Bridging the Objective Gap: A Unified Pre-Training Framework for Few-Shot Medical Image Segmentation

Shoupeng Chen[1], **Yiming Miao**[1*], **Limei Peng**[1,2], **and Pin-Han Ho**[1,3*]
[1]Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen, China
[2]Kyungpook National University, Daegu, South Korea
[3]University of Waterloo, Waterloo, ON, Canada

Abstract

Few-shot medical image segmentation relies on dense, boundary-sensitive prototype matching, yet common pre-training objectives mainly optimize global alignment or reconstruction, creating an objective gap that hurts boundary delineation and increases adaptation cost. This raises the question: *how to pre-train representations intrinsically matchable for episodic FSS while requiring minimal adaptation-induced re-organization?* We introduce a Drift–Gap diagnostic to quantify intrinsic dense-matching misalignment and adaptation-induced feature drift. Guided by this lens, we propose BOG-PRETRAIN, combining Reliability-Gated Alignment to mitigate noisy report supervision, Semantic-Guided MIM to emphasize boundary-informative regions, and Dual Consistency Regularization to stabilize episodic metric geometry. Across five benchmarks (1/4/16-shot), BOG-PRETRAIN improves mean Dice by +15.5/+15.9/+12.5 points over best priors and reduces mean HD95 by 0.9/4.6/10.1; it achieves the lowest *Drift* (0.052 vs. 0.155 baseline) and Gap_{PT} (0.256), with ablations confirming the components' complementarity.

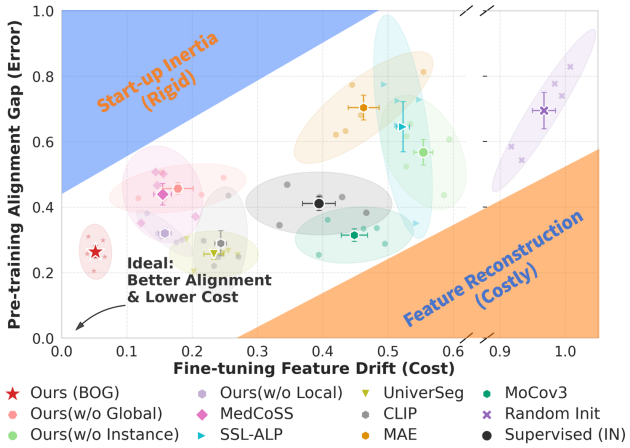

Figure 1: Bridging the objective gap: intrinsic dense-matching misalignment vs. few-shot adaptation drift. Figure 1 visualizes initializer suitability for episodic prototype-based FSS. *Drift* (x-axis) is the mean cosine distance between dense backbone features from the same query image before vs. after few-shot adaptation; lower values indicate a more stable episodic metric geometry and lower adaptation cost. Gap_{PT} (y-axis) measures intrinsic misalignment for mask-conditioned dense matching without adaptation (defined in §3.1); lower values indicate better matchability. Popular paradigms occupy different regimes, such as low-drift/high-gap "inertia" or high-drift re-organization, while ours ($\star$) moves toward the bottom-left "sweet spot".

1 Introduction

Deep learning has substantially advanced medical image segmentation, yet clinical adoption is hindered by high annotation costs [Litjens *et al.*, 2017; Ronneberger *et al.*, 2015]. Few-shot segmentation (FSS) addresses this via prototype-based matching from limited support images [Wang *et al.*, 2020; Snell *et al.*, 2017]. While recent pipelines leverage large-scale pre-training [Azizi *et al.*, 2021], standard initializers often fall short for boundary-sensitive dense matching in episodic FSS. We attribute this to an objective gap in task form: prevalent pre-training objectives emphasize global alignment or reconstruction, which misaligns with the local, pixel-precise separability required by mask-conditioned prototype matching. As a result, similarity maps can be unreliable (especially near boundaries), and few-shot fine-tuning must substantially reshape representations, leading to unstable and costly adaptation.

A good initializer should (i) be intrinsically matchable for dense prototype matching before any adaptation and (ii) remain stable under few-shot fine-tuning. We therefore introduce a Drift–Gap diagnostic (Fig. 1) that measures intrinsic dense-matching misalignment (Gap_{PT}) and adaptation-induced representation change (*Drift*). Low *Drift* indicates a well-organized episodic metric geometry, enabling effective adaptation with minimal representation re-organization; the bottom-left "sweet spot" simultaneously achieves low Gap_{PT} and low *Drift*. This diagnostic also reveals characteristic

*Corresponding authors.
†The official implementation is available at https://github.com/spchen01/BOG-Pretrain.

shifts when individual objectives are removed, suggesting that the objective gap arises from multiple mismatches that must be addressed jointly.

Concretely, we trace the objective gap to three conflicts between generic pre-training and episodic FSS. **Semantic mismatch:** Radiology reports are frequently noisy or only partially relevant [Wang *et al.*, 2022]; uniform image–text alignment can corrupt prototype semantics and weaken internal coherence. **Structural mismatch:** Standard masked image modeling optimizes reconstruction fidelity without prioritizing boundary-discriminative regions [He *et al.*, 2022], limiting boundary-sensitive delineation. **Metric instability:** Without explicit consistency constraints, same-class features can vary substantially across views, forcing large feature drift during adaptation to align support and query prototypes. We formalize these effects in §3.1 and use them to guide objective design.

To bridge these gaps, we propose BOG-PRETRAIN, a pre-training framework for episodic FSS with three complementary components. Reliability-Gated Alignment (regularized by masked language modeling [Devlin *et al.*, 2019]) down-weights unreliable report supervision to improve prototype coherence (reducing the interior mismatch). Semantic-Guided MIM biases masking and reconstruction toward boundary-informative regions to enhance boundary sensitivity (reducing the boundary mismatch). Finally, Dual Consistency Regularization combines prototype-level consistency (ProtoNCE) and same-image consistency (SameImg) to stabilize episodic metric geometry and reduce $Drift$, enabling effective adaptation with minimal representation shift. Across five benchmarks, BOG-PRETRAIN improves mean Dice by up to +15.5/+15.9/+12.5 points and reduces mean HD95 by 0.9/4.6/10.1 under 1/4/16-shot settings. Together, they move the initializer toward the "sweet spot" in Fig. 1.

