was used to assist with translating the authors' methodological designs into code and with implementation-related debugging and refinement. For thematic analysis, generative AI was used to assist in organizing and identifying recurring themes in the relevant research materials, with the resulting analyses subsequently reviewed by the authors.

We did not use generative AI tools to generate synthetic datasets, develop the theoretical models or conceptual frameworks of this work, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine research hypotheses, design or provide feedback on the research methodology or experimental design, clean or reformat datasets, support qualitative data analysis, or interpret experimental results. Where these tasks were applicable, they were carried out by the authors.

Additionally, we used generative AI tools to analyze and summarize existing literature and to source or search for relevant research information. Information, technical claims, and literature identified with AI assistance were checked by the authors against the corresponding original sources before being incorporated into the work.

All AI-assisted mathematical content was independently reviewed and verified by the authors. AI-assisted translations were checked against the intended original meaning. AI-assisted method implementations were reviewed, executed, and tested by the authors for correctness. AI-assisted thematic analyses were checked against the underlying research materials, and literature-related information was verified using the original sources. The authors reviewed all AI-assisted work and take full responsibility for the final content of this work, including its text, mathematical claims, code, analyses, and other artifacts.

REFERENCES

Srinivas Addimulam, Manzoor Anwar Mohammed, Raghunath Kashyap Karanam, Deng Ying, Rajani Pydipalli, Bhavik Patel, Mohamed Ali Shajahan, Niravkumar Dhameliya, and Vineel Mouli Natakam. Deep learning-enhanced image segmentation for medical diagnostics. *Malaysian Journal of Medical and Biological Research*, 7(2):145–152, 2020.

Talha Ahmed, Nehal Ahmed Shaikh, and Hassan Mohy-ud Din. Wave-gms: Lightweight multi-scale generative model for medical image segmentation. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6761–6765. IEEE, 2026.

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.

Maregu Assefa, Muzammal Naseer, Iyyakutti Iyappan Ganapathi, Syed Sadaf Ali, Mohamed L Seghier, and Naoufel Werghi. Dycon: Dynamic uncertainty-aware consistency and contrastive learning for semi-supervised medical image segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30850–30860, 2025.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Lev Ayzenberg, Raja Giryes, and Hayit Greenspan. Dinov2 based self supervised learning for few shot medical image segmentation. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2024.

Yunhao Bai, Duowen Chen, Qingli Li, Wei Shen, and Yan Wang. Bidirectional copy-paste for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11514–11524, 2023.

Jorge Bernal, Javier Sánchez, and Fernando Vilarino. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012.

Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pp. 205–218. Springer, 2022.

Luyang Cao, Jianwei Li, and Yinghuan Shi. Background matters: A cross-view bidirectional modeling framework for semi-supervised medical image segmentation. *IEEE Transactions on Image Processing*, 2025.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Cheng Chen, Kangneng Zhou, Zhiliang Wang, and Ruoxiu Xiao. Generative consistency for semi-supervised cerebrovascular segmentation from tof-mra. *IEEE Transactions on Medical Imaging*, 42(2):346–353, 2022.

Jiaqi Chen, Jiachen Lu, Xiatian Zhu, and Li Zhang. Generative semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7111–7120, 2023.

Hanyang Chi, Jian Pang, Bingfeng Zhang, and Weifeng Liu. Adaptive bidirectional displacement for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4070–4080, 2024.

Pierre-Henri Conze, Gustavo Andrade-Miranda, Vivek Kumar Singh, Vincent Jaouen, and Dimitris Visvikis. Current and emerging trends in medical image segmentation with deep learning. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 7(6):545–569, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Yifan Gao, Haoyue Li, Feng Yuan, Xiaosong Wang, and Xin Gao. Dino u-net: Exploiting high-fidelity dense features from foundation models for medical image segmentation. *arXiv preprint arXiv:2508.20909*, 2025.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Ali Hatamizadeh, Assaf Hoogi, Debleena Sengupta, Wuyue Lu, Brian Wilcox, Daniel Rubin, and Demetri Terzopoulos. Deep active lesion segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pp. 98–105. Springer, 2019.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of digital imaging*, 32(4):582–596, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Vlad Hondru, Florinel Alin Croitoru, Shervin Minaee, Radu Tudor Ionescu, and Nicu Sebe. Masked image modeling: A survey. *International Journal of Computer Vision*, 133(10):7154–7200, 2025.

Mandong Hu, Yi Zhong, Shuxuan Xie, Haibin Lv, and Zhihan Lv. Fuzzy system based medical image processing for brain disease prediction. *Frontiers in Neuroscience*, 15:714318, 2021.

Jiayu Huo, Vejay Vakharia, Chengyuan Wu, Ashwini Sharan, Andrew Ko, Sébastien Ourselin, and Rachel Sparks. Brain lesion synthesis via progressive adversarial variational auto-encoder. In *International workshop on simulation and synthesis in medical imaging*, pp. 101–111. Springer, 2022.

Jiayu Huo, Xi Ouyang, Sébastien Ourselin, and Rachel Sparks. Generative medical segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3851–3859, 2025.

Nabil Ibtehaz and Daisuke Kihara. Acc-unet: A completely convolutional unet model for the 2020s. In *International conference on medical image computing and computer-assisted intervention*, pp. 692–702. Springer, 2023.

Nabil Ibtehaz and M Sohel Rahman. Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural networks*, 121:74–87, 2020.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Samir Jain, Rohan Atale, Anubhav Gupta, Utkarsh Mishra, Ayan Seal, Aparajita Ojha, Joanna Jaworek-Korjakowska, and Ondrej Krejcar. Coinnet: A convolution-involution network with a novel statistical attention for automatic polyp segmentation. *IEEE Transactions on Medical Imaging*, 42(12):3987–4000, 2023.

Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*, pp. 451–462. Springer, 2019.

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

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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.

Florian Kordon, Peter Fischer, Maxim Privalov, Benedict Swartman, Marc Schnetzke, Jochen Franke, Ruxandra Lasowski, Andreas Maier, and Holger Kunze. Multi-task localization and segmentation for x-ray guided planning in knee surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 622–630. Springer, 2019.

Chenxin Li, Xinyu Liu, Wuyang Li, Cheng Wang, Hengyu Liu, Yifan Liu, Zhen Chen, and Yixuan Yuan. U-kan makes strong backbone for medical image segmentation and generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 4652–4660, 2025a.

Yuheng Li, Yizhou Wu, Yuxiang Lai, Mingzhe Hu, and Xiaofeng Yang. Meddinov3: How to adapt vision foundation models for medical image segmentation? *arXiv preprint arXiv:2509.02379*, 2025b.

Tianyu Lin, Zhiguang Chen, Zhonghao Yan, Weijiang Yu, and Fudan Zheng. Stable diffusion segmentation for biomedical images with single-step reverse process. In *International conference on medical image computing and computer-assisted intervention*, pp. 656–666. Springer, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Anglin Liu, Rundong Xue, Xu R Cao, Yifan Shen, Yi Lu, Xiang Li, Qianqian Chen, and Jintai Chen. Medsam3: Delving into segment anything with medical concepts. *arXiv preprint arXiv:2511.19046*, 2025.

Xiangde Luo, Guotai Wang, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Dimitris N Metaxas, and Shaoting Zhang. Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. *Medical Image Analysis*, 80:102517, 2022.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024a.

Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024b.

Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. Medsam2: Segment anything in 3d medical images and videos. *arXiv preprint arXiv:2504.03600*, 2025.

Juzheng Miao, Cheng Chen, Furui Liu, Hao Wei, and Pheng-Ann Heng. Caussl: Causality-inspired semi-supervised learning for medical image segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 21426–21437, 2023.

Jordi Minnema, Anne Ernst, Maureen van Eijnatten, Ruben Pauwels, Tymour Forouzanfar, Kees Joost Batenburg, and Jan Wolff. A review on the application of deep learning for ct reconstruction, bone segmentation and surgical planning in oral and maxillofacial surgery. *Dentomaxillofacial Radiology*, 51(7):20210437, 2022.

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.

