Induce to Empower: Improving Lightweight Baselines via Foundation Model Induction for Generalized Polyp Segmentation

Shivanshu Agnihotri[1], Snehashis Majhi[2], Deepak Ranjan Nayak[1], Dwarikanath Mahapatra[3], and Debesh Jha[4]

[1] Malaviya National Institute of Technology Jaipur, India
[2] Côte d'Azur University, France
[3] Khalifa University, UAE
[4] Biomedical Perception & Intelligence Lab, University of South Dakota, USA
drnayak.cse@mnit.ac.in
https://github.com/lostinrepo/Lite-Pi

Abstract. Automated polyp segmentation in colonoscopy continues to pose challenges due to substantial appearance variations and indistinct polyp boundaries. Although emerging foundation models (FMs) such as DINOv2, SAM, and OneFormer demonstrate remarkable generalization capabilities, their direct transfer to the polyp segmentation task and deployment in real-time clinical settings are difficult due to lack of large-scale labeled data and high computational demands. In addition, adopting multiple FMs together raises concerns, even though they encode complementary semantic and structural information. While lightweight models, including U-Net, PraNet, and U-Net++, are computationally efficient, they often struggle to generalize across datasets due to limited representational capacity. To address this gap, we propose **Lite**-**P**olyp**I**nductor (Lite-π), a novel foundation model induction framework that significantly enhances lightweight polyp segmentation baselines. Our proposed framework generates FM-specific prototype representations and aligns them semantically with the corresponding foundation model priors through reconstruction-based supervision. Subsequently, transformer-based fusion is introduced to highlight the polyp-relevant representations, including salient boundary information, while preserving complementary semantic cues. Extensive experiments across five polyp segmentation benchmark datasets demonstrate that Lite-π significantly improves lightweight baselines, achieving superior generalization performance with minimal computational overhead and thereby, offering a practical solution for generalized polyp segmentation. The code is available at GitHub.

Keywords: Polyp Segmentation · Foundation Model Induction · Lite-π.

1 Introduction

Colorectal cancer (CRC) is a leading cause of cancer-related mortality worldwide, with its incidence increasing across diverse age groups. Most CRC cases

originate from benign colorectal polyps, making early detection and subsequent removal crucial for effective prevention. While colonoscopy is considered the gold standard for CRC screening, it is highly operator-dependent and subject to variability among clinicians. Further, the high polyp miss rates of 6–27% [2] during routine colonoscopy underscore the pressing need for computer-aided polyp segmentation approaches to aid clinicians in accurate detection and timely intervention.

Deep learning–based automated polyp segmentation has gained significant attention and has shown promising results in recent years. Specifically, encoder-decoder Convolutional Neural Network (CNN) architectures, including U-Net and its variants [21] [30] [14] have been widely adopted, but often struggle to preserve fine boundary details. Subsequently, various models such as PraNet [11], SFA [12], MSNet [28], M^2SNet [27], and CFA-Net [29], have been proposed to address boundary-related issues and scale variations in polyps. However, these methods fall short of modeling global contextual details, resulting in suboptimal performance. On the other hand, Vision Transformer (ViT) [8] and hybrid CNN-ViT models such as PVT-Cascade [20], Polyp-PVT [7], and CTNet [26] excel at modeling global and/or local relationships, yielding rich feature representations and significantly improving polyp segmentation accuracy. Despite strong performance improvements, high computational demands and limited generalization hinder their clinical applicability. Additionally, achieving robust polyp segmentation still remains challenging due to large variations in polyp appearance, ambiguous boundaries and cross-dataset domain shifts.

