FST-SAM3: Taming SAM 3 with Frequency-Spatio-Temporal Refinement for Video Polyp Segmentation

Guanhao Wu[1], Guilian Chen[1], Huisi Wu[1(✉)], and Jin Qin[2,3]

[1] College of Computer Science and Software Engineering, Shenzhen University
[2] Centre for Smart Health, The Hong Kong Polytechnic University
[3] CAS SIAT-PolyU Multi-modal Medical Molecular Imaging Joint Laboratory, The Hong Kong Polytechnic University
hswu@szu.edu.cn

Abstract. Video polyp segmentation (VPS) is important for automated colorectal cancer screening, yet it is a very challenging task due to severe camouflage in mucosal scenes, complex intestinal anatomy, and substantial frame-to-frame appearance variations. Although Segment Anything Model 3 (SAM 3) is able to provide strong representations for segmentation tasks, directly applying it to endoscopic videos usually cannot achieve satisfactory performance owing to domain shifts, boundary ambiguities, and unstable temporal propagation. To address these challenges, we propose FST-SAM3, a novel parameter-efficient adaptation framework built on SAM 3 for VPS. FST-SAM3 introduces two innovative yet complementary components. First, we propose a Structure-Aware Frequency Enhancement (SAFE) module to strengthen structure- and boundary-sensitive cues under endoscopic appearance. Second, we develop a Spatio-Temporal Token Refinement Framework (STRF) to refine temporal information at the token level for more stable cross-frame segmentation. We freeze the SAM 3 backbone and learn only a small set of additional parameters, keeping the inference characteristics close to SAM 3 while improving segmentation performance. We conducted extensive experiments on two benchmarking datasets (SUN-SEG and CVC-612) under standardized fully automatic (0-point) and first-frame prompted (1-point at t=0) protocols. Results demonstrate that our model achieves consistent improvements over state-of-the-art approaches, including both task-specific baselines and representative SAM-based approaches. Code is available at https://github.com/GavonW/FST-SAM3.

Keywords: Video Polyp Segmentation · Segment Anything Model · Spatio-Temporal Modeling · Foundation Model Adaptation · Frequency Enhancement

✉ Corresponding Author

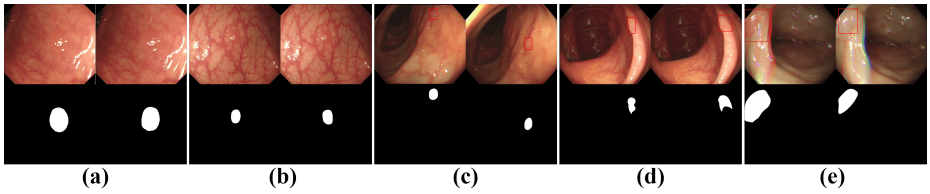

Fig. 1: Challenges in VPS. (a,b) Low-contrast camouflage in two adjacent frames. (c) Position shift, (d) shape change, and (e) scale variation across frames.

1 Introduction

Colorectal cancer (CRC) remains one of the leading causes of cancer-related mortality worldwide [35]. As the standard care tool for screening and prevention, colonoscopy enables effective detection and resection of precancerous polyps, significantly reducing incidence rates. However, in clinical practice, the effectiveness and efficacy of colonoscopy rely heavily on subjective judgments of the endoscopist, and thus the experience of the endoscopist plays an important role in the process. During prolonged procedures, factors such as operator fatigue, complex intestinal anatomy, and the subtle appearance of minute lesions contribute to a polyp miss rate approaching 20% [36]. Consequently, developing reliable automated polyp segmentation systems is essential for assisting decision-making and reducing missed diagnoses. Compared to generic video object segmentation, video polyp segmentation (VPS) exhibits several domain-specific challenges (Fig. 1). Polyps are often camouflaged with low contrast and blurred boundaries (Fig. 1 (a) and (b)), making it challenging to extract discriminative features. Endoscopic motion further induces large variations in viewpoint, scale, and morphology (Fig. 1 (c) to (e)), causing unstable cross-frame propagation and temporal drifts. These challenges complicate obtaining accurate and temporally consistent segmentation in real clinical videos.

Numerous deep learning-based methods have been proposed for polyp segmentation. Early studies [13] mainly focused on 2D images and thus cannot fully exploit spatio-temporal cues. Later works [5, 17] incorporate temporal aggregation, memory mechanisms, or multi-scale fusion to improve robustness. Nevertheless, progress is constrained by limited domain-specific data and the representational capacity of conventional backbones, especially under low-contrast and ambiguous boundaries.

Recently, SAM 3 [2] has emerged as a strong foundation model with improved representations, as well as a memory mechanism inherited from SAM 2 [34], making it a promising backbone for medical video processing tasks. However, directly adapting such large models to endoscopic VPS remains challenging due to the domain shift [18] between natural images and colonoscopy videos and the lack of explicit boundary and structure cues under the standard SAM-style downscaled encoding. Prior attempts either rely on computationally expensive full fine-tuning, or introduce external branches and adapters, which may mismatch

internal features and fail to jointly address boundary ambiguity and temporal instability.

To address these challenges, in this paper, we propose FST-SAM3, a novel SAM 3-based adaptation framework for video polyp segmentation. Our method introduces two innovative complementary components: Structure-Aware Frequency Enhancement (SAFE) on the encoder side to strengthen structure-sensitive and boundary-related representations under endoscopic appearance, and Spatio-Temporal Token Refinement (STRF) on the decoder side to stabilize cross-frame propagation via token-level temporal refinement. Importantly, we adopt a parameter-efficient adaptation strategy by freezing the SAM 3 backbone and learning a small set of additional parameters, which largely preserves the inference characteristics of SAM 3 while improving segmentation accuracy. We extensively evaluate FST-SAM3 on two benchmarking datasets, SUN-SEG and CVC-612, under standardized protocols for both fully automatic inference and SAM-style first-frame prompting. The proposed model achieves consistent improvements over state-of-the-art approaches, including both task-specific baselines and representative SAM-based approaches.

The main contributions of this work are summarized as follows:

- We propose FST-SAM3, a SAM 3-based adaptation framework for video polyp segmentation that targets low-contrast camouflage, boundary ambiguity, and temporal instability in endoscopic videos.
- We introduce SAFE and STRF to respectively enhance structure/boundary-sensitive representations and improve temporal refinement at the token level, without resorting to an additional heavy side network.
- We perform a parameter-efficient adaptation by freezing the SAM 3 backbone and training only a small number of additional parameters, achieving strong performance on SUN-SEG and CVC-612 under both fully automatic (0-point) and first-frame prompted (1-point at t=0) protocols.

