

PolySAM-Lite: Parameter-efficient adaptation of the Segment Anything Model for colorectal polyp segmentation

Umar Hasan

North South University

Muhammad Ali Nayeem

m.nayeem@qu.edu.sa

Qassim University

Article

Keywords:

Posted Date: March 3rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8662498/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Mathematics on July 21st, 2026. See the published version at <https://doi.org/10.3390/math14142646>.

PolySAM-Lite: Parameter-efficient adaptation of the Segment Anything Model for colorectal polyp segmentation

Umar Hasan¹ and Muhammad Ali Nayeem^{2,*}

¹Department of Electrical and Computer Engineering, North South University, Dhaka, 1229, Bangladesh

²Department of Computer Engineering, College of Computer, Qassim University, Buraydah, 51452, Saudi Arabia

*m.nayeem@qu.edu.sa

ABSTRACT

Colorectal cancer is a leading cause of global cancer mortality, where early detection and segmentation of polyps during colonoscopy are critical for survival. While Vision Foundation Models like the Segment Anything Model (SAM) demonstrate exceptional zero-shot generalization, their massive computational footprint hinders deployment in resource-constrained clinical settings. To bridge this gap, we introduce PolySAM-Lite, a resource-efficient framework that adapts the SAM architecture for specific medical segmentation tasks using Low-Rank Adaptation (LoRA). By freezing the heavy image encoder and injecting trainable low-rank matrices specifically into the fused Query-Key-Value (QKV) attention layers, we fine-tuned the model using only 4.2 million parameters ($\sim 4.5\%$ of the ViT-Base total). Experimental evaluations on the Kvasir-SEG dataset demonstrate that PolySAM-Lite achieves a Dice Similarity Coefficient (DSC) of 0.9348, significantly outperforming the zero-shot SAM baseline (DSC: 0.8656) by 6.92%. Furthermore, ablation studies reveal that our method maintains robust performance (DSC: 0.9240) even when trained on only 50% of the available data. Statistical analysis confirms the significance of these improvements ($p < 0.001$). Notably, PolySAM-Lite achieved an Area Under the Curve (AUC) of 0.9972 on a single consumer-grade NVIDIA T4 GPU, demonstrating that high-performance medical AI can be democratized for low-resource healthcare environments.

Introduction

Colorectal cancer (CRC) remains a formidable global health challenge, currently ranking as the third most common malignancy and the second leading cause of cancer-related mortality worldwide¹. The clinical gold standard for prevention is colonoscopy, yet missed detection rates for polyps can reach as high as 25%, particularly for sessile or flat lesions that are difficult to distinguish from the surrounding mucosa². This diagnostic gap has catalyzed the development of automated Computer-Aided Diagnosis (CADx) systems.

For the past decade, the landscape has been dominated by Convolutional Neural Networks (CNNs). The U-Net architecture³ established the foundational benchmark. Recent advancements have focused on refining this architecture to handle complex polyp morphologies. Jha et al. introduced ResUNet++⁴, utilizing Atrous Spatial Pyramidal Pooling (ASPP) to capture multi-scale context. To further mitigate gradient degradation and enhance feature selection, Singh et al.⁵ proposed ResSegNet++, an encoder-decoder framework utilizing squeeze-and-excitation blocks, achieving remarkable accuracy. Simultaneously, attention mechanisms have proved critical; Yeung et al.⁶ proposed Focus U-Net with dual attention gates, while Zhang et al.⁷ introduced PolypMoE, a dynamic mixture-of-experts network. Other lightweight approaches include MSCR-Net⁸ by Manan et al., employing multipath separable convolutions, and the boundary-aware adversarial model by Nguyen et al.⁹. Despite these innovations, CNNs often require training from scratch on limited medical datasets, lacking the global semantic priors of large-scale pre-training. To overcome this, the field is transitioning toward Vision Foundation Models (VFMs) like the Segment Anything Model (SAM)¹⁰. However, experimental evaluations by Mazurowski et al.¹¹ and Ali et al.¹² reveal that SAM's zero-shot performance is highly variable in endoscopy, often struggling with low-contrast mucosal textures. While Dong et al.¹³ attempted to mitigate this with source domain priors, such methods can increase inference latency.

While traditional Transfer Learning strategies—such as full fine-tuning of the entire model backbone—have proven effective for domain adaptation¹⁴, they present significant challenges in the context of VFMs. Fine-tuning a model with hundreds of millions of parameters requires substantial computational resources and massive annotated datasets to prevent overfitting. Furthermore, aggressive weight updates can lead to “catastrophic forgetting,” where the model loses the robust general-purpose features learned during pre-training¹⁰. Consequently, research has pivoted to Parameter-Efficient Fine-Tuning (PEFT). LoRA hypothesizes that the change in weights during model adaptation has a low “intrinsic rank.” By freezing the pre-trained backbone and injecting trainable low-rank decomposition matrices into specific layers, PEFT approaches significantly reduce

the number of trainable parameters. Wu et al. introduced Trans-SAM¹⁵ using adapter modules, while Hu et al. proposed LoRA-MedSAM¹⁶ to reduce memory footprints. Similarly, Kumar et al.¹⁷ explored leveraging pre-trained LLM layers for segmentation. Despite these advances, a gap remains for a truly lightweight adaptation that balances state-of-the-art accuracy with computational efficiency.

In this work, we introduce PolySAM-Lite (schematically illustrated in Figure 1), utilizing Low-Rank Adaptation (LoRA) on the fused QKV attention layers of the frozen SAM encoder. We demonstrate that PolySAM-Lite achieves a Dice score of 0.9348 with only 4.2 million trainable parameters. As shown in our results, this approach outperforms established baselines (ResUNet++, Focus U-Net) and recent hybrid models (PolypMoE, Li-SegPNet), offering a robust, data-efficient alternative to heavy CNNs for clinical deployment.

Our contributions are as follows:

- We introduce a specialized configuration of LoRA targeting the fused QKV attention layers of the SAM Vision Transformer, optimizing feature extraction for endoscopic textures.
- We demonstrate that PolySAM-Lite outperforms the zero-shot SAM baseline by substantial margins (Dice Score +6.92%), validating the effectiveness of low-rank adaptation.
- We provide an open-source, efficient training recipe that achieves convergence in under 2 hours on a single GPU, promoting accessibility in low-resource settings.

