Sharpening Lightweight Models for Generalized Polyp Segmentation: A Boundary Guided Distillation from Foundation Models

Shivanshu Agnihotri
Dept. of Computer Science and Engineering
Malaviya National Institute of Technology
Jaipur, India, 302017
2024rcp9057@mnit.ac.in

Snehashis Majhi
INRIA Sophia Antipolis
Côte d'Azur University
France
snehashis.majhi@inria.fr

Deepak Ranjan Nayak
Dept. of Computer Science and Engineering
Malaviya National Institute of Technology
Jaipur, India, 302017
drnayak.cse@mnit.ac.in

***Abstract*—Automated polyp segmentation is critical for early colorectal cancer detection and its prevention, yet remains challenging due to weak boundaries, large appearance variations, and limited annotated data. Lightweight segmentation models such as U-Net, U-Net++, and PraNet offer practical efficiency for clinical deployment but struggle to capture the rich semantic and structural cues required for accurate delineation of complex polyp regions. In contrast, large Vision Foundation Models (VFMs), including SAM, OneFormer, Mask2Former, and DINOv2, exhibit strong generalization but transfer poorly to polyp segmentation due to domain mismatch, insufficient boundary sensitivity, and high computational cost. To bridge this gap, we propose *LiteBounD, a Lightweight Boundary-guided Distillation* framework that transfers complementary semantic and structural priors from multiple VFMs into compact segmentation backbones. LiteBounD introduces (i) a dual-path distillation mechanism that disentangles semantic and boundary-aware representations, (ii) a frequency-aware alignment strategy that supervises low-frequency global semantics and high-frequency boundary details separately, and (iii) a boundary-aware decoder that fuses multi-scale encoder features with distilled semantically rich boundary information for precise segmentation. Extensive experiments on both seen (Kvasir-SEG, CVC-ClinicDB) and unseen (ColonDB, CVC-300, ETIS) datasets demonstrate that LiteBounD consistently outperforms its lightweight baselines by a significant margin and achieves performance competitive with state-of-the-art methods, while maintaining the efficiency required for real-time clinical use. Our code is available at GitHub repository.**

***Index Terms*—Polyp Segmentation, Boundary-Guided Distillation, Foundation Models, High-Low Frequency Modulation**

I. Introduction

Colorectal Cancer (CRC) remains a serious global health concern, ranking third among all cancer types and contributing significantly to cancer-related deaths. Most CRC cases originate from colorectal polyps- an abnormal tissue growth on the inner colon lining. Early detection and removal of such polyps is paramount for effective prevention [1]. Although colonoscopy is considered as the clinical gold standard for polyp detection, this procedure is highly operator-dependent and prone to inter-observer variability. Moreover, reported polyp miss rates of 6–27% [2] raise serious concerns, indicating the urgent need for accurate and reliable computer-aided polyp segmentation methods, thereby assisting clinicians in making precise interventions.

A plethora of encoder-decoder Convolutional Neural Network (CNN)-based models have been proposed for polyp segmentation over the past few years, including early architectures such as U-Net and its variants [3], [4]. Despite achieving satisfactory performance, these early models often struggle to capture fine-grained boundary details. Following this, several architectures such as PraNet [5], MSNet [6], SFA [7], and M^{2}SNet [8], have been introduced to enhance boundary awareness and handle scale variations of polyps. While these methods achieve notable improvements, their limited receptive fields hinder modeling of global contextual dependencies, leading to limited generalization. To this end, Vision Transformer (ViT)-based polyp segmentation models, such as CTNet [9], MCT-Net [10], PVT-Cascade [11], and Polyp-PVT [12], have been developed, owing to their ability to capture global relationships through self-attention, thereby resulting in remarkable performance improvements. However, such models typically incur high computational costs and demonstrate suboptimal generalization.

Despite these advances, inherent challenges, such as significant variations in polyp appearance and frequent occurrence of ambiguous or weak boundaries, continue to hinder robust polyp segmentation. Large-scale Vision Foundation Models (VFMs), including SAM [13], CLIP [14], OneFormer [15], MaskFormer [16], Mask2Former [17], and DINOv2 [18], have recently advanced segmentation by learning fine-grained visual representations with strong cross-domain generalization. However, their direct adaptation for polyp segmentation is constrained by insufficient domain-specific knowledge and substantial computational demands, hindering deployment in resource-constrained clinical environments. A recent work proposes SAM-Mamba [19] to improve generalization via adapter-based tuning and the Mamba-Prior module. However, it remains computationally intensive (e.g., 103M parameters

This work is supported by the Anusandhan National Research Foundation (ANRF), Government of India, under project number CRG/2023/007397 and ANRF/ARGM/2025/002890/TS.

and 423 GFLOPs), limiting its applicability in real-time clinical settings. While lightweight encoder–decoder models such as U-Net remain a de facto choice for medical segmentation, they often exhibit limited robustness to low-contrast polyps and demonstrate poor generalization. To bridge the gap between generalization and efficiency, the recently introduced Polyp-DiFoM [20] distills knowledge from multiple foundation models into a compact architecture, achieving strong generalization while maintaining efficiency. However, it still struggles to handle weak or ambiguous polyp boundaries effectively. these limitations highlight the necessity of a potential strategy that can effectively transfer the VFMs knowledge to lightweight models while preserving boundary-related cues.

