# ReDi-FM: Frozen Foundation Model for Continual Test-Time Adaptation in Medical Image Segmentation

**Jianhang Ji**[1], **Zhiming Cheng**[2], **Jianxiang Zhao**[2], **Tingyu Wang**[2], **Bingtao Ma**[2], **Yuhan Gao**[2,3], **Zuobin Ying**[1,*] and **Shuai Wang**[2,*]
[1]City University of Macau
[2]Hangzhou Dianzi University
[3]Lishui Institute of Hangzhou Dianzi University
jijianhang.ai@gmail.com, zbying@cityu.edu.mo, shuaiwang.tai@gmail.com

## Abstract

Continual test-time adaptation (CTTA) adapts a pre-trained medical segmentation model online to an unlabeled target stream whose distribution changes over time. However, most existing CTTA methods rely on pseudo-labeling and self-supervised objectives, which inevitably yield noisy supervision under domain shifts. To mitigate this limitation, we introduce off-the-shelf Vision Foundation Models (VFMs) as external knowledge sources. Zero-shot VFMs are insufficient for medical segmentation because they lack medical semantics, yet they contain rich and heterogeneous generic knowledge. To exploit such external knowledge, we propose Reciprocal Distillation with a Frozen Foundation Model (ReDi-FM), a novel framework with two core components: using structure-aware prompts from the source model to guide the frozen VFM to generate target-adapted supervision, and distilling the resulting knowledge back into the adapting model. For robust distillation, we introduce two complementary objectives: uncertainty-driven hard distillation for precise guidance in ambiguous regions, and hard class-balanced soft distillation for richer supervision of under-represented and challenging structures. A consensus-aware gate further stabilizes adaptation when the two teachers disagree. Extensive experiments on multi-domain medical segmentation benchmarks demonstrate that ReDi-FM outperforms state-of-the-art CTTA methods. Code is available at https://github.com/M4cheal/ReDi-FM.

## 1 Introduction

Medical image segmentation is essential for clinical practice, powering applications from computer-aided diagnosis to anatomical structure delineation. Recent advances in Deep Learning (DL) have driven major improvements in segmentation accuracy [Wang *et al.*, 2023a; Fang *et al.*, 2022]. However, these models are highly sensitive to domain shifts from differences in scanners, protocols, and operators, leading to significant performance drops in new clinical settings [Guan and Liu, 2021; Sankaranarayanan *et al.*, 2018]. While Domain Adaptation (DA) methods address this by leveraging abundant unlabeled target data, they yield models that remain fixed after deployment [Li *et al.*, 2024; Wang *et al.*, 2023b; Tang *et al.*, 2024]. Test-Time Adaptation (TTA) overcomes this by adapting models online using only test data [Wang *et al.*, 2020]. Yet, most TTA methods assume a static target distribution. In practice, real-world domains evolve continuously, necessitating Continual Test-Time Adaptation (CTTA) to enable models to update dynamically as the target domain changes [Wang *et al.*, 2022].

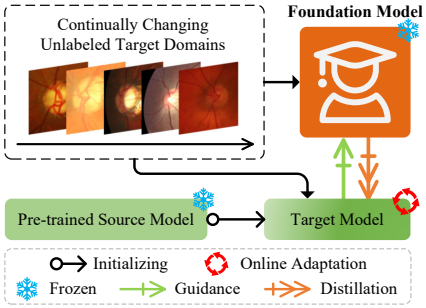


Figure 1: Motivation of ReDi-FM. Conventional CTTA relies on endogenous signals (pseudo-labels/self-supervision on unlabeled target streams), which are bounded by the source model's knowledge and prone to error accumulation. ReDi-FM injects a frozen vision foundation model (e.g., SAM) as an external knowledge source and transfers its heterogeneous knowledge via reciprocal distillation to alleviate this bottleneck.

Most CTTA methods rely on endogenous adaptation signals obtained from the model itself, such as entropy minimization, batch-statistics alignment, or pseudo-label self-training [Wang *et al.*, 2022; Chen *et al.*, 2024b; Maharana *et al.*, 2025]. Under large domain shifts, these signals are inevitably noisy and can accumulate errors over time, limiting adaptation performance. This motivates introducing an external knowledge source to complement endogenous supervision. Large-scale Vision Foundation Models (VFMs), such as the Segment Anything Model (SAM) [Kirillov *et al.*, 2023], provide rich and heterogeneous generic knowledge. This raises a natural question: *Can a frozen VFM serve as a*

*Corresponding authors.

*heterogeneous external knowledge source to improve CTTA under evolving medical domain shifts?*

To answer this question, we introduce a frozen VFM into CTTA to break the knowledge upper bound of purely endogenous supervision from the source model on unlabeled target streams (Figure 1). However, directly applying a generic frozen VFM is suboptimal: while it provides rich cross-domain generic knowledge, it lacks medical semantics. We therefore propose Reciprocal Distillation with a Frozen Foundation Model (ReDi-FM), which uses structure-aware prompting to obtain target-adapted supervision from a frozen VFM and distills the resulting generic knowledge into the target model. To robustly transfer this complementary knowledge, we introduce two distillation objectives. Uncertainty-aware hard distillation emphasizes reliable regions to mitigate noise, and hard class-balanced soft distillation provides richer supervision for under-represented and challenging structures. A consensus-aware gate further stabilizes distillation when the two teachers disagree.

