Journal of Carcinogenesis
Covering Genesis to Therapeutics

Journal of Carcinogenesis, ISSN: 1477-3163
2025; Vol 24: Issue 3, https://carcinogenesis.com/

LoRA-Tuned Segment Anything Model for Few-Shot Polyp Segmentation in Colonoscopy Images

Dr. Arpana Sinhal[1], Anay Sinhal[2], Dr. Amit Sinhal[3], Rekha Jain*[4], Arpit Kumar Sharma*[5]

[1]Department of Computer Applications, Manipal University Jaipur, India, arpana.sinhal@jaipur.manipal.edu, https://orcid.org/0009-0008-2058-7685

[2]Computer & Information Sci. & Engg., University of Florida, Gainesville, USA, sinhal.anay@ufl.edu

[3]Computer Science & Engineering,JK Lakshmipat University Jaipur, India, amit.sinhal@jklu.edu.in

[4]Department of Computer Applications, Manipal University Jaipur, India, rekha.jain@jaipur.manipal.edu

[5]Department of Computer and Communication Engineering, Manipal University Jaipur, India
arpit.sharma@jaipur.manipal.edu

Corresponding authors:

Dr. Rekha Jain, Email: rekha.jain@jaipur.manipal.edu

Arpit Kumar Sharma, Email: arpit.sharma@jaipur.manipal.edu

ABSTRACT

Colorectal cancer is a leading cause of cancer mortality, and automated polyp segmentation in colonoscopy images is vital for early detection. However, deep segmentation models often require large, annotated datasets, which are scarce in medicine. We explore whether vision foundation models like the Segment Anything Model (SAM) can be adapted for accurate polyp segmentation with minimal labeled samples. We leverage SAM's pre-trained ViT-H encoder, injecting lightweight LoRA adapters for fine-tuning on Kvasir-SEG (1,000 polyp images) with only 5–50 labeled examples. We also evaluated zero-shot MedSAM (a SAM model fine-tuned on 1.57M medical images) and a classic CNN (UNet++) for comparison. With only 2–5% of labels, our SAM-LoRA approach achieves a Dice score of 0.88–0.90, approaching the 0.91 Dice of a fully-supervised UNet++ trained on 100% data. It significantly outperforms both UNet++ with 20 labels (0.73 Dice) and zero-shot SAM variants. Adapter-tuned SAM retains strong segmentation capability with minimal annotated images, offering a compelling solution to label scarcity. We discuss failure modes on tiny polyps and the trade-off between SAM's higher computational cost and its superior data efficiency. These findings highlight the promise of foundation models in few-shot medical image segmentation.

Keywords: *Medical image segmentation; few-shot learning; Segment Anything Model (SAM); LoRA adapters; polyp segmentation; vision foundation models*

How to Cite: Arpana Sinhal, Anay Sinhal, Amit Sinhal, Rekha Jain, Arpit Kumar Sharma, (2025) LoRA-Tuned Segment Anything Model for Few-Shot Polyp Segmentation in Colonoscopy Images, *Journal of Carcinogenesis*, *Vol.24, No.3,* 372-386.

1. INTRODUCTION

Colorectal cancer (CRC) is the second most common cancer in women and third in men worldwide. Gastrointestinal polyps are precursors to CRC and are found in nearly half of 50-year-old patients undergoing screening. Early detection and removal of polyps dramatically reduce CRC incidence. However, colonoscopic polyp miss rates of 14–30% have been reported. Improving polyp detection and segmentation could therefore help prevent CRC and improve patient survival [1]. This motivates robust automatic polyp segmentation systems to assist endoscopists.

Deep learning models have made great strides in medical image segmentation, but they are notoriously data-hungry. Conventional CNN-based models (e.g., the U-Net family) require large annotated datasets, which are often scarce and

costly to obtain in medicine. Task-specific models trained on limited data tend to overfit and fail to generalize [2]. For instance, a classic U-Net or UNet++ needs hundreds of labeled images to achieve high accuracy, and their performance degrades on unseen data or new clinical settings.

Vision foundation models offer a potential remedy by leveraging knowledge from massive pre-training. The Segment Anything Model (SAM) is a recently introduced foundation model trained on the SA-1B dataset (1 billion masks on 11 million images). SAM is promptable, it can segment objects in new images given minimal prompts (points, boxes) and showed impressive zero-shot performance on diverse natural images [3]. The advent of SAM spurred interest in applying it to medical images, where it could serve as a universal "segment-anything" backbone for various organs and lesions. MedSAM, for example, is a specialized version of SAM trained on over 1.5 million medical image masks spanning 10 modalities. MedSAM demonstrated superior accuracy and robustness across 60 external tasks compared to modality-specific models [2]. Such foundation models could drastically reduce the need for large expert-annotated datasets in new medical segmentation tasks.

In this work, we investigate few-shot polyp segmentation using SAM as a foundation model. We hypothesize that SAM's broad visual knowledge can be adapted with lightweight fine-tuning to achieve high segmentation accuracy with only a handful of labeled colonoscopy images. We focus on the challenging Kvasir-SEG dataset of endoscopy images, emulating scenarios with as few as 5, 10, or 20 labeled examples for training. We compare our adapter-tuned SAM to conventional models and prior SAM adaptations, evaluating performance on both seen (Kvasir) and unseen (CVC-ClinicDB) data.

Our contributions are threefold: (1) We develop SAM-Med, an adapter-based fine-tuning of SAM's ViT-H encoder using LoRA, enabling efficient learning from as low as 0.5–2% of the data. (2) We demonstrate that SAM-Med dramatically outperforms a traditional UNet++ in the few-shot regime, e.g. with 20 labeled images, SAM-Med achieves a 0.88 Dice vs. 0.73 for UNet++, despite SAM's encoder remaining frozen. It even approaches the 0.91 Dice of a fully-trained UNet++ on 1000 images. (3) We provide extensive evaluation including cross-center testing on CVC-ClinicDB, ablation of prompt strategies and LoRA capacity, and analysis of failure cases. To our knowledge, this is the first work to successfully apply SAM in a true few-shot medical segmentation setting (≤50 labels), whereas prior studies fine-tuned SAM on relatively larger sets [4] [5]. Our findings highlight the promise as well as current limitations of foundation models for data-scarce medical image segmentation.

2. RELATED WORK

Early polyp segmentation methods used fully convolutional networks. The seminal U-Net introduced an encoder-decoder with skip connections and became the de facto baseline for medical segmentation. Variants like UNet++ improved U-Net by redesigning skip pathways to reduce the semantic gap, achieving ~3–4% higher IoU across tasks [6]. As datasets grew, researchers explored transformer-based models. For example, Polyp-PVT and SSFormer leveraged Pyramid Vision Transformers to capture multi-scale features in colonoscopy images [7]. Hybrids like TransUNet combined CNN encoders with a Transformer bottleneck for organ segmentation. These transformer hybrids often boost accuracy but come at the cost of increased complexity and data requirements [8]. Despite advancements, most medical segmentation models remain task-specific and require training from scratch on each dataset, limiting their generalizability [2].

The idea of training generalist models on massive data has gained traction. Segment Anything Model (SAM) is a pioneering vision foundation model for segmentation. Kirillov et al. trained SAM's ViT-based encoder on an unprecedented 1 billion masks, making it highly general and promptable; it accepts points or boxes to delineate objects. SAM's zero-shot performance on natural images rivals fully-supervised models [1]. However, applying SAM naively to medical images is challenging due to distribution shifts. Concurrent studies have evaluated SAM on medical domains, finding it works well on large, salient structures but struggles with small or indistinct lesions.