To this end, we propose **LiteBounD** - a Lightweight Boundary-guided Distillation framework that transfers rich semantic and structural priors from VFMs into lightweight segmentation models. Additionally, LiteBounD employs a high-low frequency modulation strategy to decompose foundation model features into global and boundary-sensitive components, which are then distilled into lightweight baselines such as U-Net, U-Net++, and PraNet. This enables precise boundary refinement while preserving semantic coherence. Extensive experiments across five benchmark datasets—Kvasir-SEG, CVC-ClinicDB, ETIS, ColonDB, and CVC-300—demonstrate that LiteBounD consistently outperforms vanilla baselines, achieving superior accuracy and cross-dataset generalization under diverse imaging conditions.

In summary, our main contributions are as follows:

- **Modular Distillation Framework:** We introduce LiteBounD, a plug-and-play distillation pipeline that infuses boundary-aware priors from VFMs (SAM, DINOv2, OneFormer) into lightweight segmentation models (U-Net, U-Net++ and PraNet), enabling efficient and scalable deployment.
- **Boundary-Aware Distillation:** We derive rich semantic and boundary-aware representations from foundation models that accurately discriminate polyp from non-polyp regions. These representations are subsequently decomposed into low- and high-frequency features using a high–low modulation strategy. The resulting features are then distilled into the baseline network, enabling precise refinement of structural details while simultaneously strengthening global semantic understanding.
- **Superior Generalization:** LiteBounD demonstrates consistent performance gains across five public datasets under seen and unseen conditions, validating its robustness and generalization capability in varied clinical scenarios.

II. Methodology

We introduce **Lightweight Boundary-guided Distillation (LiteBounD)**, a modular framework that bridges the representational richness of Vision Foundation Models (VFMs) with the efficiency of lightweight segmentation backbones. LiteBounD distills complementary semantic and structural priors from SAM, DINOv2, and OneFormer into compact models such as U-Net and U-Net++, enabling precise polyp segmentation with minimal computational overhead. As illustrated in Fig. 1, LiteBounD orchestrates a multi-stage distillation pipeline that disentangles semantic and boundary cues, modulates them via frequency decomposition, and aligns them with latent representations in the lightweight baseline model.

A. Baseline Model Architecture

We redesign the encoder of a standard U-Net to produce disentangled latent representations that separately encode semantic and boundary-specific information. Given an input endoscopic image $I \in \mathbb{R}^{H\times W\times 3}$ (with $H = W = 352$), the encoder generates multi-scale feature maps as $I \rightarrow \{f_1, f_2, f_3, f_4\}$ where $f_i \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times C_i}$ and $C_i \in \{64, 128, 256, 512\}$. At the bottleneck layer, we extract four latent vectors: $\mathbf{L1}$: global semantic embedding via 1×1 convolution on f_4, $\mathbf{L2}$: refined local semantic embedding via convolution on $\mathbf{L1}$, $\mathbf{L3}$: edge-sensitive boundary embedding via convolution on $\mathbf{L2}$, $\mathbf{L4}$: higher-order boundary context via convolution on $\mathbf{L3}$. These vectors are grouped as semantic pair $\{\mathbf{L1}, \mathbf{L2}\}$ and boundary-aware pair $\{\mathbf{L3}, \mathbf{L4}\}$, serving as alignment targets for distillation.

B. Cross-Model Feature Extraction

a) Semantic Feature Aggregation.: We extract semantic embeddings from each VFM for the input image I. Let $F_i \in \mathbb{R}^{H_i\times W_i\times C_i}$ denote the feature map from the i-th model. All feature maps are resized to a common resolution $H' \times W'$ and concatenated to yield $f^{**} \in \mathbb{R}^{H'\times W'\times C^*}$:

$$C^* = C_{\text{SAM}} + C_{\text{DINOv2}} + C_{\text{OneFormer}}. \tag{1}$$

This unified semantic tensor captures complementary global context and serves as input to the frequency-aware distillation module.

b) Boundary-Aware Feature Extraction.: To isolate boundary-sensitive cues, we construct region-specific inputs using the ground-truth mask $M \in \mathbb{R}^{H\times W}$:

$$I_{polyp} = M \otimes I, \quad I_{non-polyp} = (1 - M) \otimes I. \tag{2}$$

These inputs are passed through each VFM to obtain polyp-aware and non-polyp-aware embeddings $f^i_{polyp}, f^i_{non-polyp} \in \mathbb{R}^{H_i\times W_i\times C_i}$. After resizing and concatenation we get $f_{polyp}, f_{non-polyp} \in \mathbb{R}^{H'\times W'\times C^*}$. We then apply cross-attention between f_{polyp} (key) and $f_{non-polyp}$ (query/value) to generate boundary-enhanced features:

$$f_{BA} = \text{CrossAttn}(f_{polyp}, f_{non-polyp}). \tag{3}$$

C. Frequency-Aware Distillation

To decompose features into structural and contextual components, we apply 2D FFT to f^{**} and f_{BA}:

$$F^{\text{freq}} = \text{FFT}(f^{**}), \quad F^{\text{freq}}_{ba} = \text{FFT}(f_{BA}). \tag{4}$$

