ProMISe: Promptable Medical Image Segmentation using SAM

Jinfeng Wang*
University of Liverpool
Xi'an Jiaotong-liverpool University
Jf.Jacob.Wong@gmail.com

Sifan Song*
Massachusetts General Hospital
and Harvard Medical School
ssong25@mgh.harvard.edu

Xinkun Wang*
Xi'an Jiaotong-liverpool University
xinkun.wang21@student.xjtlu.edu.cn

Yiyi Wang*
Xi'an Jiaotong-liverpool University
yiyi.wang20@student.xjtlu.edu.cn

Yiyi Miao
Xi'an Jiaotong-liverpool University
Yiyi.Miao21@student.xjtlu.edu.cn

Jionglong Su†
Xi'an Jiaotong-liverpool University
Jionglong.Su@xjtlu.edu.cn

S. Kevin Zhou†
University of Science
and Technology of China
skevinzhou@ustc.edu.cn

Abstract

With the proposal of Segment Anything Model (SAM), fine-tuning SAM for medical image segmentation (MIS) has become popular. However, due to the large size of the SAM model and the significant domain gap between natural and medical images, fine-tuning-based strategies are costly with potential risk of instability, feature damage and catastrophic forgetting. Furthermore, some methods of transferring SAM to a domain-specific MIS through fine-tuning strategies disable the model's prompting capability, severely limiting its utilization scenarios. In this paper, we propose an Auto-Prompting Module (APM), which provides SAM-based foundation model with Euclidean adaptive prompts in the target domain. Our experiments demonstrate that such adaptive prompts significantly improve SAM's non-fine-tuned performance in MIS. In addition, we propose a novel non-invasive method called Incremental Pattern Shifting (IPS) to adapt SAM to specific medical domains. Experimental results show that the IPS enables SAM to achieve state-of-the-art or competitive performance in MIS without the need for fine-tuning. By coupling these two methods, we propose **ProMISe**, an end-to-end non-fine-tuned framework for **Pro**mptable **M**edical **I**mage **Se**gmentation. Our experiments demonstrate that both using our methods individually or in combination achieves satisfactory performance in low-cost pattern shifting, with all of SAM's parameters frozen. *Code is available at https://github.com/xinkunwang111/ProMISe*

*These authors contributed equally to this work.
†Corresponding authors

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

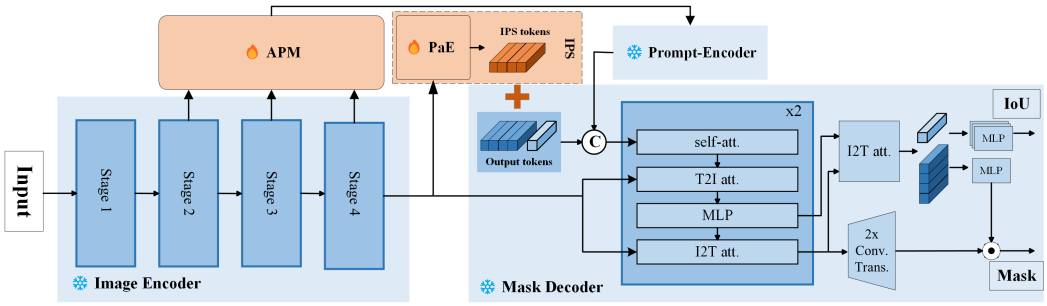

Figure 1: Overview of ProMISe. All three components of the original SAM are frozen. The **A**uto-**P**rompting **M**odule (APM) leverages features from the image encoder to predict optimal prompts in Euclidean space. The **Pa**ttern **E**mbedding (PaE) module analyzes the image embedding to extract pattern gaps between the target and source domains. The **I**ncremental **P**attern **S**hifting (IPS) tokens are added to the mask tokens of the output tokens to realize the decoder's shifting of mask patterns.

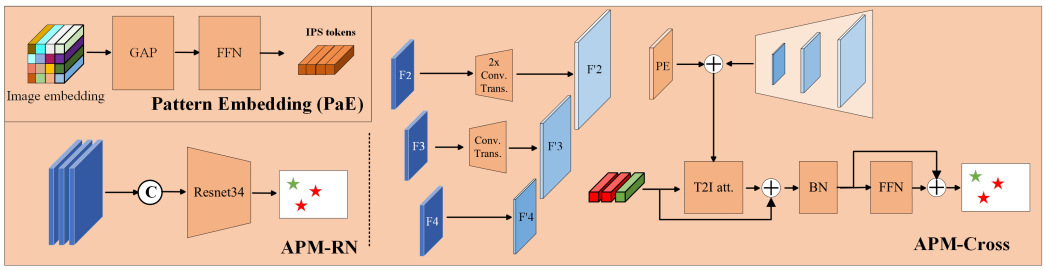

Figure 2: Detailed structure of APM and PaE (the global average pooling (GAP) and the FFN used for shape alignment and feature extraction). The APM can be implemented using various operators and modules, such as CNN and Transformer.