2 Related Work

2.1 Polyp Segmentation

Image Polyp Segmentation (IPS) has evolved from early Convolutional Neural Network (CNN) based methods [9,14,31,40] using local features, to hybrid architectures [10, 26, 39] combining CNNs with Transformers [37] for global information extraction. However, these static image-level approaches discard temporal dynamics, leading to severe inter-frame inconsistencies. Video Polyp Segmentation (VPS) addresses this by modeling spatiotemporal correlations using hybrid 2D/3D CNNs [33], normalized self-attention [19,21], or newer paradigms like the memory-enhanced SALI [17] and Mamba-based STDDNet [5].

Despite this progress, task-specific VPS models often have limited network capacity and struggle with highly camouflaged lesions [12, 25]. To address this bottleneck, recent studies adapted large vision foundation models for medical targets [27, 41]. However, bridging the domain gap between natural images and

complex endoscopic scenes [20, 42] while extending static representations to continuous temporal modeling [15, 23] remains an open challenge.

2.2 Foundation Models

Unlike task-specific models, vision foundation models use large-scale pre-training to learn robust representations. The Segment Anything Model (SAM) [24] shows strong zero-shot generalization in static image segmentation, and its successors, SAM 2 [34] and SAM 3 [2], extend this capability to the video domain. However, empirical studies [29, 42] show directly applying these models to medical images causes significant performance drops. This degradation is primarily driven by a severe domain shift characterized by low tissue contrast, ambiguous lesion boundaries, and complex mucosal textures, limiting zero-shot inference effectiveness. Consequently, general-purpose pre-training is insufficient for camouflaged endoscopic scenes, prompting research into parameter-efficient adaptation strategies for clinical applications.

2.3 SAM-based Methods

To mitigate this degradation on specialized medical data, various parameter-efficient adaptation strategies have been proposed for SAM-style architectures [3,8,16]. Many such pipelines initialize with a first-frame prompt (e.g., a point or box), introducing an interaction budget different from fully automatic methods. Early video adaptations like MediViSTA-SAM [23] and VP-SAM [15] incorporate temporal modeling but often add overhead via fixed attention schedules or external side networks. With the release of SAM 3, new methods have emerged, but they still face practical limitations in endoscopic VPS. For example, Medical SAM3 [22] requires costly full fine-tuning, zero-shot SAM 3 suffers from severe domain shift [11], and lightweight adapters like SAM3-Adapter [7] operate frame-by-frame without explicitly modeling temporal continuity.

In parallel, studies on the frequency behavior of transformer attention show that stacked attention layers tend to smooth features and suppress fine details. Methods like FDAM [6] construct complementary responses to compensate for this attenuation. Motivated by these observations, we design an adaptation framework that is parameter-efficient, preserves SAM 3 internal representations, and explicitly addresses boundary ambiguity and temporal instability in endoscopic videos. Our method introduces a Structure-Aware Frequency Enhancement (SAFE) module for structure and boundary-sensitive cues and a Spatio-Temporal Token Refinement Framework (STRF) for token-level temporal refinement, without relying on heavy side networks.

3 Method

3.1 Overall Architecture

Although the Segment Anything Model (SAM 3) shows strong zero-shot capabilities, directly transferring its large-scale pre-trained priors to video polyp

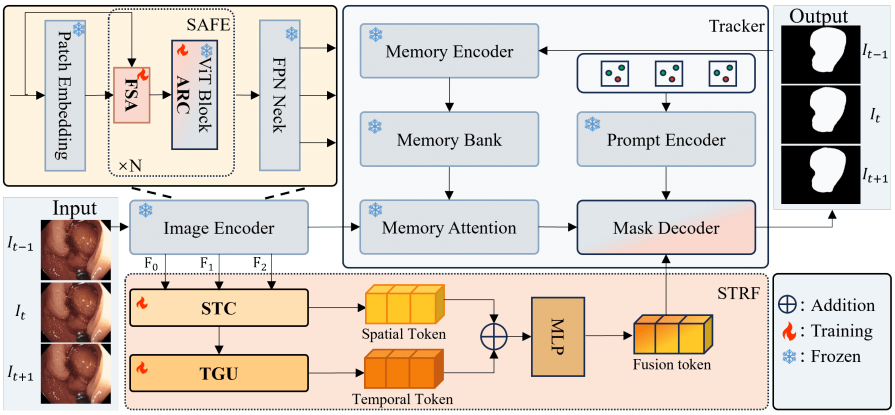

Fig. 2: Overview of FST-SAM3. We augment frozen SAM 3 with SAFE and STRF: SAFE enhances structure/boundary cues in the encoder, and STRF builds spatio-temporal tokens to guide the decoder for coherent video polyp segmentation.

segmentation (VPS) poses two fundamental challenges. First, the endoscopic domain differs substantially from natural images in texture spectra and appearance statistics, leading to severe representation shift. Second, while SAM 3 propagates information across frames via memory, it lacks fine-grained correction to prevent unreliable predictions from being carried forward. As a result, errors can accumulate over time, causing temporal jitter and progressive erosion of delicate structures.

To address these limitations, we keep the SAM 3 backbone frozen and add two lightweight modules: Structure-Aware Frequency Enhancement (SAFE) in the encoder and Spatio-Temporal Token Refinement (STRF) in the decoder, as shown in Fig. 2. SAFE reduces the appearance gap between endoscopic and natural images by strengthening structure cues and high-frequency details, whereas STRF limits error accumulation during propagation through quality-aware spatio-temporal token refinement. Given an endoscopic video, each frame is encoded with SAFE, and STRF maintains spatial and temporal tokens, fuses them into a spatio-temporal token, and replaces the original mask token in the decoder to yield temporally stable segmentation masks.