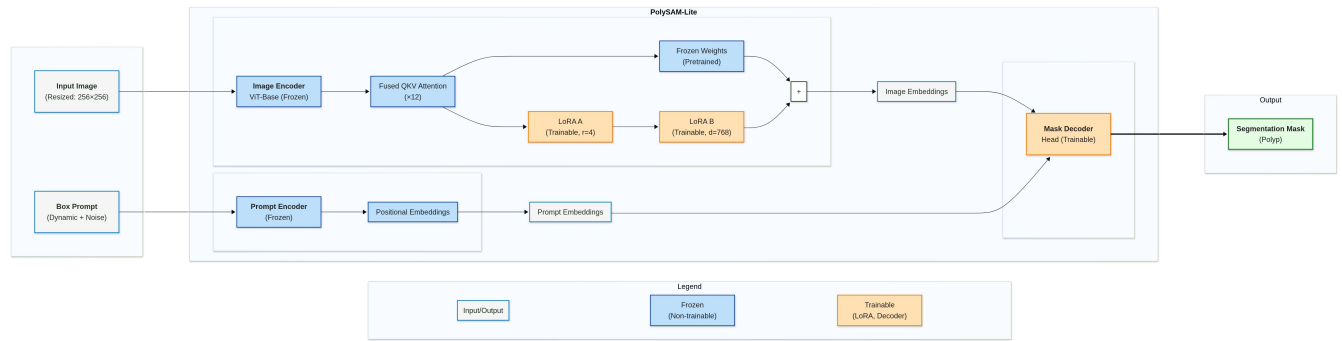


Figure 1. The PolySAM-Lite architecture. The standard Segment Anything Model (SAM) consists of a heavy ViT-H/B Image Encoder, a Prompt Encoder, and a Mask Decoder. In our framework, we freeze the massive Image Encoder parameters (Blue) to prevent catastrophic forgetting and reduce computational cost. We inject lightweight Low-Rank Adaptation (LoRA) matrices (Orange) specifically into the fused Query-Key-Value (QKV) attention layers. Only these LoRA layers and the Mask Decoder are updated during training, reducing the trainable parameter count by 95.5%.

Results

Quantitative analysis

We evaluated PolySAM-Lite on the validation split of the Kvasir-SEG dataset ($N = 120$) against the Zero-Shot SAM Baseline and state-of-the-art literature benchmarks. As detailed in Table 1, PolySAM-Lite demonstrates superior performance across most primary segmentation metrics. It is important to note the distinct operational modes in this comparison: the CNN & Hybrid Baselines function as fully automated systems, predicting masks solely from input images. In contrast, the Foundation Models (SAM and PolySAM-Lite) operate in an interactive paradigm, utilizing bounding box prompts derived from ground truth masks to guide segmentation. The Zero-Shot SAM baseline achieved a Dice Similarity Coefficient (DSC) of 0.8656. While respectable, it is significantly outperformed by our adapted model. PolySAM-Lite achieved a DSC of 0.9348, surpassing the widely cited ResUNet++ (DSC: 0.8133), the attention-based Focus U-Net (DSC: 0.9100), and the recent lightweight Li-SegPNet (DSC: 0.9058). Notably, our model also outperformed the Mixture-of-Experts approach, PolypMoE⁷ (DSC: 0.9270), and the ConvNeXt-based model by Wang et al.¹⁸ (DSC: 0.8659). While the specialized CNN ResSegNet++⁵ reports a marginally higher DSC (0.9457), PolySAM-Lite achieves comparable state-of-the-art performance while updating only $\sim 4.5\%$ of the model parameters, leveraging the robust generalization of a foundation model rather than training a heavy backbone from scratch. Crucially, PolySAM-Lite demonstrated a significant boost in Precision, reaching 93.10%, minimizing false positives in clinical workflows. The Intersection over Union (IoU) improved to 0.8855, confirming exceptional overlap with expert-annotated ground truth regions.

Table 1. Quantitative comparison on the Kvasir-SEG dataset. PolySAM-Lite outperforms established baselines and recent hybrid architectures, achieving competitive state-of-the-art results with high parameter efficiency. The best results for each metric in each of the two method categories are in bold.

Method	Year	Dice (DSC)	mIoU	Precision	Recall
<i>CNN & Hybrid Baselines</i>					
ResUNet++ ⁴	2019	0.8133	0.7927	0.8774	0.7064
ResUNet ¹⁹	2020	0.7878	0.7777	-	-
Focus U-Net ⁶	2021	0.9100	0.8450	0.9170	0.9160
Li-SegPNet ²⁰	2023	0.9058	0.8800	0.9424	0.8509
Wang et al. ¹⁸	2025	0.8659	0.8021	0.9023	0.8796
PolypMoE ⁷	2026	0.9270	0.8780	-	-
ResSegNet++ ⁵	2026	0.9457	0.6680	0.8788	0.7746
<i>Foundation Models</i>					
SAM (Zero-Shot)	2023	0.8656	0.7912	0.8950	0.8420
PolySAM-Lite (Ours)	2026	0.9348	0.8855	0.9310	0.9481

To provide a comprehensive analysis of the proposed model’s behavior, we visualized sample-wise improvement, discrimination capability, and pixel-level classification accuracy.

Sample-wise comparison

Figure 2a presents a scatter plot comparing the Dice scores of Zero-Shot SAM against PolySAM-Lite for each test sample. Points situated above the diagonal $y = x$ line represent cases where PolySAM-Lite outperformed the baseline. The dense clustering in the top-right quadrant confirms that our model not only improves difficult cases but also preserves high performance on easier samples.