Our contributions are summarized as follows:

- **Global objective bridging for episodic FSS.** We propose Reliability-Gated Alignment to robustly leverage noisy image–report data, improving prototype coherence and intrinsic matchability.
- **Local objective bridging for boundary-sensitive prediction.** We propose Semantic-Guided MIM to prioritize boundary-informative regions during masked modeling, improving boundary adherence.
- **A unified pre-training framework under an episodic diagnostic.** We introduce the Drift–Gap diagnostic and Dual Consistency Regularization to stabilize episodic metric geometry and substantially reduce adaptation drift; the resulting BOG-PRETRAIN consistently outperforms existing methods across five benchmarks in both region and boundary metrics.

2 Related Work

2.1 Few-Shot and Universal Medical Image Segmentation

Few-shot segmentation (FSS) transfers segmentation to novel classes from a few annotated support images, and many pipelines adopt prototype-based metric matching to predict query masks in a shared embedding space [Snell *et al.*, 2017]. Recent medical FSS and universal segmentation methods improve robustness via prototype refinement, self-supervised episodic training, or support-conditioned inference [Tang *et al.*, 2025; Ouyang *et al.*, 2020; Butoi *et al.*, 2023]. While these designs strengthen adaptation and generalization, they largely operate within a given representation space and thus remain bounded by the initializer: whether dense features are already separable for mask-conditioned matching and how much they must change under few-shot adaptation. This motivates studying pre-training in terms of intrinsic matchability and adaptation cost for episodic FSS (Fig. 1).

2.2 Generic Self-Supervised Pre-Training: Structural and Geometric Limitations for Episodic Dense Matching

Self-supervised learning (SSL) provides strong general-purpose initializers. Contrastive objectives such as MoCo v3 encourage view-invariant instance discrimination, typically on global embeddings [Chen *et al.*, 2021], while masked image modeling (e.g., MAE) learns via reconstructing masked patches [He *et al.*, 2022]. However, these objectives are not designed around mask-conditioned dense separability: reconstruction can preserve appearance without sharpening boundary-relevant cues, and instance-level invariance does not directly calibrate the dense metric geometry used by prototype matching. As a result, SSL/MIM initializers can exhibit intrinsic mismatch for boundary-sensitive delineation and require substantial representation re-organization during episodic fine-tuning, often appearing as elevated Gap_{PT} and/or $Drift$ regimes in Fig. 1. For episodic FSS, this highlights the need to explicitly stabilize the episodic metric geometry (e.g., via consistency constraints) so that adaptation incurs minimal drift.

2.3 Vision–Language Pre-Training: Semantic Benefits with Reliability Challenges

Vision–language pre-training (VLP) injects semantic supervision by aligning image and text representations, as exemplified by CLIP [Radford *et al.*, 2021] and medical adaptations such as MedCLIP and BioViL [Wang *et al.*, 2022; Boecking *et al.*, 2022], with continual multimodal learning explored in MedCoSS [Ye *et al.*, 2024]. For prototype-based FSS, semantics can be valuable because prototypes act as episode-specific anchors for matching. However, radiology reports are often global, noisy, or only partially relevant to localized findings, making uniform alignment unreliable and potentially weakening region-level semantic coherence. Accordingly, VLP can improve intrinsic matchability in some cases yet remain sensitive to report quality and adaptation, leading to varied trade-offs between Gap_{PT} and $Drift$ across datasets in Fig. 1.

Together, these limitations motivate our approach. We revisit pre-training for episodic, boundary-sensitive FSS from a coupled perspective of (i) semantic coherence for reliable prototypes, (ii) boundary-aware dense structure, and (iii) a stable episodic metric geometry with low adaptation drift. Building on segmentation-centric pipelines that accept

given representations [Butoi *et al.*, 2023; Tang *et al.*, 2025; Ouyang *et al.*, 2020] and generic SSL/VLP baselines [Chen *et al.*, 2021; He *et al.*, 2022; Radford *et al.*, 2021; Ye *et al.*, 2024], we develop an objective-gap-bridging pre-training framework that jointly targets these aspects under the Drift–Gap diagnostic, driving the initializer toward the bottom-left "sweet spot" with low Gap_{PT} and low $Drift$.

3 Methodology

Episodic FSS predicts a query mask via mask-conditioned prototype matching in a dense feature space, so the initializer should already support boundary-sensitive dense matching and remain stable under few-shot adaptation. We therefore aim to pre-train representations natively compatible with episodic, mask-conditioned prototype matching. Figure 2 illustrates BOG-PRETRAIN, grouping objectives by targeted mismatches to clarify global, local, and geometric interactions. We next formalize the episodic quantities defining a good initializer, then present the three components following the structure in Fig. 2.

Setup and notation. Given an image–report pair (I_i, R_i), the vision encoder E_V produces dense features $F_i \in R^{C\times H_f\times W_f}$ and a pooled embedding $v_i \in R^d$, while the text encoder E_T produces a global embedding $t_i \in R^d$ and token embeddings $T_i = \{T_{i,l}\}_{l=1}^{L}$. We use cosine similarity $\mathrm{sim}(\cdot,\cdot)$ and ℓ_2 normalization; unless specified, weighting signals are treated as constants (stop-gradient).

3.1 Bridging the Objective Gap: Episodic Diagnostics and Learning Objectives

Episodic FSS predicts a query mask via mask-conditioned prototype matching in a dense feature space. From this viewpoint, a good initializer should (i) enable accurate dense matching before any adaptation and (ii) require only lightweight representation change under few-shot fine-tuning. We instantiate these desiderata as intrinsic pre-adaptation misalignment Gap_{PT} and adaptation-induced feature change $Drift$ (Fig. 1), and use them to organize the learning objectives of BOG-PRETRAIN.

Mask-conditioned similarity map. Given a query image with dense features F and a binary foreground mask M (at image resolution), we sample Q points $\{p_j\}_{j=1}^{Q}$ from M and compute an averaged similarity heatmap:

$$S(u) = \frac{1}{Q}\sum_{j=1}^{Q} \mathrm{sim}\big(\hat{F}(u), \hat{F}(p_j)\big), \tag{1}$$

where $\hat{F}(\cdot)$ denotes ℓ_2-normalized features and $\mathrm{sim}(\cdot,\cdot)$ is cosine similarity.