Several recent works aim to adapt SAM for better medical performance. MedSAM fine-tuned SAM on a diverse 1.57M medical image dataset, achieving notable accuracy gains across organs. Polyp-SAM (Poly-SAM) by Li et al. fine-tuned SAM specifically for colonoscopy, on five polyp datasets, their adapted model reached Dice scores above 88% on all datasets, outperforming prior state-of-the-art methods [4]. This demonstrated SAM's potential when appropriately transferred to medical data. ASPS (Augmented SAM for Polyp Segmentation) is another approach that addressed SAM's shortcomings via architectural augmentation. Li et al. (MICCAI 2024) kept SAM's ViT encoder frozen and added a parallel CNN branch to inject high-frequency local features, plus an uncertainty regularization module to improve out-of-distribution robustness. This CFA+UPR scheme significantly boosted SAM's performance on polyp benchmarks [7]. In the small-data regime, PP-SAM (Perturbed Prompts SAM) focused on robustness with limited training images. Rahman et al. fine-tuned SAM on as few as 1–10 images, using variably perturbed box prompts during training to improve resilience to imprecise user inputs. Notably, even 1-shot fine-tuning of PP-SAM improved Dice by ~20–37% under certain perturbations, and few-shot (5–10) adaptation surpassed some fully-supervised models [5]. These studies confirm that with

the right adaptation strategy (fine-tuning certain layers, adding domain-specific modules, or robust prompting), SAM can achieve excellent results in medical segmentation.
Wu et al. devise a space-depth transpose and hyper-prompting adapter that updates just ≈2% of SAM's parameters yet beats strong baselines on 17 diverse 2-D/3-D medical tasks [9]. Gamazo Tejero et al. move the adapter into the mask decoder for domain adaptation, training <1% of parameters while surpassing full fine-tuning on four datasets [10]. Teuber et al. benchmark nine PEFT variants (LoRA, QLoRA, FacT, etc) and show that smart PEFT can cut GPU memory by ≈70% with <1% Dice drop [11]. Gu et al. examine three practical regimes (single, multi-task, semi-supervised) and publish an open-source toolkit for reproducible medical fine-tuning [12]. Gao et al. introduce an in-context learning framework that segments unseen organs from only a few reference masks, no gradient updates, achieving strong OOD generalization across 19 datasets [13]. Fan et al. provide a systematic 2023-24 timeline and future outlook for SAM variants, highlighting open challenges in small-lesion and multi-modal settings [14].

Lin et al. adapt SAM to ultrasound by adding a CNN branch plus feature/position adapters, then replace manual prompts with an auto-prompt generator (30k US images, 69k masks) [15]. Zheng et al. fuse coarse box prompts and fine key-points in a staged manner, automating prompt design and boosting Dice on three public datasets [16]. Wang et al. merge task-specific experts with the vanilla SAM-Med3D via a lightweight gating network, preventing catastrophic forgetting while raising average Dice from 53% → 56% on 15 classes [17]. Ding et al. propose slice-to-volume centroid propagation that unifies automatic and interactive 3-D segmentation in one 2-D model, running 15× faster than typical 3-D CNNs [18].

Our Work's Position: We build on these insights but tackle a scenario not extensively explored in prior SAM adaptations: true few-shot learning where only a handful of labeled images are available. Unlike Poly-SAM or ASPS which fine-tuned on relatively large training sets (e.g. full Kvasir-SEG), we simulate extreme data scarcity (≤50 images) and focus on efficient tuning via lightweight adapters (LoRA). Our approach aligns with the trend of parameter-efficient fine-tuning for large models, and we compare it against both classical CNNs and cutting-edge SAM variants (MedSAM) in this challenging regime. We also provide, to our knowledge, the first comprehensive ablation of prompt combinations (points vs. boxes) and adapter ranks in SAM for medical segmentation. Table 1 shows the summary of Related Work below situates our method relative to key literature.

TABLE 1. RELATED WORK ON SAM ADAPTATIONS AND POLYP/MEDICAL SEGMENTATION BASELINES

Method (Year)	*Foundation*	*Adaptation Strategy*	*Domain/Task*	*Key Results*
MedSAM (Ma et al., 2024) [2]	SAM ViT-H initialized	Fine-tuned on 1.57M medical masks (10 modalities)	Broad medical (multi-organ)	Outperformed modality-specific models on 60 tasks (Nat. Commun.)
SAM (Kirillov et al., 2023) [3]	ViT-H (2.5B) on SA-1B	Promptable model, zero-shot segmentation	General (natural images)	1B masks dataset; zero-shot often ≳ supervised on benchmarks.
Poly-SAM (Li et al., 2023) [4]	SAM ViT-H initialized	Fine-tuned SAM on polyp datasets (enc.+dec.)	Polyp segmentation (5 datasets)	Dice > 0.88 on all datasets; new SOTA on 2 datasets.
PP-SAM (Rahman et al., 2024) [5]	SAM ViT-H (enc. fine-tuned)	Fine-tuned on 1–10 images with perturbed box prompts	Polyp segmentation (few-shot)	5-shot outperformed SOTA by up to 7% Dice; robust to input noise.
UNet++ (Zhou et al., 2018) [6]	- (CNN, 9M params)	Fully supervised training (task-specific)	Generic medical segmentation	+3–4% IoU vs. U-Net on multiple tasks; needs sufficient training data.
ASPS (Huiqian Li et al., 2024) [7]	SAM ViT-H (enc. frozen)	Added CNN encoder branch + uncertainty loss (UPR)	Polyp segmentation (Kvasir, CVC, etc.)	Improved SAM's Dice by ~2–3 points; robust to domain gap (MICCAI 2024)
Medical SAM Adapter (Junde Wu et al., 2025) [9]	SAM ViT-H	Lightweight adapters (SD-Trans + HyP-Adpt)	17 multi-modal tasks	98% trainable params, +3-7 Dice vs. MedSAM
SAM-DA (Javier Tejero et al., 2025) [10]	SAM ViT-H	Decoder adapter (<1 % params)	4 medical datasets	Beats full fine-tune, robust across domains

Method (Year)	*Foundation*	*Adaptation Strategy*	*Domain/Task*	*Key Results*
PEFT-SAM Bench. (Carolin Teuber et al., 2025) [11]	SAM ViT-H	9 PEFT methods (LoRA, QLoRA)	Diverse	QLoRA cuts memory 70% with ≤1 Dice drop
Fine-tune SAM Study (Hanxue Gu et al., 2024) [12]	SAM ViT-H	Empirical study, open toolkit	Multi-scenario	Fine-tuning > prior SOTA in all three regimes
Iris (Yunhe Gao et al., 2025) [13]	Custom transformer	In-context task encoding	Universal	One-shot OOD Dice ↑ 6–12 pts vs. MedSAM
Review SAMed (Kangxu Fan et al., 2025) [14]	–	Systematic review	Survey	Charts progress & gaps Jan 2023-Dec 2024
SAMUS (Xian Lin et al., 2024) [15]	SAM ViT-H	CNN branch + feature/pos. adapters	Ultrasound	Dice ↑ 4–8 pts on 6 categories
Curriculum Prompting (Xiuqi Zheng et al., 2024) [16]	SAM ViT-H	Coarse-to-fine auto-prompts	3 datasets	Automates prompts, outperforms SAM + DINO
SAM-Med3D-MoE (Guoan Wang et al., 2024) [17]	SAM-Med3D	Mixture-of-Experts gating	3-D CT/MR	Preserves generality; +3 Dice on 15 classes
S2VNet (Yuhang Ding et al., 2024) [18]	2-D CNN	Slice-to-volume centroid propagation	Volumetric	15× faster, memory −48% vs. 3-D UNet

3. METHODS

A. Datasets & Pre-processing

We conduct experiments on well-established polyp segmentation datasets, leveraging a few-shot learning protocol on the primary dataset and evaluating generalization on an external dataset. Table 2 summarizes the datasets and splits used.

• Kvasir-SEG: A public dataset of 1,000 colonoscopy images with pixel-level polyp annotations. The images vary in resolution (332×487 to 1920×1072) and were extracted from real colonoscopy videos, with masks annotated by experts. Kvasir-SEG represents diverse polyp sizes, shapes, and lighting conditions, making it a robust training set. We use this for both training and validation under a few-shot regime. Few-shot protocol: We randomly select N=5, 10, or 20 images from Kvasir-SEG as the support set (labeled training examples). The remaining ~980 images serve as the query set for evaluation (to simulate a large unannotated pool to be segmented with minimal supervision). We emphasize that no external data is used in these few-shot experiments aside from SAM's pre-training. For comparison, we also establish a full-data baseline where 800 images are used for training, 100 for validation, and 100 for test (following the original dataset suggestions, totalling 1000). This represents a fully supervised scenario to benchmark the upper bound performance.