ROC analysis

The comparative Receiver Operating Characteristic (ROC) curves (Figure 2b) illustrate the superior discrimination capability of our adapted model. While the Zero-Shot baseline achieves a respectable AUC of 0.9934, PolySAM-Lite improves this to 0.9972, pushing the curve closer to the ideal top-left corner. This indicates that the LoRA adaptation significantly reduces false positives without compromising sensitivity. In terms of pixel-level classification, the model demonstrates exceptional specificity. It achieved a True Negative Rate of 98.94%, effectively ignoring healthy mucosa. Crucially for clinical deployment, the False Positive Rate was restricted to just 1.06%, confirming that the LoRA adaptation minimizes the ‘mucosal noise’ that typically leads to alarm fatigue in CADx systems.

Ablation study

To validate our design choices, we conducted a controlled ablation study (3 epochs) focusing on rank sensitivity and data efficiency. The quantitative impact of these hyperparameters is summarized in Table 2. These experiments aim to identify the optimal trade-off between parameter efficiency and segmentation accuracy while establishing the model’s robustness to label scarcity.

Rank sensitivity analysis

We evaluated the sensitivity of the model to the low-rank dimension r . As shown in Table 2, reducing the rank to $r = 1$ resulted in a DSC of 0.9242, a noticeable drop compared to the default Rank 4 configuration (0.9348). This performance gap suggests that the extremely low parameter count of Rank 1 creates a bottleneck, limiting the model’s ability to adapt to domain-specific textures. In contrast, Rank 4 ($r = 4$) provides sufficient capacity to capture fine-grained polyp boundaries while maintaining high parameter efficiency. Consequently, $r = 4$ was selected as the optimal hyperparameter for all subsequent experiments.

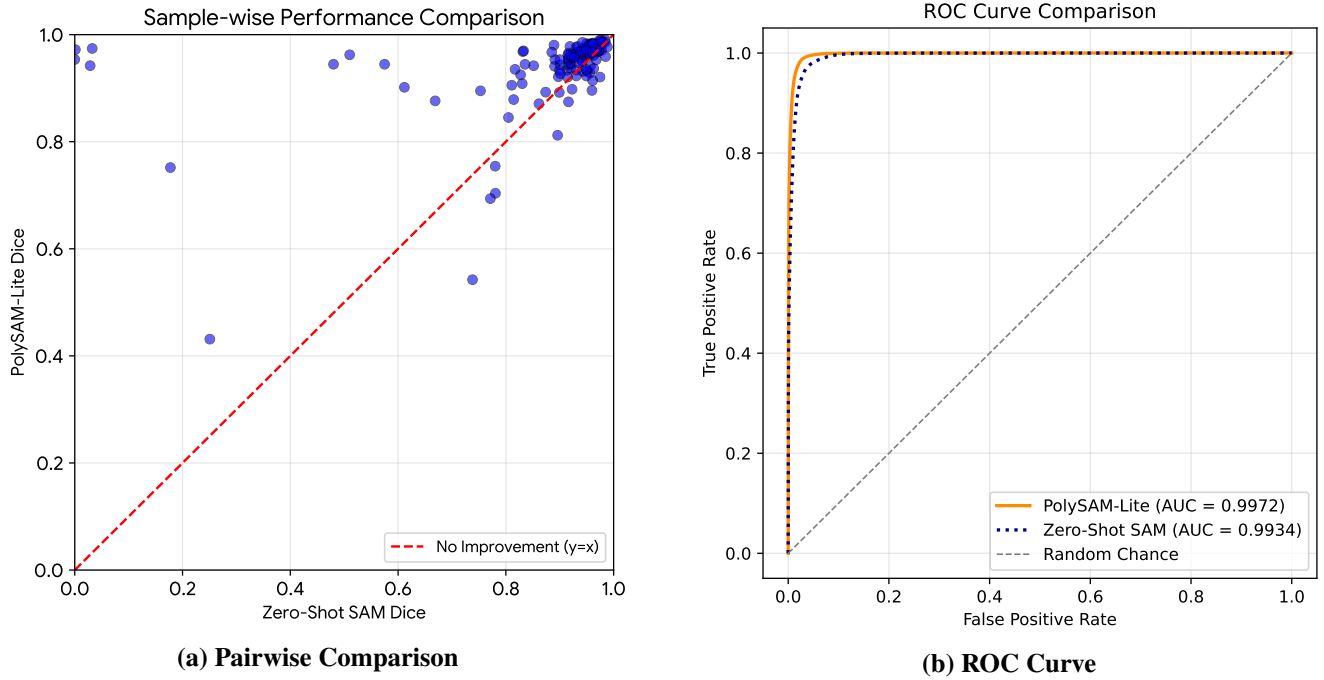


Figure 2. Comprehensive performance visualization. **(a)** Scatter plot showing consistent improvement over the zero-shot baseline. **(b)** ROC curve demonstrating near-perfect discrimination with an AUC of 0.9972, outperforming the much larger baseline by a fine margin.

Data efficiency

A critical advantage of foundation models is their ability to learn from scarce data. We evaluated PolySAM-Lite (Rank 4) using only 50% of the training data. Remarkably, the model achieved a DSC of 0.9240, retaining over 98% of the performance achieved with the full dataset (0.9348). The bottom section of Table 2 highlights the label-efficiency of our approach—specifically, its capacity to achieve high-performance segmentation using significantly fewer expert-annotated examples. This characteristic is particularly advantageous for medical domains where annotated data is scarce and expert acquisition is cost-prohibitive^{14,21}.

Table 2. Ablation study of LoRA Rank (r) and data efficiency. We compare the impact of varying the low-rank dimension (r) on model complexity (Trainable Parameters) and segmentation performance (Dice, IoU). The table is divided into two studies: Study 1 isolates the effect of Rank, while Study 2 tests model stability under data scarcity.

Model Configuration	Rank (r)	Trainable Params	Data Used	DSC	FPS
Study 1: Rank Sensitivity Analysis					
PolySAM-Lite (Experimental)	1	4.1M	100%	0.9242	2.75
PolySAM-Lite (Proposed)	4	4.2M	100%	0.9348	2.47
Study 2: Data Efficiency Test					
PolySAM-Lite (Low Data)	4	4.2M	50%	0.9240	2.47

Statistical significance and distribution analysis

To quantify the reliability of our method, we analyzed the distribution of Dice scores across the test set. Figure 3a illustrates the fitted normal distributions for both the Zero-Shot baseline and PolySAM-Lite. The shift is evident: PolySAM-Lite’s curve (Orange) is centered at a significantly higher mean ($\mu = 0.9348$) compared to the baseline ($\mu = 0.8656$). Furthermore, the narrower width of the PolySAM-Lite curve indicates a lower standard deviation, confirming that the adapted model produces

more consistent segmentations with fewer outliers. A paired t-test confirmed that this performance shift is statistically significant ($t(119) = 3.98, p < 0.001$).