In summary, our contributions are: (1) We introduce frozen vision foundation models as a heterogeneous external knowledge source for CTTA, breaking the knowledge upper bound of adaptation driven solely by the source model on unlabeled target streams. (2) We propose ReDi-FM, a reciprocal distillation framework that uses source-guided, structure-aware prompting to obtain target-adapted supervision from a frozen VFM and distills the resulting heterogeneous knowledge into the target model. (3) Extensive experiments on multi-domain medical segmentation benchmarks demonstrate that ReDi-FM outperforms prior state-of-the-art CTTA methods with improved long-term stability.

## 2 Related Work

### 2.1 Test-time Adaptation

Test-time adaptation (TTA) adapts source-pretrained models to unlabeled target data during inference. Most TTA methods use self-supervised objectives to update models online [Wang *et al.*, 2020; Zhang *et al.*, 2023]. For example, TENT [Wang *et al.*, 2020] minimizes prediction entropy to improve robustness. Other methods update batch normalization (BN) statistics to reduce distribution shifts. DomainAdaptor [Zhang *et al.*, 2023] combines source and test BN statistics and uses a generalized entropy loss for large domain gaps. However, most TTA methods assume a static target domain and perform poorly when distributions change over time.

Continual test-time adaptation (CTTA) tackles evolving target domains, where long-term adaptation leads to error accumulation and forgetting. CTTA methods are either model-based or parameter-efficient. Model-based approaches use self-supervised losses or periodic resets to combat errors [Liu *et al.*, 2024a; Wang *et al.*, 2022; Niu *et al.*, 2023; Zhang *et al.*, 2023]. CoTTA [Wang *et al.*, 2022] averages predictions and restores parts of the model. SAR [Niu *et al.*, 2023] filters noisy samples by gradient size. Parameter-efficient methods freeze most of the source model and update only small components [Chen *et al.*, 2024b; Maharana *et al.*, 2025; Liu *et al.*, 2024b]. VPTTA [Chen *et al.*, 2024b] uses lightweight prompts for BN alignment. PALM [Maharana *et al.*, 2025] updates layers selectively by gradient magnitude. Despite these advances, existing methods are fundamentally constrained by endogenous adaptation signals obtained from the model itself on unlabeled target streams. This motivates leveraging heterogeneous external knowledge beyond conventional source-driven signals.

### 2.2 Foundation Models in Medical Segmentation

Vision Foundation Models (VFMs), such as the Segment Anything Model (SAM) [Kirillov *et al.*, 2023], have recently been explored for medical image segmentation. Most approaches adapt VFMs by full fine-tuning or using lightweight adapters to improve domain-specific performance, but these methods are often computationally intensive and difficult to scale [Ma *et al.*, 2024]. Some works integrate VFMs with vision-language models like CLIP to enhance data efficiency. For example, MedCLIP-SAM [Koleilat *et al.*, 2024] and MedCLIP-SAMv2 [Koleilat *et al.*, 2025] combine BiomedCLIP and SAM for promptable zero-shot segmentation, and TTCS [Chen *et al.*, 2024a] uses BiomedCLIP prompts and self-training to improve SAM's robustness to domain shifts at test time. However, these methods typically use the foundation model as the main segmentation backbone and overlook task-specific priors in conventional models.

Another direction uses foundation models for semi-supervised segmentation [Ma *et al.*, 2025b; Konwer *et al.*, 2025]. While both ReDi-FM and these methods utilize SAM, their settings and methodologies are distinct. For example, in SynFoC [Ma *et al.*, 2025b], SAM generates high-quality pseudo-labels for the conventional model during early training, and the conventional model later corrects high-confidence predictions from SAM. In contrast, ReDi-FM uses task-specific priors from the conventional model to guide the frozen foundation model and distills adapted knowledge back into the conventional model. Importantly, these semi-supervised settings rely on labeled data and offline training cycles, whereas CTTA requires fully online adaptation on unlabeled streams without revisiting the source data.

## 3 Method

### 3.1 Problem Definition and Method Overview

Continual Test-Time Adaptation (CTTA) for medical image segmentation adapts a pre-trained source model $f_{\theta_S}$, trained on labeled source data $(\mathcal{X}_S, \mathcal{Y}_S)$, to a sequence of unlabeled target domains $\mathcal{D}_T = \{\mathcal{X}_{T_1}, \mathcal{X}_{T_2}, \ldots, \mathcal{X}_{T_n}\}$ encountered during deployment. Adaptation is performed online; each target domain is processed once, and source data are not accessible.