Liang Qiu, Jierong Cheng, Huxin Gao, Wei Xiong, and Hongliang Ren. Federated semi-supervised learning for medical image segmentation via pseudo-label denoising. *IEEE journal of biomedical and health informatics*, 27(10):4672–4683, 2023.

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

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.

Jiacheng Ruan, Mingye Xie, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. Ege-unet: an efficient group enhanced unet for skin lesion segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 481–490. Springer, 2023.

Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9(2):283–293, 2014.

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

Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J Matuszewski, Elia Bruni, Urko Sanchez, et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35:489–502, 2017.

Bethany H Thompson, Gaetano Di Caterina, and Jeremy P Voisey. Pseudo-label refinement using superpixels for semi-supervised brain tumour segmentation. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2022.

Nikhil Kumar Tomar, Debesh Jha, and Ulas Bagci. Dilatedsegnet: A deep dilated segmentation network for polyp segmentation. In *International conference on multimedia modeling*, pp. 334–344. Springer, 2023.

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.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R Zaiane. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 2441–2449, 2022a.

Wenji Wang, Qing Xia, Zhiqiang Hu, Zhennan Yan, Zhuowei Li, Yang Wu, Ning Huang, Yue Gao, Dimitris Metaxas, and Shaoting Zhang. Few-shot learning by a cascaded framework with shape-constrained pseudo label assessment for whole heart segmentation. *IEEE Transactions on Medical Imaging*, 40(10):2629–2641, 2021.

Ziheng Wang, Xiongkuo Min, Fangyu Shi, Ruinian Jin, Saida S Nawrin, Ichen Yu, and Ryoichi Nagatomi. Smeswin unet: Merging cnn and transformer for medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 517–526. Springer, 2022b.

Junde Wu, Wei Ji, Huazhu Fu, Min Xu, Yueming Jin, and Yanwu Xu. Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 6030–6038, 2024.

Haoran Xi, Chen Liu, and Xiaolin Li. Frequency prior guided matching: A data augmentation approach for generalizable semi-supervised polyp segmentation. *arXiv preprint arXiv:2508.06517*, 2025.

Zhe Xu, Yixin Wang, Donghuan Lu, Xiangde Luo, Jiangpeng Yan, Yefeng Zheng, and Raymond Kai-yu Tong. Ambiguity-selective consistency regularization for mean-teacher semi-supervised medical image segmentation. *Medical Image Analysis*, 88:102880, 2023.

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.

Sicheng Yang, Hongqiu Wang, Zhaohu Xing, Sixiang Chen, and Lei Zhu. Segdino: An efficient design for medical and natural image segmentation with dino-v3. *arXiv preprint arXiv:2509.00833*, 2025.

Moi Hoon Yap, Gerard Pons, Joan Marti, Sergi Ganau, Melcior Sentis, Reyer Zwiggelaar, Adrian K Davison, and Robert Marti. Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE journal of biomedical and health informatics*, 22(4):1218–1226, 2017.

Ling-Li Zeng, Kai Gao, Dewen Hu, Zhichao Feng, Chenping Hou, Pengfei Rong, and Wei Wang. Ss-tbn: A semi-supervised tri-branch network for covid-19 screening and lesion segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10427–10442, 2023.

Qingjie Zeng, Zilin Lu, Yutong Xie, and Yong Xia. Pick: Predict and mask for semi-supervised medical image segmentation. *International Journal of Computer Vision*, 133(6):3296–3311, 2025.

Li Zhang, Basu Jindal, Ahmed Alaa, Robert Weinreb, David Wilson, Eran Segal, James Zou, and Pengtao Xie. Generative ai enables medical image segmentation in ultra low-data regimes. *Nature Communications*, 16(1):6486, 2025.

Xiao Zhang, David Yunis, and Michael Maire. Deciphering'what'and'where'visual pathways from spectral clustering of layer-distributed neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4165–4175, 2024a.

Yichi Zhang, Zhenrong Shen, and Rushi Jiao. Segment anything model for medical image segmentation: Current applications and future directions. *Computers in Biology and Medicine*, 171:108238, 2024b.

Lei Zhu, Jun Zhou, Rick Siow Mong Goh, and Yong Liu. Advmim: Adversarial masked image modeling for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 56–66. Springer, 2025.

A SPECTRAL CLUSTERING

In this supplementary section, we provide a theoretical derivation connecting classical Normalized Cut (NCut) theory (Von Luxburg, 2007) to the layer-distributed spectral formulation introduced by Zhang *et al.* (Zhang et al., 2024a), which motivates our proposed Spectral Clustering Regularization. While conventional spectral analysis optimizes spectral embeddings with respect to affinity graphs, our formulation parameterizes the spectral representation through the task-specific network and incorporates the corresponding spectral objectives directly into end-to-end optimization.

Specifically, the spectral head predicts a continuous spectral representation from the task-specific latent features, while the multi-layer attention structures extracted from the frozen foundation model provide semantic affinity graphs. By reformulating the spectral objective as differentiable regularization terms, gradients can propagate through the spectral head and the segmentation backbone, allowing the learned latent representation to be regularized by the global affinity structures encoded by the foundation model.

In the following, we first review the classical NCut formulation and its continuous spectral relaxation. We then derive its affinity-maximization interpretation and extend it to the layer-distributed setting. Finally, we connect this formulation to the differentiable spectral losses used in our network.

A.1 NORMALIZED CUT AND SPECTRAL RELAXATION

Given an undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $N = |\mathcal{V}|$ vertices, let $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ denote the symmetric affinity matrix, where each element $A_{ij} \geq 0$ measures the similarity between vertices v_i and v_j. Following classical spectral clustering (Von Luxburg, 2007), the diagonal degree matrix is defined as: $\boldsymbol{D}_{ii} = \sum_{j=1}^{N} A_{ij}$, assuming $D_{ii} > 0$ for all vertices. The corresponding unnormalized graph Laplacian is $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$.

The objective of the Normalized Cut (NCut) problem is to partition the graph into K disjoint subgraphs $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K\}$ such that inter-cluster affinity is minimized while intra-cluster connectivity is preserved. For each partition $\mathcal{C}_j$, we define an indicator vector $\boldsymbol{h}_j \in \mathbb{R}^N$ as:

$$h_{i,j} = \begin{cases} \frac{1}{\sqrt{\mathrm{vol}(\mathcal{C}_j)}} & v_i \in \mathcal{C}_j \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where $\mathrm{vol}(\mathcal{C}_j) = \sum_{v_i \in \mathcal{C}_j} D_{ii}$ denotes the graph volume of subgraph $\mathcal{C}_j$.

By stacking all indicator vectors column-wise, we obtain the partition matrix $\boldsymbol{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_K] \in \mathbb{R}^{N\times K}$. The NCut optimization can then be equivalently expressed as the following trace minimization problem:

$$\min_{\mathcal{C}_1,\ldots,\mathcal{C}_K} \mathrm{Tr}\left(\boldsymbol{H}^\top \boldsymbol{L}\boldsymbol{H}\right) \quad \text{s.t.} \quad \boldsymbol{H}^\top \boldsymbol{D}\boldsymbol{H} = \boldsymbol{I}_K, \tag{12}$$

where $\mathrm{Tr}(\cdot)$ denotes the matrix trace operator and $\boldsymbol{I}_K$ is the $K \times K$ identity matrix.

However, since the entries of $\boldsymbol{H}$ are discrete, directly solving Eq. (12) is NP-hard. To obtain a tractable optimization problem, spectral relaxation allows $\boldsymbol{H}$ to take continuous real values. Specifically, we define $\boldsymbol{T} = \boldsymbol{D}^{1/2}\boldsymbol{H}$, which transforms Eq. (12) into:

$$\min_{\boldsymbol{T}\in\mathbb{R}^{N\times K}} \mathrm{Tr}\left(\boldsymbol{T}^\top \boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2}\boldsymbol{T}\right) \quad \text{s.t.} \quad \boldsymbol{T}^\top\boldsymbol{T} = \boldsymbol{I}_K. \tag{13}$$

The matrix

$$\boldsymbol{L}_{sym} = \boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2} \tag{14}$$

is the symmetric normalized graph Laplacian. According to the Rayleigh–Ritz theorem, the optimal solution of Eq. (13) is given by the first K eigenvectors corresponding to the smallest eigenvalues of $\boldsymbol{L}_{sym}$.