Using binary masks $\text{MH}_{\text{HFF}}, \text{ML}_{\text{LFF}} \in \{0, 1\}^{H'\times W'}$, we isolate High-Frequency Features (HFF) and Low-Frequency Features (LFF):

$$F^{-1}_{\text{LFF}} = \text{IFFT}(\text{ML}_{\text{LFF}}, F^{\text{freq}}), F^{-1}_{\text{HFF}} \quad = \text{IFFT}(\text{MH}_{\text{HFF}}, F^{\text{freq}}) \tag{5}$$

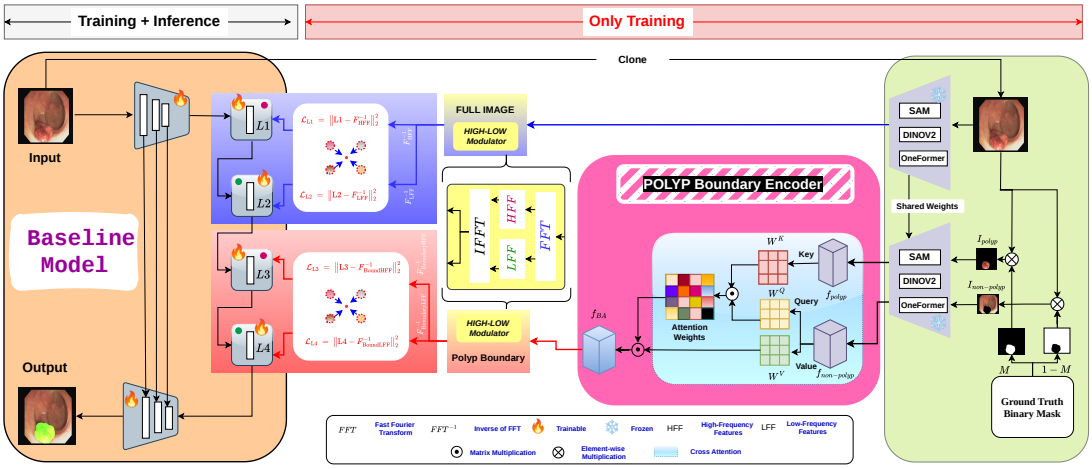


Fig. 1: Overall architecture of the proposed **LiteBounD** framework. .

$$F^{-1}_{\text{BoundLFF}} = \text{IFFT}(\text{ML}_{\text{LFF}}, F^{\text{freq}}_{ba}), F^{-1}_{\text{BoundHFF}} \quad = \text{IFFT}(\text{MH}_{\text{HFF}}, F^{\text{freq}}_{ba}) \tag{6}$$

To enable feature-level distillation, these high- and low-frequency features are mapped back to the spatial domain using a 2D inverse FFT. The resulting distillation-ready features F^{-1}_{LFF}, F^{-1}_{HFF}, F^{-1}_{BoundLFF}, and F^{-1}_{BoundHFF}, are then injected into the baseline to guide both semantic and boundary-guided learning.

D. Latent Alignment Losses

LiteBounD transfers VFM knowledge into the lightweight backbone through a dual-path latent alignment strategy that supervises semantic and structural representations separately. This design leverages the well-established observation that low-frequency features encode global semantics, while high-frequency features capture boundary-level details and rapid spatial variations. By aligning latent vectors with frequency-decomposed VFM features, LiteBounD enforces explicit representation disentanglement within the student model.

a) Semantic Alignment.: Semantic alignment transfers global contextual priors and shape-level consistency from VFMs. The semantic latent vector **L1** is aligned with the low-frequency features F^{-1}_{LFF} derived from the unified semantic embedding, while **L3** is aligned with F^{-1}_{BoundLFF} extracted from boundary-aware features. These low-frequency signals encode dominant structural patterns and coarse object geometry, which are essential for stable polyp localization under varying imaging conditions.

$$\mathcal{L}_{\text{L1}} = \frac{1}{\text{HWC}^*} \sum_{x,y,z} \left\| L1(x,y,z) - F^{-1}_{\text{LFF}}(x,y,z) \right\|_2^2 \tag{7}$$

$$\mathcal{L}_{\text{L3}} = \frac{1}{\text{HWC}^*} \sum_{x,y,z} \left\| L3(x,y,z) - F^{-1}_{\text{BoundLFF}}(x,y,z) \right\|_2^2 \tag{8}$$

This supervision encourages the student model to internalize VFM-level semantic abstraction, improving robustness to polyp shape variability and reducing false negatives.

b) Structural Alignment.: Structural alignment focuses on transferring fine-grained boundary cues that are critical for accurate delineation of polyp margins. The latent vectors **L2** and **L4** are aligned with high-frequency features F^{-1}_{HFF} and F^{-1}_{BoundHFF}, respectively. These features emphasize sharp transitions and edge discontinuities, which strongly correlate with anatomical boundaries in endoscopic imagery.

$$\mathcal{L}_{\text{L2}} = \frac{1}{\text{HWC}^*} \sum_{x,y,z} \left\| L2(x,y,z) - F^{-1}_{\text{HFF}}(x,y,z) \right\|_2^2, \tag{9}$$

$$\mathcal{L}_{\text{L4}} = \frac{1}{\text{HWC}^*} \sum_{x,y,z} \left\| L4(x,y,z) - F^{-1}_{\text{BoundHFF}}(x,y,z) \right\|_2^2 \tag{10}$$

By supervising boundary-aware latent vectors with high-frequency signals, LiteBounD enhances sensitivity to subtle contour variations and reduces over-smoothing—limitations commonly observed in CNN-based medical segmentation.