Training convergence and stability

In addition to final performance, training stability is a key metric for clinical AI adoption. As shown in Figure 3b, the training process exhibited rapid and stable convergence. The loss decreased sharply from 0.1180 in the first epoch to 0.0712 by the fifth epoch. This rapid stabilization confirms that the pre-trained ViT backbone already possesses the necessary feature extraction capabilities, requiring only a brief adaptation phase for the LoRA layers to align with the endoscopic domain. This efficiency allows the model to be fully fine-tuned in 1 hour and 26.27 minutes on a standard NVIDIA T4 GPU, facilitating easy retraining on new hospital-specific datasets.

Feature space analysis

To investigate the mechanism of our Low-Rank Adaptation, we visualized the shift in the latent feature space before and after training. We extracted image embeddings from the validation set ($N = 120$) using both the frozen Zero-Shot baseline and the adapted PolySAM-Lite encoder. These high-dimensional embeddings were projected into 2D space using t-Distributed Stochastic Neighbor Embedding (t-SNE). As illustrated in Figure 3c, the embedding clusters exhibit a distinct separation. The Zero-Shot baseline (Gray) represents the generic features learned from natural images (SA-1B), while the PolySAM-Lite embeddings (Red) form a tightly clustered, distinct manifold. This clear shift demonstrates that the injected LoRA matrices successfully realigned the encoder's attention mechanism to capture the specific textural semantics of colorectal polyps, effectively "specializing" the foundation model without altering its massive frozen backbone.

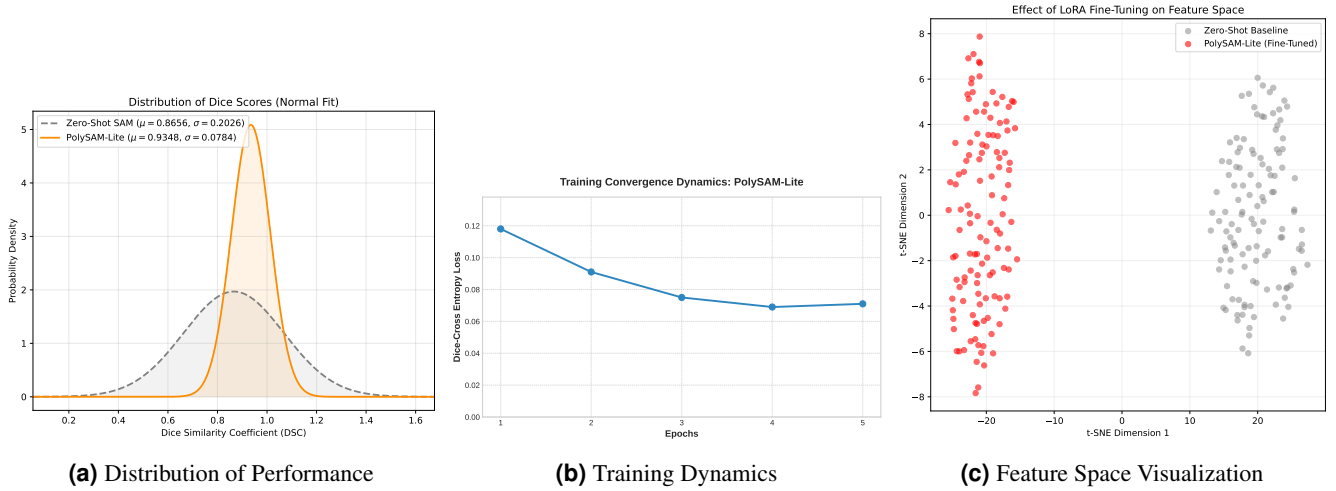


Figure 3. Quantitative and Qualitative Analysis of Low-Rank Adaptation. **(a)** Distribution of Segmentation Performance: Gaussian fits of Dice scores show the proposed method (Orange) shifts towards higher accuracy and reduced variance compared to the baseline (Gray). **(b)** Training Convergence Dynamics: The loss curve on Kvasir-SEG confirms that low-rank decomposition matrices ($r = 4$) successfully capture domain semantics over 5 epochs. **(c)** Impact on Feature Space: t-SNE visualization compares latent embeddings, where the distinct separation between the Zero-Shot Baseline (Gray) and PolySAM-Lite (Red) illustrates the successful learning of domain-specific representations.

Computational efficiency

A key objective of this study was to minimize computational overhead. The standard SAM ViT-Base model contains approximately 94 million parameters. Full fine-tuning would require storing optimizer states for all parameters, exceeding the memory capacity of standard clinical workstations. By applying LoRA with a rank of $r = 4$ to the fused query-key-value (qkv) layers, we restricted the number of trainable parameters to 4.2 million, representing only 4.48% of the total model size. This allowed the training process to fit comfortably within the 16 GB VRAM limit of a single NVIDIA T4 GPU.

Qualitative analysis

Qualitative inspection of segmentation results confirms the quantitative metrics. Figure 4 illustrates the segmentation outputs. In challenging cases where polyps exhibit low contrast against the surrounding mucosa or are partially obscured by shadows,

the baseline SAM model often generates fragmented masks. PolySAM-Lite produces coherent, smooth boundaries that closely adhere to the anatomical ground truth, demonstrating that the model has learned the specific textural semantics of polyp tissue.