1 Introduction

Recently, many large-scale foundation models [12, 24] deliver exciting performance with promising results in various domains. Among them, the latest one SAM [12], a foundation model for natural image segmentation, has attracted significant attention from researchers in the computer vision community. SAM enables interactive segmentation of targets through prompts such as points, bounding boxes, masks, and text. Medical Image Segmentation (MIS) tasks are crucial in medical image analysis. Unlike natural images, medical images contain diverse and distinctive data modalities and application scenarios. The transfer and deployment of SAM for MIS present a meaningful and promising research area.

Some studies [6, 27, 15, 9] have explored the capability of SAM in MIS in a zero-shot fashion. Unfortunately, their results demonstrate that neither points or bounding boxes guided by ground truth (GT), nor the automated methods of SAM, can achieve satisfactory and practical performance in MIS. Also, different studies have used various prompts (such as latent representations [18, 26], bounding boxes [13, 7, 25], text [25, 16], mask [7], points [7, 25]) by different strategies (*e.g.*, all-parameter fine-tuning [13], adapter [25, 4], bias-tuning [16], and LoRA [26]) to different degrees (decoder-only [14], prompt-encoder-only [18], and all components [13]), with the aim of fine-tuning SAM for transfer to the target domain. However, since SAM is a large foundation model trained on extensive datasets, these fine-tuning-based approaches incur enormous training costs and face potential risks of instability, feature damage and catastrophic forgetting. Such fine-tuning-based transfer methods typically utilize the model as a heavy pre-trained model rather than a foundation model.

In this paper, we rethink the use of each component of SAM in the following way. First, the prompt encoder, as a mapping between Euclidean and latent space, has been adequately trained during the training process of SAM. Some studies attempt to use latent representations instead of explicit Euclidean prompts to automate the SAM model, but this approach sacrifices its interactive capability

and interpretability. Second, as a foundation model trained on extensive datasets with a complex training strategy, the encoder of SAM already possesses robust feature extraction capabilities. We argue that utilizing these capabilities efficiently and properly in unseen domains is more valuable than fine-tuning them. Third, by analyzing the mask decoder, we observe that the output tokens provide prior pattern knowledge for mask prediction based on prompts. Therefore, we claim that pattern shifting of output tokens can achieve more efficient and stable domain adaptation for SAM compared to fine-tuning.

As demonstrated in Fig. 1 and Fig. 2, based on the above arguments: 1) We propose the Auto-Prompting Module (APM), which leverages the SAM framework for training and provides adaptive Euclidean prompts for SAM; 2) We propose Incremental Pattern Shifting (IPS), a novel non-invasive pattern shifting method which couples a Pattern Embedding (PaE) module with IPS tokens to cost-effectively shift the prior pattern knowledge of the mask decoder. The IPS method enables SAM to achieve state-of-the-art (SOTA) and competitive performance without the need for fine-tuning; 3) We couple the above two methods to propose the **Pro**mptable **M**edical **I**mage **Se**gmentation (**ProMISe**) framework utilizing SAM. Notably, this method is able to transfer SAM to MIS while keeping **all of SAM's parameters frozen**, resulting in significantly reduced training costs, improved stability, and the ability to retain spatial prompting capability.

2 Adaptive Prompt

2.1 Motivation

Several zero-shot studies [6, 27, 15, 9] utilize the interactive capability of SAM to investigate its potential as a large-scale foundation model for transferring to MIS tasks in an untrained manner. However, most of these approaches employ GT-based prompts, which are generated from GT masks using various different methods, such as GT-based foreground-background point sampling and noisy GT-bounding box. However, these GT-based prompt strategies fail to enable SAM to achieve practical performance in MIS tasks. We argue that this limitation is due to the coupling of prompts, image embeddings and mask patterns. In other words, without effectively utilizing image embeddings and transferring mask patterns, it remains challenging for SAM to consistently deliver satisfactory performance in unseen domains by only relying on GT-based spatial prompts.

2.2 Proposed Method

As shown in Fig. 1 and 2, we design a lightweight end-to-end module, called Auto-Prompting Module (APM), which effectively integrates the multi-level features of SAM's image encoder to predict optimal prompts in Euclidean space. By generating adaptive prompts that provide more fine-grained spatial information, the APM significantly enhances the untrained performance of SAM in the target domain. This indicates the highly expressive feature extraction capability of SAM, even in previously unseen domains.

During our research, we notice a significant performance limitation when using bounding boxes as prompts. We attribute this to the fact that bounding boxes only provide coarse-grained location information and lack fine-grained details. Thus, we mainly utilize point-based prompts in this paper. It is worth noting that our method is applicable to any form of prompts except text, and the APM can be implemented using various operators and modules (*e.g.*, ResNet34 and a simple cross-attention block).