Overall Impact. The combination of semantic and structural alignment enables LiteBounD to jointly capture global coherence and boundary precision. This dual-path distillation allows lightweight models to approximate the representational richness of VFMs while retaining real-time efficiency.

E. Boundary-Aware Decoder

The boundary-aware decoder reconstructs high-quality polyp masks by fusing multi-scale encoder features with distilled semantic and structural cues. It receives $\{f_1, f_2, f_3, f_4\}$ along with frequency-separated features, where low-frequency features F^{-1}_{LFF} guide semantic refinement in stages ($L1$, $L3$), and high-frequency features F^{-1}_{HFF} sharpen boundary-focused stages ($L2$, $L4$). This targeted fusion injects global contextual priors and boundary-sensitive structural information directly into the decoding pathway. By jointly enhancing semantic coherence and edge precision, the decoder produces accurate, pixel-level segmentation masks that effectively leverage the distilled knowledge from VFMs.

F. Multi-Phase Training

To effectively balance representation learning with knowledge transfer, we adopt a three phased training protocol which enables the model to progressively refine itself for robust and generalizable polyp segmentation. **Phase I: Task-Specific Pre-training** We first train a standard U-Net architecture from scratch using only segmentation supervision though combined segmentation loss:

$$\mathcal{L}_{\text{phase1}} = \mathcal{L}_{\text{bce}} + \mathcal{L}_{\text{dice}} \tag{11}$$

TABLE I: Quantitative comparison of LiteBounD-guided baselines against state-of-the-art methods on seen datasets.

Methods	Params (M)	FLOPs (G)	Kvasir-SEG (Seen)						CVC-ClinicDB (Seen)					
			mDice ↑	mIoU ↑	F_β^w ↑	S_α ↑	$E_\phi^{\max}$ ↑	MAE ↓	mDice ↑	mIoU ↑	F_β^w ↑	S_α ↑	$E_\phi^{\max}$ ↑	MAE ↓
State-of-the-art Methods (Without Boundary Awareness)														
SANet [21] (MICCAI 2021)	23.8	11.3	90.4	84.7	89.2	91.5	95.3	2.8	91.6	85.9	90.9	93.9	97.6	1.2
MSNet [6] (MICCAI 2021)	27.6	17.0	90.7	86.2	89.3	92.2	94.4	2.8	92.1	87.9	91.4	94.1	97.2	0.8
Polyp-PVT [12] (CAAI 2023)	25.1	10.1	91.7	86.4	91.1	92.5	95.6	2.3	93.7	88.9	93.6	94.9	98.5	0.6
M^2SNet [8] (arXiv 2023)	27.7	17.1	91.2	86.1	90.1	92.2	95.3	2.5	92.2	88.0	91.7	94.2	97.0	0.9
PVT-Cascade [11] (WACV 2023)	35.2	32.5	91.1	86.3	90.6	91.9	96.1	2.5	91.9	87.2	91.8	93.6	96.9	1.3
CTNet [9] (IEEE TCYB 2024)	44.2	32.6	91.7	86.3	91.0	92.8	95.9	2.3	93.6	88.7	93.4	95.2	98.3	0.6
SAM-Mamba [19] (WACV 2025)	103.0	423.0	92.4	87.3	94.2	93.6	96.1	2.5	94.2	88.7	94.3	95.5	98.2	0.6
State-of-the-art Methods (With Boundary Awareness)														
CFA-Net [22] (PR 2023)	25.2	55.3	91.5	86.1	90.3	92.4	96.2	2.3	93.3	88.3	92.4	95.0	98.9	0.7
MEGANet [23] (WACV 2024)	44.1	28.8	91.3	86.3	90.7	91.8	95.9	2.5	93.8	89.4	94.0	95.0	98.6	0.6
Lightweight Baseline Methods														
U-Net [3] (MICCAI 2015)	16.7	73.9	81.8	74.6	79.4	85.8	89.3	5.5	82.3	75.5	81.1	88.9	95.4	1.9
U-Net++ [24] (DLMIA 2018)	9.1	65.9	82.1	74.3	80.8	86.2	91.0	4.8	79.4	72.9	78.5	87.3	93.1	2.2
PraNet [5] (MICCAI 2020)	30.4	13.1	89.8	84.0	88.5	91.5	94.8	3.0	89.9	84.9	89.6	93.6	97.9	0.9
Lightweight Baselines Enhanced By Polyp-DiFoM [20] (WACV 2026)														
U-Net + Polyp-DiFoM	16.9	74.7	86.6	76.4	85.3	94.1	91.0	4.1	93.9	88.7	92.7	96.9	96.3	1.1
U-Net++ + Polyp-DiFoM	9.6	66.4	84.7	74.7	83.3	93.7	92.9	4.8	89.5	83.4	91.1	96.7	95.4	1.9
PraNet + Polyp-DiFoM	31.4	13.5	90.9	84.7	88.2	94.6	91.9	3.4	94.2	91.2	92.9	99.1	98.0	1.4
Lightweight Baselines Further Enhanced By LiteBounD (Ours)														
U-Net + Ours	17.0	73.1	86.9 (+5.1)	77.6 (+3.0)	86.1 (+6.7)	94.3 (+8.5)	91.7 (+2.4)	3.9 (-1.6)	94.8 (+12.5)	89.7 (+14.2)	93.4 (+12.3)	97.6 (+8.7)	97.0 (+1.6)	0.9 (-1.0)
U-Net++ + Ours	11.1	67.3	85.3 (+3.2)	75.0 (+0.7)	83.5 (+2.7)	93.9 (+7.7)	93.0 (+2.0)	4.7 (-0.1)	92.5 (+13.1)	87.8 (+14.9)	93.3 (+14.8)	98.2 (+10.9)	96.0 (+2.9)	1.6 (-0.6)
PraNet + Ours	32.8	14.8	91.9 (+2.1)	84.1 (+0.1)	88.3 (-0.2)	94.1 (+2.6)	92.2 (-2.6)	3.2 (+0.2)	95.7 (+5.8)	91.3 (+6.4)	93.4 (+3.8)	99.0 (+5.4)	97.8 (-0.1)	1.1 (+0.2)