A.2 FROM LAPLACIAN MINIMIZATION TO AFFINITY MAXIMIZATION

To connect spectral clustering with an affinity-preserving optimization objective, we reformulate the Laplacian energy using the generalized Rayleigh quotient.

For any nonzero vector $\boldsymbol{x} \in \mathbb{R}^N$, the generalized Rayleigh quotient associated with the graph Laplacian is $\mathcal{R}(\boldsymbol{x}) = \frac{\boldsymbol{x}^\top (\boldsymbol{D}-\boldsymbol{A})\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{x}}$. Substituting $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$, we obtain:

$$\mathcal{R}(\boldsymbol{x}) = \frac{\boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{x} - \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{x}} = 1 - \frac{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{x}}. \tag{15}$$

Eq. (15) shows that minimizing the normalized Laplacian energy is equivalent to maximizing the affinity-preserving term $\frac{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{x}}$, which corresponds to a generalized Rayleigh quotient.

Introducing a Lagrange multiplier λ, the optimization problem can be written as:

$$\mathcal{L} = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} - \lambda \left(\boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{x} - 1\right). \tag{16}$$

Taking derivatives with respect to $\boldsymbol{x}$ and setting them to zero yields:

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{D}\boldsymbol{x}, \tag{17}$$

which can equivalently be written as:

$$(\boldsymbol{D} - \boldsymbol{A})\boldsymbol{x} = (1 - \lambda)\boldsymbol{D}\boldsymbol{x}. \tag{18}$$

Multiplying both sides by $\boldsymbol{D}^{-1}$, we obtain:

$$\boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{x} = \mu \boldsymbol{x}, \tag{19}$$

where $\mu = \lambda$.

The matrix $\boldsymbol{M} = \boldsymbol{D}^{-1}\boldsymbol{A}$ is the random-walk transition matrix. Since $\boldsymbol{M}$ is row-stochastic, $\boldsymbol{M}\mathbf{1} = \mathbf{1}$ and $\|\boldsymbol{M}\|_\infty = 1$. Therefore, 1 is an eigenvalue and every eigenvalue satisfies $|\mu| \leq 1$. This eigenvalue bound does not generally imply $\boldsymbol{u}^\top \boldsymbol{M}\boldsymbol{u} \leq 1$ for every Euclidean unit vector $\boldsymbol{u}$, since $\boldsymbol{M}$ need not be symmetric.

Extending the formulation from a single vector to a continuous spectral embedding $\boldsymbol{X} \in \mathbb{R}^{N\times K}$, where K denotes the spectral embedding dimension, we adopt the following approximate transition-space spectral objective, following Zhang *et al.* (Zhang et al., 2024a):

$$\max_{\boldsymbol{X}} \operatorname{Tr}\left(\boldsymbol{X}^\top \boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{X}\right) \quad \text{s.t.} \quad \boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}_K. \tag{20}$$

Compared with the original Laplacian minimization formulation, Eq. (20) directly emphasizes affinity preservation under the random-walk transition operator and provides a convenient spectral objective for the subsequent multi-graph extension. It is an approximation rather than an exact reformulation of Eq. (13), since Euclidean orthogonality under $\boldsymbol{D}^{-1}\boldsymbol{A}$ differs from the degree-weighted orthogonality of NCut.

A.3 Layer-Distributed Spectral Optimization

Recent work by Zhang *et al.* (Zhang et al., 2024a) shows that semantic and structural information in deep representations is distributed across multiple network layers rather than being confined to a single feature hierarchy. Motivated by this observation, the single-graph spectral formulation can be extended to a layer-distributed setting that jointly models affinity structures extracted from different representations.

To connect this formulation with the notation used in the main text, we denote the affinity matrices extracted from L selected VFM layers as

$$\mathcal{A} = \{\boldsymbol{A}_l\}_{l=1}^{L}, \tag{21}$$

where each $\boldsymbol{A}_l \in \mathbb{R}^{N_a \times N_a}$ encodes pairwise relations among N_a spatial tokens. The task-specific network predicts a spectral embedding $\boldsymbol{z}_\theta \in \mathbb{R}^{N_e \times K}$, where K denotes the spectral embedding dimension. When the task-specific bottleneck grid and the VFM token grid differ in spatial resolution, the differentiable resampling operator $g(\cdot)$ aligns the spectral embedding with the VFM token grid, yielding $g(\boldsymbol{z}_\theta) \in \mathbb{R}^{N_a \times K}$. After resampling, each column is ℓ_2-normalized; hereafter $g(\boldsymbol{z}_\theta)$ denotes this normalized embedding.

Rather than explicitly constructing a large joint affinity matrix, the spectral structures of multiple graphs can be aggregated through an expectation-based objective. Based on the approximate random-walk spectral formulation above, the layer-distributed objective can be expressed as

$$\max_{\boldsymbol{z}_\theta} \mathbb{E}_{\boldsymbol{A}_l \in \mathcal{A}} \left[\mathrm{Tr}\left(g(\boldsymbol{z}_\theta)^\top \boldsymbol{D}_l^{-1} \boldsymbol{A}_l g(\boldsymbol{z}_\theta) \right) \right] \quad \text{s.t.} \quad g(\boldsymbol{z}_\theta)^\top g(\boldsymbol{z}_\theta) = \boldsymbol{I}_K, \tag{22}$$

where $\boldsymbol{D}_l$ denotes the degree matrix associated with affinity matrix $\boldsymbol{A}_l$.

Compared with conventional single-graph spectral clustering, Eq. (22) seeks a shared low-dimensional representation whose spectral directions preserve high-affinity structures across multiple graphs. The resulting embedding therefore captures cross-layer structural consistency rather than relying on a single representation hierarchy.

To incorporate this spectral objective into end-to-end network optimization, we relax the hard orthogonality constraint into a differentiable regularization term. For each learned spectral direction, the diagonal entries of $g(\boldsymbol{z}_\theta)^\top \boldsymbol{D}_l^{-1} \boldsymbol{A}_l g(\boldsymbol{z}_\theta)$ represent quadratic self-association scores with respect to the corresponding transition graph. Since 1 is an eigenvalue and the spectral radius of a row-normalized transition operator equals 1, we use 1 as a reference target that encourages each learned direction to preserve strong graph affinity. For a learned spectral vector that is not an eigenvector, its quadratic score is not itself an eigenvalue and need not be bounded above by 1.

Accordingly, the differentiable surrogate objective is formulated as

$$\min_{\boldsymbol{z}_\theta} \mathbb{E}_{\boldsymbol{A}_l \in \mathcal{A}} \left[\left\| \mathrm{diag}\left(g(\boldsymbol{z}_\theta)^\top \boldsymbol{D}_l^{-1} \boldsymbol{A}_l g(\boldsymbol{z}_\theta) \right) - \mathbf{1}_K \right\|_1 \right] + \left\| g(\boldsymbol{z}_\theta)^\top g(\boldsymbol{z}_\theta) - \boldsymbol{I}_K \right\|_F^2, \tag{23}$$

where $\mathbf{1}_K \in \mathbb{R}^K$ is an all-ones vector, $\|\cdot\|_1$ denotes the ℓ_1 norm, and $\|\cdot\|_F$ denotes the Frobenius norm.

The first term encourages the learned spectral directions to preserve high-affinity structures across the distribution of semantic graphs, while the second term discourages collapse among different spectral components and promotes non-redundant spectral directions. Importantly, matching these quadratic association scores to the spectral reference value 1 should be interpreted as a differentiable surrogate for encouraging dominant affinity-consistent structures, rather than as an assumption that all K eigenvalues of an individual affinity graph are exactly equal to 1.