3.2 Structure-Aware Frequency Enhancement

Polyp regions in colonoscopic videos are often camouflaged and low-contrast. As a result, directly applying SAM 3 to this domain can lead to a noticeable representation gap, which commonly appears as over-smoothed boundaries and missing fine structures. We introduce Structure-Aware Frequency Enhancement (SAFE). As shown in Fig. 3, SAFE inserts a Frequency-Spectrum Adapter (FSA) and an Attention Residual Compensation (ARC) module into the SAM 3 image

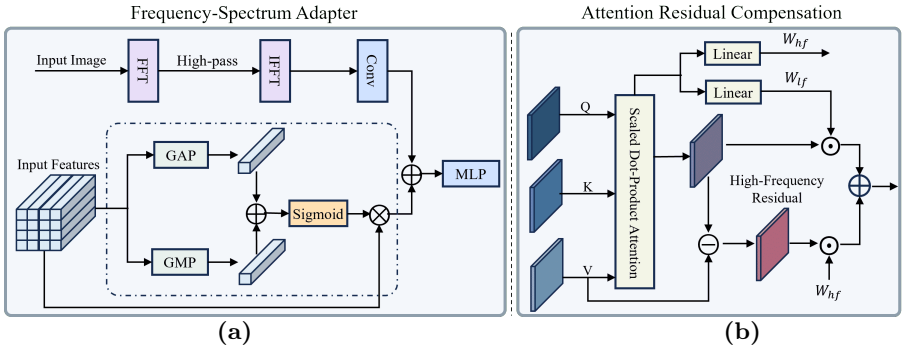

Fig. 3: Architecture of SAFE. (a) Frequency-Spectrum Adapter (FSA) for frequency-based enhancement. (b) Attention Residual Compensation (ARC) for residual compensation.

encoder to strengthen structure cues and preserve fine details that are otherwise weakened by attention.

Frequency-Spectrum Adapter (FSA). FSA adopts a dual-branch design to inject boundary-related high-frequency cues while preserving the semantics of SAM 3 features. For the i-th encoder stage, let $\mathbf{F}_i \in \mathbb{R}^{C_i \times H_i \times W_i}$ denote the stage feature. We compute a high-frequency residual feature $\mathbf{F}_i^{hf}$ from the input frame in the frequency domain. Specifically, given $\mathbf{I} \in \mathbb{R}^{3 \times H \times W}$, we compute its channel-wise frequency representation $z = \mathrm{fft}(\mathbf{I})$ (with $\mathrm{ifft}(\cdot)$ denoting the inverse transform), shift low-frequency coefficients to the center of the spectrum, and apply a binary square high-pass mask $\mathbf{M}_{hp} \in \{0,1\}^{H \times W}$:

$$\mathbf{M}_{hp}(u,v) = \begin{cases} 0, & \left|u - \frac{W}{2}\right| < \ell(r) \ \wedge \ \left|v - \frac{H}{2}\right| < \ell(r), \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

Here $\ell(r)$ controls the side length of the masked low-frequency square and is chosen such that the masked area accounts for a fraction r of the spectrum (i.e., $4\ell(r)^2 \approx rHW$). The high-frequency response is recovered by

$$\mathbf{I}^{hf} = \left|\mathrm{ifft}(z \odot \mathbf{M}_{hp})\right|. \tag{2}$$

Following EVP [28] and our empirical validation, we set $r = 0.25$. In practice, we transform $\mathbf{I}^{hf}$ into a multi-scale high-frequency pyramid through a hierarchical patch-embedding pipeline, yielding stage-wise features $\mathbf{F}_i^{hf} \in \mathbb{R}^{C_i \times H_i \times W_i}$.

The recalibration branch reweights channels to improve robustness under endoscopic appearance variations. We form the gate from two complementary statistics: global average pooling (GAP) summarizes the overall activation of each channel, while global max pooling (GMP) highlights its most salient responses, which often align with structural cues such as boundaries. Their sum, followed by a sigmoid, yields a channel-wise mask that recalibrates the feature

responses, producing the recalibrated feature $\mathbf{F}_i^{rc}$:

$$\mathbf{F}_i^{rc} = \mathbf{F}_i \odot \sigma(\mathrm{GAP}(\mathbf{F}_i) + \mathrm{GMP}(\mathbf{F}_i)), \tag{3}$$

where $\odot$ denotes channel-wise multiplication. We then merge the two branches through an MLP residual adapter:

$$\mathbf{F}_i' = \mathbf{F}_i + \mathrm{MLP}_{up}\Big(\mathrm{GELU}\Big(\mathrm{MLP}_{down}\Big(\mathbf{F}_i^{hf} + \mathbf{F}_i^{rc}\Big)\Big)\Big). \tag{4}$$

Attention Residual Compensation (ARC). Due to the inherent low-pass filtering behavior of standard Transformer blocks, stacked attention layers tend to preserve low-frequency semantics while weakening boundary-level details. ARC addresses this issue inside encoder attention blocks by restoring high-frequency residuals with data-dependent modulation. Let $\mathbf{Y}$ denote the scaled dot-product attention output, and let $\mathbf{V}$ denote the value matrix. We define the high-frequency residual as $\mathbf{V}_{hf} = \mathbf{V} - \mathbf{Y}$.

ARC conditions the modulation on the current attention response. We pass $\mathbf{Y}$ through StarReLU to obtain a dynamic feature $\mathbf{F}_{dy} = \mathrm{StarReLU}(\mathbf{Y})$, then predict low- and high-frequency weights with two linear projections:

$$\mathbf{W}_{lf} = \mathrm{Linear}_{lf}(\mathbf{F}_{dy}), \quad \mathbf{W}_{hf} = \mathrm{Linear}_{hf}(\mathbf{F}_{dy}). \tag{5}$$

The refined attention output is computed as:

$$\tilde{\mathbf{Y}} = \mathbf{Y} + \gamma_{lf}(\mathbf{W}_{lf} \odot \mathbf{Y}) + \gamma_{hf}(\mathbf{W}_{hf} \odot \mathbf{V}_{hf}), \tag{6}$$

where $\odot$ denotes element-wise multiplication and γ_{lf}, γ_{hf} are learnable scaling parameters. Through FSA and ARC, SAFE injects high-frequency cues and recalibrates channel responses, while also recovering high-frequency residuals attenuated by attention, helping the encoder retain sharp boundaries and fine structures.

3.3 Spatio-Temporal Token Refinement Framework

SAM 3 supports memory-based propagation, yet its temporal updates can be brittle on camouflaged polyps with weak boundaries. Minor errors may persist in the propagated state and gradually compound, leading to jitter and boundary drift over time. We therefore introduce the Spatio-Temporal Token Refinement Framework (STRF), which modifies the decoder input by adding a small set of tokens, without changing the frozen SAM 3 backbone. STRF constructs a spatial token as a structural anchor and updates a temporal token to summarize history; the two are fused into a *fusion token* that replaces the mask token in the SAM 3 decoder, improving cross-frame coherence and boundary quality.