Foundation Models (FMs), such as SAM [17], DINOv2 [18], OneFormer [15], CLIP [19], MaskFormer [6] and Mask2Former [5], have significantly advanced image segmentation by learning rich visual representations from large-scale data that facilitates strong cross-domain generalization. However, their direct application to polyp segmentation is challenging due to data scarcity, high computational overhead, increased memory demands, and insufficient domain-specific knowledge. To mitigate these issues, SAM-Mamba [10] has been proposed, introducing adapter-based tuning via a Mamba-Prior module to improve generalization performance. Following this, SAM-MaGuP [9] has recently been introduced, incorporating a Mamba-guided boundary prior along with a 1D–2D Mamba adapter to address the weak-boundary challenge. However, these approaches incur substantial computational overhead, requiring approximately 103–106M parameters and 423–431 GFLOPs, which limits their real-time clinical deployment. Polyp-DiFoM [1], a distillation-based framework, mitigates deployment challenges by transferring enriched representations from multiple FMs into lightweight segmentation baselines. Although this framework significantly improves the performance of lightweight baselines, their performance and generalizability still fall short of meeting real-time clinical demands. This limitation stems from representational redundancies and task-bias introduced when leveraging multiple FMs.

Motivated by the above analysis, we propose **Lite-Polyp**Inductor (Lite-π), a novel framework, that introduces **Unified Foundation Model Inductor**

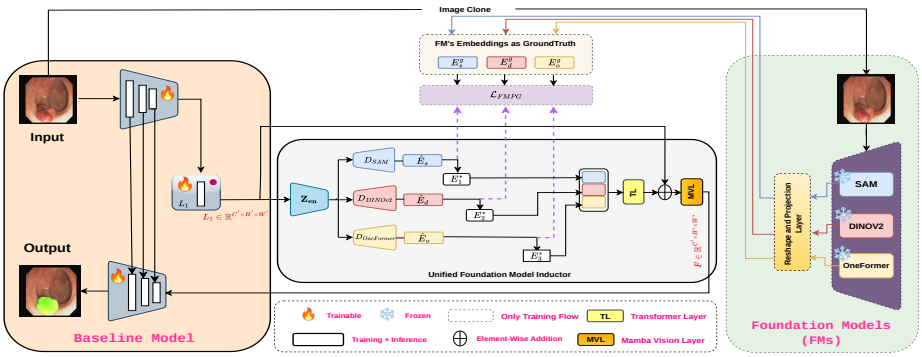


Fig. 1: Overall pipeline of the proposed **Lite-π** framework. It incorporates a novel inductor module which generates synthetic prototype embeddings and performs semantic alignment with structural priors from foundation models to enhance representation capability of lightweight segmentation baselines.

(UFMI) module to improve the performance of lightweight segmentation baselines (U-Net, U-Net++, and PraNet) with minimal additional computational overhead. Unlike conventional knowledge transfer approaches, Lite-π induces FM-specific synthetic prototype embeddings from the intermediate representations of a lightweight network and semantically aligns them with structural priors derived from corresponding FMs. Further, it selectively amplifies polyp-relevant representations, enabling enhanced global contextual dependency modeling and fine-grained feature refinement, thereby improving segmentation performance.Our framework is evaluated on five benchmark datasets, demonstrating significant performance gains and improved generalization over lightweight models, achieving near state-of-the-art (SoTA) performance.

Our **contributions** are three-fold: (1) We introduce Lite-π that leverages three FMs (SAM, DINOv2, and OneFormer) to inject task-specific and rich semantic representations into lightweight models for improved polyp segmentation. (2) We propose a novel inductor module, which generates prototype representations from the lightweight encoder and enforces semantic alignment with structural priors extracted from three FMs via a reconstruction loss, enriching feature representations of lightweight models while ensuring their alignment is relevant to polyp segmentation. (3) We demonstrate consistent performance improvements across five benchmark datasets with strong generalization capabilities with low computational overhead, making it well-suited for real-world clinical applications.

2 Methodology

Our proposed Lite-π framework consists of three key components: a **Baseline-Specific Encoder**, a **Unified Foundation Model Inductor (UFMI)**, and

a **Foundation Model Induced Decoder**, as illustrated in Figure 1. The Lite-π aims to effectively transfer crucial visual semantics and relevant task-specific information from three powerful FMs into lightweight models.