In our implementation, the semantic graphs are obtained from the self-attention matrices of the frozen DINOv3 encoder. When attention is normalized row-wise over precisely the retained spatial tokens, or renormalized after other tokens are removed, each $\boldsymbol{A}_l$ is a random-walk transition operator, such that $\boldsymbol{D}_l^{-1} \boldsymbol{A}_l = \boldsymbol{A}_l$. If the extracted attention submatrix is not renormalized, its degree matrix must instead be retained in Eq. (23). Under the row-normalization condition, Eq. (23) reduces to

$$\min_{\boldsymbol{z}_\theta} \mathbb{E}_{\boldsymbol{A}_l \in \mathcal{A}} \left[\left\| \mathrm{diag}\left(g(\boldsymbol{z}_\theta)^\top \boldsymbol{A}_l g(\boldsymbol{z}_\theta) \right) - \mathbf{1}_K \right\|_1 \right] + \left\| g(\boldsymbol{z}_\theta)^\top g(\boldsymbol{z}_\theta) - \boldsymbol{I}_K \right\|_F^2. \tag{24}$$

The first term corresponds to the affinity-based eigenvalue approximation objective used in the main text, while the second imposes orthogonality among the learned spectral vectors. With column-wise ℓ_2 normalization, the squared Frobenius term contains only off-diagonal Gram-matrix entries; the main text specifies the corresponding losses as per-mode and off-diagonal averages with explicit weights. This establishes the connection between classical graph spectral relaxation, layer-distributed foundation-model affinity priors, and the differentiable spectral regularization employed by our segmentation network.

B Training Algorithm

We provide the training procedures of PoPNet under supervised and semi-supervised settings. Algorithm 1 presents the fully supervised training procedure, and Algorithm 2 extends it to unlabeled data, where spectral regularization remains label-free and predicted masks are used only as spatial anchors for ROI construction.

For compact notation, $\mathrm{Fuse}_\theta(\cdot)$ denotes the gated VFM feature adaptation in Eq. 1, including prior dropout with probability p. $\mathrm{ROI}(\cdot)$ constructs the bounding-box ROI, and $\mathrm{Mask}(\cdot, r)$ randomly masks a ratio r of patches within the ROI to obtain $\boldsymbol{M}_{\mathrm{pixel}}$. $\mathrm{GenForward}_\theta(\cdot)$ denotes the generation

forward pass through the shared encoder–decoder and generation head, with stop-gradient applied to the spectral embedding. All losses denote averages over the indicated samples, and $[\cdot;\cdot]$ denotes batch-wise concatenation.

Algorithm 1 Supervised Training of PoPNet

Input: Labeled dataset $\mathcal{D}_{\mathrm{L}}$, trainable parameters θ, frozen VFM F^{DINO}, mask ratio r, prior-drop probability p, warm-up epoch T_{w}, loss weights $\lambda_{\mathrm{ce}}, \lambda_{\mathrm{dice}}, \lambda_{\mathrm{eig}}, \lambda_{\mathrm{ortho}}, \lambda_{\mathrm{gen}}$
Output: Optimized parameters θ^*

for each epoch t **do**
  for each batch $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}_{\mathrm{L}}$ **do**
    Extract encoder feature $\boldsymbol{z}_\theta^e \leftarrow f_\theta^{\mathrm{enc}}(\boldsymbol{x})$.
    Extract VFM query feature and affinity matrices $(\boldsymbol{z}_q, \mathcal{A}) \leftarrow F^{\mathrm{DINO}}(\boldsymbol{x})$.
    Obtain fused representation $\boldsymbol{z}_\theta^g \leftarrow \mathrm{Fuse}_\theta(\boldsymbol{z}_\theta^e, \boldsymbol{z}_q; p)$.
    Predict spectral embedding $\boldsymbol{z}_\theta \leftarrow f_\theta^{\mathrm{sp}}(\boldsymbol{z}_\theta^g)$.
    Apply structural feedback $\boldsymbol{z}_\theta^m \leftarrow \boldsymbol{z}_\theta^g + f_\theta^{\mathrm{sm}}(\boldsymbol{z}_\theta) \odot \boldsymbol{z}_\theta^g$.
    Generate segmentation prediction $\hat{\boldsymbol{y}}_\theta \leftarrow f_\theta^{\mathrm{seg}}(f_\theta^{\mathrm{dec}}(\boldsymbol{z}_\theta^m))$.
    Compute segmentation loss $\ell_\theta^{\mathrm{seg}} \leftarrow \lambda_{\mathrm{ce}}\ell_\theta^{\mathrm{ce}}(\hat{\boldsymbol{y}}_\theta, \boldsymbol{y}) + \lambda_{\mathrm{dice}}\ell_\theta^{\mathrm{dice}}(\hat{\boldsymbol{y}}_\theta, \boldsymbol{y})$.
    Compute spectral regularization $\ell_\theta^{\mathrm{sc}} \leftarrow \lambda_{\mathrm{eig}}\ell_\theta^{\mathrm{eig}}(\boldsymbol{z}_\theta, \mathcal{A}) + \lambda_{\mathrm{ortho}}\ell_\theta^{\mathrm{ortho}}(\boldsymbol{z}_\theta)$.
    Construct ROI mask $\boldsymbol{M}_{\mathrm{pixel}} \leftarrow \mathrm{Mask}(\mathrm{ROI}(\boldsymbol{y}), r)$.
    Generate masked input $\tilde{\boldsymbol{x}} \leftarrow \boldsymbol{x} \odot (1 - \boldsymbol{M}_{\mathrm{pixel}})$.
    if $t > T_{\mathrm{w}}$ **then**
      Reconstruct masked regions $\hat{\boldsymbol{x}}_\theta^{\mathrm{gen}} \leftarrow \mathrm{GenForward}_\theta(\tilde{\boldsymbol{x}}, F^{\mathrm{DINO}})$.
      Compute ROI masked generation loss $\ell_\theta^{\mathrm{gen}}$ according to Eq. 8.
    else
      Set $\ell_\theta^{\mathrm{gen}} \leftarrow 0$.
    end if
    Optimize $\ell_\theta^{\mathrm{total}} \leftarrow \ell_\theta^{\mathrm{seg}} + \ell_\theta^{\mathrm{sc}} + \lambda_{\mathrm{gen}}\ell_\theta^{\mathrm{gen}}$.
    Update θ.
  end for
end for
return θ^*

C Experiment Setup

C.1 Dataset

Supervised Experiment Setup. We evaluate supervised segmentation on five public medical image datasets: BUS (Yap et al., 2017), BUSI (Al-Dhabyani et al., 2020), GlaS (Sirinukunwattana et al., 2017), HAM10000 (Tschandl et al., 2018), and Kvasir-Instrument (Jha et al., 2021). BUS and BUSI contain 163 and 647 breast ultrasound images, respectively. For GlaS, we follow its official split of 85 training and 80 testing images. HAM10000 contains 10,015 dermatoscopic images, while Kvasir-Instrument contains 590 endoscopic images for surgical instrument segmentation. For datasets without an official test split, we follow the evaluation protocol of Huo et al. (2025), using 80% of the data for training and 20% for testing; the training subset is further divided into training and validation sets at a 7:1 ratio. All input images are resized to 224×224, and the ROI bounding-box margin used by the generative branch is set to 8 pixels.

We compare PoPNet with nine representative discriminative segmentation methods: U-Net (Ronneberger et al., 2015), MultiResUNet (Ibtehaz & Rahman, 2020), ACC-UNet (Ibtehaz & Kihara, 2023), nnUNet (Isensee et al., 2021), EGE-UNet (Ruan et al., 2023), SwinUNet (Cao et al., 2022), SME-SwinUNet (Wang et al., 2022b), UCTransNet (Wang et al., 2022a), and Dino-UNet (Gao et al., 2025). We further include five representative generative segmentation methods: MedSegDiff-V2 (Wu et al., 2024), SDSeg (Lin et al., 2024), GSS (Chen et al., 2023), GMS (Huo et al., 2025), and WAVE-GMS (Ahmed et al., 2026).