Spatial Token Construction (STC). We build a compact spatial token s_t that serves as a structure-aware reference for the decoder. Unlike the mask token used in standard decoders, s_t is derived from multi-scale encoder features and thus summarizes complementary cues across depth and resolution. In particular,

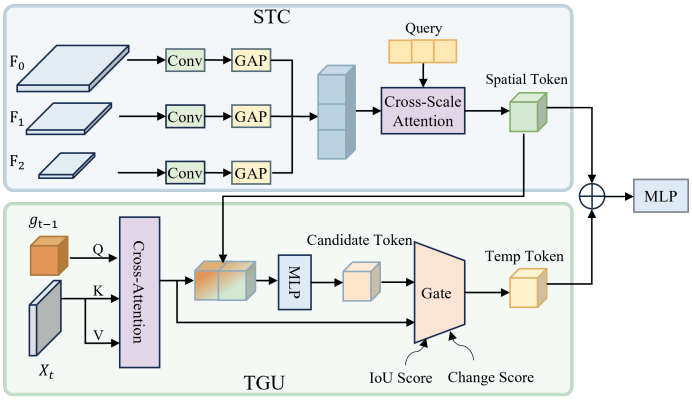

Fig. 4: Architecture of STRF, consisting of Spatial Token Construction (STC) and Quality-Aware Temporal Token-Guided Update (TGU).

it aggregates early-stage features that retain fine textures and edges together with late-stage features that carry stronger semantics, making the resulting token more sensitive to fine-grained structures.

Concretely, we take K=3 encoder feature maps from different stages $\{\mathbf{F}_{t,k}\}_{k=1}^{K}$ and project each map to D=256 with a 1×1 convolution, followed by Global Average Pooling:

$$\mathbf{z}_{t,k} = \mathrm{GAP}\big(\mathrm{Conv}_{1\times1}^{(k)}(\mathbf{F}_{t,k})\big) \in \mathbb{R}^{1\times D}, \qquad k = 1, \ldots, K. \tag{7}$$

We concatenate the scale tokens as $\mathbf{Z}_t = [\mathbf{z}_{t,1}, \ldots, \mathbf{z}_{t,K}] \in \mathbb{R}^{K\times D}$. A lightweight multi-head attention module with h=4 heads aggregates $\mathbf{Z}_t$ using a learnable query $\mathbf{q} \in \mathbb{R}^{1\times D}$:

$$\mathbf{a}_t = \mathrm{MHA}(\mathbf{q}, \mathbf{Z}_t, \mathbf{Z}_t) \in \mathbb{R}^{1\times D}. \tag{8}$$

Finally, we apply a linear projection followed by LayerNorm to obtain the spatial token:

$$s_t = \mathrm{LN}(\mathrm{Linear}(\mathbf{a}_t)) \in \mathbb{R}^{D}. \tag{9}$$

Overall, STC condenses multi-scale encoder cues into a single token that can be directly consumed by the decoder, providing a stable structural anchor for subsequent temporal refinement.

Quality-Aware Temporal Token-Guided Update (TGU). Given s_t, TGU updates the temporal state while reducing error accumulation under inter-frame variations. Let $\mathbf{X}_t \in \mathbb{R}^{B\times HW\times C}$ denote the output of SAM 3 memory attention at frame t. We use the previous temporal token g_{t-1} as the Query to attend to $\mathbf{X}_t$ (Key/Value), yielding an updated historical token $\tilde{g}_t \in \mathbb{R}^{B\times1\times C}$.

To control the write-back to the temporal state for the next frame, we introduce a quality-aware gate based on two scalar signals derived from the current prediction. The first is a prediction-quality score from the decoder's IoU head:

$$q_t = \max\big(\hat{\mathbf{i}}_t\big), \tag{10}$$

where $\hat{\mathbf{i}}_t$ is the IoU-head output over multiple mask candidates at frame t. The second is an inter-frame variation score that reflects prediction instability:

$$\Delta_t = \begin{cases} 0, & t = 0, \\ \dfrac{1}{HW} \displaystyle\sum_{u=1}^{H} \sum_{v=1}^{W} |P_t(u,v) - P_{t-1}(u,v)| \cdot P_t(u,v), & t > 0, \end{cases} \tag{11}$$

where $P_t = \sigma(M_t)$ is the foreground probability map of the selected mask candidate at frame t, chosen by the highest IoU-head score. We refer to Δ_t as the *change score*; larger values indicate more abrupt prediction changes, where smaller candidate updates are preferred.

The gate coefficient is computed as:

$$\alpha_t = \sigma(w_q q_t - w_\Delta \Delta_t)\,, \tag{12}$$

where σ is the Sigmoid function and w_q, w_Δ are learnable scalars. We form a candidate update token by combining the current spatial token with the updated historical token:

$$c_t = \mathrm{MLP}\big([s_t, \tilde{g}_t]\big), \tag{13}$$

and update the temporal token as:

$$g_t = (1 - \alpha_t)\tilde{g}_t + \alpha_t c_t. \tag{14}$$

The updated token g_t is carried to frame $t{+}1$ as a compact historical prior.

For the current frame, we fuse the spatial and historical tokens to form a fusion token:

$$f_t = \mathrm{MLP}(s_t + \tilde{g}_t), \tag{15}$$

which replaces the original mask token in the SAM 3 decoder to predict the polyp mask at frame t. STRF combines STC and TGU to improve video propagation stability. STC summarizes multi-scale structure cues into a spatial anchor, and TGU updates the temporal state with quality-aware gating, which reduces drift and keeps boundaries consistent over time.

3.4 Loss Functions

To train FST-SAM3, we optimize a joint objective consisting of a structure-aware segmentation loss and a temporal consistency regularization.