Objective Gap (Gap_{PT}) and its decomposition. To avoid threshold tuning (which can be dataset-, class-, and episode-dependent and introduces an extra hyperparameter), we binarize the heatmap by a size-matched top-k rule, where $k = |M|$. This threshold-free binarization controls the predicted foreground size and focuses the evaluation on the ranking quality of the similarity map. Let $\hat{M} = \mathrm{TopK}(S, |M|)$ keep the top-$|M|$ pixels of S. We define the pre-training alignment gap as

$$Gap_{\text{PT}} = 1 - \mathrm{Dice}\big(\hat{M},\, M\big). \tag{2}$$

To localize the mismatch, we split M into an interior region and a boundary band using morphology at image resolution: $M_{\text{in}} = \mathrm{Erode}(M, r)$ and $M_{\text{bd}} = \mathrm{Dilate}(M, r) - M_{\text{in}}$. We then measure region-specific gaps as

$$Gap_x = 1 - \mathrm{Dice}\big(\hat{M},\, M_x\big), \quad x \in \{\text{in}, \text{bd}\}. \tag{3}$$

Here Gap_{in} reflects *semantic coherence* inside the target region (prototype reliability), while Gap_{bd} reflects *boundary-sensitive* dense separability for precise delineation.

Feature Drift (adaptation cost). Let F^{pt} and F^{ft} be the dense features of the *same* query image extracted by the pre-trained backbone and the few-shot adapted backbone, respectively. We quantify the adaptation-induced change as the mean cosine distance over spatial locations:

$$Drift = 1 - \frac{1}{N}\sum_{u=1}^{N} \mathrm{sim}\big(\hat{F}^{\text{pt}}(u), \hat{F}^{\text{ft}}(u)\big), \tag{4}$$

where $N = H_f W_f$. Low $Drift$ indicates a stable episodic metric geometry, enabling effective adaptation without large representation re-organization.

Learning objectives guided by the episodic lens. The above quantities clarify *what* an initializer must provide for episodic dense matching: reliable semantics (low Gap_{in}), boundary-sensitive structure (low Gap_{bd}), and a stable metric geometry (low $Drift$). Following the roadmap in Fig. 2, we next introduce three components that target these properties from complementary angles: reliability-aware global alignment (§3.2), boundary-aware local modeling (§3.3), and dual consistency regularization (§3.4).

3.2 Reliability-Gated Global Alignment for Reducing Interior Gap

Reliability-gated global alignment. As analyzed in §3.1, a large interior gap (Gap_{in}) indicates insufficient semantic coherence inside the target region, which weakens prototype reliability even before episodic adaptation. Paired reports can inject global semantics to tighten this coherence, but clinical narratives are often noisy, incomplete, or only partially relevant; enforcing uniform image–report alignment can therefore propagate misleading gradients and distort the semantic anchors that prototypes depend on. We address this by making global alignment *selective*: each image–report pair contributes proportionally to its estimated reliability.

Concretely, for each pair (I_i, R_i) we compute a reliability score $\gamma_i \in [0, 1]$ by fusing complementary signals of alignment confidence, linguistic coherence (via an auxiliary MLM signal), report informativeness, and batch-wise consistency:

$$\gamma_i = \sigma\Big(\alpha \sum_{k=1}^{4} w_k\, \mathcal{Z}(u_i^{(k)})\Big), \quad u_i^{(k)} \in \{p_i, m_i, r_i, c_i\}, \tag{5}$$

where $\mathcal{Z}(\cdot)$ denotes standardization, and p_i, m_i, r_i, c_i are the image–text cosine similarity, MLM token-prediction confidence, TF-IDF report informativeness score, and batch-wise prototype consistency, respectively. We stop gradients through γ_i so that the gate only modulates optimization.

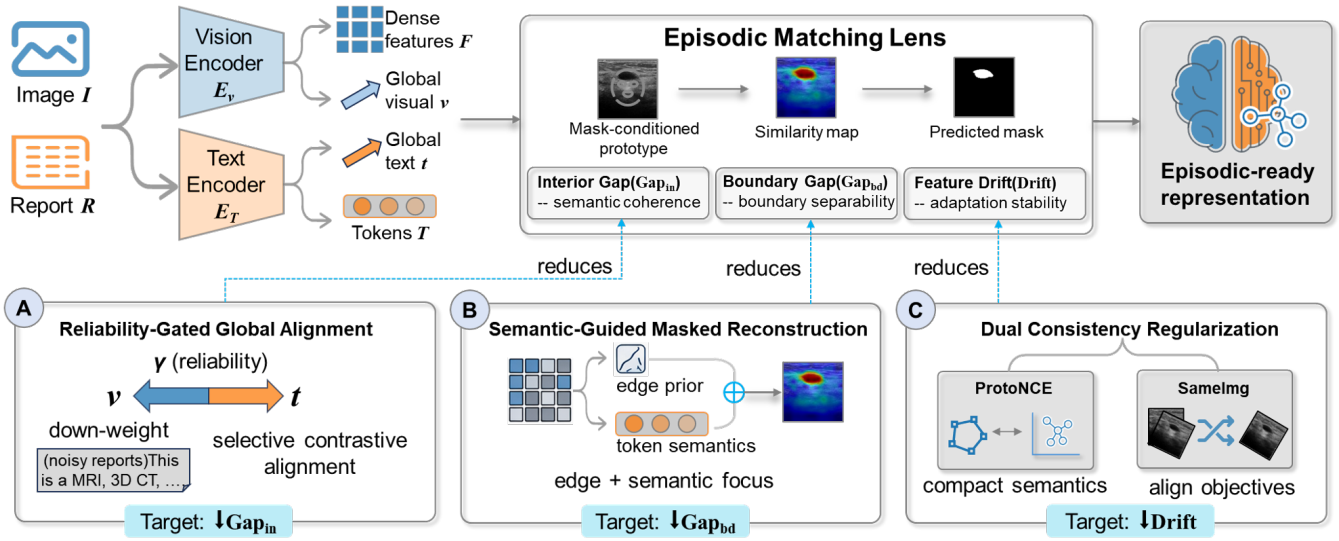

Figure 2: Overview of BOG-PRETRAIN under the Drift–Gap episodic lens. From an image–report pair (I, R), E_V outputs dense F and pooled v, and E_T outputs global t and tokens T. Episodic prototype matching induces three diagnostics (Gap_{in}, Gap_{bd}, and $Drift$), which we target with: (A) Reliability-Gated Global Alignment (contrastive, weighted by γ) to reduce Gap_{in}; (B) Semantic-Guided Masked Reconstruction (edge + token semantics) to reduce Gap_{bd}; (C) Dual Consistency Regularization (ProtoNCE + SameImg) to reduce $Drift$. The three terms form a unified pre-training loss, producing an episodic-ready representation.