Semi-Supervised Experiment Setup. We conduct semi-supervised evaluations across six public polyp segmentation benchmarks: Kvasir (Jha et al., 2019), CVC-ClinicDB (Bernal et al., 2015),

Algorithm 2 Semi-Supervised Training of PoPNet

Input: Labeled dataset $\mathcal{D}_{\mathrm{L}}$, unlabeled dataset $\mathcal{D}_{\mathrm{U}}$, trainable parameters θ, frozen VFM F^{DINO}, mask ratio r, prior-drop probability p, warm-up epoch T_{w}, loss weights $\lambda_{\mathrm{ce}}, \lambda_{\mathrm{dice}}, \lambda_{\mathrm{eig}}, \lambda_{\mathrm{ortho}}, \lambda_{\mathrm{gen}}$
Output: Optimized parameters θ^*

for each epoch t **do**
  for each mixed mini-batch $\mathcal{B}_{\mathrm{L}} \subset \mathcal{D}_{\mathrm{L}}$ and $\mathcal{B}_{\mathrm{U}} \subset \mathcal{D}_{\mathrm{U}}$ **do**
    Let $(\boldsymbol{x}_{\mathrm{L}}, \boldsymbol{y}_{\mathrm{L}})$ and $\boldsymbol{x}_{\mathrm{U}}$ denote the labeled and unlabeled batch tensors, and combine $\boldsymbol{x} \leftarrow [\boldsymbol{x}_{\mathrm{L}}; \boldsymbol{x}_{\mathrm{U}}]$.
    All samples: extract encoder features and VFM priors $(\boldsymbol{z}_\theta^e, \boldsymbol{z}_q, \mathcal{A}) \leftarrow (f_\theta^{\mathrm{enc}}(\boldsymbol{x}), F^{\mathrm{DINO}}(\boldsymbol{x}))$.
    All samples: obtain fused representation $\boldsymbol{z}_\theta^g \leftarrow \mathrm{Fuse}_\theta(\boldsymbol{z}_\theta^e, \boldsymbol{z}_q; p)$.
    All samples: predict spectral embedding $\boldsymbol{z}_\theta \leftarrow f_\theta^{\mathrm{sp}}(\boldsymbol{z}_\theta^g)$ and modulated feature $\boldsymbol{z}_\theta^m \leftarrow \boldsymbol{z}_\theta^g + f_\theta^{\mathrm{sm}}(\boldsymbol{z}_\theta) \odot \boldsymbol{z}_\theta^g$.
    All samples: predict segmentation masks $\hat{\boldsymbol{y}}_\theta \leftarrow f_\theta^{\mathrm{seg}}(f_\theta^{\mathrm{dec}}(\boldsymbol{z}_\theta^m))$, and split them into $\hat{\boldsymbol{y}}_{\theta,\mathrm{L}}$ and $\hat{\boldsymbol{y}}_{\theta,\mathrm{U}}$.
    Labeled only: compute $\ell_\theta^{\mathrm{seg}} \leftarrow \lambda_{\mathrm{ce}}\ell_\theta^{\mathrm{ce}}(\hat{\boldsymbol{y}}_{\theta,\mathrm{L}}, \boldsymbol{y}_{\mathrm{L}}) + \lambda_{\mathrm{dice}}\ell_\theta^{\mathrm{dice}}(\hat{\boldsymbol{y}}_{\theta,\mathrm{L}}, \boldsymbol{y}_{\mathrm{L}})$.
    All samples: compute $\ell_\theta^{\mathrm{sc}} \leftarrow \lambda_{\mathrm{eig}}\ell_\theta^{\mathrm{eig}}(\boldsymbol{z}_\theta, \mathcal{A}) + \lambda_{\mathrm{ortho}}\ell_\theta^{\mathrm{ortho}}(\boldsymbol{z}_\theta)$.
    Construct labeled ROI $\boldsymbol{M}_{\mathrm{box,L}} \leftarrow \mathrm{ROI}(\boldsymbol{y}_{\mathrm{L}})$ and unlabeled ROI $\boldsymbol{M}_{\mathrm{box,U}} \leftarrow \mathrm{ROI}(\hat{\boldsymbol{y}}_{\theta,\mathrm{U}})$.
    All samples: sample ROI masks $\boldsymbol{M}_{\mathrm{pixel}} \leftarrow \mathrm{Mask}([\boldsymbol{M}_{\mathrm{box,L}}; \boldsymbol{M}_{\mathrm{box,U}}], r)$ and generate $\tilde{\boldsymbol{x}} \leftarrow \boldsymbol{x} \odot (1 - \boldsymbol{M}_{\mathrm{pixel}})$.
    if $t > T_{\mathrm{w}}$ **then**
      All samples: reconstruct masked regions $\hat{\boldsymbol{x}}_\theta^{\mathrm{gen}} \leftarrow \mathrm{GenForward}_\theta(\tilde{\boldsymbol{x}}, F^{\mathrm{DINO}})$.
      All samples: compute $\ell_\theta^{\mathrm{gen}}$ according to Eq. 8.
    else
      Set $\ell_\theta^{\mathrm{gen}} \leftarrow 0$.
    end if
    Optimize $\ell_\theta^{\mathrm{total}} \leftarrow \ell_\theta^{\mathrm{seg}} + \ell_\theta^{\mathrm{sc}} + \lambda_{\mathrm{gen}}\ell_\theta^{\mathrm{gen}}$.
    Update θ.
  end for
end for
return θ^*

CVC-ColonDB (Jain et al., 2023), ETIS (Silva et al., 2014), CVC-300 (Bernal et al., 2012), and BKAI (Tomar et al., 2023). Following the evaluation protocol established in (Xi et al., 2025), our experimental design assesses two critical capabilities. First, we evaluate standard in-domain semi-supervised segmentation performance on the Kvasir, CVC-ClinicDB, CVC-ColonDB, and ETIS datasets. Second, we assess zero-shot domain generalization on the CVC-300 and BKAI datasets, which are strictly excluded from all training and validation phases to represent completely unseen domains. Detailed dataset partition configurations are summarized in Table 6. For data preprocessing, all input images are uniformly resized to 256×256 pixels. Furthermore, during the generative phase, the spatial margin of the ROI bounding box is empirically set to 8 pixels.

Table 6: Details of the datasets used in our experiments, including total sample size and the specific split for training, validation, and testing. Note that CVC-300 and BKAI are exclusively used as unseen datasets for evaluating zero-shot generalization.

Dataset	Sample size	Train	Validation	Test
Kvasir	1000	700	100	200
CVC-ClinicDB	612	400	62	150
CVC-ColonDB	380	230	50	100
ETIS	196	100	32	64
CVC-300	60	–	–	60
BKAI	1000	–	–	1000

Implementation Details. The proposed framework was implemented in PyTorch and trained end-to-end on a single NVIDIA A100 GPU (80GB). Network parameters were optimized using the Adam

optimizer with an initial learning rate of 5×10^{-4}. Training was conducted for 1000 epochs, with a 20-epoch warm-up before activating the generative objective. The overall batch size was set to 8 for all experiments. Under the semi-supervised setting, each mini-batch contained 4 labeled and 4 unlabeled images, maintaining a 1:1 sampling ratio.

The loss weights were set to $\lambda_{\text{ce}} = 1$, $\lambda_{\text{dice}} = 2$, $\lambda_{\text{eig}} = 1$, $\lambda_{\text{ortho}} = 1$, and $\lambda_{\text{gen}} = 1$. We applied random flipping, random rotation, color jittering, and shift-scale-rotation as data augmentation during training. For ROI construction, we identify connected target components from the ground-truth segmentation mask for labeled images and from the predicted mask for unlabeled images. A tight bounding box is constructed around each component and expanded outward by an 8-pixel margin; the union of the resulting boxes defines the ROI mask M_{box}. Within the ROI, the default masking patch size was set to 4. To accommodate differences in spatial granularity across datasets, the patch size was adjusted to 2 for GlaS and 6 for Kvasir-Instrument.