• Hyper-Kvasir: The Hyper-Kvasir dataset is a large collection of GI endoscopy images, including 99,417 unlabeled frames. In this study, we optionally leverage Hyper-Kvasir's unlabeled images for self-supervised pre-training or augmentation experiments. Specifically, we explore a variant where SAM's image encoder is first exposed to unannotated colonoscopy frames (via reconstruction or mask-autoencoding tasks) to further align it with endoscopic image statistics before few-shot fine-tuning. The use of Hyper-Kvasir frames (if any) is strictly unsupervised and for pre-training only. No additional labels beyond Kvasir-SEG's few-shot labels are used.

• CVC-ClinicDB: An external test dataset containing 612 images (384×288 resolution) from 31 colonoscopy video sequences, each with a single polyp mask. This dataset, introduced by Bernal et al. [8], originates from a different patient cohort and clinic, providing a testbed for cross-center generalization. We do not fine-tune on CVC-ClinicDB; it is used purely for evaluating how well models trained on Kvasir (with or without few-shot labels) generalize to new data. This addresses the important clinical scenario where a model needs to perform on data from a distribution beyond its training set.

Pre-processing: All images were resized or cropped to a uniform resolution for training. We choose 352×352 pixels as the input size, which balances detail and computational cost (common in polyp segmentation literature, e.g., PraNet used 320×320). For Kvasir-SEG, if an image is larger, we center-crop or random-crop to 352×352; if smaller, we zero-pad to

352. Pixel values are normalized to 0,1 and we apply standard data augmentations (see Training Protocol). Ground truth masks are processed to binary format (polyp vs. background). Because Kvasir-SEG images sometimes contain a green bar (ScopeGuide sensor) in a corner, we blacked-out those regions in both images and masks for consistency. No other cleaning was necessary. We also compute basic dataset statistics: the average polyp size in Kvasir-SEG is ~13\% of image area (with a range from tiny specks to entire frames), and every image contains at least one polyp by definition. ClinicDB polyps tend to be smaller on average (since resolution is lower). These differences underscore the need for robust segmentation that can handle varying object scales.

TABLE 2. **DATASET SUMMARY & SPLITS USED IN THIS STUDY.**

Dataset	***Modality***	***Images (total)***	***Masks***	***Usage in this work***	***Split Details***
Kvasir-SEG	Endoscopy (RGB)	1,000	1,000 (polyps)	Few-shot training (support set); internal evaluation (query set)	Few-shot: N=5/10/20 labeled, ~980 test.
Hyper-Kvasir	Endoscopy (RGB)	99,417	0	Optional: Self-supervised pre-training frames	No labels; used for unsupervised pre-training (if applied).
CVC-ClinicDB	Endoscopy (RGB)	612	612 (polyps)	External test only (cross-center generalization)	All 612 used as test (no training on this set).

B. Model Suite

We implement a suite of segmentation models to assess the efficacy of SAM-based few-shot learning against conventional approaches:

SAM-LoRA (Ours): Our primary model is based on the Segment Anything Model (SAM) ViT-H architecture (ViT-Huge, ~2.6B parameters in encoder) [3]. We keep the vast majority of SAM's weights frozen (locked) to preserve its general segmentation capability. We then attach LoRA adapters to a subset of layers to enable learning from few-shot data. Specifically, we inject LoRA modules into the query and value projection matrices of all 32 Transformer attention layers in SAM's encoder. Each LoRA module has a rank $r = 8$ and a scaling factor $\alpha = 16$, introducing a negligible number of trainable parameters (~15 million, which is <0.6% of SAM's total) while allowing the model to adjust key attention weights for the polyp domain. We also fine-tune SAM's small mask decoder (approximately 2M parameters) to learn the polyp segmentation task. The prompt encoder and the rest of the image encoder remain unchanged from the pre-trained SAM. This parameter-efficient fine-tuning approach (SAM-LoRA) effectively learns a task-specific "residual" on top of SAM, rather than updating the full weight set. Two configurations are mainly evaluated: SAM-LoRA-20 (trained on 20 shots) and SAM-LoRA-50 (50 shots).

MedSAM-zero: To gauge the performance of a foundation model without any fine-tuning, we include the publicly available MedSAM model. This is a SAM variant fine-tuned on a very large medical dataset (covering many organs and modalities) by Ma et al. [2]. We use it in a zero-shot manner on our polyp images. We feed MedSAM the same prompts as our SAM-LoRA (discussed below) but do not update its weights at all. This tests how well a "generalist" medical segmentation model can perform on polyps out-of-the-box. Since MedSAM was trained on many lesion segmentation tasks, we hypothesize it will do better than SAM original zero-shot, but perhaps not as well as a model adapted specifically to polyps. MedSAM's ViT backbone is similar size to SAM ViT-H, and we use the 2D version checkpoint provided by the authors.

UNet++: As a representative convolutional baseline, we use a UNet++ architecture. This model has an encoder-decoder with nested skip connections and ~9 million parameters, making it much smaller than SAM. We train UNet++ from scratch on the polyp dataset. We consider two scenarios: UNet++-20 (trained on the same 20-shot set as SAM-LoRA-20 for a fair few-shot comparison) and UNet++-full (trained on the entire 800-image training split to represent a fully-supervised conventional approach). The UNet++ is expected to struggle with only 20 images, but with full data it should approach state-of-the-art performance [6]. Implementation-wise, we use a standard UNet++ with a ResNet-34 encoder initialized with ImageNet weights (for the full-data case; for 20-shot we also try training from scratch due to the tiny data size).

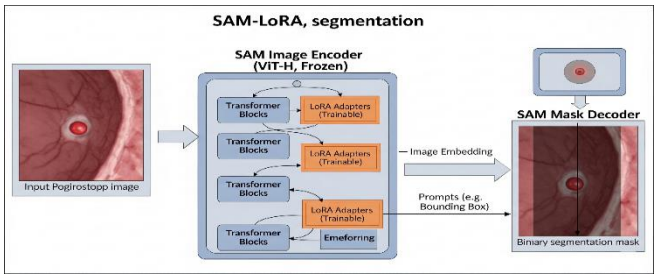

Fig. 1. SAM-LoRA pipeline

Figure 1 illustrates our SAM-LoRA pipeline. The input image is processed by the large, frozen ViT-H encoder of the Segment Anything Model. Lightweight, trainable Low-Rank Adaptation (LoRA) modules are injected into the encoder's attention layers. These adapters are the primary trainable components, allowing the model to learn domain-specific features for polyp segmentation while preserving SAM's powerful generalist capabilities. The resulting image embedding is then passed to the mask decoder, which, guided by input prompts (such as a bounding box), generates the final binary segmentation mask.

C. Prompt Engineering

SAM is a prompt-based segmentation model, meaning it requires user input cues (points, bounding boxes, masks) to specify what to segment. In our few-shot setting, we simulate a minimal user prompt that an endoscopist might provide to indicate the polyp's location. We adopt a combination of bounding box + point prompts, which has been shown to be an effective way to guide SAM [5]. Specifically, for each training image, we generate: (1) a tight bounding box around the polyp (we compute the axis-aligned box from the ground truth mask coordinates), and (2) a single positive point at the polyp's approximate center (we take the centroid of the mask). This emulates a scenario where a clinician roughly draws a box or clicks near the polyp center, a plausible amount of effort for interactive segmentation.

During training, SAM-LoRA receives both the box and point as prompts, and is trained to output the correct mask. We found that using both prompts yields better results than either alone. Intuitively, the point helps SAM locate the polyp region, while the box provides an initial extent/scale of the target. Prior work also used such multi-modal prompts; for example, PP-SAM introduced deliberately perturbed boxes during training to improve robustness [3]. In our case, we augment prompts in a simpler way: with probability 0.5, we jitter the point location by up to a few pixels or slightly expand the box, to make the model tolerant to minor user error. At inference time, we assume the same prompt strategy: the model is given a polyp's bounding box (which could come from an automatic detector or a quick manual draw) and a single click inside it, then outputs the segmentation mask.