Figure 2 gives an overview of the ReDi-FM framework. The target model $f_{\theta_T}$, initialized from the source model, is updated for each target image with guidance from a frozen vision foundation model $f_{\theta_{FM}}$ (e.g., SAM). Our approach consists of two main steps: generating target-adapted segmentation supervision using structure-aware prompts, and distilling knowledge from SAM into $f_{\theta_T}$ via uncertainty-driven hard distillation and hard class-balanced soft distillation.

### 3.2 Target-Adapted Supervision Generation

To overcome the knowledge limitations of conventional models and achieve task-specific, target-adaptive supervision, we

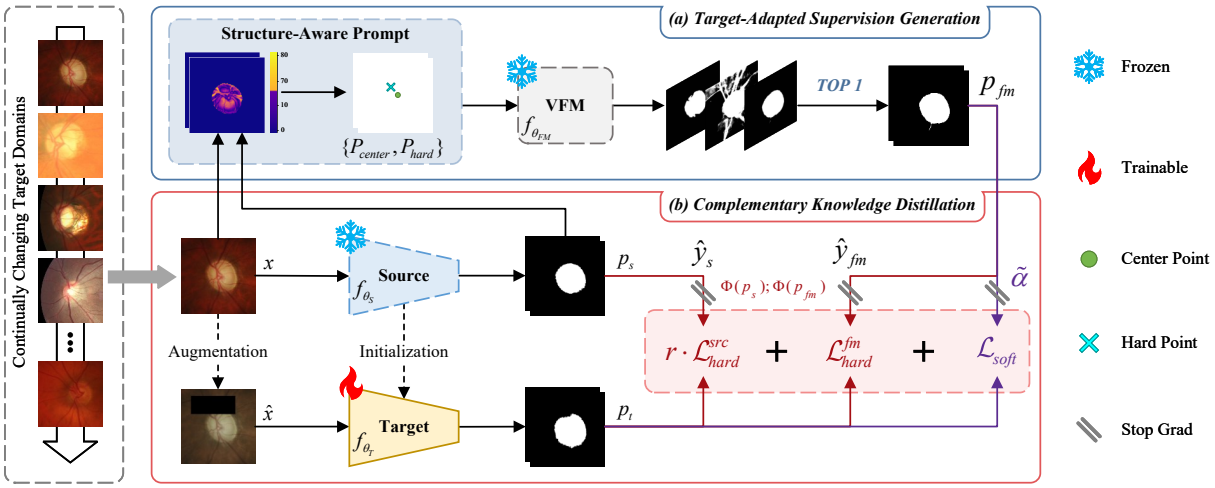


Figure 2: Overview of ReDi-FM. Given each target image $x$, we first perform (a) target-adapted supervision generation. Structure-aware prompts $P_{center}$ and $P_{hard}$ are extracted from the source model's pseudo-label $\hat{y}_s$ and provided to a frozen Vision Foundation Model (VFM, e.g., SAM) to generate candidate masks. The mask with the highest Dice similarity is selected as $\hat{y}_{fm}$. Then we perform (b) complementary knowledge distillation. An augmented image $\hat{x}$ is generated from $x$ and used as input to the target model $f_{\theta_T}$. The target model is updated by uncertainty-driven hard distillation $\mathcal{L}_{hard}$ with pixel-wise confidence weights $\Phi(p_s)$ and $\Phi(p_{fm})$, and hard class-balanced soft distillation $\mathcal{L}_{soft}$ with adaptive class weights $\tilde{\alpha}$. A consensus score $r$ gates the contributions of the source and VFM teachers under disagreement.

leverage semantic priors from the source model to guide the frozen foundation model. Given a target image $x \in \mathcal{D}_T$ with spatial dimensions $H \times W$, the source model $f_{\theta_S}$ produces a probability map $p_s = f_{\theta_S}(x)$, from which a binary pseudo-label $\hat{y}_s$ is generated by thresholding at $\tau$:

$$\hat{y}_s = \mathbb{I}[p_s > \tau], \tag{1}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function.

Directly using $\hat{y}_s$ as a mask/box prompt may inherit its shape errors under domain shift. We therefore use sparse point prompts to provide only coarse localization. To represent the main structure of each class, we extract the center point $P^c_{center}$, defined as the centroid of the largest connected foreground component in $\hat{y}^c_s$. While the center point provides a strong structural prior, it may miss complex or atypical regions, especially when noisy pseudo-labels are present. To address this limitation, we introduce an additional structure-aware prompt: the hard point $P^c_{hard}$. For each class $c$, we first compute the class prototype $z^c$ as the mean intensity of all foreground pixels:

$$z^c = \frac{\sum_v x^v \cdot \hat{y}^{c,v}_s}{\sum_v \hat{y}^{c,v}_s}. \tag{2}$$