where $\mathcal{L}_{\text{bce}}$ and $\mathcal{L}_{\text{dice}}$ denote the binary cross-entropy and dice losses, respectively. **Phase II: Integrated Distillation** In this phase, we activate the distillation modules, allowing the network to integrate both segmentation targets and rich semantic, boundary-aware priors extracted from the foundation models. The overall objective is expressed as:

$$\mathcal{L}_{\text{phase2}} = \lambda_1 \left(\mathcal{L}_{\text{bce}} + \mathcal{L}_{\text{dice}}\right) + \lambda_2 \mathcal{L}_{\text{L1}} + \lambda_3 \mathcal{L}_{\text{L2}} + \lambda_4 \mathcal{L}_{\text{L3}} + \lambda_5 \mathcal{L}_{\text{L4}} \quad (12)$$

where, we set $\lambda_1 = 0.6$ and $\lambda_2 = \lambda_3 = \lambda_4 = \lambda_5 = 0.1$ to prioritize semantic learning through segmentation while encouraging structural consistency via distillation. **Phase III: Targeted Mask Distillation** To stabilize the learned representations and increase decoding capability, we freeze the encoder in the final phase and optimize only the decoder and distillation pathways. The loss remains same as Phase II, enabling focused refinement of mask while reducing overfitting.

III. Experiments

A. Datasets and Performance Metrics

Datasets: To evaluate the generalization and robustness of the proposed framework, we conduct experiments on five standard public benchmarks: Kvasir-SEG [25], CVC-ClinicDB [26], ETIS [27], CVC-ColonDB [6], and EndoScene [28]. To ensure a fair comparison with recent state-of-the-art methods, we adopt the same data split protocol as in [5]. In total, 1450 images (900 from Kvasir-SEG and 550 from CVC-ClinicDB) are used for training and the remaining 100 images from Kvasir-SEG and 62 images from CVC-ClinicDB are kept for testing. Additionally, to rigorously assess the generalization capability, three unseen datasets are considered: ETIS (196 images), CVC-ColonDB (380 images), and CVC-300 (60 images). **Metrics:** The performance of our model, as well as state-of-the-art methods, is quantitatively assessed using six standard metrics: the mean IoU (mIoU), mean Dice (mDice), Structure-measure (S_α), weighted F-measure (F_β^w), Enhanced-alignment measure ($E_\phi^{\max}$), and Mean Absolute Error (MAE).

B. Implementation Details

The model is implemented using PyTorch framework on a NVIDIA Tesla V100 GPU (32 GB). All input images are rescaled to a resolution of 352×352 pixels. For augmentation, we apply random horizontal/vertical flipping and multi-scale resizing with factors $\{0.75, 1.0, 1.25\}$. Additionally, we perform training using the Adam optimizer with an initial learning rate of 1×10^{-4} and default momentum parameters. In our work, training follows the three-phase strategy as described in Section II-F for a total of 120 epochs (Phase I: Epochs 1–40, Phase II: Epochs 41–80 and Phase III: Epochs 81–120).

IV. Results Analysis

A. Quantitative Comparison

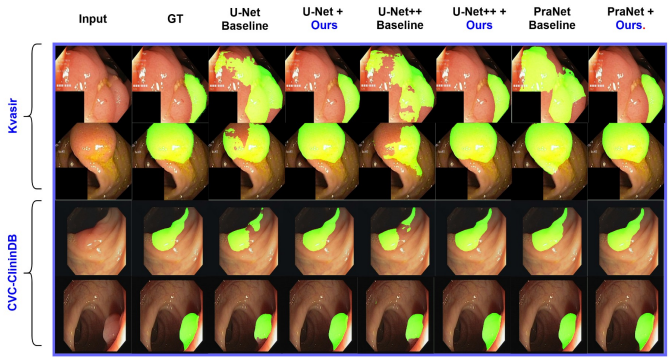
To evaluate the effectiveness of the proposed LiteBounD, we choose three standard lightweight segmentation baselines: U-Net [3], U-Net++ [24] and PraNet [5].