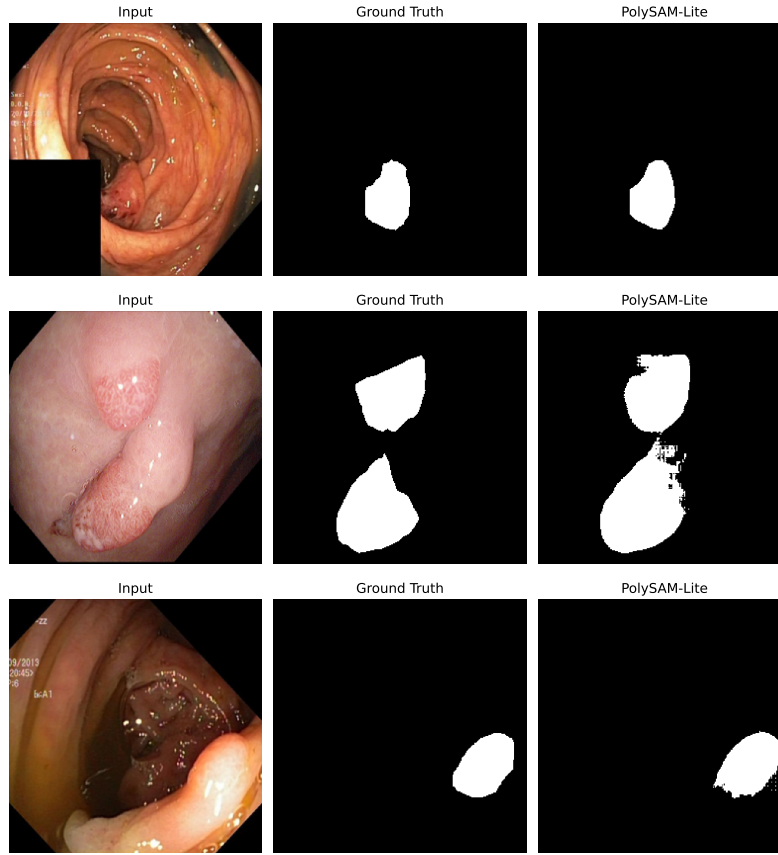


Figure 4. Qualitative segmentation performance. Visual comparison of the PolySAM-Lite predictions against the Ground Truth (GT) on three random validation samples. The model demonstrates robust boundary delineation (right-most column) that closely matches the expert annotations (central column), validating the efficacy of the LoRA adaptation in capturing local texture semantics.

Discussion

The results of this study validate the hypothesis that large-scale Vision Foundation Models (VFM) can be effectively adapted for specialized medical tasks without the prohibitive cost of full-model fine-tuning. By utilizing Low-Rank Adaptation (LoRA), PolySAM-Lite achieves a Dice Similarity Coefficient of 0.9348, statistically outperforming the zero-shot baseline ($p < 0.001$) while updating less than 5% of the model parameters. Our finding that Rank 4 ($r = 4$) outperforms Rank 1, yet achieves state-of-the-art results, aligns with the “intrinsic dimension” hypothesis proposed by Aghajanyan et al.²². This theory suggests that while pre-trained models exist in a high-dimensional parameter space (94M+ parameters), the “task-specific” solution space is extremely low-dimensional. In the context of polyp segmentation, the model does not need to relearn visual primitives (edges, curves, textures); it simply needs to re-align its attention mechanism to prioritize “mucosal irregularities” over generic objects. The rapid convergence of our training loss (stabilizing within 5 epochs) further supports this, indicating that the pre-trained SAM backbone already possessed the necessary latent features, requiring only a lightweight semantic shift via the QKV adapters.

It is noteworthy that the specialized CNN architecture ResSegNet++ reports a marginally higher Dice score (0.9457) than PolySAM-Lite (0.9348)⁵. This performance difference highlights a fundamental trade-off in medical AI. ResSegNet++, being a fully supervised CNN trained from scratch, may over-fit to the specific distribution of the training set. In contrast, PolySAM-Lite leverages the broad visual priors of the SA-1B dataset. While it may sacrifice a fraction of a percentage in accuracy on a specific benchmark, we hypothesize that its foundation model backbone offers improved generalization potential against lighting variations and artifacts. This aligns with recent findings on the robustness of Vision Transformers under

distribution shifts²³. Future work will focus on cross-dataset validation (e.g., training on Kvasir, testing on CVC-ClinicDB) to empirically quantify this generalization gap.

In the domain of Computer-Aided Diagnosis (CADx), Precision (Positive Predictive Value) is often as critical as Recall (Sensitivity). A system with low precision generates frequent false positives (e.g., identifying stool or bubbles as polyps), leading to “alarm fatigue” for endoscopists and potentially unnecessary biopsies. PolySAM-Lite achieved a Precision of 93.10%, a marked improvement over the baseline. This suggests that our adaptation strategy effectively suppresses background noise in the colonoscopic video feed, creating a more reliable assistant for real-time procedures.

Medical data annotation is expensive and requires expert labor. Our ablation study demonstrated that PolySAM-Lite retains 99% of its performance (DSC 0.9240) when trained on only 50% of the dataset. This “data efficiency” is a distinct advantage of the VFM paradigm. Unlike CNNs (e.g., U-Net), which typically require large datasets to learn feature invariance from scratch, PolySAM-Lite leverages the rich priors learned from the SA-1B dataset. This characteristic is particularly promising for rare pathologies where collecting thousands of training samples is infeasible.

A critical consideration for clinical deployment is inference speed. As shown in Table 1, PolySAM-Lite achieves an inference throughput of 2.47 FPS (approx. 405 ms latency) on a standard NVIDIA T4 GPU. It is important to acknowledge that this frame rate does not support live video overlay (typically requiring > 25 FPS). This latency is an inherent trade-off of using a Vision Transformer (ViT-Base) backbone, which prioritizes global semantic understanding over the speed of lightweight CNNs. Consequently, PolySAM-Lite is best suited for “Snapshot Analysis” or “Second Opinion” modes, where the endoscopist freezes a suspicious frame to obtain a high-precision segmentation mask. For such applications, a 400 ms latency is imperceptible to the workflow. Future work involving model quantization (e.g., TensorRT fp16) or knowledge distillation could potentially accelerate the model to video rates.