3 Incremental Pattern Shifting

3.1 Rethink the mask decoder

As mentioned above, many studies based on fine-tuning strategies have been proposed to improve the performance of SAM in MIS tasks. However, inappropriate fine-tuning is likely to weaken the original parameter distribution of SAM. As such, it becomes difficult to sufficiently adjust to the optimal target distribution, especially in medical datasets with distinctive domains, making fine-tuning SAM unstable and inefficient. From a broader perspective, a simple fine-tuning strategy for pattern shifting

is regarded as the model as a heavily pre-trained model rather than a foundation model, resulting in degraded model capabilities.

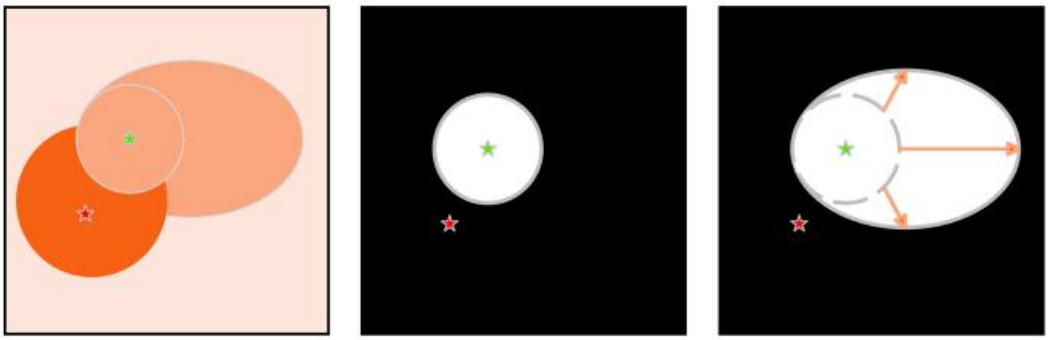

Figure 3: Theoretical illustration of IPS. Left: Image with point prompts; Middle: Output mask from vanilla SAM; Right: Output mask from SAM with IPS. Arrows represent patterns shifting.

After rethinking the components of the SAM, we observe that the SAM has a very efficient mask decoder that refers to DETR [2] and MaskFormer [5]. In this study, we find that in the mask decoder, output tokens initially receive the location information of the point of interest (PoI) or region of interest (RoI) from the prompt encoder through self-attention. Subsequently, as shown in Fig. 1, output tokens extract semantic information from the image embedding through cross-attention, forming semantic patterns.

3.2 Proposed Method

Based on these findings, we argue that with a large amount of training, the output tokens can be regarded as the *pattern information* acquired from the source domain, enabling SAM to generate predicted masks based on PoI or RoI. Therefore, assuming the image encoder of SAM is sufficiently powerful and well-trained, transferring SAM to MIS can be equivalent to transferring the mask prediction patterns of SAM to MIS.

$$[IoU, Mask] = \text{MaskDecoder}(ImageEmbedding, T_{Pattern}, T_{Prompts}) \tag{1}$$

$$T_{Pattern} = \text{Concat}(T_{IoU}, (T_{Mask} + T_{IPS})) \tag{2}$$

$$T_{IPS} = \text{FFN}(\text{GAP}(ImageEmbedding)) \tag{3}$$

With the above arguments, we propose a method called Incremental Pattern Shifting (IPS). As shown in Fig. 1, we extract the *pattern shifting information* (IPS tokens) from the image embedding using a lightweight PaE module, and then non-invasively shift the patterns of the mask decoder by adding the IPS tokens to the mask tokens (Eqs. 1-3). In Fig. 3, the mask decoder, when provided with the same PoI, generates mask predictions that are more compatible with the target domain after pattern shifting. As an experimental validation of Fig. 3, the right two columns of Fig. 4 (SAM and IPS) show that the transfer of the segmentation pattern of SAM using only IPS is solid. Notably, our experiments show the effectiveness of IPS, utilizing GT-based prompts for both training and inference, in enabling SAM to achieve SOTA or competitive performance in MIS tasks.

4 ProMISe framework

As shown in Fig. 1, IPS does not conflict with the proposed APM which provides prompts for end-to-end SAM transfer and inference. Deploying non-invasive APM, PaE, and IPS tokens to

Table 1: Comparison of the performance enhancement brought to SAM by adaptive prompts provided by APMs (RN and Cross). [a] Trainable Parameters.

Benchmarks		Kvasir			EndoScene			ColonDB		
Methods	TP[a]	mDice	mIoU	MAE↓	mDice	mIoU	MAE↓	mDice	mIoU	MAE↓
U-Net [17]	-	0.818	0.746	0.055	0.710	0.627	0.022	0.512	0.444	0.061
ResUNet++ [10]	-	0.821	0.743	0.048	0.707	0.624	0.018	0.483	0.410	0.064
SAM-5P	-	0.750	0.645	0.104	0.656	0.582	0.139	0.569	0.482	0.215
SAM-16P	-	0.719	0.620	0.140	0.692	0.613	0.118	0.548	0.467	0.228
APM-RN-5P	21.7M	0.741	0.645	0.060	0.781	0.689	0.016	0.594	0.501	0.056
APM-RN-16P	21.7M	**0.797**	**0.706**	**0.046**	0.789	0.694	**0.013**	0.595	0.502	**0.051**
APM-Cross-5P	44.3M	0.749	0.648	0.074	0.711	0.619	0.059	0.511	0.433	0.165
APM-Cross-16P	44.3M	0.789	0.697	0.051	**0.794**	**0.699**	0.018	**0.616**	**0.517**	0.059