Next, for each foreground pixel, we calculate its intensity deviation from the prototype:

$$d_{c,v} = |x^v - z^c|. \tag{3}$$

The hard point is then defined as the foreground pixel with the maximum deviation:

$$P^c_{hard} = \arg\max_v d_{c,v}. \tag{4}$$

The set of structure-aware prompts, $\mathcal{P} = \{P^c_{center}, P^c_{hard}\}_{c=1}^{C}$, is provided to the frozen foundation model $f_{\theta_{FM}}$ along with the target image $x$, i.e., $f_{\theta_{FM}}(x, \mathcal{P})$, which generates candidate segmentation masks. Among these candidates, we select the mask $\hat{y}_{fm}$ that achieves the highest Dice similarity with the source pseudo-label:

$$\hat{y}_{fm} = \arg\max_{m \in \mathcal{M}} Dice(m, \hat{y}_s), \tag{5}$$

where $\mathcal{M}$ is the set of candidate masks and $Dice(\cdot,\cdot)$ denotes the Dice coefficient. We use $\hat{y}_s$ only as a coarse anchor to select the intended mask among candidates, rather than enforcing its boundary shape. The selected mask $\hat{y}_{fm}$ and its associated probabilistic output $p_{fm}$ serve as target-adapted supervision for subsequent distillation.

### 3.3 Complementary Knowledge Distillation

The frozen foundation model (e.g., SAM) provides valuable external knowledge for the target domain, but its predictions alone are often insufficient for accurate, domain-specific medical segmentation. To effectively transfer the generated supervision to the trainable target model $f_{\theta_T}$, we propose two complementary distillation objectives: an uncertainty-driven hard distillation and a hard class-balanced soft distillation.

**Uncertainty-Driven Hard Distillation.** Directly using hard pseudo-labels from the source model or SAM can result in errors due to domain shift and prediction noise. To enhance robustness, we introduce an uncertainty-weighted hard distillation objective, where pixel-wise confidence is used to adaptively weight the supervision signal. Following [Yang *et al.*, 2024], the confidence weighting function $\Phi(p)$ for a predicted probability $p \in [0, 1]$ is defined as:

$$\Phi(p^i) = \gamma - \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(p^i - \mu)^2}{2\sigma^2}\right), \tag{6}$$

where $\gamma = 1.3$, $\mu = 0.5$, and $\sigma^2 = 0.1$. This assigns higher weights to confident predictions (probabilities near 0 or 1) and lower weights to uncertain pixels.

Pixel-wise confidence weights are computed for the outputs of both the source model ($p_s$) and SAM ($p_{fm}$), resulting in $\Phi(p_s^i)$ and $\Phi(p_{fm}^i)$. These weights are incorporated into a segmentation loss combining weighted binary cross-entropy and Dice terms. Let $p_t = f_{\theta_T}(\hat{x})$ denote the target model's probability map with $\hat{x}$ an augmented version of $x$, $\hat{y}_s$ the source pseudo-label, and $\hat{y}_{fm}$ the binary mask from SAM. The hard distillation losses are defined as:

$$\mathcal{L}_{hard}^{src} = \mathcal{L}_{seg}^{w}(p_t, \hat{y}_s; \Phi(p_s)), \tag{7}$$

$$\mathcal{L}_{hard}^{fm} = \mathcal{L}_{seg}^{w}(p_t, \hat{y}_{fm}; \Phi(p_{fm})), \tag{8}$$

where $\mathcal{L}_{seg}^{w}(p, \hat{y}; w)$ applies a pixel-wise weight map $w$ (here $w = \Phi(\cdot)$) to both the BCE and Dice terms. For convenience, we also define the ungated sum of the two branches as

$$\mathcal{L}_{hard}^{base} = \mathcal{L}_{hard}^{src} + \mathcal{L}_{hard}^{fm}. \tag{9}$$

This adaptive weighting allows all pixels to contribute, while emphasizing those with higher confidence, thus mitigating the negative impact of noisy or ambiguous supervision.

**Hard Class-Balanced Soft Distillation.** Soft distillation leverages the full probabilistic outputs of the teacher model (SAM), offering richer supervisory signals and greater stability under noisy pseudo-labels. However, medical image segmentation is often hampered by severe class imbalance, as foreground structures typically occupy only a small portion of the image. To address this challenge, we propose a hard class-balanced soft distillation objective that dynamically emphasizes both difficult and under-represented classes. For each class $c$, we compute an adaptive class weight $\eta^c$ that jointly accounts for the spatial prevalence of class $c$ in SAM's prediction and the degree of disagreement between the target and SAM outputs:

$$\eta^c = \frac{\sum_v \mathbb{I}[\hat{y}_{fm}^{c,v} = 1] + \sum_v \mathbb{I}[\hat{y}_t^{c,v} \neq \hat{y}_{fm}^{c,v}]}{2 \times H \times W}, \tag{10}$$

where $H \times W$ denotes the image resolution, and $\hat{y}_t = \mathbb{I}[p_t > \tau]$ is the binarized prediction from the target model.