2.1 Baseline-Specific Encoder

Let $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ be the input colonoscopy image, which is fed to a lightweight encoder (U-Net, U-Net++, PraNet) consisting of a sequence of convolution layers, resulting in hierarchical feature maps $\{f_1, f_2, \ldots, f_n\}$. To extract a latent representation (L_1), we apply 1×1 convolution over the deepest feature map (f_n) of the encoder.

$$L_1 = \text{Conv}_{1\times1}(f_n) \in \mathbb{R}^{C^{'} \times H^{'} \times W^{'}} \tag{1}$$

These representations (L_1) serve as the base for the foundation model induction process.

2.2 Unified Foundation Model Inductor (UFMI)

The key innovation of our framework lies in the UFMI module, which refines the latent feature representations of lightweight baselines by generating FM prototypes relevant to polyps and performing semantic alignment with structural priors from foundation models (SAM, DINOv2, and OneFormer), thereby improving polyp segmentation.

Foundation Model Prototype Generation (FMPG) It generates prototype embeddings from L_1 corresponding to each foundation model: $\hat{E}_{\text{s}} \in \mathbb{R}^{C^* \times H^* \times W^*}$ for SAM, $\hat{E}_{\text{d}} \in \mathbb{R}^{C^* \times H^* \times W^*}$ for DINOv2 and $\hat{E}_{\text{o}} \in \mathbb{R}^{C^* \times H^* \times W^*}$ for OneFormer. This approach is introduced to selectively learn crucial priors essential for polyp segmentation while suppressing noise and redundancies from original FM embeddings. To achieve this, we adopt a lightweight encoder–decoder framework consisting of a single encoder followed by three parallel decoders. The encoder $\mathbf{Z_{en}}$ uses two 3×3 convolutional layers to project the latent embeddings to a low-dimensional latent space. While the decoder $D_i(.)$ employs a 3×3 convolution followed by a 1×1 convolution to project the resultant embeddings to FM-specific representations, while preserving the dimensionality. Finally, a 1×1 convolution is applied to match the dimensionality of the ground truth-embeddings generated from foundation models.

$$\mathbf{Z_{en}} = Conv_{3\times3}(Conv_{3\times3}(\mathbf{L}_1)) \tag{2}$$

$$\hat{E}_{\text{i}} = Conv_{1\times1}(D_i(\mathbf{Z_{en}})), i \in \{s, d, o\} \tag{3}$$

FM-based Semantic Feature Extraction We leverage three powerful FMs (SAM, DINOv2 and OneFormer) to extract semantically enriched feature embeddings $e_i \in \mathbb{R}^{c \times h \times w}$ from the input image I. These embeddings are then rescaled using bilinear interpolation followed by a 1×1 convolution to generate ground-truth embeddings $E_i^g \in \mathbb{R}^{C^* \times H^* \times W^*}$ while ensuring consistent spatial dimensions with $\hat{E}_{\text{i}}$.

Training Objective for FMPG Training of FMPG ensures better generation of synthetic prototypes specific to polyp characteristics. This is achieved through feature alignment between reconstructed embeddings ($\hat{E}_{\mathrm{s}}, \hat{E}_{\mathrm{d}}, \hat{E}_{\mathrm{o}}$) and their corresponding ground-truth embeddings (E_s^g, E_d^g, E_o^g) generated from FMs. The training objective for this stage is defined as follows.

$$\mathcal{L}_{\mathrm{FMPG}} = \sum_{i \in \{\mathrm{s,d,o}\}} \frac{1}{C^*} \sum_{k=1}^{C^*} \left(\hat{E}_i - E_i^g \right)^2 \tag{4}$$