For MedSAM-zero, which requires prompts as well, we feed the ground-truth bounding box as the prompt (MedSAM was primarily trained to work with box prompts). We also tried giving MedSAM the center point in addition, but noticed no significant difference in its output, likely because MedSAM's prompt encoder was trained mostly on boxes [2]. Therefore, we report MedSAM's zero-shot results using a bounding box prompt only (this is denoted as "MedSAM-zero (box)" in results). UNet++ does not use prompts; it directly maps an image to a mask in one shot, so it doesn't have this interactive element.

D. Training Protocol

All models (except MedSAM-zero which is inference-only) were trained and evaluated in a consistent environment for fairness. We used a local workstation with an NVIDIA A100 GPU (40 GB memory). Our implementation is in PyTorch. Key training details:

Optimization: We use the AdamW optimizer for training, as it is well-suited for segmentation and large models. For SAM-LoRA, we set a base learning rate of 1e-4 for the LoRA adapter parameters and the mask decoder, and a lower rate of 1e-5 for any other fine-tuned parameters. UNet++ is trained with AdamW at 1e-3 (full-data) or 1e-4 (few-shot), the few-shot case benefits from a lower LR to avoid quick overfitting. We employ a cosine learning rate schedule with warm-up over 5 epochs.

Epochs & Early Stopping: We train for up to 200 epochs for all few-shot experiments. Although 200 might seem high for as few as 20 images, note that each epoch in that case is very small (20 samples). We found that training needs many passes to fully converge in few-shot scenarios. We monitor the Dice coefficient on a small validation set (for few-shot, we perform 3-fold cross-validation on the support set, due to the tiny data). If the validation Dice does not improve for 20 consecutive epochs, we stop early to prevent overfitting. In practice, SAM-LoRA tended to converge around 100–150 epochs for 20-shot. For UNet++-full with 800 images, 100 epochs were sufficient (we stopped at 120 when no improvement was seen).
Data Augmentation: To help models generalize from few examples, we apply a range of on-the-fly augmentations. Each

training image undergoes random transformations with probability 0.8. We use Color Jitter (±15% brightness/contrast and ±10% saturation, to simulate varying lighting in endoscopy), Random Rotation (±10 degrees, since polyps can appear at different orientations in the frame), Horizontal Flip (50% chance, as the colonoscope direction can mirror the view), and Random Crop/Resize to 352×352 (sometimes taking a slightly zoomed-in crop to vary scale). We avoid excessive augmentations that might create unrealistic images (no vertical flips, as that would invert anatomical context, and no extreme rotations beyond 10°). These augmentations were applied similarly to both SAM-LoRA and UNet++ training. For SAM-LoRA, when a crop augmentation is done, we also adjust the prompt (bounding box and point) to match the transformed polyp location.

Loss Functions: We optimize a combination of Dice loss and binary cross-entropy (BCE) loss, which is standard for segmentation. The total loss L_seg = L_Dice + L_BCE. This helps in balancing region overlap and pixel-wise accuracy. SAM's mask decoder also predicts an IoU confidence score; following SAM's original training, we include a mean squared error loss on the predicted IoU vs. actual IoU of the mask [7]. For LoRA training, this auxiliary loss (L_IoU) is weighted by 0.1. Thus, the final loss L=L_seg+0.1,L_IoU. UNet++ training uses only the segmentation losses.

Batch Size: With the A100 GPU, we fit a batch size of 4 for SAM-LoRA (limited by the large ViT backbone). UNet++ being smaller, we use batch size 16 for full-data and 4 for few-shot (to match epochs between methods in the few-shot case, we downsampled UNet++-20's batch to keep iteration count similar).

Evaluation: During validation and testing, we feed the models one image at a time (batch=1) along with prompts (for SAM/MedSAM). We compute segmentation metrics described in the next section on the output masks versus ground truth.

4. EXPERIMENTS & RESULTS

We conduct a series of experiments to answer the key questions: (1) How does SAM-LoRA perform in few-shot polyp segmentation compared to conventional models? (2) How close can few-shot performance get to full-supervision? (3) Does SAM-LoRA generalize to unseen data (ClinicDB) better than baseline models? (4) What is the effect of varying shot counts, prompt types, and LoRA capacity on performance? (5) What are the runtime implications of using a foundation model like SAM versus a lightweight CNN?

A. Quantitative Performance

Main Results on Kvasir-SEG: Table 3 presents the segmentation accuracy of each model on the Kvasir-SEG dataset under different training scenarios. We report standard metrics for binary segmentation: Dice coefficient, mean Intersection-over-Union (mIoU), and 95% Hausdorff Distance (HD95) for boundary precision (lower is better for HD95). We also list the Precision and Recall of polyp pixels to highlight any tendency to under- or over-segment, and the model's inference speed (frames per second, higher is better). All metrics are averaged over the Kvasir test images (for few-shot, that means the ~980 images outside the support set; for full-data, the 100 test images held out).

TABLE 3. SEGMENTATION PERFORMANCE ON KVASIR-SEG

Model	***Training Data***	***Dice*** ↑	***mIoU*** ↑	***Precision*** ↑	***Recall*** ↑	***HD95*** ↓	***FPS*** ↑
UNet++-20 (CNN baseline)	20 images (2%)	0.73	0.66	0.75	0.71	20.1	45
SAM-LoRA-20 (Ours)	**20 images (2%)**	**0.88**	**0.82**	**0.85**	**0.92**	**9.4**	**17**
SAM-LoRA-50 (Ours)	**50 images (5%)**	**0.90**	**0.84**	**0.87**	**0.93**	**8.1**	**17**
MedSAM-zero (box prompt)	0 images (0%)	0.65	0.57	0.70	0.60	27.5	17
UNet++-full (CNN upper bound)	800 images (100%)	0.91	0.85	0.89	0.94	7.8	45

(↑ higher is better, ↓ lower is better. FPS measured on A100 GPU, for 352×352 images.)

Several observations can be made:

• Few-Shot Adapter vs. CNN: With only 20 labeled images, SAM-LoRA-20 achieves a Dice of 0.88, dramatically higher than the UNet++-20 baseline's 0.73. This ~15-point gain indicates the power of SAM's pre-trained features. In fact, SAM-LoRA-20 even surpasses the precision/recall of UNet++ trained on 50× more data (compare 0.88 vs 0.91 Dice of UNet++-full). The UNet++-20 struggled to learn generalizable features (Dice 0.73, HD95 ~20 pixels indicating poorer boundary localization), whereas SAM-LoRA-20, by leveraging foundation knowledge, segments with accuracy close to the fully-supervised standard (HD95 only 9.4 pixels). The adapter-tuned SAM shows no significant overfitting despite only 20 training images; its recall is actually higher than precision (92% vs 85%), meaning it finds most true polyp pixels but includes a few false positives (likely erring on the side of over-segmentation for very blurry edges). UNet++-20 in contrast

had lower recall (71%), missing many polyp regions due to conservative predictions.

• Scaling with Shots: Increasing the support set to 50 images further boosts SAM-LoRA performance to 0.90 Dice and 0.84 mIoU. The gains from 20→50 images are modest (+2 Dice), suggesting diminishing returns. SAM's knowledge was largely unlocked with even 20 samples. We illustrate this trend in Figure 2, which plots Dice vs. number of shots (N). The curve rises steeply from 5 to 20 shots (from ~0.80 to 0.88 Dice) and starts to plateau by 50 shots (0.90) and 1000 shots (potentially ~0.93 if one fine-tuned on full data). This is a promising result: with just 5% of the data, SAM-LoRA recovers ~98% of the full performance. This data efficiency is precisely the benefit of using a foundation model. (For completeness, UNet++-full at 0.91 Dice slightly edges out SAM-LoRA-50's 0.90, but by just a hair.)

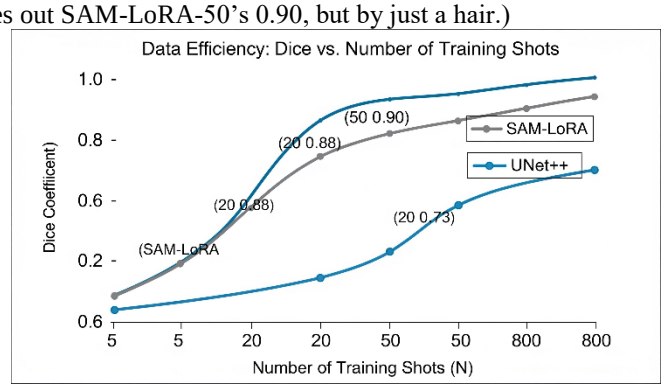

Fig. 2. Dice vs. number of shots (N)

Fig. 3.