Performance on Seen Datasets: Results on the seen datasets (Kvasir-SEG and CVC-ClinicDB) are summarized in Table I. Our framework consistently outperforms all baseline architectures with minimal additional parameters and FLOPs. For the U-Net backbone, the Dice score improves by +5.1% (81.8% → 86.9%) on Kvasir-SEG and +12.5% (79.4% → 94.8%) on CVC-ClinicDB, with only a 0.3M parameter increase (16.7M → 17.0M). For U-Net++, it achieves gains of +3.2% and +13.1% on Kvasir-SEG and CVC-ClinicDB, respectively. For PraNet, it yields 95.7% Dice and 91.3% mIoU on CVC-ClinicDB, surpassing existing SOTA methods while maintaining high efficiency. Further, we compare LiteBounD with Polyp-DiFoM [20] under a fair three-foundation-model (3F) setting. As shown in Table I, LiteBounD consistently outperforms Polyp-DiFoM, achieving +3% Dice and +4.4% mIoU gains on CVC-ClinicDB (U-Net++) and +1% Dice on Kvasir-SEG (PraNet). These results highlight the effectiveness of boundary-guided distillation in improving performance with fewer VFMs.

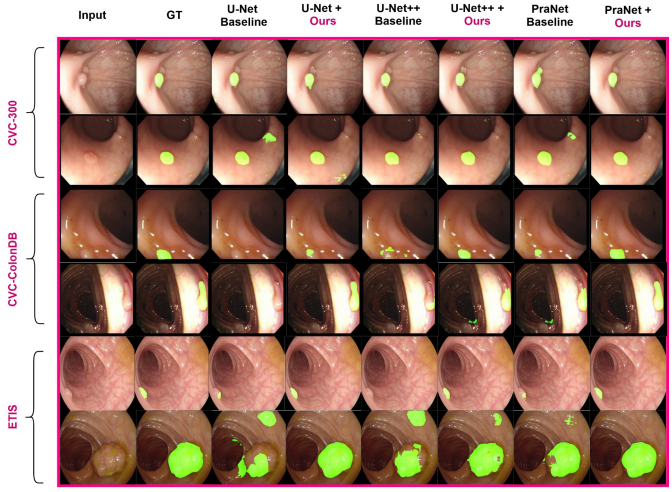
TABLE II: Quantitative comparison of LiteBounD-guided baselines against state-of-the-art methods on unseen datasets.

Methods	CVC-300 (Unseen)						CVC-ColonDB (Unseen)						ETIS (Unseen)					
	mDice ↑	mIoU ↑	F_β^w ↑	S_α ↑	$E_\phi^{\max}$ ↑	MAE ↓	mDice ↑	mIoU ↑	F_β^w ↑	S_α ↑	$E_\phi^{\max}$ ↑	MAE ↓	mDice ↑	mIoU ↑	F_β^w ↑	S_α ↑	$E_\phi^{\max}$ ↑	MAE ↓
State-of-the-art Methods (Without Awareness																		
SANet [21]	88.8	81.5	85.9	92.8	97.2	0.8	75.3	67.0	72.6	83.7	87.8	4.3	75.0	65.4	68.5	84.9	89.7	1.5
MSNet [6]	86.9	80.7	84.9	92.5	94.3	1.0	75.5	67.8	73.7	83.6	88.3	4.1	71.9	66.4	67.8	84.0	83.0	2.0
Polyp-PVT [12]	90.0	83.3	88.4	93.5	97.3	0.7	80.8	72.7	79.5	86.5	91.3	3.1	78.7	70.6	75.0	87.1	90.6	1.3
M^2SNet [8]	90.3	84.2	88.1	93.9	96.5	0.9	75.8	68.5	73.7	84.2	86.9	3.8	74.9	67.8	71.2	84.6	87.2	1.7
PVT-Cascade [11]	89.2	82.4	87.3	93.2	95.9	0.9	78.1	71.0	77.9	85.5	89.6	3.1	78.6	71.2	75.9	87.2	89.6	1.3
CTNet [9]	90.8	84.4	89.4	97.5	97.5	0.6	81.3	73.4	80.1	87.4	91.5	2.7	81.0	73.4	77.6	88.6	91.3	1.4
SAM-Mamba [19]	92.0	86.1	88.8	94.6	98.1	0.6	85.3	77.1	85.6	89.8	93.3	1.7	84.8	78.2	85.5	91.6	93.3	1.0
State-of-the-art Methods (With Boundary Awareness)																		
CFA-Net [22]	89.3	82.7	93.8	87.5	97.8	0.8	74.3	66.5	72.8	83.5	89.8	3.9	73.2	65.5	69.3	84.5	89.2	1.4
MEGANet [23]	89.9	83.4	88.2	93.5	96.9	0.7	79.3	71.4	77.9	85.4	89.5	4.0	73.9	66.5	70.2	83.6	85.8	3.7
Lightweight Baseline Methods																		
U-Net [3]	71.0	62.7	68.4	84.3	87.6	2.2	51.2	44.4	49.8	71.2	77.6	6.1	39.8	33.5	36.6	68.4	74.0	3.6
U-Net++ [24]	70.7	62.4	68.7	83.9	89.8	1.8	48.3	41.0	46.7	69.1	76.0	6.4	40.1	34.4	39.0	68.3	77.6	3.5
PraNet [5]	87.1	79.7	84.3	92.5	97.2	1.0	70.9	64.0	69.6	81.9	86.9	4.5	62.8	56.7	60.0	79.4	84.1	3.1
Lightweight Baselines Enhanced with Polyp-DiFoM [20]																		
U-Net + PolypDiFoM	82.3	73.3	74.7	89.4	86.9	1.6	68.3	57.4	60.9	81.4	77.1	4.7	54.3	42.4	47.1	76.6	70.1	2.6
U-Net++ + PolypDiFoM	77.8	70.9	75.9	92.7	90.1	1.5	67.7	54.1	64.0	87.1	80.0	5.1	51.8	42.7	49.0	80.1	72.0	3.2
PraNet + PolypDiFoM	87.4	79.9	87.1	95.9	97.9	0.8	74.0	64.3	73.0	88.4	85.9	3.9	71.5	58.2	65.6	87.0	83.5	2.2
Lightweight Baselines Further Enhanced By LiteBounD																		
U-Net + Ours	84.3 (+13.3)	75.1 (+12.4)	75.7 (+7.3)	91.3 (+7.0)	87.1 (-0.5)	1.3 (-0.9)	68.8 (+17.6)	58.2 (+13.8)	61.6 (+11.8)	81.6 (+10.4)	77.3 (-0.3)	4.8 (-1.3)	54.9 (+15.1)	43.4 (+9.9)	48.2 (+11.6)	77.3 (+8.9)	72.9 (-1.1)	2.5 (-1.1)
U-Net++ + Ours	83.2 (+12.5)	73.3 (+10.9)	77.1 (+8.4)	93.3 (+9.4)	89.3 (-0.5)	1.4 (-0.4)	68.9 (+20.6)	57.7 (+16.7)	65.1 (+18.4)	88.2 (+19.1)	81.0 (+5.0)	4.8 (-1.6)	52.7 (+12.6)	43.0 (+8.6)	50.1 (+11.1)	80.7 (+12.4)	75.6 (-2.0)	3.2 (-0.3)
PraNet + Ours	89.7 (+2.6)	81.7 (+2.0)	87.7 (+3.4)	96.5 (+4.0)	97.3 (+0.1)	0.7 (-0.3)	75.0 (+4.1)	66.7 (+2.7)	71.0 (+1.4)	90.3 (+8.4)	86.6 (-0.3)	4.1 (-0.4)	71.2 (+8.4)	60.1 (+3.4)	67.1 (+7.1)	88.3 (+8.9)	84.3 (+0.2)	2.0 (-1.1)