We then re-weight a standard symmetric image–text contrastive objective by γ_i:

$$\mathcal{L}_{\text{G-ITC}} = \frac{1}{2} \sum_{d \in \{I \to T,\, T \to I\}} \frac{\sum_{i=1}^{B} \text{sg}(\gamma_i)\, \mathcal{L}_{\text{ITC}}^{d}(i)}{\sum_{i=1}^{B} \text{sg}(\gamma_i)}, \quad (6)$$

where $\mathcal{L}_{\text{ITC}}^{d}(i)$ is the standard InfoNCE term for direction d and sample i, and $\text{sg}(\cdot)$ denotes stop-gradient. By attenuating unreliable pairs, the global semantic geometry becomes less sensitive to report noise, yielding more reliable prototype anchors and directly reducing Gap_{in}.

3.3 Local Structural Modeling via Semantic-Guided Masked Reconstruction

Semantic-guided masked reconstruction. A high boundary gap (Gap_{bd}) in §3.1 indicates that dense features are not sufficiently boundary-discriminative for episodic delineation. Masked reconstruction can strengthen local structure, but uniform masking tends to allocate learning to generic textures rather than boundary-critical regions. We instead make reconstruction *spatially selective*, concentrating supervision on edge- and semantics-informative locations to sharpen boundary sensitivity.

Text-conditioned semantic relevance map. At feature resolution, we compute a semantic relevance map $S_i \in [0,1]^{H_f \times W_f}$ by correlating dense visual features with informative report tokens. Let $\phi_v(\cdot)$ and $\phi_t(\cdot)$ be lightweight projections, and $\mathcal{K}_i$ denote tokens whose TF-IDF weight exceeds the per-report median. We define

$$S_i(u) = \max_{l \in \mathcal{K}_i} \text{sim}\big(\phi_v(F_i(u)), \phi_t(T_{i,l})\big), \quad (7)$$

followed by per-image normalization to $[0,1]$. Text features are detached when forming S_i, so the map serves as a masking prior rather than an additional supervision target.

Edge–semantic fusion with reliability-gated guidance. We compute an edge prior $E_i \in [0,1]^{H_f \times W_f}$ from image gradients and fuse it with semantics as $W_i = (1-\beta)E_i + \beta S_i$. Crucially, we gate this guidance by the global reliability γ_i from §3.2: when reports are unreliable, text-guided masking should not shape local structure; when reports are reliable, reconstruction should focus on boundary- and semantic-rich regions. Given a random/block mask M_i^{blk} and a curriculum $\eta_t \in [0,1]$ increasing over training, we form a soft mask

$$M_i = (1 - \text{sg}(\gamma_i)\eta_t)\, M_i^{\text{blk}} + \text{sg}(\gamma_i)\eta_t\, W_i. \quad (8)$$

Masked reconstruction objective. Let $\hat{y}_i(u)$ and $y_i(u)$ be the prediction and target at location u. We apply a standard reconstruction loss on masked regions, aggregated by the soft mask:

$$\mathcal{L}_{\text{SG-MIM}} = \frac{1}{B} \sum_{i=1}^{B} E_{u \sim M_i}\big[\ell_{\text{rec}}(\hat{y}_i(u), y_i(u))\big]. \quad (9)$$

When γ_i is high, masking concentrates on W_i and forces recovery of edge/semantic details; when γ_i is low, it reverts to M_i^{blk}. This reliability-gated, boundary-aware reconstruction directly targets Gap_{bd} by sharpening boundary-sensitive dense separability.

3.4 Dual Consistency Regularization for Stabilizing Metric Geometry

Reducing Gap_{in} and Gap_{bd} improves intrinsic matchability; to further suppress adaptation-induced drift ($Drift$), we introduce Dual Consistency Regularization that enforces stable

metric geometry via complementary prototype- and instance-level constraints.

Prototype-level consistency (ProtoNCE). We maintain a prototype bank $\{c_k\}_{k=1}^{K}$ in the text embedding space, capturing semantic modes inferred from reports. Each sample is assigned to its nearest prototype in text space, $k(i) = \arg\max_k \mathrm{sim}(t_i, c_k)$; prototypes are updated via EMA with momentum m=0.97 outside the computational graph, so a single noisy text embedding contributes only $\sim$3% per step and is further diluted by batch averaging. We then pull the visual embedding v_i toward its assigned prototype while pushing it away from others:

$$\mathcal{L}_{\text{Proto}} = \frac{1}{B}\sum_{i=1}^{B} -\log \frac{\exp\big(\mathrm{sim}(v_i, c_{k(i)})/\tau_p\big)}{\sum_{k=1}^{K}\exp\big(\mathrm{sim}(v_i, c_k)/\tau_p\big)}. \quad (10)$$

This encourages compact semantic clusters and reduces the degrees of freedom of the embedding space, yielding a more stable episodic metric.

Instance-level consistency (SameImg). When optimizing multiple objectives (global alignment, local reconstruction, prototype consistency), the same image may receive conflicting gradients. We directly enforce that embeddings of the *same* image produced under different views/pathways remain consistent. Let $v_i^{(a)}$ and $v_i^{(b)}$ be two pooled embeddings of the same image (e.g., from different augmentations or objective pathways), we apply a cosine pull with stop-gradient:

$$\mathcal{L}_{\text{SI}} = \frac{1}{B}\sum_{i=1}^{B}\Big(1 - \mathrm{sim}\big(\hat{v}_i^{(a)}, \mathrm{sg}(\hat{v}_i^{(b)})\big)\Big), \quad (11)$$

where $\hat{v}$ denotes ℓ_2-normalized embeddings. This term suppresses cross-objective interference by forcing a unified representation for the same content.