C.2 Segmentation Loss Function

Given an input image defined on the spatial domain $\Omega \subset \mathbb{R}^{H \times W}$ with $N = H \times W$ pixels and C semantic classes, the network prediction is denoted as $\hat{\boldsymbol{y}}_\theta \in [0,1]^{N \times C}$, where $\hat{\boldsymbol{y}}_\theta^{(i,c)}$ represents the predicted probability that pixel i belongs to class c. The corresponding one-hot ground-truth label is denoted by $\boldsymbol{y} \in \{0,1\}^{N \times C}$.

To jointly optimize pixel-wise classification accuracy and region-level structural consistency, the overall segmentation objective is formulated as

$$\ell_\theta^{\text{seg}} = \lambda_{\text{ce}}\, \ell_\theta^{\text{ce}}(\hat{\boldsymbol{y}}_\theta, \boldsymbol{y}) + \lambda_{\text{dice}}\, \ell_\theta^{\text{dice}}(\hat{\boldsymbol{y}}_\theta, \boldsymbol{y}), \tag{25}$$

where λ_{ce} and λ_{dice} denote the weights for the Cross-Entropy loss and Dice loss, respectively.

Cross-Entropy Loss. The Cross-Entropy (CE) loss supervises pixel-wise semantic classification and is defined as

$$\ell_\theta^{\text{ce}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} \boldsymbol{y}^{(i,c)} \log \hat{\boldsymbol{y}}_\theta^{(i,c)}. \tag{26}$$

Dice Loss. To alleviate the class-imbalance issue commonly encountered in medical image segmentation, the Dice loss is further introduced to maximize the overlap between predicted masks and ground-truth regions:

$$\ell_\theta^{\text{dice}} = 1 - \frac{1}{C-1} \sum_{c=1}^{C-1} \frac{2 \sum_{i=1}^{N} \hat{\boldsymbol{y}}_\theta^{(i,c)} \boldsymbol{y}^{(i,c)} + \epsilon}{\sum_{i=1}^{N} \hat{\boldsymbol{y}}_\theta^{(i,c)} + \sum_{i=1}^{N} \boldsymbol{y}^{(i,c)} + \epsilon}, \tag{27}$$

where ϵ is a small smoothing constant for numerical stability. Following common practice, the background category is excluded from Dice computation.

The combination of CE loss and Dice loss provides complementary supervision from both pixel-level semantic discrimination and region-level overlap optimization, leading to more robust segmentation performance under foreground-background imbalance.

C.3 Evaluation Metrics

To comprehensively evaluate segmentation performance from both region-overlap and boundary-quality perspectives, we adopt Dice Similarity Coefficient (DSC), Intersection over Union (IoU), and 95% Hausdorff Distance (HD95) as evaluation metrics.

Dice Similarity Coefficient (DSC). DSC measures the overlap consistency between the predicted segmentation mask $\boldsymbol{X}$ and the ground-truth mask $\boldsymbol{Y}$, which is defined as

$$\text{DSC} = \frac{2|\boldsymbol{X} \cap \boldsymbol{Y}|}{|\boldsymbol{X}| + |\boldsymbol{Y}|}. \tag{28}$$

The DSC score ranges from 0 to 1, where a higher value indicates better segmentation overlap.

Intersection over Union (IoU). IoU, also known as the Jaccard Index, evaluates the ratio between the intersection and union regions of the predicted and ground-truth masks:

$$\mathrm{IoU} = \frac{|\boldsymbol{X} \cap \boldsymbol{Y}|}{|\boldsymbol{X} \cup \boldsymbol{Y}|} = \frac{|\boldsymbol{X} \cap \boldsymbol{Y}|}{|\boldsymbol{X}| + |\boldsymbol{Y}| - |\boldsymbol{X} \cap \boldsymbol{Y}|}. \tag{29}$$

Compared with DSC, IoU provides a stricter evaluation criterion by more heavily penalizing false-positive and false-negative predictions.

95% Hausdorff Distance (HD95). To further assess boundary-level segmentation quality, we employ the 95th percentile Hausdorff Distance (HD95). Let $\partial \boldsymbol{X}$ and $\partial \boldsymbol{Y}$ denote the boundary point sets of the predicted and ground-truth masks, respectively. HD95 is formulated as

$$\mathrm{HD95}(\partial \boldsymbol{X}, \partial \boldsymbol{Y}) = \max\left(\mathcal{H}_{95}(\partial \boldsymbol{X}, \partial \boldsymbol{Y}), \mathcal{H}_{95}(\partial \boldsymbol{Y}, \partial \boldsymbol{X})\right), \tag{30}$$

where

$$\mathcal{H}_{95}(A, B) = \mathrm{Percentile}_{95}\left(\left\{\min_{b \in B} d(a, b) \mid a \in A\right\}\right), \tag{31}$$

and $d(\cdot, \cdot)$ denotes the Euclidean distance between two boundary points.

Unlike overlap-based metrics such as DSC and IoU, HD95 mainly evaluates boundary alignment accuracy and is more sensitive to contour prediction errors. A lower HD95 value indicates better boundary consistency between the predicted segmentation and the ground truth.

Following common experimental observations in medical image segmentation, performance fluctuations within approximately $0.5\% \sim 3\%$ may occur under different parameter settings, random seeds, or training environments across datasets.

D Additional Semi-Supervised and Zero-Shot Results

Following the same evaluation protocol and data splits as in the main paper, we provide additional results under 5% and 20% labeled source data to complement the 10% labeled setting reported in Table 2. We compare PoPNet with the same ten semi-supervised methods, including URPC (Luo et al., 2022), AC-MT (Xu et al., 2023), CauSSL (Miao et al., 2023), BCPNet (Bai et al., 2023), UniMatch (Yang et al., 2023), ABD (Chi et al., 2024), DyCON (Assefa et al., 2025), CVBM (Cao et al., 2025), PICK (Zeng et al., 2025), and FPGM (Xi et al., 2025). Kvasir, CVC-ClinicDB, CVC-ColonDB, and ETIS are used for semi-supervised segmentation, while CVC-300 and BKAI remain completely excluded from training and validation for zero-shot evaluation. Detailed dataset splits, training configurations, and evaluation protocols are provided in Section C.

Semi-Supervised Segmentation. Table 7 shows that PoPNet consistently outperforms all competing methods under both 5% and 20% labeled settings. With only 5% labeled data, PoPNet already improves over the strongest baseline by 3.82/5.81 points in DSC/IoU on Kvasir and by 9.91/10.33 points on ETIS, while reducing HD95 by 10.19 and 14.90, respectively. The gains remain substantial when the labeled ratio increases to 20%; for example, PoPNet improves DSC by 7.90 points on CVC-ClinicDB, 6.40 points on CVC-ColonDB, and 13.33 points on ETIS over the best previous results. The boundary improvements are particularly pronounced on the more challenging datasets, where HD95 is reduced from 35.30 to 16.60 on CVC-ColonDB and from 33.04 to 13.76 on ETIS. Together with the 10% results reported in Table 2, these results show that the advantage of PoPNet persists across a wide range of annotation budgets rather than being tied to a specific supervision ratio. The consistently large gains in both region-overlap and boundary metrics are also consistent with the complementary roles of spectral clustering regularization for global structural learning and ROI Masked Generation for local target-aware refinement.

Zero-Shot Generalization. Table 8 further evaluates zero-shot generalization under 5% and 20% labeled source data. With only 5% labeled data, PoPNet improves over the strongest baselines by 1.36/1.10 points in DSC/IoU on CVC-300 and by 7.92/8.86 points on BKAI, while reducing HD95 from 18.73 to 11.07 and from 34.69 to 23.01, respectively. With 20% labeled data, the gains on CVC-300 become smaller as competing methods already achieve strong overlap scores, yet PoPNet still obtains the best DSC, IoU, and HD95. More notably, the advantage on BKAI remains substantial, improving DSC/IoU from 69.53/60.94 to 79.26/71.87 and reducing HD95 from 28.23 to 15.05. Together with the 10% results in Table 2, these results show that PoPNet maintains consistent

Table 7: Additional semi-supervised segmentation results with 5% and 20% labeled data. Results of PoPNet are highlighted in green, and the best previous results among all baselines are highlighted in red.