SAM allows automation and transfer of original SAM to MIS tasks. Our proposed IPS method effectively facilitates segmentation pattern shifting. However, using GT-based prompts for pattern transfer does not allow end-to-end training process. In addition, GT-based prompt sampling may introduce randomness, thereby posing challenges to the stability of the training and evaluation process. Therefore, we couple the IPS method with APM to achieve end-to-end pattern shifting for SAM to target medical domains. Moreover, favored by the preservation of Euclidean spatial form prompts, this ProMISe framework can handle both automatic and manual prompts during inference while retaining its interpretability.

5 Experiments

5.1 Experimental Setup

We individually conduct experiments in two modalities: endoscopy and dermoscopy. For endoscopy experiments, we follow the experimental setups of PraNet [8], DuAT [21], and SSFormer [23]. We extract 1450 images from Kavsir [11] and CVC-ClinicDB [19] as the training set. Tests are then conducted on Kvasir, EndoScene [20], CVC-ColonDB [1], and ETIS [22], using Mean Absolute Error (MAE), mean Dice, and mean IoU as metrics. For dermatoscopy, we use ISIC2018 for training and testing, with mean Dice and mean IoU as metrics. We utilize ViT-B as the image encoder for all SAM-related methods in the experiments. In addition, we apply three prompt settings in training and inference, namely 3/5/16 points (1/2/8 positive + 2/3/8 negative, 3/5/16P, respectively). All GT-based prompt points used in this paper were obtained by random sampling from GT masks. For fair comparisons, the prompt points provided to the models are exactly the same in each experimental setting.

We implement our methods in pytorch, using one TESLA A100 (80G) GPU. Adam optimizer and a learning rate of 0.00001 are employed. The training period is 200 epochs. Our loss is a combination of Dice and BCE loss. The IPS method requires approximately 22 GPU hours and 25 GPU hours for training in polyps and ISIC, respectively. In addition, APM-cross requires approximately 28 GPU hours and 30 GPU hours for training in polyps and ISIC. The training periods for APM-RN are 21 GPU hours and 25 GPU hours.

5.2 Adaptive Prompt

As described in Sect. 2, to study the actual performance of vanilla SAM in MIS, we train and test two types of APMs, *i.e.*, ResNet34 (RN) and a one-layer cross-attention Transformer module (Cross), on polyp benchmarks. As shown in Tab. 1, the adaptive prompts generated by both APMs effectively improve the performance of SAM in the polyp segmentation task, achieving a level comparable to the baseline MIS segmentation model. Furthermore, we observe a significant reduction in MAE by increasing the number of point prompts with more provided fine-grained information, which suggests that our method effectively utilizes the boundary sensitivity of vanilla SAM.

5.3 Pattern Shifting

In Tab. 2 and 3, we train the proposed IPS method (PaE with 4 IPS tokens) using GT-based prompt points. We perform tests using the same GT-based point prompts for SAM-related models