To further concentrate learning on hard regions, we define a pixel-wise difficulty mask $\bar{y}^{c,i}$ for each class $c$ and pixel $i$. The mask $\bar{y}^{c,i}$ is set to 1 if either (i) the source model predicts class $c$ at pixel $i$, or (ii) the source model's prediction at $i$ exhibits high uncertainty (i.e., high entropy); otherwise, $\bar{y}^{c,i} = 0$. This mechanism leverages both structure-aware priors and the current model's uncertainty, ensuring that challenging or ambiguous pixels are adaptively up-weighted during training. The per-pixel adaptive weight is then given by:

$$\alpha^{c,i} = \eta^c \cdot \mathbb{I}[\bar{y}^{c,i} = 1] + (1 - \eta^c) \cdot \mathbb{I}[\bar{y}^{c,i} = 0], \tag{11}$$

where $\eta^c$ is the adaptive class weight. The final soft distillation loss is given by:

$$\mathcal{L}_{soft} = -\frac{1}{H \times W} \sum_i \sum_c \alpha^{c,i}\, p_t^{c,i} \log p_{fm}^{c,i}, \tag{12}$$

where $p_t^{c,i}$ and $p_{fm}^{c,i}$ denote the predicted probabilities for class $c$ at pixel $i$ from the target model and SAM, respectively. By emphasizing pixels identified by the source model or those with high uncertainty, this loss improves adaptation to domain shifts and enhances robustness in segmenting small or under-represented structures, which are prevalent in medical images.

**Consensus-Aware Gating.** The source model provides task-specific semantics but is sensitive to domain shift, while the frozen VFM offers a generic cross-domain structural prior. When they disagree, enforcing both signals can amplify errors in unlabeled CTTA. We therefore introduce a consensus-aware gate. We measure their agreement by

$$r = \text{Dice}(\hat{y}_s, \hat{y}_{fm}), \quad r \in [0, 1]. \tag{13}$$

We use $r$ to gate the shift-sensitive source supervision while keeping the VFM branch active:

$$\mathcal{L}_{hard} = \mathcal{L}_{hard}^{fm} + r \cdot \mathcal{L}_{hard}^{src}. \tag{14}$$

For soft distillation, we anneal the pixel-wise re-weighting toward uniform weights under disagreement:

$$\tilde{\alpha}^{c,i} = r \cdot \alpha^{c,i} + (1 - r) \cdot 1. \tag{15}$$

When $r$ is small, the source and VFM disagree and the source-guided hard weighting in $\alpha^{c,i}$ may over-emphasize unreliable regions. Annealing $\alpha^{c,i}$ toward 1 tempers this hard weighting and moves $\mathcal{L}_{soft}$ toward uniform weighting. We replace $\alpha^{c,i}$ with $\tilde{\alpha}^{c,i}$ in $\mathcal{L}_{soft}$.

**Total Objective.** The overall objective is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{hard} + \mathcal{L}_{soft}. \tag{16}$$

# 4 Experiments

## 4.1 Datasets and Evaluation Metrics

**OD/OC Segmentation Dataset.** We benchmark on five public fundus image datasets for Optic Disc (OD) and Optic Cup (OC) segmentation: RIM-ONE-r3 [Fumero *et al.*, 2011] (Domain A, 159 images), REFUGE [Orlando *et al.*, 2020] (Domain B, 400 images), ORIGA [Zhang *et al.*, 2010] (Domain C, 650 images), REFUGE-Validation/Test [Orlando *et al.*, 2020] (Domain D, 800 images), and Drishti-GS [Sivaswamy *et al.*, 2014] (Domain E, 101 images). All images are resized to $512 \times 512$ pixels. Performance is evaluated using the Dice similarity coefficient (DSC).

**Prostate Segmentation Dataset.** For cross-domain prostate segmentation, we utilize T2-weighted MRI slices from six medical centers: BIDMC (Domain A, 261 slices), BMC (Domain B, 384 slices), HK (Domain C, 158 slices), I2CVB (Domain D, 468 slices), RUNMC (Domain E, 421 slices), and UCL (Domain F, 175 slices) [Bloch *et al.*, 2015; Lemaître *et al.*, 2015; Litjens *et al.*, 2014; Liu *et al.*, 2020]. Each slice is resampled to $384 \times 384$ pixels. Results are reported in terms of DSC.