Overall effect on $Drift$. ProtoNCE compacts the semantic geometry, while SameImg aligns the optimization signals across objectives/views. Together, they yield a tight and consistent metric space, which substantially reduces $Drift$ during few-shot adaptation.

3.5 Overall Objective and Conceptual Summary

Overall objective. We jointly optimize all components of BOG-PRETRAIN with:

$$\mathcal{L} = \mathcal{L}_{\text{G-ITC}} + \lambda_{\text{mim}}\mathcal{L}_{\text{SG-MIM}} + \lambda_{\text{cons}}\mathcal{L}_{\text{Cons}}, \quad (12)$$

where

$$\mathcal{L}_{\text{Cons}} = \lambda_{\text{proto}}\mathcal{L}_{\text{Proto}} + \lambda_{\text{si}}\mathcal{L}_{\text{SI}}. \quad (13)$$

We optionally add $\lambda_{\text{mlm}}\mathcal{L}_{\text{MLM}}$ to regularize the text encoder and to provide the linguistic-coherence cue in Eq. (5).

Conceptual summary under the Drift–Gap lens. BOG-PRETRAIN is organized around the episodic diagnostics in §3.1. The reliability-gated global alignment $\mathcal{L}_{\text{G-ITC}}$ makes semantic supervision selective under noisy reports, improving interior coherence and lowering Gap_{in}. The semantic- and edge-guided masked reconstruction $\mathcal{L}_{\text{SG-MIM}}$ concentrates local supervision on boundary-critical regions, sharpening dense separability and lowering Gap_{bd}. Finally, dual consistency regularization ($\mathcal{L}_{\text{Proto}}$ and $\mathcal{L}_{\text{SI}}$) stabilizes the metric geometry by enforcing compact semantics and suppressing cross-objective/view interference, which directly reduces Feature Drift during few-shot adaptation.

Together, these objectives align the pre-trained representation with the demands of episodic, mask-conditioned matching, driving the initializer toward the low-gap, low-drift "sweet spot" in Fig. 1.

4 Experiments

4.1 Experimental Setup

Pre-training data. We pre-train on **MedPix 2.0** [Siragusa *et al.*, 2025], a comprehensive multimodal archive containing $\sim$59K images paired with structured clinical case reports. The rich linguistic context in this dataset provides diverse semantic supervision, which is critical for optimizing our Reliability-Gated Alignment module.

Downstream evaluation. We evaluate on five benchmarks [Cheng *et al.*, 2015; Al-Dhabyani *et al.*, 2020; Byra *et al.*, 2020; Jha *et al.*, 2020; Chowdhury *et al.*, 2020; Rahman *et al.*, 2021; Gong *et al.*, 2021]. Following MedCLIP-SAMv2 [Koleilat *et al.*, 2025], we adopt their splits for **Brain**, **Lung**, and **Breast** tumors. Notably, Breast Tumor adopts a cross-dataset setting (train on BUSI [Al-Dhabyani *et al.*, 2020], test on UDIAT [Byra *et al.*, 2020]). For **Thyroid** [Gong *et al.*, 2021], we use the official split, and for **Kvasir-SEG** [Jha *et al.*, 2020], a random 70/30 partition.

Architecture and protocol. We use an ImageNet-initialized **Swin-S** [Liu *et al.*, 2021] with an **ASPP** head [Chen *et al.*, 2017]. Results are reported as mean **Dice** and **HD95** over 1,000 episodes under **1/4/16-shot** settings. We pre-train for 200 epochs using AdamW ($lr = 5e^{-4}$, weight decay 0.05, batch size 128); total cost is $\sim$3.7 GPU·h on 1×RTX4090 with 62.96M parameters ($\approx$1.7× the wall-clock of a CLIP baseline on the same backbone). The text encoder is discarded after pre-training, so inference cost matches an image-only Swin-S.

Diagnostic metrics. To probe pre-training quality beyond end-task scores, we report (i) **Objective Gap** ($Gap_{\text{in}}/Gap_{\text{bd}}$) that decomposes intrinsic misalignment into semantic coherence (*Interior*) and boundary delineation (*Boundary*) without adaptation, and (ii) **Feature Drift** measured by cosine distance between pre-trained and adapted prototype features after few-shot support.

4.2 Main Results: Comparison with SOTA

Table 1 reports few-shot segmentation results (1/4/16-shot). In line with our *Objective Gap* formulation, we interpret these gains through the intrinsic dense-matching misalignment and adaptation cost. As summarized in Table 2 and visualized in Fig. 1, BOG-PRETRAIN moves the initializer toward the desired bottom-left regime: it achieves the lowest Gap_{PT} *before* adaptation and the lowest $Drift$ *during* adaptation among all compared methods, confirming that the initializer is both intrinsically matchable and geometrically stable without requiring costly representation re-organization.

The downstream improvements align with the *decomposed* gap analysis established in §3.1 and validated in §4.3. First,