Transformer and Mamba-based Fusion To emphasize the polyp-relevant features and extract complementary features from FMs, we first concatenate the aligned embeddings (E_1^*, E_2^*, E_3^*) and subsequently employ a 1×1 convolution to obtain a compact representation $\mathbf{x}$. Next, $\mathbf{x}$ is processed through a transformer layer $TL(.)$, which employs self-attention to focus the most informative representations of FMs by modeling cross-correlations among them. To preserve fine-grained structural information, a skip connection is incorporated from the latent representation L_1. Finally, the representations are fed to a Mamba Vision Layer (MVL) [13] to further enhance the modeling of global contextual relationships with linear complexity, critical for precise polyp segmentation.

$$F = \mathrm{MVL}\left(\mathrm{TL}(\mathbf{x}) + L_1\right) \tag{5}$$

2.3 Foundation Model Induced Decoder

The refined representation F, enriched with crucial priors from multiple FMs, including boundary cues and semantic awareness, is forwarded to the decoder of the lightweight baseline. The decoder receives two feature streams: (i) the hierarchical encoder feature maps and (ii) the refined foundation-induced embeddings. The decoding pathway follows the native design of each backbone (e.g., progressive upsampling in U-Net), and the final segmentation map is produced using a convolutional head.

2.4 Final Training Objective

We start by training the baseline model using a combined Binary Cross-Entropy (BCE) and Dice loss for stable segmentation and then introduce a progressive weighted training strategy that effectively balances segmentation supervision with FMPG loss. The total training loss $\mathcal{L}_{\mathrm{total}}$ is defined as:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{seg}} + \lambda_p \cdot \mathcal{L}_{\mathrm{FMPG}}, \tag{6}$$

$$\mathcal{L}_{\mathrm{seg}} = \mathcal{L}_{\mathrm{dice}} + \mathcal{L}_{\mathrm{bce}}, \tag{7}$$

where, λ_p indicates the weight factor, which is progressively increased during training (0.1 for first 30 epochs, 0.3 from 30 to 60 epochs, and 0.5 in later epochs), allowing the model to leverage rich prior knowledge while progressively refining its spatial representations.

Table 1: Quantitative comparison of baselines with Lite-π against SoTA methods on seen datasets.

Methods	Params (M)	FLOPs (G)	Kvasir-SEG (Seen)						CVC-ClinicDB (Seen)					
			mDice↑	mIoU↑	F_β^w↑	S_α↑	E_ϕ^{max}↑	$\mathcal{M}$↓	mDice↑	mIoU↑	F_β^w↑	S_α↑	E_ϕ^{max}↑	$\mathcal{M}$↓
State-of-the-art methods														
SANet [25] (MICCAI'21)	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 [28] (MICCAI'21)	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 [7] (CAAI'23)	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
PVT-Cascade [20] (WACV'23)	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
CFA-Net [29] (PR'23)	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
CTNet [26] (TCYB'24)	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
MEGANet [4] (WACV'24)	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
SAM-Mamba [10] (WACV'24)	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
SAM-MaGuP [9] (MICCAI'25)	106.0	431.0	94.7	89.0	95.1	94.2	98.1	1.6	95.3	91.3	95.8	96.3	98.8	0.5
Lightweight baselines														
U-Net [21] (MICCAI'15)	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++ [30] (DLMIA'18)	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 [11] (MICCAI'20)	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 with Polyp-DiFoM (WACV'26) [1]														
Polyp-DiFoM (U-Net)	16.8	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
Polyp-DiFoM (U-Net++)	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
Polyp-DiFoM (PraNet)	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 empowered with Lite-π (Ours)														
Lite-π (U-Net)	19.6	74.3	87.4 (+5.6)	80.2 (+5.6)	86.1 (+6.7)	92.1 (+6.3)	91.4 (+2.1)	4.0 (-1.5)	94.9 (+12.6)	90.1 (+14.6)	94.5 (+13.4)	96.6 (+7.7)	98.9 (+3.5)	0.9 (-1.0)
Lite-π (U-Net++)	10.1	66.6	87.6 (+5.5)	81.5 (+7.2)	87.3 (+6.5)	93.9 (+7.7)	92.4 (+1.4)	3.8 (-1.0)	93.3 (+13.9)	87.7 (+14.8)	93.1 (+14.6)	96.7 (+9.4)	90.7 (-3.4)	1.0 (-1.2)
Lite-π (PraNet)	34.9	13.9	92.9 (+3.1)	87.7 (+3.7)	88.5 (0.0)	93.7 (+2.2)	93.1 (-1.7)	2.7 (-0.3)	95.4 (+5.5)	91.7 (+6.8)	94.9 (+5.3)	97.8 (+4.2)	99.1 (+1.2)	0.8 (-0.1)