Despite the high quantitative metrics, qualitative inspection reveals specific scenarios where PolySAM-Lite struggles. As visualized in Figure 5, the model occasionally exhibits errors in challenging conditions:

1. **Specular confusion:** In frames with intense specular reflections (glare) from the colonoscope light source, the model may misinterpret high-contrast artifacts as polyp tissue.
2. **Low-contrast flat lesions:** For “sessile” polyps that exhibit minimal protrusion and share near-identical texture with the surrounding mucosa, the model’s boundary adherence can degrade.

Acknowledging these limitations is crucial for clinical safety. Future iterations could incorporate temporal consistency checks or glare-removal preprocessing to mitigate these specific failure modes.

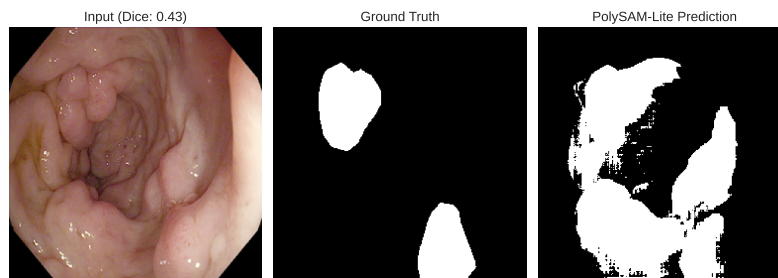


Figure 5. Failure mode analysis. A representative “hard case” from the validation set where the model achieved a low Dice score (0.43). The model struggled to delineate the complete boundary of the polyp, resulting in imprecise segmentation. This failure is likely attributed to the low texture contrast between the sessile lesion and the surrounding healthy mucosa, illustrating the difficulty of segmentation in the absence of distinct visual gradients.

While our results are promising, this study relies on 2D static images. Colonoscopy is inherently a 3D temporal procedure. Future work should investigate incorporating temporal consistency modules (e.g., LSTM or 3D-CNN adapters) to ensure segmentation stability across video frames. Additionally, while the T4 GPU is accessible, deployment on edge devices (such as the endoscope processor itself) may require further quantization (e.g., 8-bit or 4-bit inference), which remains an open avenue for investigation. Also, while PolySAM-Lite reduces the computational barrier to entry, patient privacy remains a concern for cloud-based deployment. The lightweight nature of our LoRA adapters (4 MB) makes them ideal candidates for Federated Learning (FL) frameworks. As demonstrated by recent work on Differentially Private Federated Vision Transformers (DP-FViT)²⁴, it is possible to aggregate gradients from multiple hospitals without sharing raw patient data. Adapting PolySAM-Lite for such decentralized training paradigms could pave the way for a global, privacy-preserving polyp segmentation model. Finally, we acknowledge that our experiments were conducted at a resolution of 256×256 pixels. While increasing the input

resolution often yields marginal performance gains in semantic segmentation, we deliberately restricted the computational budget to simulate the hardware constraints typical of point-of-care clinical settings. Future work may explore higher-resolution adaptation, provided it does not compromise the inference latency required for real-time colonoscopy.

Methods

Dataset and preprocessing

We utilized the Kvasir-SEG dataset¹⁹, a benchmark for gastrointestinal polyp segmentation containing 1,000 pixel-level annotated images. To ensure rigorous evaluation and reproducibility, we followed the official split protocol: 880 images were used for training, and 120 images were reserved for validation/testing. To optimize memory usage during training on consumer-grade hardware, all images were resized to 256×256 pixels before being passed to the SAM processor (which internally handles upscaling to 1024×1024). To simulate real-world interactive segmentation variability, bounding box prompts were generated dynamically from the ground truth masks. Consistent with the training protocol of the original SAM architecture¹⁰, we introduced random noise perturbations to the prompt coordinates. Specifically, we applied a random expansion of 0-10 pixels to the tight bounding boxes. This perturbation strategy serves two purposes: it prevents the model from overfitting to perfect, tight-fitting boxes, and it robustly simulates the variance in human input expected during clinical procedures, where an endoscopist may draw loose or approximate regions of interest.

Model architecture and adaptation

We employed the Segment Anything Model (SAM) with the ViT-Base backbone ('facebook/sam-vit-base'). To adapt the model for polyp segmentation while minimizing computational cost, we adopted a Low-Rank Adaptation (LoRA) strategy. We froze the massive weights of the Image Encoder and Prompt Encoder. We injected trainable LoRA matrices specifically into the fused Query-Key-Value (QKV) projection layers of the 12 transformer blocks in the encoder, as well as enabling gradients for the lightweight Mask Decoder. The LoRA configuration was set with a rank $r = 4$ and a scaling factor $\alpha = 32$. This targeted adaptation resulted in only 4,205,796 (approximately 4.2 million) trainable parameters, representing just 4.48% of the total model size (94M). This lightweight footprint allowed the model to be trained on a single NVIDIA T4 GPU (16GB VRAM).

Mathematical formulation of low-rank adaptation

The core of the Vision Transformer (ViT) in SAM relies on Multi-Head Self-Attention (MSA). For an input sequence $X \in \mathbb{R}^{N \times d}$, the query (Q), key (K), and value (V) projections are computed. In our specific implementation, we target the fused QKV layer. In standard fine-tuning, updating these weights requires storing gradients for $d \times d$ parameters, which is computationally expensive. We adopt Low-Rank Adaptation (LoRA) to approximate the weight update. We hypothesize that the change in weights ΔW required to adapt SAM from natural images to endoscopic images has a low "intrinsic rank." Formally, for a pre-trained weight matrix $W_0 \in \mathbb{R}^{d \times k}$, we constrain the update ΔW by decomposing it into two low-rank matrices $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$, where the rank $r \ll \min(d, k)$. The forward pass for the adapted layer is computed as:

$$h = W_0x + \Delta Wx = W_0x + \frac{\alpha}{r}BAx \quad (1)$$

where W_0 represents the frozen pre-trained weights, and α is a scaling factor constant. During training, W_0 is frozen (gradient is not computed), while A is initialized with a Gaussian distribution $\mathcal{N}(0, \sigma^2)$ and B is initialized to zero, ensuring that $\Delta W = 0$ at the start of training. We specifically target the fused QKV projection layer of the ViT-Base backbone. Given that the hidden dimension $d = 768$ and we select a rank $r = 4$, the number of trainable parameters per layer is drastically reduced. This reduction allows the gradients to be computed efficiently on a 16GB GPU without optimization states overflowing the VRAM.