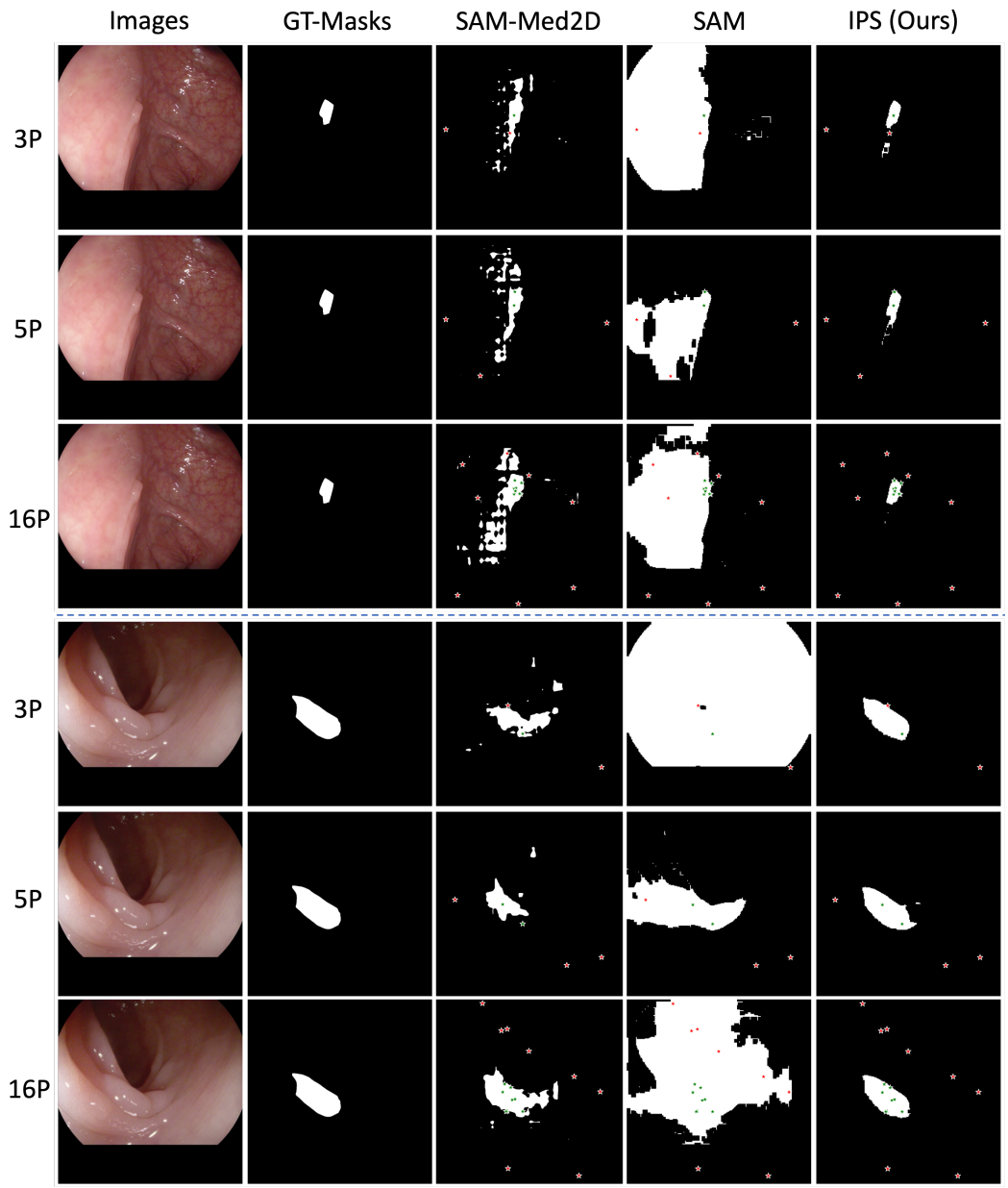

Figure 4: Comparison of the performance with different prompt number in endoscopy datasets. The 3P, 5P and 16P contain 1, 2, and 8 positive, as well as 2, 3, and 8 negative points, respectively.

and compare our method (SAM with IPS) to SOTA methods, SAM and SAM-Med2D. Following the mainstream validation frameworks in the field of polyp segmentation [8, 23]. the Kvasir and EndoScene benchmarks are used to validate the transfer performance of the proposed method, while the ColonDB and ETIS benchmarks are used to validate the generalization ability of the proposed method due to the domain gap with the training set.

As shown in Tab. 2, our proposed IPS significantly improves the performance of SAM on familiar benchmarks, achieving mDice improvements of 10%-21% in Kvasir, 19%-29% in EndoScene, and 18%-37% in ISIC2018. Compared to SAM-Med2D, which is trained on a large-scale medical dataset and utilizes a complex training strategy, our method also achieves a remarkable improvement.

Experimental results in Tab. 3 indicate that the IPS method adequately leverages SAM's powerful generalization and feature expression capabilities. Even in the unfamiliar and challenging benchmarks

Table 2: Quantitative comparison of our IPS, SOTA methods, and other SAM-based methods. Best-in-class results are bolded, while second-best results are underlined. The 3P|5P and 3P|16P represent training our IPS on 3 GT-based prompt points and testing them with 5 and 16 ones. [a] Number of Trainable Parameters. [b] Number of Endoscopy/Dermoscopy images is utilized in the corresponding training set. * Some test set images from the corresponding dataset may be seen in the training process.

Benchmarks			Kvasir		EndoScene		ISIC2018	
Methods	TP[a]	Images[b]	mDice	mIoU	mDice	mIoU	mDice	mIoU
SOTA Methods								
U-net [17]	-	1450/2594	0.818	0.746	0.710	0.627	0.855	0.785
PraNet [8]	-	1450/2594	0.898	0.840	0.835	0.797	0.875	0.787
TransUNet [3]	-	1450/2594	0.913	0.857	**0.893**	0.660	0.880	0.809
SSFormer [23]	-	1450/2594	**0.926**	**0.874**	0.887	**0.821**	**0.919**	**0.861**
SAM-based Methods								
SAM-3P [12]	-	-	0.589	0.471	0.513	0.414	0.489	0.367
SAM-5P	-	-	0.750	0.645	0.656	0.582	0.687	0.569
SAM-16P	-	-	0.719	0.620	0.692	0.613	0.738	0.624
Med2D-3P [7]	184.5M	5838/7935	*0.821	*0.735	0.697	0.597	*0.872	*0.803
Med2D-5P	184.5M	5838/7935	*0.822	*0.735	0.722	0.623	*0.884	*0.813
Med2D-16P	184.5M	5838/7935	*0.832	*0.748	0.727	0.620	*0.893	*0.823
IPS-3P	1.3M	1450/2594	0.797	0.704	0.806	0.718	0.854	0.762
IPS-3P|5P	1.3M	1450/2594	0.821	0.732	0.804	0.716	0.865	0.774
IPS-3P|16P	1.3M	1450/2594	0.843	0.752	0.815	0.721	0.874	0.785
IPS-5P	1.3M	1450/2594	0.855	0.772	0.854	0.764	0.889	0.808
IPS-16P	1.3M	1450/2594	**0.902**	**0.835**	**0.888**	**0.810**	**0.915**	**0.847**