Structure-Aware Segmentation Loss. Endoscopic polyps often exhibit blurred boundaries and low contrast, where uniform pixel-wise losses can underemphasize boundary regions. Motivated by structure-aware objectives used in polyp segmentation [13], we adopt a loss that up-weights hard boundary pixels. Specifically, the predicted mask P_t from the mask decoder is supervised against the ground-truth mask Y_t using a combination of weighted Binary Cross-Entropy (wBCE) and weighted Intersection over Union (wIoU):

$$\mathcal{L}_{\mathrm{seg}}(P_t, Y_t) = \mathcal{L}_{\mathrm{wBCE}}(P_t, Y_t) + \mathcal{L}_{\mathrm{wIoU}}(P_t, Y_t). \tag{16}$$

Temporal Consistency Loss. To reduce temporal jitter and encourage smooth propagation of historical context, we regularize the latent temporal token maintained by STRF. We penalize abrupt token fluctuations using a margin-based cosine distance:

$$\mathcal{L}_{\text{temp}}(g_t, g_{t-1}) = \max\Big(1 - \cos(g_t, g_{t-1}) - \delta,\, 0\Big), \tag{17}$$

where δ is a tolerance margin. This loss encourages the temporal token to evolve smoothly across adjacent frames.

Final Loss. The final training objective is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}}(P_t, Y_t) + \lambda\, \mathcal{L}_{\text{temp}}(g_t, g_{t-1}). \tag{18}$$

We choose $\lambda = 0.1$ so that $\mathcal{L}_{\text{temp}}$ acts as a mild regularizer relative to $\mathcal{L}_{\text{seg}}$, and set $\delta = 0.3$ to tolerate small token variations while penalizing abrupt changes.

4 Experiments

4.1 Experimental Setups

Datasets. We evaluate on two widely used colonoscopy video segmentation benchmarks, SUN-SEG [21] and CVC-612 [1]. SUN-SEG contains 49,136 frames from 285 sequences and provides an official split into a training set (19,544 frames/112 sequences), SUN-SEG-Easy (17,070/119), and SUN-SEG-Hard (12,544/54). CVC-612 contains 612 frames from 31 sequences; we split it at the sequence level into train/val/test with a 6:2:2 ratio to avoid frame-level leakage. For SUN-SEG, 20% of the training sequences are reserved for validation.

Evaluation metrics. We employ four standard metrics: structure measure (S_α), enhanced-alignment measure (E_ϕ), weighted F-measure (F_β^w), and Dice score.

Implementation details. FST-SAM3 is implemented in PyTorch and trained on a single NVIDIA GeForce RTX 3090 GPU. All frames are resized to 352×352, and the clip length is set to $L = 6$. We train for 15 epochs with batch size 2 using AdamW, with the learning rate decayed from 1×10^{-4} to 1×10^{-5}. In Table 1, **PE** denotes the frozen perception encoder of SAM 3. For a controlled comparison, all SAM-based methods use the same first-frame prompt: one positive point at t=0 sampled from the ground-truth foreground, with a fixed random seed to keep prompt locations identical across methods. Task-specific methods are evaluated in their fully automatic setting without prompts.

4.2 Comparison with State-of-the-art Methods

To demonstrate the superiority of our proposed method, we compare our method with several state-of-the-art (SOTA) methods, including task-specific methods ZoomNeXt [32], MAST [4], PNS+ [21], SALI [17], and STDDNet [5], and SAM-based methods SAM3 [2], CamSAM2 [43], MedSAM2 [30], DAPSAM [38], and VP-SAM [15].

Table 1: Quantitative comparison under two standardized protocols: *Fully Automatic (0-point init)* and *First-frame Prompted (1-point at t=0)*.

Method	Year	Backbone	SUN-SEG-Easy				SUN-SEG-Hard				CVC-612			
			S_α	E_ϕ	F_β^w	Dice	S_α	E_ϕ	F_β^w	Dice	S_α	E_ϕ	F_β^w	Dice
Fully Automatic (0-point init)														
PNS+	2022	Res2Net-50	85.64	87.91	77.24	82.03	84.17	86.54	74.89	78.86	94.60	96.43	89.51	92.89
MAST	2024	PVTv2-B2	85.20	91.58	78.20	79.77	86.25	92.39	78.79	81.05	93.35	96.67	89.07	91.37
STDDNet	2025	PVTv2-B2	89.84	93.43	85.08	85.86	88.75	92.66	83.32	84.59	93.97	95.56	87.78	90.39
ZoomNeXt	2024	PVTv2-B5	88.25	92.02	80.56	86.89	87.64	90.79	80.21	86.50	94.52	97.61	92.29	93.05
SALI	2024	PVTv2-B5	89.16	93.42	83.03	85.26	88.69	92.98	81.30	84.26	94.41	97.30	90.93	93.02
Ours (0-point init)	2026	PE	**92.12**	**95.00**	**86.47**	**89.53**	**92.17**	**94.53**	**85.40**	**89.39**	**95.31**	**98.07**	**92.31**	**94.40**
First-frame Prompted (1-point at t=0)														
SAM3	2025	PE	79.49	80.13	65.98	72.01	81.83	82.49	67.96	75.17	87.38	88.96	76.55	74.81
CamSAM2	2025	Hiera-T	86.11	91.41	81.06	82.75	86.41	90.05	78.15	80.66	91.40	95.89	89.14	90.12
MedSAM2	2024	ViT-B	87.14	90.36	79.26	81.88	84.69	88.19	74.94	77.42	92.56	94.23	85.20	88.56
DAPSAM	2024	ViT-B	82.56	88.61	72.04	73.46	82.67	89.85	72.89	73.18	91.76	95.62	87.75	90.68
VP-SAM	2024	ViT-B	90.89	94.26	84.73	87.56	88.93	92.13	82.98	85.28	93.26	95.75	89.63	92.33
Ours (1-point at t=0)	2026	PE	**92.50**	**95.36**	**87.39**	**90.12**	**92.54**	**95.30**	**87.01**	**90.01**	**95.47**	**98.26**	**92.58**	**94.50**

Table 2: Ablation of SAFE on SUN-SEG (Dice and F_β^w).

FSA	ARC	Easy		Hard	
		F_β^w	Dice	F_β^w	Dice
		80.83	84.00	78.16	82.24
✓		84.18	86.03	82.40	84.82
	✓	83.19	85.57	83.17	86.33
✓	✓	**86.47**	**89.53**	**85.40**	**89.39**

Table 3: Ablation of STRF on SUN-SEG (Dice and F_β^w).