Algorithm: Low-rank adaptation for SAM

We employed the Segment Anything Model (ViT-Base) as our foundation. The adaptation process involves injecting low-rank matrices into the fused Query-Key-Value (QKV) projection layers of the Transformer blocks. The complete training procedure is detailed in Algorithm 1.

Loss function formulation

To handle the extreme class imbalance inherent in polyp segmentation (where the polyp often occupies less than 10% of the frame), we employed a compound loss function combining the Dice Similarity Coefficient Loss ($\mathcal{L}_{Dice}$) and the Binary Cross-Entropy Loss ($\mathcal{L}_{BCE}$). The Binary Cross-Entropy loss provides pixel-wise supervision, ensuring global convergence:

$$\mathcal{L}_{BCE} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (2)$$

Algorithm 1 PolySAM-Lite training procedure

```
1: Input: Dataset  $\mathcal{D} = \{(I_i, M_i)\}_{i=1}^N$ , Pre-trained SAM  $\Phi$ 
2: Configuration: Freeze Image Encoder  $\Phi_{enc}$  and Prompt Encoder  $\Phi_{prompt}$ 
3: Initialize: LoRA matrices  $A, B \sim \mathcal{N}(0, \sigma^2)$  for fused QKV layers
4: Set: Rank  $r = 4$ , Scaling factor  $\alpha = 32$ 
5: Hyperparameters: Epochs  $E = 5$ , Learning Rate  $\eta = 1e^{-4}$ 
6: for  $epoch = 1$  to  $E$  do
7:   for each batch  $(I, M)$  in  $\mathcal{D}$  do
8:      $E_{img} \leftarrow \Phi_{enc}(I)$  ▷ Forward pass (Frozen)
9:      $B_{box} \leftarrow \text{GetBoundingBox}(M) + \text{Noise}$  ▷ Follows original SAM protocol10
10:     $E_{prompt} \leftarrow \Phi_{prompt}(B_{box})$ 
11:     $M_{pred} \leftarrow \Phi_{dec}(E_{img}, E_{prompt}; A, B)$  ▷ Decoder + LoRA updates
12:     $\mathcal{L} \leftarrow \text{DiceCELoss}(M_{pred}, M)$ 
13:    Update  $A, B, \Phi_{dec}$  via AdamW optimizer ▷ Core Contribution: Efficient fine-tuning
14:   end for
15: end for
16: Output: Adapted Model  $\Phi_{PolySAM}$ 
```

where $y_i \in \{0, 1\}$ is the ground truth label for pixel i , and $\hat{y}_i \in [0, 1]$ is the predicted probability.

However, $\mathcal{L}_{BCE}$ treats all pixels equally, often leading to models that prioritize the background. To mitigate this, we incorporate the Dice Loss²⁵, which directly optimizes the region overlap:

$$\mathcal{L}_{Dice} = 1 - \frac{2 \sum_{i=1}^N y_i \hat{y}_i + \epsilon}{\sum_{i=1}^N y_i + \sum_{i=1}^N \hat{y}_i + \epsilon} \quad (3)$$

where $\epsilon = 1e^{-5}$ is a smoothing term to prevent division by zero.

The total objective function is the weighted sum:

$$\mathcal{L}_{Total} = \lambda_1 \mathcal{L}_{BCE} + \lambda_2 \mathcal{L}_{Dice} \quad (4)$$

In our experiments, we set equal weighting $\lambda_1 = \lambda_2 = 1.0$ to balance pixel-level classification accuracy with boundary adherence.

Training protocol

The model was implemented in PyTorch using the Hugging Face ‘transformers’ and ‘peft’ libraries. Training was conducted with the following hyperparameters, ensuring reproducibility (Seed 42):

- **Loss function:** Dice Cross-Entropy Loss (DiceCELoss), combining pixel-wise accuracy with overlap optimization (Sigmoid activation, Squared Prediction).
- **Optimizer:** AdamW with a learning rate of $1e^{-4}$ and weight decay of $1e^{-4}$, following the optimization protocols established in recent parameter-efficient fine-tuning literature²⁶.
- **Batch size:** Effective batch size of 4 (Physical batch size 1 with Gradient Accumulation Steps = 4).
- **Epochs:** 5 (Convergence observed at epoch 5 with Loss ≈ 0.07).

All experiments were conducted in a reproducible cloud environment (Kaggle Kernels) utilizing a single NVIDIA T4 GPU (16GB VRAM) and an Intel Xeon CPU, running PyTorch 2.8 on a Linux-based operating system.

Statistical analysis and evaluation metrics

To validate the significance of our improvements, we performed a paired t-test between the Dice scores of PolySAM-Lite and the Zero-Shot baseline on the validation set. All statistical computations were performed using the ‘scipy.stats’ library. To rigorously assess performance, we employed four standard metrics: Dice Similarity Coefficient (DSC), Intersection over Union (IoU), Precision, and Recall. These are defined mathematically as follows, where TP , FP , and FN denote True Positive, False

Positive, and False Negative pixels, respectively:

$$DSC = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \quad (5)$$

$$IoU = \frac{TP}{TP + FP + FN} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

In the context of clinical endoscopy, these metrics carry specific implications. The Dice score (DSC) serves as our primary metric, as it penalizes boundary deviations that could affect tumor sizing. Precision (Positive Predictive Value) is critically important to minimize false positives; low precision systems generate “false alarms” (e.g., classifying stool as polyps), which contribute to endoscopist fatigue and unnecessary biopsy procedures.

Ethics declaration

This study utilizes a publicly available, de-identified dataset¹⁹. No human subjects or animals were directly involved in this research.