Table 3: Generalization performance comparison of our IPS, SOTA methods, and other SAM-based methods. Best-in-class results are bolded, while second-best results are underlined.

Benchmarks			ColonDB		ETIS	
Methods	TP[a]	Images[b]	mDice	mIoU	mDice	mIoU
SOTA Methods						
U-net [17]	-	1450/2594	0.512	0.444	0.398	0.335
PraNet [8]	-	1450/2594	0.712	0.640	0.628	0.567
TransUNet [3]	-	1450/2594	**0.781**	**0.699**	0.731	0.624
SSFormer [23]	-	1450/2594	0.772	0.697	**0.767**	**0.698**
SAM-based Methods						
SAM-3P [12]	-	-	0.447	0.356	0.464	0.381
SAM-5P	-	-	0.569	0.482	0.541	0.472
SAM-16P	-	-	0.548	0.467	0.524	0.455
Med2D-3P [7]	184.5M	5838/7935	0.689	0.588	0.633	0.524
Med2D-5P	184.5M	5838/7935	0.686	0.576	0.677	0.571
Med2D-16P	184.5M	5838/7935	0.685	0.575	0.622	0.514
IPS-3P	1.3M	1450/2594	0.724	0.618	0.659	0.569
IPS-3P|5P	1.3M	1450/2594	0.761	0.655	0.727	0.626
IPS-3P|16P	1.3M	1450/2594	0.783	0.678	0.763	0.674
IPS-5P	1.3M	1450/2594	0.819	0.716	0.801	0.704
IPS-16P	1.3M	1450/2594	**0.874**	**0.789**	**0.854**	**0.770**

of ColonDB and ETIS, IPS enables SAM to achieve SOTA performance. Using SSFormer as a baseline, we obtain up to 10% and 9% improvement in ColonDB and ETIS, respectively. Moreover, compared to vanilla SAM, IPS can achieving mDice improvements of 25%-33% in ColonDB, 20%-33% in ETIS. It is promising that the above performance improvement depends only on a tiny number of trainable parameters (1.3M). Those demonstrate that the IPS can efficiently transfer SAM to the target task and achieve SOTA-level performance with powerful generalization ability. This makes the low-cost rapid deployment of SAM promising.

As demonstrated in Fig. 4 and Tab. 2 and 3, SAM and SAM-Med2D struggle to handle exposures and insignificant small-volume targets well, even when provided with more fine-grained guidance through an increased number of points (*i.e.* more prompt points). In contrast, our proposed IPS significantly optimizes its predicted masks through non-invasive pattern shifting based on the same prompts (Fig. 4). Furthermore, the results in Tab. 2 and 3 demonstrate that even if the number of

Table 4: Performance of SAM with ProMISe (Cross and RN) is tested using GT-base prompts. [a] Number of Trainable Parameters. [b] Number of Endoscopy/Dermoscopy images is utilized in the corresponding training set. * Some test set images from the corresponding dataset may be seen in the training process.

Benchmarks			Kvasir		EndoScene		ISIC2018	
Methods	TP[a]	Images[b]	mDice	mIoU	mDice	mIoU	mDice	mIoU
U-Net	-	1450/2594	0.818	0.746	0.710	0.627	0.855	0.785
ResUNet++	-	1450/2594	0.821	0.743	0.707	0.624	0.809	0.729
SAM-5P	-	-	0.750	0.645	0.656	0.582	0.687	0.569
SAM-16P	-	-	0.719	0.620	0.692	0.613	0.738	0.624
Med2D-5P	184.5M	5838/7935	*0.822	*0.735	0.722	0.623	*0.884	*0.813
Med2D-16P	184.5M	5838/7935	*0.832	*0.748	0.727	0.620	*0.893	*0.823
Cross-5P	45.6M	1450/2594	0.834	0.744	0.788	0.687	0.742	0.631
Cross-16P	45.6M	1450/2594	0.858	0.777	0.768	0.661	0.819	0.705
RN-5P	23.0M	1450/2594	0.776	0.673	0.705	0.587	0.798	0.688
RN-16P	23.0M	1450/2594	0.846	0.759	0.803	0.735	0.878	0.791

Table 5: Generalization performance of SAM with ProMISe (Cross and RN) is tested using GT-base prompts.