STC	TGU		Easy		Hard	
	no-gate	Full	F_β^w	Dice	F_β^w	Dice
			83.92	86.39	82.18	85.58
✓			86.01	88.42	83.59	86.43
✓	✓		86.33	89.01	84.70	88.21
✓		✓	**86.47**	**89.53**	**85.40**	**89.39**

Comparison with SOTA methods. As shown in Table 1, our proposed method achieves significant performance improvements across all metrics compared to SOTA task-specific methods. This superiority is primarily attributed to our SAFE and STRF modules, which effectively capture high-frequency structural details and maintain spatio-temporal consistency. Furthermore, our approach consistently outperforms recent SAM-based methods. This indicates that we successfully adapt the powerful representation capabilities of SAM 3 to the complex colonoscopy video domain. By robustly disentangling highly camouflaged polyps and exploiting valuable temporal dynamics, our method demonstrates exceptional accuracy and robustness.

Visual Comparison with SOTA. Fig. 5 compares FST-SAM3 with representative task-specific and SAM-based baselines on SUN-SEG without prompts. Competing methods often miss parts of the polyp or bleed into background regions, leading to jagged boundaries under low contrast, specular highlights, and complex folds. FST-SAM3 produces cleaner masks with more accurate boundaries. The gain aligns with our design: SAFE enhances boundary cues in the encoder, and STRF refines spatio-temporal tokens in the decoder to keep propagation stable across frames.

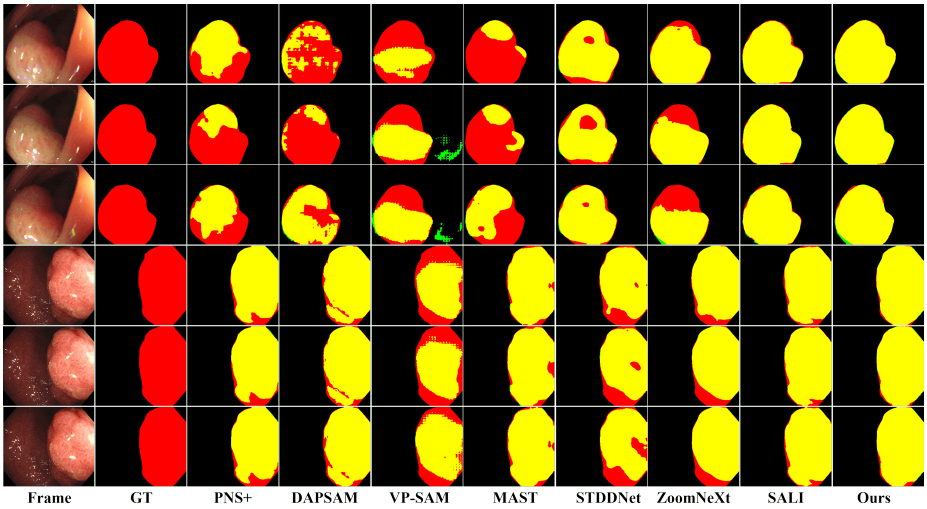

Fig. 5: Visual comparison with SOTA methods on SUN-SEG. Red, green and yellow represent the GT, prediction and their overlapping regions, respectively. (w/o prompts)

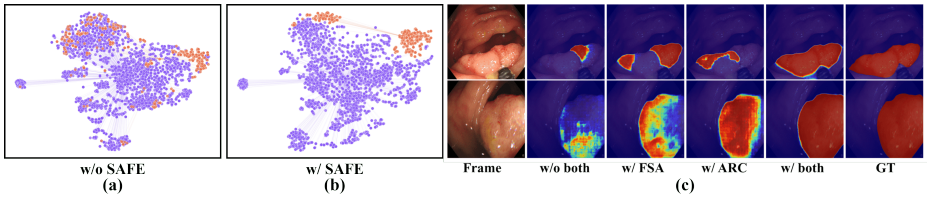

Fig. 6: Visualization analyses for SAFE on SUN-SEG. (a) t-SNE of image embeddings w/o SAFE. (b) t-SNE of image embeddings w/ SAFE. (c) Visual comparison of image embeddings across SAFE components.

4.3 Ablation Studies

Effectiveness of SAFE. Table 2 validates SAFE and its components. Adding FSA brings clear gains, consistent with frequency-based enhancement improving ambiguous boundaries. ARC further improves performance, suggesting that attention-level compensation complements structure enhancement. Fig. 6(a,b) shows reduced foreground–background overlap in the embedding space after SAFE, and Fig. 6(c) indicates stronger structure responses with suppressed background activations under camouflage.

Effectiveness of STRF. Table 3 reports progressive ablations for STRF. STC improves accuracy, indicating that multi-scale features provide a stronger spatial anchor. For TGU, the ungated update improves propagation, and the quality-aware gate yields consistent additional gains, supporting its role in suppressing noisy write-back. Fig. 7(a) further evaluates temporal stability using the CDF of per-sequence Dice variance (computed from frame-wise Dice within each test

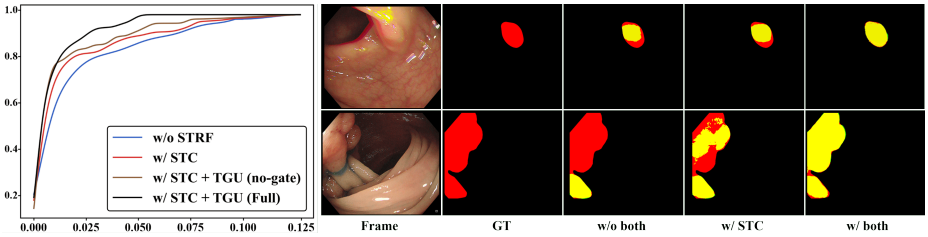

Fig. 7: Analyses of STRF on SUN-SEG. (a) CDF of per-sequence Dice variance (lower is better). (b) Qualitative visualization of STRF ablations.

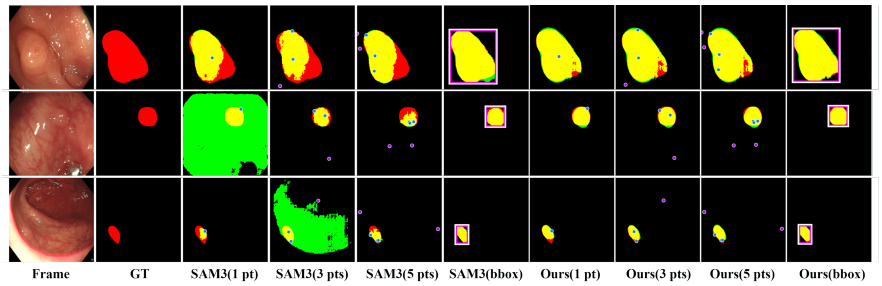

Fig. 8: Comparison under different manual prompts. Red, green, and yellow denote the ground truth, prediction, and overlap. Blue points are positive and purple points are negative.

sequence); the full STRF achieves the most favorable curve. Fig. 7(b) provides qualitative examples.