Labeled	Method	Kvasir			CVC-ClinicDB			CVC-ColonDB			ETIS		
		DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓
5%	URPC	48.56	36.64	76.79	34.48	25.89	57.87	12.80	8.77	53.62	22.68	17.86	53.79
	AC-MT	42.20	30.21	85.63	36.25	27.22	73.71	20.03	13.50	76.74	25.40	19.50	89.40
	CauSSL	43.89	32.23	80.94	31.01	23.35	62.04	10.73	6.78	57.51	24.11	18.00	74.65
	BCPNet	42.07	30.80	82.07	26.51	20.02	61.96	10.57	7.00	60.26	6.44	4.20	64.21
	UniMatch	74.57	64.48	42.79	67.24	59.24	41.31	45.34	38.13	55.57	45.44	37.77	46.28
	ABD	44.14	30.97	79.60	36.31	25.52	69.54	21.83	14.38	79.79	23.48	15.71	75.49
	DyCON	65.85	55.12	54.14	49.75	42.53	49.59	40.77	33.24	46.84	36.04	28.87	48.25
	CVBM	48.35	37.74	69.55	32.10	25.07	57.11	21.41	15.97	53.05	12.17	8.65	87.20
	PICK	59.42	46.63	73.40	47.25	36.26	65.87	35.40	25.92	90.18	26.22	19.17	81.96
	FPGM	80.66	71.34	36.80	78.01	70.47	27.41	63.13	54.63	40.79	51.07	43.69	39.64
	PoPNet	**84.48**	**77.15**	**26.61**	**83.38**	**75.87**	**24.60**	**66.49**	**56.89**	**25.49**	**60.98**	**54.02**	**24.74**
20%	URPC	65.06	55.05	46.24	55.59	47.03	40.32	38.62	31.29	43.50	32.35	26.74	34.01
	AC-MT	59.32	48.56	54.68	49.48	40.54	44.08	33.93	26.60	54.62	29.70	24.00	37.17
	CauSSL	60.32	49.35	50.72	52.82	44.29	50.13	33.14	25.79	53.26	23.93	19.31	42.55
	BCPNet	61.19	50.38	52.91	52.38	43.38	51.03	35.75	28.18	65.01	23.74	19.04	33.04
	UniMatch	80.88	71.35	36.05	75.42	67.92	27.83	68.93	59.34	41.85	53.00	45.31	40.28
	ABD	45.09	32.51	62.38	36.66	26.84	62.40	25.65	18.18	70.00	24.94	17.98	78.51
	DyCON	79.11	69.30	42.19	74.20	65.38	36.51	59.34	50.22	54.76	51.27	42.52	55.76
	CVBM	72.22	61.82	43.10	64.75	56.16	39.37	44.84	37.45	49.08	38.58	31.45	50.21
	PICK	74.94	64.39	46.27	67.62	59.21	31.71	53.77	45.10	55.63	53.69	43.78	53.04
	FPGM	84.65	76.65	29.79	82.48	75.19	25.45	72.19	63.15	35.30	58.83	51.08	42.49
	PoPNet	**88.72**	**82.22**	**20.32**	**90.38**	**84.44**	**16.14**	**78.59**	**70.83**	**16.60**	**72.16**	**66.21**	**13.76**

transferability across different source-domain annotation budgets, with particularly strong gains on BKAI. The improvements in both overlap and boundary metrics are consistent with the proposed global structural regularization and target-aware local refinement producing representations that remain effective on unseen domains.

E Multi-Seed Stability

To assess the sensitivity of PoPNet to random initialization, we conduct five independent runs with random seeds $\{41, 42, 43, 44, 45\}$ under the same configurations used in the main experiments. All preprocessing, augmentation, optimization, training schedules, hyperparameters, and evaluation protocols are kept fixed, with only the random seed varied. We report the mean and standard deviation under both fully supervised and 20%-labeled source-data settings.

E.1 Fully Supervised Setting

Table 9 reports the results over five independent runs on the five fully supervised datasets. PoPNet exhibits very small variation across all metrics: the standard deviation of DSC remains below 0.31 on every dataset, while that of IoU remains below 0.39 and HD95 below 0.72. The particularly low variation on GlaS, Kvasir-Instrument, and HAM10000 indicates that the fully supervised performance is largely insensitive to random initialization.

E.2 Semi-Supervised and Zero-Shot Settings

We further evaluate multi-seed stability with 20% labeled source data. Table 10 reports results on the four in-domain semi-supervised datasets and the two held-out datasets used for zero-shot evaluation. Compared with the fully supervised setting, the variation is moderately larger under limited supervision, particularly on CVC-ColonDB and ETIS, where the DSC standard deviations reach 1.13 and 1.71, respectively. Nevertheless, the overall fluctuations remain limited across datasets and metrics. For the held-out datasets, CVC-300 shows standard deviations of only $0.77/0.53/0.56$ in DSC/IoU/HD95, while BKAI remains within 1.26 across all metrics, indicating that the zero-shot performance is also only weakly affected by random initialization.

Table 8: Additional zero-shot generalization results on held-out CVC-300 and BKAI with 5% and 20% labeled source data. No labels from the two target datasets are used during training or validation. Results of PoPNet are highlighted in green, and the best previous results among all baselines are highlighted in red.

Source Labeled	Method	CVC-300			BKAI		
		DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓
5%	URPC	3.33	1.95	18.73	29.97	21.71	72.38
	AC-MT	29.33	20.86	71.31	24.97	17.40	83.21
	CauSSL	22.04	12.20	22.56	27.14	18.88	76.14
	BCPNet	8.34	4.75	27.34	20.83	14.86	73.46
	UniMatch	64.17	57.18	33.54	55.62	46.33	46.95
	ABD	15.67	10.53	41.91	30.45	20.68	77.25
	DyCON	60.81	53.79	26.66	48.64	40.05	48.73
	CVBM	27.37	20.44	32.78	27.87	21.22	94.52
	PICK	61.43	50.97	58.79	39.36	29.21	80.05
	FPGM	82.46	75.42	21.00	65.84	56.64	34.69
	PoPNet	**83.82**	**76.52**	**11.07**	**73.76**	**65.50**	**23.01**
20%	URPC	51.95	44.07	20.89	47.15	39.05	39.30
	AC-MT	30.04	24.18	24.62	37.80	30.36	41.82
	CauSSL	36.45	29.89	37.78	41.32	33.10	42.88
	BCPNet	46.31	39.03	48.83	45.29	36.97	59.53
	UniMatch	83.52	76.31	20.85	60.61	52.10	35.65
	ABD	39.30	27.81	65.23	31.93	22.74	59.70
	DyCON	72.95	66.21	12.60	61.78	52.68	42.03
	CVBM	67.70	58.53	36.62	51.91	43.58	41.26
	PICK	73.84	64.84	37.90	53.06	44.18	45.23
	FPGM	86.80	80.89	8.14	69.53	60.94	28.23
	PoPNet	**87.59**	**80.91**	**7.37**	**79.26**	**71.87**	**15.05**

Table 9: Multi-seed stability of PoPNet under fully supervised settings. Results are reported as mean $\pm$ standard deviation over five random seeds $\{41, 42, 43, 44, 45\}$.

Dataset	DSC↑	IoU↑	HD95↓
BUS	92.54 ± 0.13	86.32 ± 0.12	4.52 ± 0.23
BUSI	83.66 ± 0.31	75.50 ± 0.39	18.12 ± 0.72
GlaS	92.53 ± 0.09	86.65 ± 0.13	12.68 ± 0.35
Kvasir-Instrument	96.74 ± 0.11	93.87 ± 0.18	4.84 ± 0.29
HAM10000	95.01 ± 0.06	91.03 ± 0.05	8.06 ± 0.04

Table 10: Multi-seed stability of PoPNet with 20% labeled source data. Kvasir, CVC-ClinicDB, CVC-ColonDB, and ETIS are evaluated under semi-supervised segmentation, while CVC-300 and BKAI are held out for zero-shot evaluation. Results are reported as mean $\pm$ standard deviation over five random seeds $\{41, 42, 43, 44, 45\}$.