Benchmarks			ColonDB		ETIS	
Methods	TP[a]	Images[b]	mDice	mIoU	mDice	mIoU
U-Net	-	1450/2594	0.512	0.444	0.398	0.335
ResUNet++	-	1450/2594	0.483	0.410	0.401	0.344
SAM-5P	-	-	0.569	0.482	0.541	0.472
SAM-16P	-	-	0.548	0.467	0.524	0.455
Med2D-5P	184.5M	5838/7935	0.686	0.576	0.677	0.571
Med2D-16P	184.5M	5838/7935	0.685	0.575	0.622	0.514
Cross-5P	45.6M	1450/2594	0.744	0.636	0.678	0.578
Cross-16P	45.6M	1450/2594	0.732	0.626	0.691	0.591
RN-5P	23.0M	1450/2594	0.664	0.547	0.605	0.508
RN-16P	23.0M	1450/2594	0.735	0.624	0.673	0.566

prompt points used in training does not match the number of points used in testing, the IPS can still handle these information gaps well.

5.4 ProMISe framework

To avoid random sampling of prompts and thus achieve end-to-end pattern shifting, we propose the ProMISe framework, which couples the APM and IPS for training with adaptive prompts. When tested using GT-based prompts, the end-to-end pattern shifting of ProMISe with both APMs (Cross and RN) significantly improves SAM's performance to a practical and competitive level in MIS (Tab. 4 and 5). Importantly, ProMISe maintains its interpretability using both adaptive/GT-based Euclidean prompts and keeps all of SAM's parameters frozen, resulting in a more practical and applicable approach for real clinical scenarios.

5.5 Multi-Modality Experiments

To evaluate the multi-modality training potential of our proposed method, we input both endoscopy and dermoscopy together as the training set. Additionally, to verify the stability of our method in multi-modality training, we applied five different random seeds to provide GT-based prompt points during the evaluation. As shown in Fig. 5, the performance of multi-modality training differs only slightly from the results of the corresponding single-modality training (green dashed lines) and is significantly higher than MedSAM-2D (red lines). These experimental results indicate that our IPS method has the potential not only for rapid and low-cost pattern shifting to a single medical domain but also for multiple specific modalities.

Furthermore, the error bars (standard deviation) indicate that our method has excellent robustness on both Dice and IoU metrics, as the GT-based prompt points generated from five different random seeds produce relatively consistent results. This is particularly meaningful in clinical scenarios, as

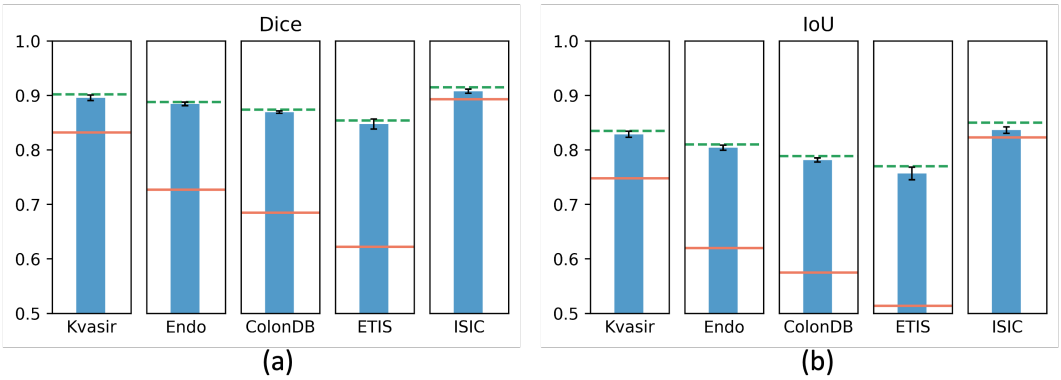

Figure 5: Multi-modality training performance of pattern shifting with 16 points (mDice scores). The corresponding results of MedSAM-2D and our proposed IPS are represented by red lines and green dashed lines, respectively, as shown in Table 2 and 3. The blue bars represent the performance of IPS training with both endoscopy and dermoscopy images.

Table 6: Ablation study for pattern shifting with 16 points (mDice scores). Mask tokens refer to unfreezing the mask tokens in SAM's mask decoder. The optimal setting is in bold. [a] Number of Trainable Parameters.

Mask tokens	IPS tokens	PaE	Frozen SAM	TP[a]	Kvasir	EndoScene	ColonDB	ETIS
✓			✗	1.02K	0.581	0.796	0.756	0.726
	✓		✓	1.02K	0.862	0.817	0.802	0.778
	✓	✓	✓	1.29M	**0.902**	**0.888**	**0.874**	**0.854**

different clinicians may provide preferred prompt points. Using the method proposed in this paper, they can receive similar results.