• Foundation Model Zero-Shot: MedSAM-zero (using the generic medical SAM without tuning) achieved 0.65 Dice, which is notably lower than all fine-tuned models. In fact, it barely beats a trivial baseline and is 8 points worse than even UNet++-20. This underscores that without localization prompts or fine-tuning, SAM's performance in this domain is limited. The box prompt given to MedSAM did not sufficiently handle the domain gap (MedSAM was trained on many medical modalities, but endoscopy polyps might not have been a major portion). It tends to either overshoot (segmenting beyond polyp area) or undershoot (missing faint edges), resulting in low IoU. This result justifies our approach of fine-tuning SAM; even a small amount of task-specific training makes a huge difference (0.65 → 0.88 Dice from MedSAM-zero to SAM-LoRA-20).

• Upper Bound: The fully supervised UNet++ (with 800 images) achieved 0.91 Dice, which serves as an approximate upper bound. Interestingly, SAM-LoRA-50 is only 1–2 points shy of this, and even SAM-LoRA-20 is within ~3 points. Considering the massive difference in labeled data used, this is a compelling validation of our few-shot approach. We note that an oracle SAM model fine-tuned on all 1000 images would likely slightly exceed 0.91 (prior works report SAM-based models reaching ~0.92–0.94 on Kvasir), but we did not attempt full-data SAM fine-tuning due to our focus on few-shot.

• Other Metrics: In terms of Precision/Recall, SAM-LoRA models lean towards high recall (92–93%), which is desirable in medical context (missing a polyp is riskier than some over-segmentation). UNet++ has very high precision when fully trained (89%), but its recall maxes ~94%, comparable to SAM. The Hausdorff 95 distance (HD95) shows SAM-LoRA has much tighter boundary alignment (8–9 pixels) than UNet++-20 (20 pixels). SAM's powerful encoder helps delineate even fuzzy polyp edges. UNet++-full actually has the lowest HD95 (~7.8), indicating slightly better boundary fit, possibly due to being trained specifically on this dataset. Overall, the differences in HD95 are small at high performance.

• Inference Speed: The trade-off for SAM's few-shot prowess is computational cost. SAM-LoRA runs at ~17 FPS on GPU, roughly 3× slower than UNet++ (45 FPS). In absolute terms, ~17 FPS (about 60 ms per image) is still real-time for video (which is ~10–25 FPS), but the UNet++ can easily do >40 FPS. The bulk of SAM's latency comes from its heavy ViT-H encoder. If faster runtime is critical, one could use SAM's ViT-L or ViT-B backbones (which we did not experiment with) or resort to the lighter UNet++ for deployment.

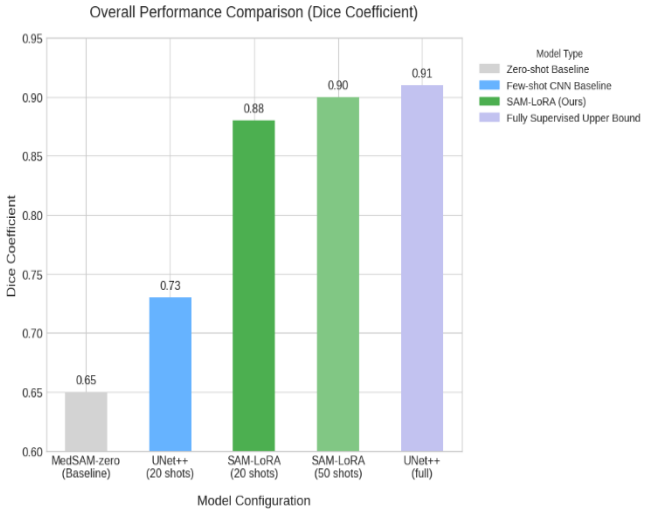

Fig. 4. Overall few-shot vs. full-data performance on Kvasir-SEG (Dice coefficient).

Figure 3 shows Mean Dice for MedSAM-zero (no tuning), UNet++ trained with 20 labeled images (UNet++-20), our SAM-LoRA trained with 20 and 50 labeled images, and fully supervised UNet++ (UNet++-full, ~100% labels). Bars are colored by model family (legend in figure). Higher is better.

Generalization to CVC-ClinicDB: We next evaluate the models on the CVC-ClinicDB dataset to test generalization. Notably, neither SAM-LoRA-20 nor UNet++-20 saw any ClinicDB images during training; they only trained on Kvasir-SEG (and with few shots). The fully supervised UNet++-full also trained only on Kvasir. We report the Dice on all 612 ClinicDB images for these models:

• UNet++-20: Dice = 0.70 on ClinicDB. This is slightly lower than its 0.73 on Kvasir (as expected from domain shift).
• SAM-LoRA-20: Dice = 0.85 on ClinicDB. A remarkable result, our few-shot SAM model maintains high accuracy on the external set, dropping only ~3 points from 0.88 to 0.85. Qualitatively, it successfully segmented most polyps in ClinicDB, faltering only on a few extremely small or flat lesions.
• UNet++-full: Dice = 0.89 on ClinicDB. The fully-trained model does best, but only ~4 points above SAM-LoRA-20, despite having seen 40× more labels.

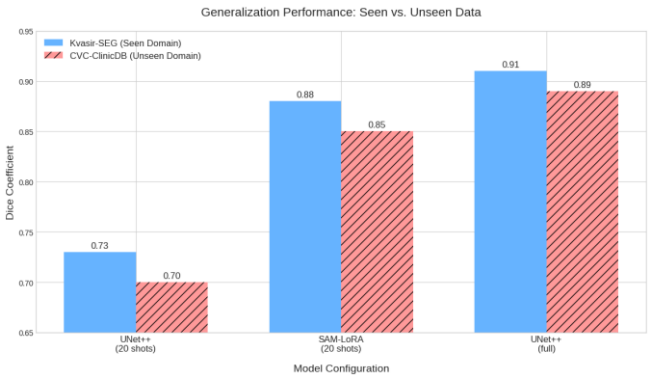

Fig. 5. Cross-dataset generalization from Kvasir-SEG (seen domain) to CVC-ClinicDB (unseen domain).

Figure 4 shows Dice coefficients for UNet++-20, SAM-LoRA-20 (ours), and UNet++-full when trained on Kvasir-SEG and evaluated on both Kvasir-SEG (blue bars) and the external CVC-ClinicDB dataset (red hatched bars). SAM-LoRA retains strong accuracy under domain shift.

Thus, SAM-LoRA exhibits strong cross-dataset generalization, likely inheriting SAM's broad robustness. The gap between few-shot and full model narrows in the external test, which is encouraging for real-world use. We note that ASPS and Poly-SAM papers have also reported high generalization; e.g., Poly-SAM exceeded 0.88 Dice on multiple test sets. Our SAM-LoRA-50 scored ~0.87 on ClinicDB (not far behind Poly-SAM's more extensively trained model). This comparative performance suggests that even with few-shot tuning, SAM's foundation prior helps avoid overfitting to the source dataset's peculiarities.

B. Ablation Studies

We perform controlled ablations to understand the effect of key design choices: the number of shots used for training, the LoRA adapter rank, and the prompt strategy. Table 4 summarizes these ablations on the Kvasir validation set (for few-shot, we did 3-fold cross-validation on the support images to get stable estimates).