3 Experiment and Results

Datasets: We evaluate the effectiveness of our framework on five benchmark polyp segmentation datasets: Kvasir-SEG [16], CVC-ClinicDB [3], ETIS [22], CVC-ColonDB [23], and EndoScene [24]. We use 1,450 images (900 from Kvasir-SEG and 550 from CVC-ClinicDB) for training, and the remaining 100 and 62 images for testing, following [11], thereby ensuring a fair comparison with SOTA models. To comprehensively assess the generalization capability, we further evaluate it on three additional datasets: CVC-ColonDB (380 images), ETIS (196 images), and CVC-300 (test set of EndoScene) (60 images).

Performance Metrics: To ensure a fair comparison, we assess the performance of our approach as well as existing methods using six benchmark evaluation metrics such as mean Dice (mDice), mean IoU (mIoU), F-measure (F_β^w), S-measure (S_α), E-measure (E_ϕ^{max}), and mean absolute error ($\mathcal{M}$), as adopted in [11].

Implementation Details: Our model is implemented in the PyTorch 2.7.1 framework and trained on an NVIDIA Tesla V100 GPU (32GB) with CUDA 11.8. All input images are uniformly resized to 352×352 pixels, and different data augmentation strategies, including multi-scale scaling with factors $\{0.75, 1.0, 1.25\}$ and random vertical/horizontal flipping, are adopted. The model training is performed for 100 epochs using the Adam optimizer with a batch size of 6 and an initial learning rate of 1×10^{-4}.

3.1 Quantitative Comparison

We evaluate the proposed Lite-π framework with three lightweight baselines: U-Net [21], U-Net++ [30], and PraNet [11]. As shown in Table 1, Lite-π consistently improves segmentation performance with minimal increase in parameters and FLOPs. On **seen datasets**, Lite-π boosts mDice and mIoU across all baselines, achieving gains of +6-14% on CVC-ClinicDB and +3-5% on Kvasir-SEG,

Table 2: Quantitative comparison of baselines with Lite-π against SoTA methods on unseen datasets.