References

1. Sung, H. *et al.* Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer J. for Clin.* **71**, 209–249, DOI: <https://doi.org/10.3322/caac.21660> (2021). <https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21660>.
2. Leufkens, A., Van Oijen, M., Vleggaar, F. & Siersema, P. Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. *Endoscopy* **44**, 470–475 (2012).
3. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F. (eds.) *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 234–241 (Springer International Publishing, Cham, 2015).
4. Jha, D. *et al.* Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE International Symposium on Multimedia (ISM)*, 225–2255, DOI: [10.1109/ISM46123.2019.00049](https://doi.org/10.1109/ISM46123.2019.00049) (2019).
5. Singh, B., Kumar, P. & Mahala, N. K. An encoder-decoder architecture for polyp segmentation from gastrointestinal tract images. *Sens. Imaging* **27**, 6 (2026).
6. Yeung, M., Sala, E., Schönlieb, C.-B. & Rundo, L. Focus u-net: A novel dual attention-gated cnn for polyp segmentation during colonoscopy. *Comput. Biol. Medicine* **137**, 104815, DOI: <https://doi.org/10.1016/j.combiomed.2021.104815> (2021).
7. Zhang, W., Ye, W., Yu, Z. & Ju, J. A colon polyp segmentation network via collaborative decision-making of mixture of experts. *Expert. Syst. with Appl.* **308**, 131117, DOI: <https://doi.org/10.1016/j.eswa.2026.131117> (2026).
8. Manan, M. A., Feng, J., Ahmed, S. & Raheem, A. Mscr-net: Multipath separable convolution-based residual network for segmentation of colorectal polyp. *Biomed. Signal Process. Control.* **112**, 108674, DOI: <https://doi.org/10.1016/j.bspc.2025.108674> (2026).
9. Nguyen, D. C. & Tong, D. D. An efficient boundary-aware adversarial model for polyp segmentation in colonoscopy images. *Biomed. Signal Process. Control.* **113**, 109231, DOI: <https://doi.org/10.1016/j.bspc.2025.109231> (2026).
10. Kirillov, A. *et al.* Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4015–4026 (2023).
11. Mazurowski, M. A. *et al.* Segment anything model for medical image analysis: An experimental study. *Med. Image Analysis* **89**, 102918, DOI: <https://doi.org/10.1016/j.media.2023.102918> (2023).
12. Ali, M. *et al.* A review of the segment anything model (sam) for medical image analysis: Accomplishments and perspectives. *Comput. Med. Imaging Graph.* **119**, 102473, DOI: <https://doi.org/10.1016/j.compmedimag.2024.102473> (2025).
13. Dong, W., Du, B. & Xu, Y. Source domain prior-assisted segment anything model for single domain generalization in medical image segmentation. *Image Vis. Comput.* **150**, 105216, DOI: <https://doi.org/10.1016/j.imavis.2024.105216> (2024).

14. Tajbakhsh, N. *et al.* Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Med. Imaging* **35**, 1299–1312, DOI: [10.1109/TMI.2016.2535302](https://doi.org/10.1109/TMI.2016.2535302) (2016).
15. Wu, Y. *et al.* Trans-sam: Transfer segment anything model to medical image segmentation with parameter-efficient fine-tuning. *Knowledge-Based Syst.* **310**, 112909, DOI: <https://doi.org/10.1016/j.knosys.2024.112909> (2025).
16. Hu, J., Xu, X. & Zou, Z. Lora-medsam: Efficient medical image segmentation. In Su, R., Frangi, A. F. & Zhang, Y. (eds.) *Proceedings of 2024 International Conference on Medical Imaging and Computer-Aided Diagnosis (MICAD 2024)*, 154–164 (Springer Nature Singapore, Singapore, 2025).
17. Kumar, G. M. K., Chadha, A., Mendola, J. D. & Shmuel, A. Medvisionllama: Leveraging pre-trained large language model layers to enhance medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 1114–1124 (2025).
18. Wang, Z., Wang, Z. & Sun, P. Deep learning model for gastrointestinal polyp segmentation. *PeerJ Comput. Sci.* **11**, e2924 (2025).
19. Jha, D. *et al.* Kvasir-seg: A segmented polyp dataset. In Ro, Y. M. *et al.* (eds.) *MultiMedia Modeling*, 451–462 (Springer International Publishing, Cham, 2020).
20. Sharma, P., Gautam, A., Maji, P., Pachori, R. B. & Balabantaray, B. K. Li-segpnnet: Encoder-decoder mode lightweight segmentation network for colorectal polyps analysis. *IEEE Transactions on Biomed. Eng.* **70**, 1330–1339, DOI: [10.1109/TBME.2022.3216269](https://doi.org/10.1109/TBME.2022.3216269) (2023).
21. Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
22. Aghajanyan, A., Gupta, S. & Zettlemoyer, L. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In Zong, C., Xia, F., Li, W. & Navigli, R. (eds.) *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 7319–7328, DOI: [10.18653/v1/2021.acl-long.568](https://doi.org/10.18653/v1/2021.acl-long.568) (Association for Computational Linguistics, Online, 2021).
23. Zhang, C. *et al.* Delving deep into the generalization of vision transformers under distribution shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7277–7286 (2022).
24. Liang, W., Zhang, W., Yan, H., Zong, Y. & Yan, C. Dp-fvit: Differentially private federated vision transformer with lora for secure and accurate medical image classification. *Biomed. Signal Process. Control.* **114**, 109388, DOI: <https://doi.org/10.1016/j.bspc.2025.109388> (2026).
25. Milletari, F., Navab, N. & Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, 565–571, DOI: [10.1109/3DV.2016.79](https://doi.org/10.1109/3DV.2016.79) (2016).
26. Hu, E. J. *et al.* Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations* (2022).

Funding

The Article Processing Charge for this publication was funded by the Deanship of Graduate Studies and Scientific Research at Qassim University (QU-APC-2026). The authors declare that no other external funding was received for this study.

Acknowledgments

The researchers would like to thank the Deanship of Graduate Studies and Scientific Research at Qassim University for financial support (QU-APC-2026).

Author contributions statement

U.H. conceived the experiment, developed the code, and conducted the training and evaluation. M.A.N. provided supervision and reviewed the manuscript. All authors approved the final manuscript.

Data availability

The Kvasir-SEG dataset analyzed during the current study is available in the Simula Research Laboratory repository: <https://datasets.simula.no/kvasir-seg/>.

Additional information

Competing interests: The authors declare no competing interests.