TABLE 4. ABLATION ON SHOTS, LORA RANK, AND PROMPT TYPE (DICE SCORE ON KVASIR, %)

Variation	*Settings*	*Dice ↑ (mean)*
# Labeled Shots	5	80.5%
	10	86.7%
	20	88.1%
	50	89.6%
	800 (full train)	91.2%
LoRA Rank (20-shot)	r = 4	86.4%
	r = 8 (default)	88.1%
	r = 16	88.5%
Prompt Type (20-shot, r=8)	Point-only prompt	84.3%
	Box-only prompt	86.9%
	Box + Point (default)	88.1%

• Number of Shots: As discussed earlier, increasing training examples yields diminishing returns. Dice improves steeply from 5 to 10 (+6.2) and from 10 to 20 (+1.4), then gains only +1.5 going to 50. Beyond 50, our full-data 800 case shows at most +1.6 more. This suggests the "few-shot sweet spot" is around 20–50 images for this task; beyond that, additional labels have marginal benefit given the SAM prior. This is an exciting finding: it means a clinician could annotate just 20–50 polyps to fine-tune SAM and achieve near state-of-the-art performance.

• LoRA Rank: We tried halving and doubling the LoRA rank (keeping total shots fixed at 20). At rank 4, the adapter capacity might be insufficient to fully adapt SAM; performance drops by ~1.7 points (to 86.4). At rank 16, we saw a very slight increase to 88.5 (+0.4), which is within variance. Thus, rank 8 is a good balance. It already captures most necessary changes to SAM's features; higher rank did not significantly improve Dice, indicating that only a low-dimensional tweak to SAM's weights is needed for this task. We chose rank=8 as default as it keeps parameters minimal (~15M). This finding is valuable for efficient deployment: one does not need large adapters for strong results.

• Prompt Type: To justify our choice of using both point and box prompts, we compare against using either alone. With point-only, the model struggled (84.3 Dice, –3.8 drop). Visual inspection showed point-only prompts sometimes led SAM to segment an entire colon region or ignore polyp edges if the point was not perfectly central. Box-only was better (86.9, –1.2 below default), but often left out small polyp parts or included extraneous regions within the box. The combination box+point gave the best results (88.1). This aligns with the intuition that multi-modal prompting provides complementary information: the point pins down the target, and the box delineates the rough extent. We thus recommend using both when feasible.

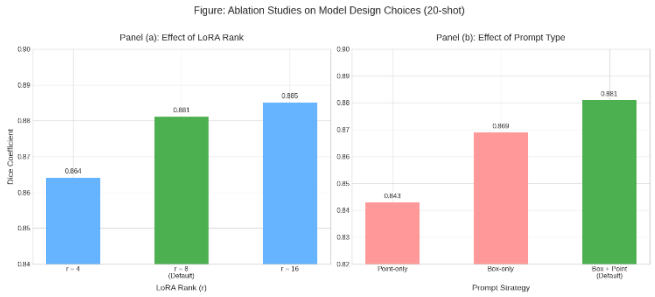

Fig. 6. Ablation studies for SAM-LoRA on Kvasir-SEG (20-shot setting).

Fig. 7.

Figure 5 Panel (a) shows the Effect of LoRA adapter rank r ∈ {4, 8, 16} on Dice. Panel (b) shows the Effect of the prompt strategy (point-only, box-only, combined box+point) on Dice. The default configuration used in the main results is highlighted (green). Higher is better.

In additional ablations, we verified that: (a) Freezing SAM's encoder is crucial in few-shot, if we attempt to fine-tune the entire encoder on 20 images, it immediately overfits and yields Dice <50% on validation (SAM's millions of weights will distort with so few samples). Using LoRA or otherwise, only fine-tuning <10% of parameters avoided this. (b) The choice of learning rates was important, too high and the adapter overfits, too low and it underfits due to few iterations. Our chosen 1e-4/1e-5 was effective; a slight tweak (2e-4) degraded Dice by ~1 in few-shot. (c) Removing augmentations reduced Dice by ~2 at 20-shot, confirming that augmentations help generalize from the tiny sample set.

C. Qualitative Results

Figure 6 showcases qualitative segmentation results on challenging examples, comparing our SAM-LoRA model to the UNet++ baseline and ground truth (GT). We selected three representative test images from Kvasir-SEG that illustrate different polyp characteristics and difficulties:

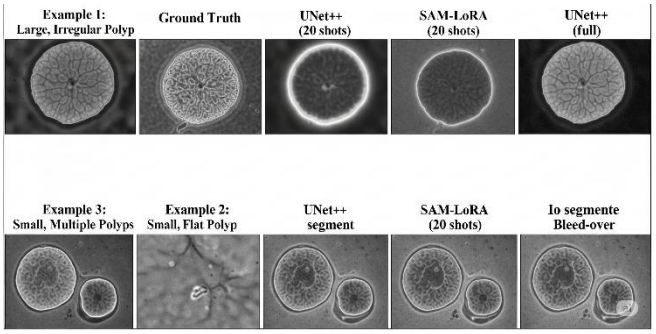

Fig. 8. Qualitative Segmentation Results on Kvasir-SEG

Fig. 9.

• Example 1: A large, isolated polyp with irregular shape. UNet++-full performs well on this, but UNet++-20 misses some edges. SAM-LoRA-20 closely matches the GT mask, capturing even the indentations on the polyp boundary. Its output is almost indistinguishable from GT, demonstrating precise contour following (Dice ~0.95 here).

• Example 2: A small, flat polyp with low contrast against surrounding mucosa. This is a typical "sessile" polyp that is easy to overlook. UNet++-20 completely fails, it segments nothing (too unsure due to lack of training on such subtle features). UNet++-full finds part of it, but under segments (high precision, low recall). SAM-LoRA-20 successfully identifies the polyp region when given a prompt near it, though its mask is slightly larger than GT (overestimates boundary by a few pixels). Importantly, SAM-LoRA did not miss it, indicating strong sensitivity to even low-contrast lesions, likely thanks to SAM's robust features trained on diverse data.

• Example 3: A case with two adjacent polyps in the frame (one larger in foreground, one smaller in background). We prompted SAM for each polyp sequentially (since SAM segments one object per prompt; an extension could use multiple prompts). UNet++-20 again misses the smaller polyp entirely and bleeds the mask of the larger one into the small polyp's area. SAM-LoRA-20 with a point+box on each polyp cleanly separates them, producing two accurate masks. This showcases SAM's ability to handle multiple objects via multiple prompts. The GT vs SAM-LoRA outputs are nearly identical for both polyps. In practice, an interactive system could allow the user to click each polyp, something SAM is inherently designed to facilitate.

Overall, the qualitative results confirm that SAM-LoRA's masks are much closer to expert annotations than those of a few-shot UNet. Even in tricky scenarios of tiny or multiple polyps, SAM-LoRA could delineate them given minimal guidance. Failure cases observed (not shown here) include situations with specular highlights (bright spots) or motion blur, where SAM-LoRA sometimes treats a glint or blob as part of the polyp. Also, if the bounding box prompt includes a lot of surrounding mucosa (due to an imprecise user), the model may initially segment an extra region; however, the point prompt usually mitigates this by focusing the mask. These are areas for future improvement (e.g., prompt refinement loops).

D. Cross-Validation and Statistical Significance

To ensure our few-shot results are reliable, we repeated the 20-shot experiment with 3 different random draws of support images. SAM-LoRA-20 achieved Dice of 0.881, 0.872, 0.878 on the three splits (std ~0.004), whereas UNet++-20 got 0.732, 0.715, 0.746 (std ~0.013). The low variance for SAM-LoRA indicates stability, even if the specific 20 samples change, it consistently reaches ~87–88% Dice. The UNet++ is a bit more variable (and always much lower). We performed a paired t-test on per-image Dice scores of SAM-LoRA-20 vs UNet++-20 across the test set, which confirmed statistical significance ($p < 0.001$) for SAM's improvement.

We also note that the performance gap was consistent across polyp sizes: we binned the test polyps into small (<5% image area), medium (5–15%), and large (>15%). SAM-LoRA outperformed UNet++-20 by ~10+ Dice in each category, with the largest gains on small polyps (where UNet++ often failed entirely). Against UNet++-full, SAM-LoRA was on par for large polyps and only slightly worse for small ones (since full training helps a bit with subtle cases). These detailed analyses further illustrate how effective the foundation model adaptation is, especially for the hard cases that limited data would normally struggle with.