Methods	CVC-300 (Unseen)						CVC-ColonDB (Unseen)						ETIS (Unseen)					
	mDice↑	mIoU↑	F_β^w↑	S_α↑	E_ϕ^{max}↑	$\mathcal{M}$↓	mDice↑	mIoU↑	F_β^w↑	S_α↑	E_ϕ^{max}↑	$\mathcal{M}$↓	mDice↑	mIoU↑	F_β^w↑	S_α↑	E_ϕ^{max}↑	$\mathcal{M}$↓
State-of-the-art methods																		
SANet [25]	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 [28]	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 [7]	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
PVT-Cascade [20]	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
CFA-Net [29]	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
CTNet [26]	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
MEGANet [4]	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
SAM-Mamba [10]	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
SAM-MaGuP [9]	92.7	88.0	90.1	-	-	0.5	85.9	78.5	86.2	-	-	1.7	85.4	78.9	86.2	-	-	1.0
Lightweight baselines																		
U-Net [21]	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++ [30]	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 [11]	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 with Polyp-DiFoM (WACV'26) [1]																		
Polyp-DiFoM (U-Net)	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
PolypDiFoM (U-Net++)	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
PolypDiFoM (PraNet)	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 empowered with Lite-π (Ours)																		
Lite-π (U-Net)	85.3	76.5	85.2	87.2	93.7	1.2	71.4	60.8	71.3	83.2	85.2	4.4	55.6	42.9	51.9	76.4	71.7	2.9
	(+14.3)	(+13.8)	(+16.8)	(+2.9)	(+6.1)	(-1.0)	(+20.2)	(+16.4)	(+21.5)	(+12.0)	(+7.6)	(-1.7)	(+15.8)	(+9.4)	(+15.3)	(+8.0)	(-2.3)	(-0.7)
Lite-π (U-Net++)	83.5	74.8	83.3	89.9	92.1	1.4	72.7	61.6	72.5	87.4	86.3	4.7	55.1	44.7	54.7	76.8	72.9	3.2
	(+12.8)	(+12.4)	(+14.6)	(+6.0)	(+2.3)	(-0.4)	(+24.4)	(+20.6)	(+25.8)	(+18.3)	(+10.3)	(-1.7)	(+15.0)	(+10.3)	(+15.7)	(+8.5)	(-4.7)	(-0.3)
Lite-π (PraNet)	89.4	83.6	89.8	97.1	95.4	0.6	75.9	70.1	76.7	88.4	89.1	3.8	72.9	61.6	66.2	80.8	84.9	1.5
	(+2.3)	(+3.9)	(+5.5)	(+4.6)	(-1.8)	(-0.4)	(+5.0)	(+6.1)	(+7.1)	(+6.5)	(+2.2)	(-0.7)	(+10.1)	(+4.9)	(+6.2)	(+1.4)	(+0.8)	(-1.6)

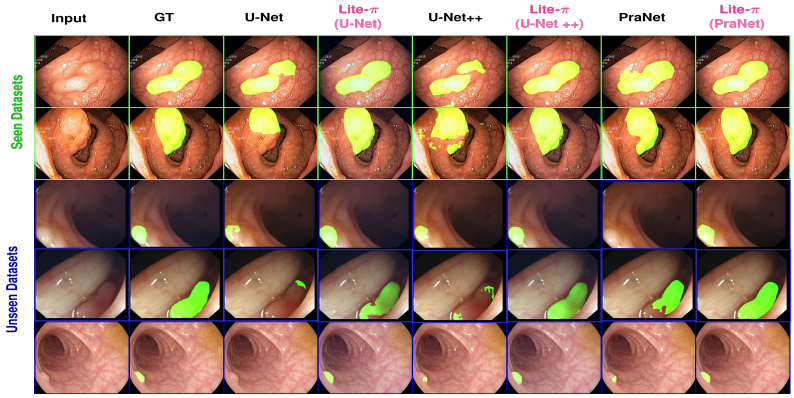
while reducing MAE. On **unseen datasets**, it yields strong generalization performance gains, improving mDice scores by nearly +15% for baselines U-Net and U-Net++. Compared to Polyp-DiFoM [1], it delivers superior performance across all five datasets. For a fair comparison, the results of Polyp-DiFoM [1] have been reproduced under the 3-FM configuration, which incorporates the same three FMs. Despite having significantly fewer parameters and FLOPs, Lite-π enables lightweight baselines to achieve performance on par with recent FM-driven methods such as SAM-Mamba [10] and SAM-MaGuP [9], which exceed 100M parameters. Additionally, our framework achieves an average latency of 17.5 ms and an average throughput of 62.6 FPS, with peak GPU memory usage of 326 MB, thereby establishing a new benchmark for real-time polyp segmentation. It is worth noting that all compared methods were evaluated using the same dataset split, preprocessing pipeline, and input resolution as the proposed approach, ensuring a fair comparison.