Few-shot	Method	Brain Tumors (n=600)		Breast Tumors (n=113)		Kvasir (n=300)		Lung X-ray (n=957)		Thyroid (n=614)	
		Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓
16-shot	SSL-ALPNet	$43.05_{31.53}$	$68.38_{116.85}$	$49.32_{25.42}$	$33.09_{22.84}$	$53.85_{24.77}$	$70.33_{11.84}$	$83.50_{6.29}$	$12.71_{12.51}$	$40.26_{23.97}$	$65.21_{39.25}$
	UniverSeg	$50.67_{36.07}$	$79.79_{125.36}$	$64.36_{25.30}$	$34.15_{30.96}$	$47.63_{24.38}$	$42.19_{17.28}$	$90.96_{7.47}$	$16.82_{16.36}$	$50.72_{25.02}$	$\mathbf{34.38}_{16.58}$
	MedCoSS	$56.37_{34.15}$	$79.54_{73.00}$	$67.47_{25.86}$	$88.01_{95.35}$	$43.59_{18.55}$	$108.77_{86.01}$	$91.71_{6.30}$	$11.73_{16.16}$	$34.67_{15.01}$	$202.89_{131.48}$
	Ours	$\mathbf{58.28}_{27.90}$	$\mathbf{50.56}_{128.63}$	$\mathbf{70.32}_{22.01}$	$\mathbf{23.99}_{25.12}$	$\mathbf{79.81}_{18.25}$	$\mathbf{34.65}_{29.79}$	$\mathbf{93.46}_{5.73}$	$\mathbf{9.27}_{10.15}$	$\mathbf{64.83}_{26.39}$	$38.16_{46.33}$
4-shot	SSL-ALPNet	$34.14_{36.70}$	$107.47_{156.02}$	$41.81_{32.50}$	$77.85_{67.58}$	$40.22_{25.94}$	$76.17_{26.34}$	$74.76_{6.12}$	$19.32_{13.41}$	$36.46_{26.44}$	$78.24_{39.18}$
	UniverSeg	$38.58_{33.94}$	$73.42_{120.92}$	$36.85_{25.73}$	$60.48_{72.60}$	$42.19_{23.30}$	$43.55_{16.66}$	$86.44_{10.03}$	$20.57_{18.64}$	$37.94_{25.68}$	$\mathbf{49.00}_{18.32}$
	MedCoSS	$44.06_{26.76}$	$142.84_{79.10}$	$53.17_{28.50}$	$136.25_{90.43}$	$33.76_{14.96}$	$141.42_{97.02}$	$87.63_{12.92}$	$24.92_{23.06}$	$20.76_{15.20}$	$236.91_{114.23}$
	Ours	$\mathbf{46.60}_{29.96}$	$\mathbf{64.54}_{134.15}$	$\mathbf{59.35}_{26.28}$	$\mathbf{51.87}_{69.97}$	$\mathbf{73.77}_{18.74}$	$\mathbf{38.27}_{28.28}$	$\mathbf{90.12}_{5.71}$	$\mathbf{12.23}_{12.09}$	$\mathbf{51.48}_{21.67}$	$57.12_{33.21}$
one-shot	SSL-ALPNet	$12.28_{28.55}$	$294.45_{264.31}$	$34.93_{30.73}$	$155.43_{182.06}$	$30.70_{27.19}$	$83.54_{30.97}$	$64.65_{6.63}$	$28.04_{10.17}$	$37.74_{19.32}$	$84.31_{46.38}$
	UniverSeg	$13.03_{20.31}$	$290.83_{273.57}$	$20.16_{21.43}$	$132.31_{113.04}$	$23.02_{19.71}$	$43.76_{17.02}$	$77.72_{11.50}$	$23.91_{16.78}$	$36.16_{19.01}$	$\mathbf{36.29}_{14.30}$
	MedCoSS	$15.14_{17.07}$	$\mathbf{222.14}_{47.10}$	$30.43_{24.69}$	$171.25_{98.09}$	$29.37_{17.28}$	$209.92_{103.19}$	$80.95_{10.58}$	$34.33_{17.23}$	$19.93_{14.55}$	$249.21_{119.70}$
	Ours	$\mathbf{16.85}_{26.25}$	$276.84_{241.94}$	$\mathbf{47.67}_{30.74}$	$\mathbf{109.70}_{187.14}$	$\mathbf{60.30}_{19.74}$	$\mathbf{63.30}_{31.24}$	$\mathbf{86.32}_{6.19}$	$\mathbf{17.15}_{11.00}$	$\mathbf{46.90}_{22.04}$	$55.75_{26.67}$
sup.	nnU-Net	$87.06_{13.29}$	$12.03_{15.91}$	$80.40_{22.04}$	$28.80_{15.29}$	$89.56_{14.75}$	$30.66_{26.26}$	$98.15_{1.36}$	$16.62_{5.91}$	$83.70_{16.67}$	$27.30_{43.75}$

Table 1: Performance comparison on five medical segmentation datasets under 1-, 4-, and 16-shot settings. Results are reported as mean_{std}. Boldface indicates the best performance (Dice ↑, HD95 ↓).

Method	*Drift* ↓	Gap_{PT} ↓
MoCo v3 [Chen *et al.*, 2021]	$0.448_{0.045}$	$0.315_{0.043}$
MAE [He *et al.*, 2022]	$0.463_{0.053}$	$0.704_{0.086}$
CLIP [Radford *et al.*, 2021]	$0.244_{0.021}$	$0.289_{0.086}$
SSL-ALPNet [Ouyang *et al.*, 2020]	$0.523_{0.023}$	$0.646_{0.171}$
UniverSeg [Butoi *et al.*, 2023]	$0.233_{0.034}$	$0.263_{0.042}$
MedCoSS [Ye *et al.*, 2024]	$0.155_{0.031}$	$0.439_{0.074}$
Ours (BOG-PRETRAIN)	$\mathbf{0.052}_{0.012}$	$\mathbf{0.256}_{0.035}$

Table 2: Drift–Gap diagnostics averaged over five datasets (Brain Tumors, Breast Tumors, Kvasir, Lung X-ray, Thyroid/tn3k). Values are mean_{std} across datasets (lower is better).

higher Dice indicates more coherent prototype anchors, consistent with a reduced interior semantic gap (Gap_{in}): by down-weighting unreliable report semantics, the pre-trained embedding becomes less sensitive to noisy/global text, leading to stronger overlap across modalities and annotation settings (e.g., large Dice gains on Kvasir and the cross-dataset Breast Tumors benchmark). Second, we emphasize HD95 because it reflects boundary adherence that Dice can miss; our boundary-aware pre-training reduces the boundary gap (Gap_{bd}), which is most evident on irregular structures (e.g., Kvasir shows a pronounced HD95 drop), matching the boundary-gap trends in §4.3.

We note that on **Brain Tumors** (1-shot) the HD95 improvement is less consistent. With a single support mask and highly irregular tumor contours, boundary refinement remains strongly support-dependent and HD95 is sensitive to small outlier regions, a known challenge even for supervised segmenters, while BOG-PRETRAIN still yields the best Dice, indicating more reliable localization under extreme label scarcity. Finally, the gap to nnU-Net reflects the inherent cost–performance trade-off between fully-supervised dense annotation and few-shot generalization: our goal is to improve the efficiency frontier under scarce supervision, and BOG-PRETRAIN consistently advances both overlap and boundary quality among few-shot methods, making it a strong candidate for annotation-scarce clinical scenarios.

4.3 Mechanism Analysis: Decoupling the Gains

To move beyond performance tables and interpret *why* our method works, we leverage the proposed diagnostic metrics to decouple the contributions of Global, Local, and Consistency modules.