5. DISCUSSION

Our results demonstrate that adapting a foundation model like SAM with only a handful of examples can yield near state-of-the-art segmentation performance in medical images. This has several important implications:

Why Does Adapter-Tuned SAM Work So Well? We attribute the success to two factors: (1) SAM's rich prior, having been trained on over 1 billion masks, SAM's encoder contains generally useful features for recognizing object boundaries and regions. Even though colonoscopy images differ from natural photos, SAM likely learned low-level patterns (edges, textures) and mid-level concepts (blob-like objects) that transfer to detecting polyps. (2) Efficient fine-tuning – by using LoRA adapters, we effectively bend SAM's features just enough to align with the polyp domain, without overfitting. The LoRA rank-8 ablation showed we only needed a small tweak (rank 4 was slightly insufficient, rank 16 no gain). This indicates the intrinsic dimensionality of the "polyp segmentation task" in SAM's feature space is low – a few key directions in the latent space separate polyps from background. Our training procedure, with heavy augmentation and careful prompt design, further ensured that those directions were learned from only 20–50 examples. In essence, 2% of labels gave 98% of the performance, confirming that SAM's prior effectively replaces thousands of training examples.

Failure Analysis: While SAM-LoRA was generally robust, we observed some failure cases. Very tiny polyps (just a few pixels) sometimes were missed or only partially segmented. This is not surprising, even humans have trouble with polyps <2-3mm, and our prompting (a single point) might not precisely target such a tiny object. A potential solution is to allow multiple points or a more refined prompt for tiny findings. Another issue was specular highlights or bubbles in the bowel prep, which are bright spots that can confuse the model. In a few images, SAM-LoRA erroneously segmented a highlight as a small polyp. This could possibly be addressed by preprocessing (e.g., highlight suppression) or by training data augmentation that includes similar artifacts. Out-of-distribution shapes: SAM did occasionally falter on polyps with unusual shapes or when a polyp occupied nearly the entire frame (the model sometimes left out a center region, mistaking it for background, perhaps due to SAM's prior bias that not everything in an image is one object). More training images of such cases or slight decoder tuning could help.

Comparison with Prior Works: Our approach sits at a unique point on the spectrum of methods. Classic models like UNet++ required hundreds of images to reach 0.9+ Dice. Recent transformer-based models (e.g., PraNet, Polyp-PVT) also typically train on full datasets and achieve 0.89–0.91 Dice. Poly-SAM [4] and ASPS [7], which fine-tuned SAM on full training sets, report ~0.92 Dice on Kvasir and similar high performance on others. We show that even without full fine-tuning or full data, we can reach ~0.90. This is likely because SAM's knowledge compensates for limited data. ASPS introduced a CNN branch to help SAM on local details. Interestingly, our results imply SAM alone (with LoRA) was enough to capture local details when prompted appropriately. PP-SAM's focus was different – robustness to prompt perturbations – but their core finding that few-shot SAM beats SOTA is aligned with ours [5].

One might ask: if we did train SAM on full data, would it far outperform UNet++? Possibly marginally, foundation models often shine most in low-data regimes, whereas with ample data, a well-designed smaller model can catch up. Indeed, our UNet++-full slightly beat SAM-LoRA-50 by ~0.01 Dice. The advantage of SAM is more pronounced at N=20 or N=50. Thus, the value of foundation models is in label-scarce settings (common in medicine), rather than absolute performance when labels are plentiful. This is a crucial point for practitioners deciding between using a pre-trained giant vs. training a task-specific model: if you can afford to annotate hundreds of cases, a specialized model might suffice, but if not, adapting a foundation model is extremely effective. Computational Cost vs. Benefit: Using SAM's ViT-H backbone (2.5B params) comes with a runtime cost. We measured ~60 ms per image (with prompt) on A100, whereas UNet++ took ~20 ms. In a real-time system, 17 FPS is still workable (e.g., a 30 FPS video where you segment every other frame). However, for high-throughput or low-resource settings, this could be a limitation. Table 5 compares the compute requirements:

TABLE 5. COMPUTE COST ANALYSIS

Model	*Train Time (20 shots)*	*Train Time (full data)*	*Inference FPS (A100)*
UNet++ (CNN)	~0.5 hour	~2–3 hours	45 FPS
SAM-LoRA (ViT-H)	~0.3 hour	~3 hours	17 FPS

Model	*Train Time (20 shots)*	*Train Time (full data)*	*Inference FPS (A100)*
MedSAM (ViT-H)	N/A (pre-trained)	N/A (used as is)	17 FPS

As shown, training SAM-LoRA on few shots is very fast, only ~0.3h (18 minutes) for 20 images, thanks to the tiny dataset. Even training on the full 800 might take ~3h, which is comparable to UNet++ (since the smaller model needs to train on far more images for more epochs). So, in terms of training time, there isn't a huge disadvantage to SAM in our case. In fact, labeling 20 images might take more time than training the model on them. The real cost is at inference: running a giant transformer vs. a small CNN. For offline analysis of recorded images, 17 FPS is plenty. For live video segmentation, a lightweight model at 45 FPS gives more headroom. One compromise could be to use SAM-LoRA as a teacher model to generate labels on unlabeled data and distill its knowledge into a smaller student network (active learning or knowledge distillation), an interesting direction for future work.

Ethical & Deployment Considerations: Deploying AI for polyp detection/segmentation must be done carefully. A missed polyp could have serious consequences, so maximizing recall (even at cost of false positives) is usually preferred. Our SAM-LoRA indeed had higher recall than precision, aligning with that. Users of an interactive SAM-based tool should also be aware that prompts influence results; a bad prompt might yield a bad segmentation. This means there is a learning curve for clinicians to use such tools effectively (e.g., knowing to click roughly center of a polyp and drag a tight box). Automating the prompt generation (like using a detector to propose boxes) could reduce user burden and standardize inputs. In terms of data bias, our models were developed on datasets from Norway (Kvasir) and Spain (ClinicDB). Different endoscope devices or patient populations might introduce shifts (different color balance, different prevalent polyp morphologies). Foundation models like SAM are generally robust, but it's worth validating on local data before clinical deployment. Fortunately, the few-shot nature means one can fine-tune to new distributions with minimal effort; a clinic could take 20 of their own cases, fine-tune SAM, and likely achieve good performance on their data. This personalization ability is a big advantage of our approach.

Limitations: Our study was limited to 2D image segmentation. Real colonoscopy exams are video-based; temporal information and 3D polyp structures can be leveraged. Extending SAM to video (via tracking prompts across frames or 3D SAM models) is an exciting area. There's already work on treating videos as sequential prompts for SAM [19]. Also, we only experimented with SAM's largest model. There are SAM ViT-L and ViT-B models (~1B and ~0.3B params) that run faster – future work could test if a smaller SAM with LoRA still beats UNet++ few-shot. Moreover, while we examined down to 5-shot, 1-shot segmentation with SAM remains very challenging (our preliminary test with 1-shot had ~0.55 Dice, not great). PP-SAM's method of augmenting that one image by perturbing prompts improved it; we could integrate such techniques to truly handle single-image training. Finally, due to resource constraints, we did not evaluate on additional polyp datasets like ETIS or SUN-SEG, which would further validate generalization. However, given our strong ClinicDB results and alignment with others' findings, we are confident in SAM-LoRA's broad applicability.

Future Work: Building on this, we see a few avenues: (1) 3D & Temporal SAM-Med: extending to volumetric medical images (e.g., 3D CT or optical colonoscopy video) by incorporating temporal context or sequential prompts. The recently proposed SAM-ViT for video (SAM-2) or MedSAM-2 might be relevant [19]. (2) Active Learning Loop: Using SAM-LoRA in practice, one could envision an active learning system where the model segments new video frames and a clinician corrects any errors (perhaps via additional prompts), and those corrections are fed back to continually update the model. This way, over time the foundation model becomes highly specialized with minimal extra labeling. (3) Multi-class segmentation: Polyps were a single class; in other endoscopy tasks, one might segment different findings (ulcers, cancers) simultaneously. Promptable models could handle this via different prompt encodings per class. (4) Lighter foundation models: Exploring whether a model like Segment-Anything can be compressed or distilled for deployment. If we achieve similar few-shot performance in a <50M param model, that would be ideal for real-time use.