3.2 Qualitative Comparison

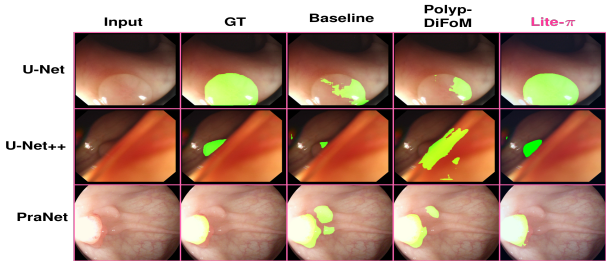
Fig. 2a shows Lite-π consistently achieves more accurate segmentation on both seen and unseen datasets, particularly in challenging scenarios where baseline models struggle, such as tiny polyps, weak boundaries, and low-contrast regions. U-Net produces highly inconsistent results, often missing polyps entirely or yielding false positives, whereas Lite-π refines them and obtains better predictions consistently. Comparison with Polyp-DiFoM [1] highlights significant improvements in segmentation results with all baselines (Fig. 2b). While Polyp-DiFoM improves baseline predictions, it still fails to accurately detect polyps in several cases. These results demonstrate the significance of the UFMI module in enhancing the segmentation capability of lightweight baselines.

3.3 Ablation Study

UFMI is designed as a unified module in which tightly coupled components jointly operate to perform semantic prototype induction and fusion. Therefore,



(a) Qualitative comparison on seen and unseen datasets, demonstrating the ability of Lite-π to accurately segment polyps of varying appearances.



(b) Comparison with Polyp-DiFoM [1], showing Lite-π's superior segmentation results in complex polyp structures.

Fig. 2: Qualitative results with a comparison against Polyp-DiFoM [1].

we assess the benefit of combining multiple FMs by progressively incorporating three FMs (SAM, DINOv2, and OneFormer) through ablation experiments. Table 3 shows that the best performance is obtained with the 3-FM configuration for all three baselines. In contrast to prior approaches, which exhibit performance saturation with progressive FM addition, our Lite-π framework supports scalable multi-FM integration without significant computational burden.

Table 3: Ablation study showing the effectiveness of incorporating multiple FMs into Lite-π.

Baseline	SAM	DINOv2	OneFormer	Kvasir	CVC-ClinicDB	CVC-300	CVC-ColonDB	ETIS
U-Net	✓	-	-	81.9	89.1	75.6	60.6	45.0
	✓	✓	-	85.7	92.1	80.7	67.7	50.9
	✓	✓	✓	**87.4**	**94.9**	**85.3**	**71.4**	**55.6**
U-Net++	✓	-	-	82.2	90.1	73.9	60.1	46.0
	✓	✓	-	84.6	92.1	81.3	67.1	52.1
	✓	✓	✓	**87.6**	**93.3**	**83.5**	**72.7**	**55.1**
PraNet	✓	-	-	89.5	93.4	86.7	74.6	69.2
	✓	✓	-	91.0	94.6	87.9	**76.1**	**73.7**
	✓	✓	✓	**92.9**	**95.4**	**89.4**	75.9	72.9

4 Conclusion

We present Lite-π, a new induction-based approach that effectively transfers structural and semantic cues from three FMs into a lightweight segmentation baseline. By synthesizing prototype representations and aligning them with the corresponding FMs through a reconstruction objective, followed by transformer- and Mamba-based feature refinement, our method facilitates efficient learning of high-level contextual and boundary-aware cues without incurring additional complexity. Extensive experiments across five polyp segmentation benchmarks demonstrate that Lite-π consistently enhances segmentation accuracy, robustness, and generalization while maintaining low computational overhead. Lite-π offers a practical and scalable solution for real-time clinical deployment where large foundation models are impractical.

Acknowledgments. D. R. Nayak is supported by the Anusandhan National Research Foundation (ANRF), Govt. of India (grant number ANRF/ARGM/2025/002890/TS and CRG/2023/007397). D. Jha is supported in part by the U.S. Department of Education (P116Z240151 to the University of South Dakota and the South Dakota School of Mines & Technology). The views expressed are those of the author(s) and do not necessarily represent the official views of the U.S. Department of Education.

Disclosure of Interests. The authors declare that they have no competing interests.