Performance on Unseen Datasets: On unseen datasets (CVC-300, CVC-ColonDB, and ETIS), as shown in Table II, U-Net demonstrates substantial improvements, particularly on CVC-300, where the Dice score increases from 71.0% to 84.3% (+13.3%) and mIoU from 62.7% to 75.1% (+12.4%). Similar gains are observed on other challenging datasets, highlighting enhanced generalization capability and the effectiveness of learning robust, transferable features. In addition, LiteBounD consistently outperforms PolypDiFoM, with notable gains such as +5.4% Dice on CVC-300 (U-Net++) and +2.4% mIoU for PraNet. These improvements highlight the effectiveness of our model in enhancing generalization.



(a) Seen datasets: Kvasir-SEG and CVC-ClinicDB.



(b) Unseen datasets: CVC-300, CVC-ColonDB, and ETIS.

Fig. 2: Qualitative comparison on seen and unseen datasets.

B. Qualitative Comparison

To further assess the effectiveness of our approach, we present qualitative comparisons between baseline models and their LiteBounD-enhanced counterparts across both seen and unseen datasets. As illustrated in Fig. 2 baseline architectures often under-perform in challenging scenarios—such as weak boundaries, small-scale polyps, or low-contrast regions, where predicted masks either miss subtle structures or include spurious edge artifacts. In contrast, LiteBounD variants consistently produce sharper boundaries and more complete polyp delineations. For instance, the U-Net baseline exhibits highly inconsistent behavior: in some cases, it completely fails to detect the polyp, while in others it incorrectly segments non-polyp regions as foreground. In contrast, our model delivers substantially improved segmentation accuracy, producing more consistent and reliable predictions across challenging samples. U-Net++ and PraNet can detect polyps but often produce coarse boundaries while our boundary-aware mechanism refines edge representations, producing sharp, precise, and anatomically accurate polyp boundaries.

C. Ablation Study

Polyp-DiFoM [20] has previously shown that progressively incorporating task-specific and semantically complementary foundation models leads to consistent improvements in segmentation performance. Specifically, the 1F (SAM only) and 2F (SAM + DINOv2) settings were shown to yield comparatively limited gains, indicating that richer multi-model supervision is necessary to fully exploit the benefits of distillation. Motivated by these findings, we directly adopt the 3F setting (SAM, DINOv2, and OneFormer) as our baseline distillation configuration, avoiding the less effective 1F and 2F variants. The primary objective of this ablation is to examine

TABLE III: Effectiveness of Boundary-Guided Distillation

Boundary-Guidance	Seen		Unseen		
	Kvasir	CVC-ClinicDB	CVC-300	CVC-ColonDB	ETIS
U-Net					
✗	86.6	93.9	82.3	68.3	54.3
✓	**86.9**	**94.8**	**84.3**	**68.8**	**54.9**
U-Net++					
✗	**84.7**	89.5	77.8	67.7	51.8
✓	84.1	**92.5**	**83.2**	**68.9**	**52.7**
PraNet					
✗	90.9	94.2	87.4	74.0	**71.5**
✓	**91.9**	**95.7**	**89.7**	**75.0**	71.2

the role of boundary awareness in challenging polyp segmentation scenarios. We perform experiments with and without the Boundary-Guided Distillation mechanism. As reported in Table III, the inclusion of mechanism consistently leads to improved performance, yielding notable gains in both Dice and IoU scores across baselines. These results confirm the effectiveness of explicitly modeling boundary information.