Impact of Prompting Strategies. Fig. 8 compares SAM 3 and our method under different first-frame prompt types and budgets (1/3/5 points and box prompts). Under sparse point prompts, our method produces more coherent masks with fewer boundary defects, suggesting reduced reliance on precise first-frame interactions. Fig. 9 varies the number of automatically generated positive points provided only at t=0; performance shows diminishing returns as more points are added, and remains strong even with a single point.

Impact of Input Clip Length. Fig. 10 studies the input clip length L. Increasing L from 1–3 to a moderate window improves the median Dice on SUN-SEG-Easy and SUN-SEG-Hard. Performance peaks around L=6–7, while longer clips bring little benefit and can slightly reduce accuracy, likely due to accumulated noise and increased misalignment under rapid motion.

4.4 Discussion and Limitations

FST-SAM3 adapts SAM 3 to colonoscopy videos with a small number of additional parameters. As shown in Table 4, the frozen SAM 3 backbone dominates

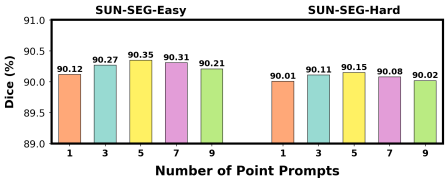

Fig. 9: Ablation on the number of point prompts.

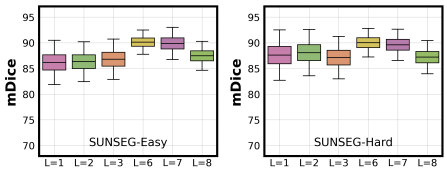

Fig. 10: Ablation on input clip length L.

Table 4: Efficiency comparison of FST-SAM3 vs. SAM 3 on SUN-SEG-Hard (single RTX 3090, batch size = 1).

Method	Dice	Total Params	GFLOPs	FPS
SAM 3 (frozen)	75.17	473.0M	945.32	24.59
FST-SAM3 (Ours)	90.01	476.7M	995.14	16.43
Ours overhead	**+14.84**	**+3.7M**	**+49.82**	**-8.16**

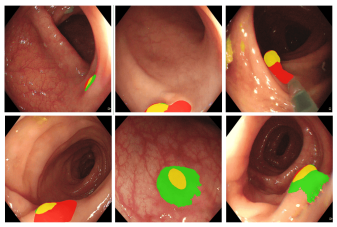

Fig. 11: Failure cases. Red, green, and yellow represent the GT, prediction, and overlapping regions.

both compute and parameter count. With an overhead of +3.7M parameters and +49.82 GFLOPs, FST-SAM3 improves Dice on SUN-SEG-Hard from 75.17 to 90.01, while FPS drops from 24.59 to 16.43 (RTX 3090, batch size = 1). This trade-off suggests that our benefit comes from lightweight adaptation on top of a frozen backbone, rather than reducing backbone complexity.

Failure cases are shown in Fig. 11. The model can fail on flat, highly camouflaged lesions whose texture is close to surrounding mucosa, where both boundary cues and motion cues are weak. Occlusions from instruments and severe truncation can further disrupt temporal context and degrade boundaries. Future work will consider illumination-robust features and explicit occlusion handling.

5 Conclusion

We present FST-SAM3, a parameter-efficient SAM 3 adaptation framework for video polyp segmentation. SAFE enhances boundary cues in the encoder, and STRF refines spatio-temporal information at the token level to stabilize cross-frame propagation. By freezing SAM 3 and training only a small set of additional parameters, FST-SAM3 keeps an inference profile close to SAM 3 while improving segmentation quality. Experiments on SUN-SEG and CVC-612 under standardized 0-point and first-frame 1-point (t=0) protocols show consistent improvements over task-specific and SAM-based baselines. Future work includes improving robustness under severe occlusions and illumination changes, and exploring more efficient token update and memory schemes to reduce runtime overhead while maintaining temporal stability.

Acknowledgements

This work was supported partly by National Natural Science Foundation of China (No. 62273241), Natural Science Foundation of Guangdong Province, China (No. 2024A1515011946), the Shenzhen Research Foundation for Basic Research, China (No. JCYJ20250604181940054), and the grant of Collaborative Research Fund under Hong Kong Research Grants Council (project no C5055-24G).

References

1. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized medical imaging and graphics **43**, 99–111 (2015)
2. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
3. Chen, C., Miao, J., Wu, D., Zhong, A., Yan, Z., Kim, S., Hu, J., Liu, Z., Sun, L., Li, X., et al.: Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. Medical Image Analysis **98**, 103310 (2024)
4. Chen, G., Yang, J., Pu, X., Ji, G.P., Xiong, H., Pan, Y., Cui, H., Xia, Y.: Mast: Video polyp segmentation with a mixture-attention siamese transformer. arXiv preprint arXiv:2401.12439 (2024)
5. Chen, G., Wu, H., Qin, J.: Stddnet: Harnessing mamba for video polyp segmentation via spatial-aligned temporal modeling and discriminative dynamic representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21364–21373 (2025)
6. Chen, L., Gu, L., Fu, Y.: Frequency-dynamic attention modulation for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22620–22632 (2025)
7. Chen, T., Cao, R., Yu, X., Zhu, L., Ding, C., Ji, D., Chen, C., Zhu, Q., Xu, C., Mao, P., et al.: Sam3-adapter: Efficient adaptation of segment anything 3 for camouflage object segmentation, shadow detection, and medical image segmentation. arXiv preprint arXiv:2511.19425 (2025)
8. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sam-med2d. arXiv preprint arXiv:2308.16184 (2023)
9. Cheng, M., Kong, Z., Song, G., Tian, Y., Liang, Y., Chen, J.: Learnable oriented-derivative network for polyp segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 720–730. Springer (2021)
10. Dong, B., Wang, W., Fan, D.P., Li, J., Fu, H., Shao, L.: Polyp-pvt: Polyp segmentation with pyramid vision transformers. arXiv preprint arXiv:2108.06932 (2021)
11. Dong, W., Yu, J., Huang, Y., Wang, H., Zhu, L., Chung, A., Ren, H., Bai, L.: More than segmentation: Benchmarking sam 3 for segmentation, 3d perception, and reconstruction in robotic surgery. arXiv preprint arXiv:2512.07596 (2025)
12. Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J., Shao, L.: Camouflaged object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2777–2787 (2020)

13. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 263–273. Springer (2020)
14. Fang, Y., Chen, C., Yuan, Y., Tong, K.y.: Selective feature aggregation network with area-boundary constraints for polyp segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 302–310. Springer (2019)
15. Fang, Z., Liu, Y., Wu, H., Qin, J.: Vp-sam: Taming segment anything model for video polyp segmentation via disentanglement and spatio-temporal side network. In: European Conference on Computer Vision. pp. 367–383. Springer (2024)
16. Gong, S., Zhong, Y., Ma, W., Li, J., Wang, Z., Zhang, J., Heng, P.A., Dou, Q.: 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable tumor segmentation. Medical Image Analysis **98**, 103324 (2024)
17. Hu, Q., Yi, Z., Zhou, Y., Peng, F., Liu, M., Li, Q., Wang, Z.: Sali: Short-term alignment and long-term interaction network for colonoscopy video polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 531–541. Springer (2024)
18. Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., et al.: Segment anything model for medical images? Medical Image Analysis **92**, 103061 (2024)
19. Ji, G.P., Chou, Y.C., Fan, D.P., Chen, G., Fu, H., Jha, D., Shao, L.: Progressively normalized self-attention network for video polyp segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 142–152. Springer (2021)
20. Ji, G.P., Fan, D.P., Xu, P., Cheng, M.M., Zhou, B., Van Gool, L.: Sam struggles in concealed scenes–empirical study on" segment anything". arXiv preprint arXiv:2304.06022 (2023)
21. Ji, G.P., Xiao, G., Chou, Y.C., Fan, D.P., Zhao, K., Chen, G., Van Gool, L.: Video polyp segmentation: A deep learning perspective. Machine Intelligence Research **19**(6), 531–549 (2022)
22. Jiang, C., Ding, T., Song, C., Tu, J., Yan, Z., Shao, Y., Wang, Z., Shang, Y., Han, T., Tian, Y.: Medical sam3: A foundation model for universal prompt-driven medical image segmentation. arXiv preprint arXiv:2601.10880 (2026)
23. Kim, S., Jin, P., Chen, C., Kim, K., Lyu, Z., Ren, H., Kim, S., Liu, Z., Zhong, A., Liu, T., Li, X., Li, Q.: MediViSTA: Medical video segmentation via temporal fusion SAM adaptation for echocardiography. arXiv preprint arXiv:2309.13539 (2024)
24. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
25. Le, T.N., Cao, Y., Nguyen, T.C., Le, M.Q., Nguyen, K.D., Do, T.T., Tran, M.T., Nguyen, T.V.: Camouflaged instance segmentation in-the-wild: Dataset, method, and benchmark suite. IEEE Transactions on Image Processing **31**, 287–300 (2021)
26. Li, S., Sui, X., Luo, X., Xu, X., Liu, Y., Goh, R.: Medical image segmentation using squeeze-and-expansion transformers. arXiv preprint arXiv:2105.09511 (2021)
27. Li, Y., Hu, M., Yang, X.: Polyp-sam: Transfer sam for polyp segmentation. In: Medical imaging 2024: computer-aided diagnosis. vol. 12927, pp. 749–754. SPIE (2024)
28. Liu, W., Shen, X., Pun, C.M., Cun, X.: Explicit visual prompting for low-level structure segmentations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19434–19445 (2023)

29. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature communications **15**(1), 654 (2024)
30. Ma, J., Yang, Z., Kim, S., Chen, B., Baharoon, M., Fallahpour, A., Asakereh, R., Lyu, H., Wang, B.: Medsam2: Segment anything in 3d medical images and videos. arXiv preprint arXiv:2504.03600 (2025)
31. Nguyen, T.C., Nguyen, T.P., Diep, G.H., Tran-Dinh, A.H., Nguyen, T.V., Tran, M.T.: Ccbanet: Cascading context and balancing attention for polyp segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 633–643. Springer (2021)
32. Pang, Y., Zhao, X., Xiang, T.Z., Zhang, L., Lu, H.: Zoomnext: A unified collaborative pyramid network for camouflaged object detection. IEEE transactions on pattern analysis and machine intelligence **46**(12), 9205–9220 (2024)
33. Puyal, J.G.B., Bhatia, K.K., Brandao, P., Ahmad, O.F., Toth, D., Kader, R., Lovat, L., Mountney, P., Stoyanov, D.: Endoscopic polyp segmentation using a hybrid 2d/3d cnn. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 295–305. Springer (2020)
34. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
35. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **71**(3), 209–249 (2021)
36. Van Rijn, J.C., Reitsma, J.B., Stoker, J., Bossuyt, P.M., Van Deventer, S.J., Dekker, E.: Polyp miss rate determined by tandem colonoscopy: a systematic review. Official journal of the American College of Gastroenterology| ACG **101**(2), 343–350 (2006)
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
38. Wei, Z., Dong, W., Zhou, P., Gu, Y., Zhao, Z., Xu, Y.: Prompting segment anything model with domain-adaptive prototype for generalizable medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 533–543. Springer (2024)
39. Wu, H., Zhao, Z., Wang, Z.: Meta-unet: Multi-scale efficient transformer attention unet for fast and high-accuracy polyp segmentation. IEEE Transactions on Automation Science and Engineering **21**(3), 4117–4128 (2023)
40. Wu, H., Zhao, Z., Zhong, J., Wang, W., Wen, Z., Qin, J.: Polypseg+: A lightweight context-aware network for real-time polyp segmentation. IEEE Transactions on Cybernetics **53**(4), 2610–2621 (2022)
41. Zhang, Y., Zhou, T., Wang, S., Liang, P., Zhang, Y., Chen, D.Z.: Input augmentation with sam: Boosting medical image segmentation with segmentation foundation model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 129–139. Springer (2023)
42. Zhou, T., Zhang, Y., Zhou, Y., Wu, Y., Gong, C.: Can sam segment polyps? arXiv preprint arXiv:2304.07583 (2023)
43. Zhou, Y., Li, Y., Fu, Y., Benini, L., Konukoglu, E., Sun, G.: Camsam2: Segment anything accurately in camouflaged videos. arXiv preprint arXiv:2503.19730 (2025)