**Polyp Segmentation Dataset.** Polyp segmentation is evaluated on four datasets: BKAI-IGH-NeoPolyp [Ngoc Lan *et al.*, 2021] (Domain A, 1000 images), CVC-ClinicDB [Bernal *et al.*, 2015] (Domain B, 612 images), ETIS-LaribPolypDB [Silva *et al.*, 2014] (Domain C, 196 images), and Kvasir-Seg [Jha *et al.*, 2019] (Domain D, 1000 images). All images are resized to $352 \times 352$ pixels. We report DSC for region overlap, structure similarity ($S_\alpha$) for structural consistency, and enhanced alignment measure ($E_\phi^{\max}$) for boundary alignment.

| Methods | Venue | OD/OC Segmentation | | | | | | Prostate Segmentation | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Do. A | Do. B | Do. C | Do. D | Do. E | Average | Do. A | Do. B | Do. C | Do. D | Do. E | Do. F | Average |
| | | *DSC* | *DSC* | *DSC* | *DSC* | *DSC* | *DSC* ↑ | *DSC* | *DSC* | *DSC* | *DSC* | *DSC* | *DSC* | *DSC* ↑ |
| TTCS | BIBM'24 | 49.38 | 44.92 | 40.09 | 46.52 | 44.42 | 45.07 | 9.20 | 22.44 | 23.91 | 37.45 | 33.51 | 28.63 | 25.86 |
| MedCLIP-SAM | MICCAI'24 | 37.42 | 38.74 | 28.47 | 44.23 | 36.25 | 37.02 | 8.74 | 8.93 | 8.76 | 5.68 | 8.29 | 8.80 | 8.20 |
| MedCLIP-SAMv2 | MIA'25 | 62.83 | 61.89 | 63.10 | 61.84 | 61.91 | 62.31 | 37.26 | 38.37 | 35.48 | 28.38 | 34.85 | 36.31 | 35.11 |
| Source Only | - | 53.59 | 74.01 | 67.24 | 57.39 | 71.17 | 64.68 | 51.02 | 58.79 | 26.88 | 24.15 | 60.42 | 53.33 | 45.77 |
| TENT | ICLR'21 | 69.81 | 75.78 | 70.85 | 47.22 | **76.16** | 67.96 | 63.74 | 52.89 | 4.51 | **44.51** | 58.58 | 6.61 | 38.47 |
| CoTTA | CVPR'22 | 57.11 | 58.45 | 47.16 | 70.59 | 52.14 | 57.09 | 42.11 | 48.13 | 30.50 | 30.84 | 47.74 | 38.33 | 39.61 |
| SAR | ICLR'23 | 70.13 | 76.82 | 69.05 | 62.53 | 69.73 | 69.65 | 62.67 | 61.42 | 49.73 | 38.04 | 62.50 | 60.55 | 55.82 |
| VPTTA | CVPR'24 | 66.70 | 75.01 | 68.66 | 59.76 | 71.73 | 68.37 | 65.74 | 61.71 | 42.12 | 30.10 | 62.66 | 59.48 | 53.64 |
| PALM | AAAI'25 | 69.75 | **77.37** | 68.71 | 57.74 | 73.30 | 69.37 | 64.44 | 61.90 | 45.07 | 40.27 | 62.93 | 49.16 | 53.96 |
| SURGEON | CVPR'25 | 69.89 | 76.99 | 69.54 | 62.45 | 69.43 | 69.66 | 62.82 | 61.44 | 49.74 | 38.02 | 62.47 | 60.40 | 55.82 |
| **ReDi-FM (Ours)** | This paper | **73.11** | 75.69 | **73.31** | **73.48** | 72.37 | **73.59** | **66.10** | **66.44** | **49.89** | 41.24 | **73.50** | **73.69** | **61.81** |

Table 1: Performance comparison on OD/OC and Prostate segmentation tasks. The source model is ResUNet-34. 'Do. X' denotes using domain X as the source and reporting the mean performance over all remaining target domains; 'Average' further averages over all source domains. Best and second-best results are in bold and underline, respectively.

## 4.2 Implementation Details

All models are implemented in PyTorch 2.2.2 on a single NVIDIA RTX 3090 GPU. To ensure statistical reliability, we conduct experiments with three random seeds and report the mean. For OD/OC and prostate segmentation, we use ResUNet-34 [Chen *et al.*, 2024b] as the source model. For the more challenging polyp segmentation task, we adopt Segformer-B0 [Xie *et al.*, 2021]. During continual test-time adaptation, each test image is adapted with a single online iteration (batch size = 1). All methods, including ReDi-FM and all baselines, are evaluated under identical experimental settings to ensure a fair comparison. For deployment of ReDi-FM, we use the Adam optimizer for all tasks, with a learning rate of 0.0001 for OD/OC and prostate segmentation, and 0.00001 for polyp segmentation. The pseudo-label threshold $\tau$ is set to 0.5. The ViT-H variant of the Segment Anything Model (SAM) is used as the frozen foundation model in all experiments. Data augmentation includes color jittering and random erasing. For MedCLIP-SAM/MedCLIP-SAMv2/TTCS, we follow the official inference protocol and prompts provided by the authors.