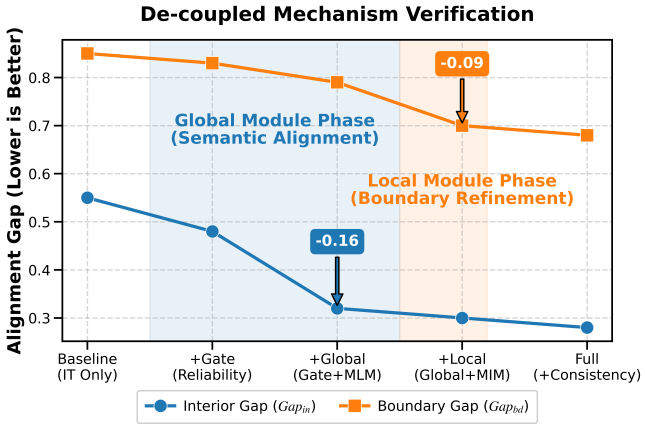

Figure 3: **Step-wise Gap Reduction.** Adding the Global module drastically reduces the Interior Gap (Semantic), whereas the Local module specifically targets the Boundary Gap (Structural), confirming the spatial decoupling of our objectives.

1. Spatially Decoupled Improvements (Global vs. Local). As visualized in Figure 3, the ablation reveals a clear functional separation: Reliability-Gated Alignment delivers a large reduction in Gap_{in}, confirming that global contrastive learning aggregates object-interior pixels into coherent semantic clusters; once Gap_{in} saturates, adding SG-MIM triggers a secondary drop specifically in Gap_{bd}, proving the local objective captures high-frequency boundary details overlooked by global objectives. Figure 4 further validates this

Objective	Dice ↑	HD95 ↓	Gap_{in} ↓	Gap_{bd} ↓	*Drift* ↓
MoCo v3	63.25	46.00	0.33	0.79	0.448
MAE	65.33	38.73	0.45	0.71	0.463
CLIP	64.61	49.99	0.30	0.85	0.244
Ours	**75.24**	**29.37**	**0.28**	**0.69**	**0.052**

Table 3: Pre-training objective comparison (16-shot avg over datasets). We report downstream performance and episodic diagnostics (lower is better for gaps/drift).

Variant	Dice ↑	HD95 ↓	Gap_{in} ↓	Gap_{bd} ↓	*Drift* ↓
Full (G+L+C)	**75.24**	**29.37**	**0.28**	**0.69**	**0.052**
w/o Global	72.06	34.48	0.31	0.70	0.101
w/o Local	70.29	43.23	0.27	0.76	0.147
w/o Cons.	72.72	34.49	0.32	0.74	0.215

Table 4: Component ablation study (avg over datasets). "Global" mainly targets Gap_{in}, "Local" targets Gap_{bd}, and "Cons." primarily reduces *Drift*.

via strong correlations (r>0.9) between Gap_{in} and Dice and between Gap_{bd} and HD95, suggesting that the two diagnostic axes are complementary and neither alone suffices to predict overall segmentation quality.

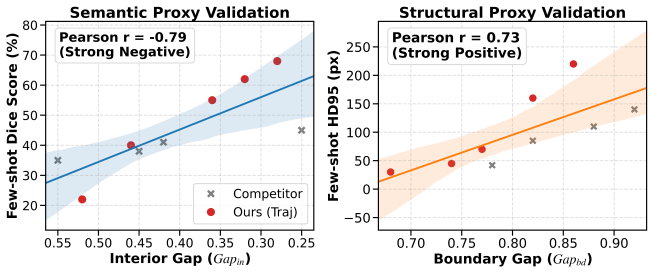

Figure 4: **Metric Validation.** The strong negative correlation between Gap_{in} and Dice, and positive correlation between Gap_{bd} and HD95, validates that our diagnostic metrics are faithful proxies for downstream performance.

2. Efficiency and Stability (The Role of Consistency). Projecting models onto the Drift–Gap plane (Fig. 5) reveals an orthogonal two-phase trajectory: Phase 1 (Global + Local) drives a vertical descent that closes the Gap, while Phase 2 (Consistency) drives a horizontal shift that reduces *Drift* by over 80% (from ∼0.48 to 0.08) without further changing Gap, indicating that the two improvement axes are largely independent. This shows that consistency constraints do not merely fine-tune accuracy but fundamentally restructure the feature space to resist few-shot adaptation shifts, enabling effective generalization even under extreme label scarcity.

4.4 Ablation Studies

Table 3 links end-task performance to our decomposed Objective Gap. Compared with generic SSL/MIM/VLP objectives, BOG-PRETRAIN achieves the best Dice/HD95 while lowering both Gap_{in} and Gap_{bd}, indicating improved intrinsic matchability for mask-conditioned dense prototype

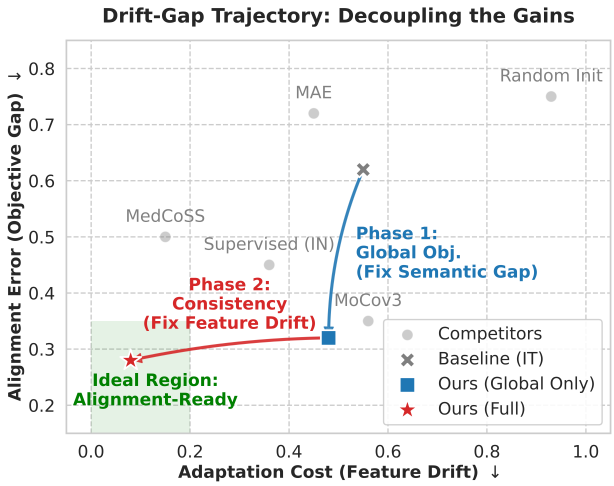

Figure 5: **Drift-Gap Trajectory.** The evolution of our method exhibits an orthogonal improvement path: Phase 1 reduces the Objective Gap via Global/Local alignment, while Phase 2 (Consistency) slashes Feature Drift, shifting the model into the efficiency-optimal ideal region.

matching. Notably, BOG-PRETRAIN also yields the lowest *Drift*, suggesting that its embedding geometry requires minimal re-organization during few-shot fine-tuning.