Dataset	DSC↑	IoU↑	HD95↓
Kvasir	88.48 ± 0.56	82.30 ± 0.33	19.84 ± 0.78
CVC-ClinicDB	90.54 ± 0.41	84.75 ± 0.27	16.96 ± 0.59
CVC-ColonDB	78.00 ± 1.13	71.04 ± 1.05	21.25 ± 0.85
ETIS	72.37 ± 1.71	66.91 ± 1.12	13.49 ± 1.03
CVC-300	87.00 ± 0.77	80.82 ± 0.53	7.19 ± 0.56
BKAI	79.29 ± 1.08	71.89 ± 1.26	15.41 ± 1.02

Overall Stability. Across both fully supervised and limited-label settings, PoPNet shows consistently low variation over independent random initializations. The increase in variance under limited supervision is expectedly more pronounced on the harder datasets, but remains modest overall, while the held-out zero-shot results also remain stable across seeds. These results indicate that the reported performance is not strongly dependent on a particular random initialization.

F Sensitivity to the Spectral Embedding Dimension K

In PoPNet, K denotes the dimensionality of the learned spectral embedding $\boldsymbol{z}_\theta \in \mathbb{R}^{N_e \times K}$. Each column defines a spectral vector over the bottleneck locations and is jointly regularized by the cross-layer VFM affinity graphs. Therefore, K controls the capacity of the structural abstraction rather than the number of segmentation classes.

A smaller K restricts the number of complementary structural components that can be represented, whereas increasing K provides a higher-capacity spectral space. However, an excessively large K may introduce redundant spectral vectors despite the orthogonality constraint and also increases the representation dimension. We therefore evaluate $K \in \{8, 16, 32, 64\}$ under the fully supervised setting.

Table 11: Sensitivity analysis of the number of spectral modes K under the fully supervised setting. The U-Net baseline does not use spectral regularization. K denotes the dimensionality of the learned spectral embedding. The best result for each dataset is highlighted in **bold**.

Method	BUS			BUSI			GlaS			Kvasir-Instrument		
	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓	DSC↑	IoU↑	HD95↓
U-Net (Baseline)	81.50	70.77	17.68	72.27	63.00	35.42	87.99	80.01	18.45	93.82	89.23	8.71
PoPNet ($K = 8$)	92.05	85.54	5.40	83.33	74.48	18.60	92.07	85.91	14.43	96.49	93.48	5.68
PoPNet ($K = 16$)	92.30	85.92	5.01	83.60	75.54	**16.85**	92.23	86.25	14.30	96.36	93.27	4.86
PoPNet ($K = 32$)	**92.48**	**86.36**	**4.84**	83.85	75.68	17.64	**92.54**	**86.74**	**12.49**	**96.71**	**93.86**	**4.12**
PoPNet ($K = 64$)	91.91	85.32	6.31	**83.95**	**75.98**	17.24	92.38	86.43	13.10	96.69	93.84	4.28

Sensitivity Analysis. Table 11 shows that PoPNet consistently outperforms the U-Net baseline across all tested values of $K \in \{8, 16, 32, 64\}$, indicating that its performance gains are not sensitive to a specific spectral embedding dimension. The overall variation across different K values is modest. Performance generally improves from $K = 8$ to $K = 32$, while further increasing K to 64 yields only marginal changes and slightly degrades several metrics. Among the tested settings, $K = 32$ provides the most consistent overall performance across datasets and is therefore used as the default.

G Computational Efficiency

We further evaluate the computational efficiency of PoPNet on the BUS dataset using a single NVIDIA A100 GPU. Table 12 reports the trainable parameters, GFLOPs, peak GPU memory, training time per epoch, and inference latency per image.

Table 12: Computational efficiency of PoPNet on the BUS dataset using a single NVIDIA A100 GPU. Training time is measured in seconds per epoch, and inference latency is measured in milliseconds per image.

Method	Trainable Params (M)	GFLOPs	Peak Memory (MB)	Training Time (s/epoch)	Latency (ms/image)
PoPNet	61.08	135.51	5929.86	3.47	14.44

Efficiency Analysis. PoPNet contains 61.08M trainable parameters and requires 135.51 GFLOPs, with a peak GPU memory usage of 5929.86 MB. On a single NVIDIA A100 GPU, training takes 3.47 seconds per epoch, while inference requires 14.44 ms per image. These results demonstrate that the proposed spectral clustering regularization and ROI masked generation can be incorporated with practical computational cost, while introducing no additional inference-time optimization procedure.

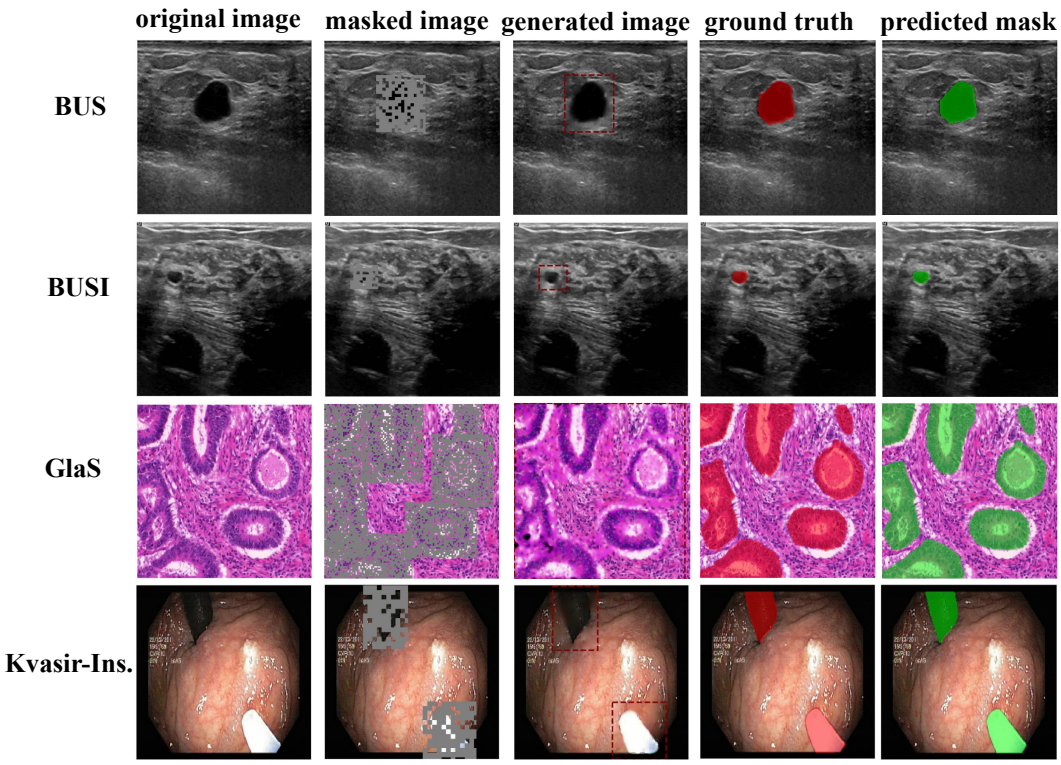

Figure 3: Qualitative visualization of ROI masked generation and the corresponding segmentation results across four medical imaging datasets. The red dashed boxes indicate the reconstructed regions of interest (ROI).

H VISUALIZATION

Generation Performance Visualization. Figure 3 visualizes the ROI masked generation results together with the corresponding segmentation predictions across four datasets.

As shown in the second and third columns, 75% of the patches within the target-aware ROI are masked, while the generative branch reconstructs the missing content from the remaining local context. Across different imaging modalities, the reconstructed regions preserve the major target structures and surrounding appearance, indicating that the auxiliary generation objective encourages the network to capture informative local context within the segmentation-relevant region.

The last two columns further show that the predicted masks are generally well aligned with the ground truth in both target extent and boundary shape, including cases with small or irregular regions. These qualitative results are consistent with the motivation of ROI masked generation: concentrating reconstruction supervision on target-relevant regions promotes fine-grained local representation learning that is complementary to the segmentation objective.