## 4.3 Experimental Results

We benchmark ReDi-FM against a comprehensive set of baselines, including: (1) the 'Source Only' model (trained on the source domain and directly evaluated on the target domain without adaptation), (2) three zero-shot foundation models, MedCLIP-SAM [Koleilat *et al.*, 2024], MedCLIP-SAMv2 [Koleilat *et al.*, 2025], and TTCS [Chen *et al.*, 2024a], and (3) several state-of-the-art CTTA methods: TENT [Wang *et al.*, 2020] and SAR [Niu *et al.*, 2023] (entropy-based), CoTTA [Wang *et al.*, 2022] (pseudo-labeling), parameter-efficient methods VPTTA [Chen *et al.*, 2024b] and PALM [Maharana *et al.*, 2025], and SURGEON [Ma *et al.*, 2025a] (using dynamic activation sparsity). For fair comparison, we re-implement all CTTA baselines with the same backbone and follow exactly the same test-time protocol as ReDi-FM. Following the standard CTTA evaluation protocol, each domain is used in turn as the source, and the model is evaluated on all remaining target domains. Our results are averaged over three random seeds. For each source domain, the remaining target domains are streamed in a fixed order (A→B→ $\cdots$) for all methods.

**OD/OC Segmentation.** Table 1 presents segmentation results across five domains. Most CTTA methods outperform the 'Source Only' baseline, confirming the benefit of online adaptation under domain shifts. ReDi-FM achieves the best overall performance, surpassing both zero-shot foundation models and prior CTTA methods. Zero-shot baselines such as MedCLIP-SAM and TTCS leverage foundation priors but do not adapt to the target stream, which limits robustness across domains and can underperform 'Source Only'. The ReDi-FM framework generates target-adaptive supervision and transfers heterogeneous external knowledge via reciprocal distillation. This enables robust segmentation across all domains: ReDi-FM remains competitive on relatively easy domains and achieves the best performance on more challenging ones (e.g., Domain D).

**Prostate Segmentation.** Table 1 reports cross-domain prostate segmentation results, a task characterized by high inter-domain variability and frequent model overconfidence. Due to excessive noise, entropy-based methods (TENT) and pseudo-labeling methods (CoTTA) can be unstable in this setting and may even underperform the 'Source Only' baseline. Methods that control error accumulation (e.g., SAR) and parameter-efficient adaptation (e.g., VPTTA and PALM) are generally more robust, yet their adaptation signal is still constrained by relying only on the source model and unlabeled target data, which can be insufficient under large domain shifts. Zero-shot foundation models such as MedCLIP-SAM and TTCS also perform poorly, highlighting the limits of generic knowledge without online adaptation. In contrast, ReDi-FM distills external knowledge into the target model, surpassing the knowledge constraints of both the source and the unlabeled target stream. As a result, ReDi-FM improves the best prior CTTA method by 13.14 $DSC$ points on Domain F and achieves an overall gain of 16.04 $DSC$ points over 'Source Only'.

| Methods | Venue | Domain A | | | Domain B | | | Domain C | | | Domain D | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $DSC$ | $E_\phi^{\max}$ | $S_\alpha$ | $DSC$ | $E_\phi^{\max}$ | $S_\alpha$ | $DSC$ | $E_\phi^{\max}$ | $S_\alpha$ | $DSC$ | $E_\phi^{\max}$ | $S_\alpha$ | $DSC\uparrow$ | $E_\phi^{\max}\uparrow$ | $S_\alpha\uparrow$ |
| TTCS | BIBM'24 | 12.57 | 39.63 | 49.52 | 39.73 | 70.26 | 63.18 | 25.08 | 54.84 | 56.14 | 35.49 | 67.49 | 62.02 | 28.22 | 58.05 | 57.72 |
| MedCLIP-SAM | MICCAI'24 | 46.35 | 63.39 | 58.81 | 42.61 | 62.09 | 58.36 | 44.31 | 62.96 | 58.63 | 36.45 | 57.71 | 54.54 | 42.43 | 61.54 | 57.59 |
| MedCLIP-SAMv2 | MIA'25 | 35.51 | 63.28 | 56.29 | 37.65 | 67.88 | 59.44 | 39.31 | 66.09 | 60.03 | 40.05 | 67.37 | 63.27 | 38.13 | 66.15 | 59.76 |
| Source Only | - | 74.22 | 83.82 | 81.37 | 70.41 | 82.67 | 80.02 | 68.46 | 83.05 | 77.47 | 76.55 | 88.34 | 84.56 | 72.41 | 84.47 | 80.86 |
| TENT | ICLR'21 | 71.14 | 81.82 | 79.23 | 66.67 | 79.79 | 77.18 | 67.16 | 83.07 | 76.24 | 71.38 | 83.35 | 81.06 | 69.09 | 82.01 | 78.43 |
| CoTTA | CVPR'22 | 50.41 | 72.04 | 65.26 | 42.76 | 60.22 | 60.47 | 49.00 | 76.33 | 64.22 | 49.13 | 67.29 | 66.70 | 47.83 | 68.97 | 64.16 |
| SAR | ICLR'23 | 72.17 | 82.72 | 79.81 | 65.57 | 78.53 | 76.44 | 66.78 | 82.90 | 76.02 | 71.18 | 83.24 | 80.90 | 68.93 | 81.85 | 78.29 |
| PALM | AAAI'25 | 74.43 | 84.22 | 81.20 | 67.46 | 79.76 | 77.72 | 68.37 | **83.77** | 77.07 | 72.90 | 84.28 | 82.05 | 70.79 | 83.01 | 79.51 |
| SURGEON | CVPR'25 | 72.26 | 82.84 | 79.91 | 65.55 | 78.53 | 76.39 | 66.59 | 82.87 | 75.88 | 71.02 | 83.11 | 80.87 | 68.86 | 81.84 | 78.27 |
| **ReDi-FM (Ours)** | This paper | **77.47** | **87.02** | **82.78** | **73.97** | **85.65** | **81.75** | **69.56** | 82.48 | **78.49** | **78.35** | **90.02** | **85.22** | **74.84** | **86.29** | **82.06** |