V. Conclusion

This paper introduces LiteBounD, a novel and effective boundary-guided distillation framework that leverages the rich representational capacity of large-scale vision foundation models to boost the performance of lightweight medical image segmentation architectures. By systematically generating strong semantic features from foundation models, converting them into rich boundary-aware representations, and fusing them into lightweight baselines, LiteBounD substantially enhances the accuracy, robustness, and generalizability of these baselines, while preserving their computational efficiency. Extensive evaluations conducted on five widely used polyp segmentation benchmarks demonstrate that LiteBounD achieves comparable or better performance than recent state-of-the-art methods with significantly lower computational overhead. Overall, LiteBounD enables real-time polyp segmentation with high precision in weak-boundary cases, without heavy computation.

References

[1] E. Morgan, M. Arnold *et al.*, "Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from globocan," *Gut*, vol. 72, no. 2, pp. 338–344, 2023.

[2] S. B. Ahn, D. S. Han, J. H. Bae, T. J. Byun, J. P. Kim, and C. S. Eun, "The miss rate for colorectal adenoma determined by quality-adjusted, back-to-back colonoscopies," *Gut and liver*, vol. 6, no. 1, p. 64, 2012.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[4] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, and H. D. Johansen, "Resunet++: An advanced architecture for medical image segmentation," in *2019 IEEE international symposium on multimedia (ISM)*. IEEE, 2019, pp. 225–2255.

[5] D.-P. Fan, G.-P. Ji *et al.*, "Pranet: Parallel reverse attention network for polyp segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 263–273.

[6] X. Zhao, L. Zhang, and H. Lu, "Automatic polyp segmentation via multi-scale subtraction network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 120–130.

[7] Y. Fang, C. Chen, Y. Yuan, and K.-y. Tong, "Selective feature aggregation network with area-boundary constraints for polyp segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 302–310.

[8] X. Zhao, H. Jia, Y. Pang, L. Lv, F. Tian, L. Zhang, W. Sun, and H. Lu, "M^2 snet: Multi-scale in multi-scale subtraction network for medical image segmentation," *arXiv preprint arXiv:2303.10894*, 2023.

[9] B. Xiao, J. Hu, W. Li, C.-M. Pun, and X. Bi, "Ctnet: Contrastive transformer network for polyp segmentation," *IEEE Transactions on Cybernetics*, vol. 54, no. 9, pp. 5040–5053, 2024.

[10] N. Chakraborti and D. R. Nayak, "Mct-net: a lightweight multiscale convolutional transformer network for polyp segmentation," in *2024 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2024, pp. 2944–2950.

[11] M. M. Rahman and R. Marculescu, "Medical image segmentation via cascaded attention decoding," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 6222–6231.

[12] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao, "Polyp-pvt: Polyp segmentation with pyramid vision transformers," *CAAI Artificial Intelligence Research*, vol. 2, p. 9150015, 2023.

[13] A. Kirillov, E. Mintun *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[14] A. Radford, J. W. Kim *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PmLR, 2021, pp. 8748–8763.

[15] J. Jain, J. Li, M. T. Chiu *et al.*, "Oneformer: One transformer to rule universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2989–2998.

[16] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 864–17 875, 2021.

[17] B. Cheng, I. Misra, A. G. Schwing *et al.*, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2022, pp. 1290–1299.

[18] M. Oquab, T. Darcet, T. Moutakanni *et al.*, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2024, featured Certification.

[19] T. K. Dutta, S. Majhi, D. R. Nayak, and D. Jha, "Sam-mamba: Mamba guided sam architecture for generalized zero-shot polyp segmentation," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 4655–4664.

[20] S. Agnihotri, S. Majhi, D. R. Nayak, and D. Jha, "From sam to dinov2: Towards distilling foundation models to lightweight baselines for generalized polyp segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026, pp. 1757–1766.

[21] J. Wei, Y. Hu, R. Zhang, Z. Li, S. K. Zhou, and S. Cui, "Shallow attention network for polyp segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 699–708.

[22] T. Zhou, Y. Zhou, K. He, C. Gong, J. Yang, H. Fu, and D. Shen, "Cross-level feature aggregation network for polyp segmentation," *Pattern Recognition*, vol. 140, p. 109555, 2023.

[23] N.-T. Bui, D.-H. Hoang *et al.*, "Meganet: Multi-scale edge-guided attention network for weak boundary polyp segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7985–7994.

[24] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *International Workshop on Deep Learning in Medical Image Analysis*. Springer, 2018, pp. 3–11.

[25] D. Jha, P. H. Smedsrud *et al.*, "Kvasir-seg: A segmented polyp dataset," in *International Conference on Multimedia Modeling*. Springer, 2019, pp. 451–462.

[26] J. Bernal, F. J. Sánchez *et al.*, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.

[27] J. Silva, A. Histace *et al.*, "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer," *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, no. 2, pp. 283–293, 2014.

[28] D. Vázquez, J. Bernal *et al.*, "A benchmark for endoluminal scene segmentation of colonoscopy images," *Journal of Healthcare Engineering*, vol. 2017, no. 1, p. 4037190, 2017.