5.6 Ablation Study

Apart from the differences in effects brought by the individual and combined application of IPS and APM discussed in Sect. 5.2 - 5.4, we also find that modifications to mask tokens can significantly enhance the pattern-shifting ability of the SAM-based model. As demonstrated in Tab. 6, adding IPS tokens to mask tokens significantly improve the performance, and the PaE module is indispensable for the IPS method to achieve SOTA results. Using IPS tokens alone represents initialization of the tokens rather than being generated from the PaE. Thus, removing the PaE provides an extremely lightweight option.

6 Conclusions

In this paper, we propose a novel adaptive prompt generation module, Auto-Prompting Module (APM), which improves the transfer-free performance of SAM in the target domain by generating optimal Euclidean prompts. In addition, we propose Incremental Pattern Shifting (IPS), which enables non-fine-tuned pattern shifting to improve SAM's performance in unfamiliar domains, achieving SOTA and competitive results. Furthermore, we couple IPS with APM to propose the ProMISe framework, which can realizes end-to-end pattern shifting to improve training efficiency and stability. We conducted experiments in endoscopic and dermoscopic benchmark datasets to demonstrate the usefulness and promise of our proposed methods. Benefiting from IPS, increasing prompt point number results in significant performance gains for SAM, which may be further extended to include scrawl, sketch or coarse prompts. More importantly, our results prove that a fine-tuning-based approach is not necessarily optimal for utilizing a foundation model like SAM.

7 Limitations and future works

Although IPS has achieved promising performance in medical image segmentation tasks in endoscopic and dermoscopic modalities, the medical modalities involved in this paper still need to be increased to validate the effect of IPS in the entire medical image segmentation domain. Moreover, while our proposed APM approach can effectively improve the performance of transfer-free SAM in medical image segmentation, it performs sub-optimally when coupled with IPS. This suggests that we must trade between the end-to-end framework and performance.

Based on the above limitations, we plan to extend the ProMISe approach to as many medical image modalities (including 2D and 3D) as possible. In order to improve the reliability of ProMISe, we plan to optimize the APM method while attempting a lightweight, non-invasive shifting of the representation encoder. Moreover, to improve the extensibility and practicality of ProMISe and IPS, we plan to enrich further the prompt types, such as bounding boxes, scribble, sketch, and text.

References

[1] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño. 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] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[3] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[4] T. Chen, L. Zhu, C. Ding, R. Cao, S. Zhang, Y. Wang, Z. Li, L. Sun, P. Mao, and Y. Zang. Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more. *arXiv preprint arXiv:2304.09148*, 1(2):5, 2023.

[5] B. Cheng, A. Schwing, and A. Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021.

[6] D. Cheng, Z. Qin, Z. Jiang, S. Zhang, Q. Lao, and K. Li. Sam on medical images: A comprehensive study on three prompt modes. *arXiv preprint arXiv:2305.00035*, 2023.

[7] J. Cheng, J. Ye, Z. Deng, J. Chen, T. Li, H. Wang, Y. Su, Z. Huang, J. Chen, L. Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023.

[8] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 263–273. Springer, 2020.

[9] S. He, R. Bao, J. Li, P. E. Grant, and Y. Ou. Accuracy of segment-anything model (sam) in medical image segmentation tasks. *arXiv preprint arXiv:2304.09324*, 2023.

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

[11] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. De Lange, D. Johansen, and H. D. Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, pages 451–462. Springer, 2020.

[12] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[13] Y. Li, M. Hu, and X. Yang. Polyp-sam: Transfer sam for polyp segmentation. In *Medical Imaging 2024: Computer-Aided Diagnosis*, volume 12927, pages 759–765. SPIE, 2024.

[14] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[15] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023.

[16] J. N. Paranjape, N. G. Nair, S. Sikder, S. S. Vedula, and V. M. Patel. Adaptivesam: Towards efficient tuning of sam for surgical scene segmentation. *arXiv preprint arXiv:2308.03726*, 2023.

[17] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[18] T. Shaharabany, A. Dahan, R. Giryes, and L. Wolf. Autosam: Adapting sam to medical images by overloading the prompt encoder. *arXiv preprint arXiv:2306.06370*, 2023.

[19] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9:283–293, 2014.

[20] N. Tajbakhsh, S. R. Gurudu, and J. Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015.

[21] F. Tang, Z. Xu, Q. Huang, J. Wang, X. Hou, J. Su, and J. Liu. Duat: Dual-aggregation transformer network for medical image segmentation. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 343–356. Springer, 2023.

[22] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzal, and A. Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017, 2017.

[23] J. Wang, Q. Huang, F. Tang, J. Meng, J. Su, and S. Song. Stepwise feature fusion: Local guides global. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 110–120. Springer, 2022.

[24] X. Wang, X. Zhang, Y. Cao, W. Wang, C. Shen, and T. Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023.

[25] J. Wu, R. Fu, H. Fang, Y. Liu, Z. Wang, Y. Xu, Y. Jin, and T. Arbel. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.

[26] K. Zhang and D. Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023.

[27] T. Zhou, Y. Zhang, Y. Zhou, Y. Wu, and C. Gong. Can sam segment polyps? *arXiv preprint arXiv:2304.07583*, 2023.