Table 4 attributes improvements to specific components under the Drift–Gap lens. Removing **Global** increases Gap_{in} (0.28→0.31) and reduces Dice (75.24→72.06), consistent with reliability-gated alignment protecting prototype anchors from noisy report semantics. Removing **Local** markedly increases Gap_{bd} (0.69→0.76) and causes the largest HD95 degradation (29.37→43.23), confirming that boundary-aware structural modeling is essential for closing the boundary gap and sharpening contours. Finally, removing **Consistency** produces the largest *Drift* increase (0.052→0.215) with comparatively smaller gap changes, showing that consistency regularization primarily stabilizes episodic metric geometry rather than merely improving intrinsic alignment.

5 Conclusion

We proposed BOG-PRETRAIN, a unified pre-training framework that bridges the *objective gap* between representation learning and episodic few-shot medical segmentation. BOG-PRETRAIN combines (i) reliability-gated image–text alignment to mitigate noisy reports, (ii) semantic- and edge-guided masked reconstruction to enhance boundary-sensitive structure, and (iii) consistency regularization to stabilize episodic metric geometry. Experiments on five public benchmarks show consistent improvements over strong baselines in both Dice and HD95, especially in low-shot regimes where labels are scarce. Moreover, our Drift–Gap diagnostics and mechanism analyses show that Gap_{in}, Gap_{bd}, and feature drift decrease step-wise with each component and correlate strongly with downstream performance. We hope BOG-PRETRAIN encourages objective-driven pre-training designs that narrow the pre-training–to–few-shot segmentation mismatch for accurate and stable label-efficient medical segmentation.

Acknowledgments

This work was supported by the Shenzhen Science and Technology Program (RCBS20231211090723042).

References

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

[Azizi *et al.*, 2021] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, Vivek Natarajan, and Mohammad Norouzi. Big self-supervised models advance medical image classification. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 3458–3468. IEEE, 2021.

[Boecking *et al.*, 2022] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie L. Hyland, Maria Wetscherek, Tristan Naumann, Aditya V. Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision-language processing. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI*, volume 13696 of *Lecture Notes in Computer Science*, pages 1–21. Springer, 2022.

[Butoi *et al.*, 2023] Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R. Sabuncu, John V. Guttag, and Adrian V. Dalca. Universeg: Universal medical image segmentation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 21381–21394. IEEE, 2023.

[Byra *et al.*, 2020] Michal Byra, Piotr Jarosik, Aleksandra Szubert, Michael Y. Galperin, Haydee Ojeda-Fournier, Linda K. Olson, Mary O'Boyle, Christopher Comstock, and Michael P. André. Breast mass segmentation in ultrasound with selective kernel u-net convolutional neural network. *Biomed. Signal Process. Control.*, 61:102027, 2020.

[Chen *et al.*, 2017] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.

[Chen *et al.*, 2021] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9620–9629. IEEE, 2021.

[Cheng *et al.*, 2015] Jun Cheng, Wei Huang, Shuangliang Cao, Ru Yang, Wei Yang, Zhaoqiang Yun, Zhijian Wang, and Qianjin Feng. Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PloS one*, 10(10):e0140381, 2015.

[Chowdhury *et al.*, 2020] Muhammad Enamul Hoque Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al-Emadi, Mamun Bin Ibne Reaz, and Mohammad Tariqul Islam. Can AI help in screening viral and COVID-19 pneumonia? *IEEE Access*, 8:132665–132676, 2020.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[Gong *et al.*, 2021] Haifan Gong, Guanqi Chen, Ranran Wang, Xiang Xie, Mingzhi Mao, Yizhou Yu, Fei Chen, and Guanbin Li. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In *18th IEEE International Symposium on Biomedical Imaging, ISBI 2021, Nice, France, April 13-16, 2021*, pages 257–261. IEEE, 2021.

[He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15979–15988. IEEE, 2022.

[Jha *et al.*, 2020] 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 Yong Man Ro, Wen-Huang Cheng, Junmo Kim, Wei-Ta Chu, Peng Cui, Jung-Woo Choi, Min-Chun Hu, and Wesley De Neve, editors, *MultiMedia Modeling - 26th International Conference, MMM 2020, Daejeon, South Korea, January 5-8, 2020, Proceedings, Part II*, volume 11962 of *Lecture Notes in Computer Science*, pages 451–462. Springer, 2020.

[Koleilat *et al.*, 2025] Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Medclip-samv2: Towards universal text-driven medical image segmentation. *Medical Image Anal.*, 106:103749, 2025.

[Litjens *et al.*, 2017] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Anal.*, 42:60–88, 2017.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC,*

Canada, October 10-17, 2021, pages 9992–10002. IEEE, 2021.

[Ouyang *et al.*, 2020] Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIX*, volume 12374 of *Lecture Notes in Computer Science*, pages 762–780. Springer, 2020.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[Rahman *et al.*, 2021] Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas M. Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al-Máadeed, Susu M. Zughaier, Muhammad Salman Khan, and Muhammad Enamul Hoque Chowdhury. Exploring the effect of image enhancement techniques on COVID-19 detection using chest x-ray images. *Comput. Biol. Medicine*, 132:104319, 2021.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells III, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015.

[Siragusa *et al.*, 2025] Irene Siragusa, Salvatore Contino, Massimo La Ciura, Rosario Alicata, and Roberto Pirrone. Medpix 2.0: A comprehensive multimodal biomedical data set for advanced ai applications with retrieval augmented generation and knowledge graphs. *Data Science and Engineering*, pages 1–17, 2025.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4077–4087, 2017.

[Tang *et al.*, 2025] Song Tang, Shaxu Yan, Xiaozhi Qi, Jianxin Gao, Mao Ye, Jianwei Zhang, and Xiatian Zhu. Few-shot medical image segmentation with high-fidelity prototypes. *Medical Image Anal.*, 100:103412, 2025.

[Wang *et al.*, 2020] Haochen Wang, Xudong Zhang, Yutao Hu, Yandan Yang, Xianbin Cao, and Xiantong Zhen. Few-shot semantic segmentation with democratic attention networks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIII*, volume 12358 of *Lecture Notes in Computer Science*, pages 730–746. Springer, 2020.

[Wang *et al.*, 2022] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3876–3887. Association for Computational Linguistics, 2022.

[Ye *et al.*, 2024] Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, Qi Wu, and Yong Xia. Continual self-supervised learning: Towards universal multi-modal medical data representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 11114–11124. IEEE, 2024.