References

1. Agnihotri, S., Majhi, S., Nayak, D.R., Jha, D.: 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. pp. 1757–1766 (2026)
2. Ahn, S.B., Han, D.S., Bae, J.H., Byun, T.J., Kim, J.P., Eun, C.S.: The miss rate for colorectal adenoma determined by quality-adjusted, back-to-back colonoscopies. Gut and liver **6**(1), 64 (2012)
3. Bernal, J., Sánchez, F.J., et al.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics **43**, 99–111 (2015)
4. Bui, N.T., Hoang, D.H., 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. pp. 7985–7994 (2024)
5. Cheng, B., Misra, I., Schwing, A.G., et al.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 1290–1299 (2022)
6. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. Advances in Neural Information Processing Systems **34**, 17864–17875 (2021)
7. Dong, B., Wang, W., Fan, D.P., Li, J., Fu, H., Shao, L.: Polyp-pvt: Polyp segmentation with pyramid vision transformers. CAAI Artificial Intelligence Research **2**, 9150015 (2023)

8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
9. Dutta, T.K., Majhi, S., Nayak, D.R., Jha, D.: Mamba guided boundary prior matters: a new perspective for generalized polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 380–391. Springer (2025)
10. Dutta, T.K., Majhi, S., Nayak, D.R., Jha, D.: SAM-Mamba: Mamba guided SAM architecture for generalized zero-shot polyp segmentation. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4655–4664. IEEE (2025)
11. Fan, D.P., Ji, G.P., et al.: Pranet: Parallel reverse attention network for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 263–273. Springer (2020)
12. Fang, Y., Chen, C., Yuan, Y., Tong, K.y.: Selective feature aggregation network with area-boundary constraints for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 302–310. Springer (2019)
13. Hatamizadeh, A., Kautz, J.: Mambavision: A hybrid mamba-transformer vision backbone. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 25261–25270 (June 2025)
14. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.W., Wu, J.: Unet 3+: A full-scale connected unet for medical image segmentation. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 1055–1059. Ieee (2020)
15. Jain, J., Li, J., Chiu, M.T., et al.: Oneformer: One transformer to rule universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2989–2998 (2023)
16. Jha, D., Smedsrud, P.H., et al.: Kvasir-seg: A segmented polyp dataset. In: International Conference on Multimedia Modeling. pp. 451–462. Springer (2019)
17. Kirillov, A., Mintun, E., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
18. Oquab, M., Darcet, T., Moutakanni, T., et al.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
19. Radford, A., Kim, J.W., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PmLR (2021)
20. Rahman, M.M., Marculescu, R.: Medical image segmentation via cascaded attention decoding. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6222–6231 (2023)
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
22. Silva, J., Histace, A., et al.: 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)
23. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE Transactions on Medical Imaging **35**(2), 630–644 (2015)

24. Vázquez, D., Bernal, J., et al.: A benchmark for endoluminal scene segmentation of colonoscopy images. Journal of Healthcare Engineering **2017**(1), 4037190 (2017)
25. Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K., Cui, S.: Shallow attention network for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 699–708. Springer (2021)
26. Xiao, B., Hu, J., Li, W., Pun, C.M., Bi, X.: Ctnet: Contrastive transformer network for polyp segmentation. IEEE Transactions on Cybernetics **54**(9), 5040–5053 (2024)
27. Zhao, X., Jia, H., Pang, Y., Lv, L., Tian, F., Zhang, L., Sun, W., Lu, H.: M^2 snet: Multi-scale in multi-scale subtraction network for medical image segmentation. arXiv preprint arXiv:2303.10894 (2023)
28. Zhao, X., Zhang, L., Lu, H.: Automatic polyp segmentation via multi-scale subtraction network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 120–130. Springer (2021)
29. Zhou, T., Zhou, Y., He, K., Gong, C., Yang, J., Fu, H., Shen, D.: Cross-level feature aggregation network for polyp segmentation. Pattern Recognition **140**, 109555 (2023)
30. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: International Workshop on Deep Learning in Medical Image Analysis. pp. 3–11. Springer (2018)