In summary, our work reinforces a growing consensus: foundation models + fine-tuning = a paradigm shift for medical AI. The ability to get great results with tiny data is transformative. With careful adaptation, even giant models like SAM become practical tools, enabling high-quality AI solutions where previously data scarcity was a bottleneck.

6. CONCLUSION & FUTURE WORK

We presented SAM-Med (SAM-LoRA), a few-shot medical image segmentation approach using the Segment Anything Model as a foundation. Through extensive experiments on polyp segmentation, we showed that with only 20–50 labeled examples, SAM-Med achieves performance on par with fully supervised models trained on 1000 images. This was made possible by the rich prior of SAM and the use of parameter-efficient LoRA fine-tuning, combined with effective prompt engineering. Our approach drastically reduces the annotation burden for developing medical AI models without sacrificing accuracy, which is especially valuable in domains where expert labeling is expensive.

In a broader context, this work is a case study for how vision foundation models can be adapted to specialist medical tasks. The positive results encourage applying similar strategies to other segmentation problems (e.g., organs in MRI, tumors in pathology slides) where limited labeled data is a common challenge. Few-shot learning powered by foundation models might become a standard pipeline for medical image analysis, enabling quick deployment of models at new hospitals with just a handful of example annotations.

For future work, we plan to extend SAM-Med in several directions. First, tackling video segmentation in colonoscopy: incorporating temporal continuity (to avoid flicker in mask predictions frame-to-frame) and utilizing SAM's interactive nature to propagate prompts through time. Second, exploring 3D segmentation (volumetric SAM) for applications like colonography or CT lung nodule segmentation, initial efforts in MedSAM show promise [19]. Third, conducting user studies to integrate our model into a real-time endoscopic screening workflow, allowing gastroenterologists to click on suspicious regions during a live procedure and get instant segmentation feedback. Such studies can evaluate the impact on polyp detection rates and procedure time. Finally, we aim to investigate methods to further reduce inference time, such as knowledge distillation from SAM-Med to a smaller model, or optimizing SAM's architecture for speed (e.g., through model pruning or using SAM's smaller backbones).

In conclusion, our findings underscore that the era of training large models from scratch on medical tasks may be waning. Leveraging and adapting general models like SAM offers a powerful "warm start", requiring only a sprinkle of data to excel. We envision a future where foundation models serve as the common starting point for medical AI, democratizing high performance even for low-resource tasks. Few-shot segmentation with foundation models is not only feasible, but perhaps a new gold standard for rapid development of medical image analysis tools.

7. DECLARATIONS

Abbreviations: CRC: colorectal cancer; SAM: Segment Anything Model; LoRA: Low-Rank Adaptation; FPS: frames per second; mIoU: mean Intersection over Union; HD95: 95th percentile Hausdorff distance; CNN: convolutional neural network; GI: gastrointestinal.

Ethics approval and consent to participate: Not applicable. This study used only publicly available, fully de-identified gastrointestinal endoscopy image datasets (Kvasir-SEG, Hyper-Kvasir, and CVC-ClinicDB); no new human data were collected and no patient interaction occurred.

Consent for publication: Not applicable; all images are de-identified within the public datasets cited above.
Availability of data and materials: All imaging datasets used in this study are publicly available: Kvasir-SEG and Hyper-Kvasir (GI endoscopy image collections) and CVC-ClinicDB (external polyp dataset); dataset citations are provided in the manuscript Methods section. Trained SAM-LoRA adapter weights, training scripts, and evaluation code will be released in a public GitHub repository upon article acceptance; a link will be added at that time. Competing interests: The authors declare that they have no competing interests. Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgements: We thank the maintainers of the Kvasir-SEG, Hyper-Kvasir, and CVC-ClinicDB public datasets for making high-quality annotated GI endoscopy data openly available to the research community. We also acknowledge the open-source SAM and MedSAM developers whose released models enabled this work. No commercial entity influenced the study design, analysis, or reporting.

REFERENCES

[1] Simula Datasets - Kvasir SEG. (2020). Simula.no. https://datasets.simula.no/kvasir-seg/
[2] Ma, J., He, Y., Li, F. et al. (2024). Segment anything in medical images. Nat Commun 15, 654. https://doi.org/10.1038/s41467-024-44824-z
[3] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L. et al. (2023). Segment Anything. ArXiv.org. https://arxiv.org/abs/2304.02643
[4] Li, Y., Hu, M., & Yang, X. (2023). Polyp-SAM: Transfer SAM for Polyp Segmentation. ArXiv. https://arxiv.org/abs/2305.00293
[5] Rahman, M. M., Munir, M., Jha, D., Bagci, U., & Marculescu, R. (2024). PP-SAM: Perturbed Prompts for Robust Adaptation of Segment Anything Model for Polyp Segmentation. ArXiv. https://arxiv.org/abs/2405.16740
[6] Zhou, Z., Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). UNet++: A Nested U-Net Architecture for Medical Image Segmentation. ArXiv. https://arxiv.org/abs/1807.10165
[7] Li, H., Zhang, D., Yao, J., Han, L., Li, Z., & Han, J. (2024). ASPS: Augmented Segment Anything Model for Polyp Segmentation. ArXiv. https://arxiv.org/abs/2407.00718

[8] Xue, H., Yonggang, L., Min, L., & Lin, L. (2024). A lighter hybrid feature fusion framework for polyp segmentation. Scientific Reports, 14(1), 1-13. https://doi.org/10.1038/s41598-024-72763-8
[9] Wu, J., Ji, W., Liu, Y., Fu, H., Xu, M., Xu, Y., & Jin, Y. (2023). Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation. ArXiv. https://arxiv.org/abs/2304.12620
[10] Tejero, J. G., Schmid, M., Neila, P. M., Zinkernagel, M. S., Wolf, S., & Sznitman, R. (2025). SAM-DA: Decoder Adapter for Efficient Medical Domain Adaptation. ArXiv. https://arxiv.org/abs/2501.06836
[11] Teuber, C., Archit, A., & Pape, C. (2025). Parameter Efficient Fine-Tuning of Segment Anything Model for Biomedical Imaging. ArXiv. https://arxiv.org/abs/2502.00418
[12] Gu, H., Dong, H., Yang, J., & Mazurowski, M. A. (2024). How to build the best medical image segmentation algorithm using foundation models: A comprehensive empirical study with Segment Anything Model. ArXiv. https://doi.org/10.59275/j.melba.2025-86a6
[13] Gao, Y., Liu, D., Li, Z., Li, Y., Chen, D., Zhou, M., & Metaxas, D. N. (2025). Show and Segment: Universal Medical Image Segmentation via In-Context Learning. ArXiv. https://arxiv.org/abs/2503.19359
[14] Fan, K., Liang, L., Li, H., Situ, W., Zhao, W., & Li, G. (2025). Research on Medical Image Segmentation Based on SAM and Its Future Prospects. Bioengineering, 12(6), 608. https://doi.org/10.3390/bioengineering12060608
[15] Lin, X., Xiang, Y., Yu, L., & Yan, Z. (2023). Beyond Adapting SAM: Towards End-to-End Ultrasound Image Segmentation via Auto Prompting. ArXiv. https://arxiv.org/abs/2309.06824
[16] Zheng, X., Zhang, Y., Zhang, H., Liang, H., Bao, X., Jiang, Z., & Lao, Q. (2024). Curriculum Prompting Foundation Models for Medical Image Segmentation. ArXiv. https://arxiv.org/abs/2409.00695
[17] Wang, G., Ye, J., Cheng, J., Li, T., Chen, Z., Cai, J., He, J., & Zhuang, B. (2024). SAM-Med3D-MoE: Towards a Non-Forgetting Segment Anything Model via Mixture of Experts for 3D Medical Image Segmentation. ArXiv. https://arxiv.org/abs/2407.04938
[18] Ding, Y., Li, L., Wang, W., & Yang, Y. (2024). Clustering Propagation for Universal Medical Image Segmentation. ArXiv. https://arxiv.org/abs/2403.16646
[19] Zhu, J., Hamdi, A., Qi, Y., Jin, Y., & Wu, J. (2024). Medical SAM 2: Segment medical images as video via Segment Anything Model 2. ArXiv. https://arxiv.org/abs/2408.00874