Table 2: Performance comparison on polyp segmentation. The source model is Segformer-B0. Best and second-best results are shown in bold and underline, respectively.

| Time | | OD/OC Segmentation → | | | | | | Prostate Segmentation → | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Round | | 1 | 2 | 3 | Average | Gain | Perform. | 1 | 2 | 3 | Average | Gain | Perform. |
| Methods | Venue | $DSC$ | $DSC$ | $DSC$ | $DSC\uparrow$ | $DSC\uparrow$ | Degra.$\downarrow$ | $DSC$ | $DSC$ | $DSC$ | $DSC\uparrow$ | $DSC\uparrow$ | Degra.$\downarrow$ |
| Source Only | - | 64.68 | 64.68 | 64.68 | 64.68 | - | - | 45.77 | 45.77 | 45.77 | 45.77 | - | - |
| TENT | ICLR'21 | 67.96 | 60.24 | 55.07 | 61.09 | -3.59 | 6.87 | 38.47 | 14.67 | 6.16 | 19.77 | -26.00 | 18.70 |
| CoTTA | CVPR'22 | 57.09 | 37.45 | 29.38 | 41.31 | -23.37 | 15.78 | 39.61 | 12.04 | 8.38 | 20.01 | -25.76 | 19.60 |
| SAR | ICLR'23 | 69.65 | 69.58 | 69.48 | 69.57 | +4.89 | 0.08 | 55.82 | 55.92 | 55.98 | 55.91 | +10.14 | -0.09 |
| VPTTA | CVPR'24 | 68.37 | 67.90 | 67.70 | 67.99 | +3.31 | 0.38 | 53.64 | 52.86 | 52.48 | 52.99 | +7.23 | 0.64 |
| PALM | AAAI'25 | 69.37 | 68.19 | 67.08 | 68.22 | +3.53 | 1.16 | 53.96 | 41.52 | 39.09 | 44.86 | -0.91 | 9.10 |
| SURGEON | CVPR'25 | 69.66 | 69.64 | 69.60 | 69.63 | +4.95 | **0.02** | 55.82 | 55.76 | 55.59 | 55.72 | +9.96 | 0.09 |
| **ReDi-FM (Ours)** | This paper | **73.59** | **73.17** | **72.74** | **73.17** | **+8.48** | 0.43 | **61.81** | **64.85** | **65.33** | **64.00** | **+18.23** | **-2.19** |

Table 3: Performance comparison in a long-term CTTA on the OD/OC segmentation and the Prostate segmentation. Gain indicates the improvement over the 'Source Only'. 'Perform. Degra.' is computed as Round 1 − Average (negative indicates improvement over rounds).

**Polyp Segmentation.** We evaluate ReDi-FM on polyp segmentation using Segformer-B0 as the source model, while VPTTA is excluded because it relies on batch normalization. Table 2 presents results across four domains. In this task, several CTTA methods perform worse than the 'Source Only' baseline, reflecting the challenges of segmenting small, hidden, and highly variable polyps, which increases the risk of negative transfer during adaptation. Zero-shot foundation models also perform poorly, underscoring the need for task-specific adaptation. Despite these difficulties, ReDi-FM consistently achieves the best overall performance across metrics. These results indicate that reciprocal distillation can inject complementary foundation knowledge into the target model and reduce negative transfer on challenging polyp datasets, which is not achieved by prior CTTA methods.

**Comparison in a Long-Term CTTA.** We evaluate the long-term robustness of ReDi-FM through multi-round adaptation on the OD/OC and prostate segmentation benchmarks (Table 3). In each experiment, the model is sequentially adapted to target domains over three rounds, simulating repeated real-world domain shifts. Results show that most baseline CTTA methods suffer from performance drift as adaptation proceeds. Entropy minimization (TENT) and pseudo-labeling (CoTTA) exhibit clear degradation across rounds, while update-filtering methods (e.g., SAR) and activation-sparsity (SURGEON) improve stability but with limited gains. In contrast, ReDi-FM achieves the highest average performance and the largest overall gains, while remaining stable across rounds. Notably, on prostate segmentation ReDi-FM further improves with successive adaptation cycles (negative degradation), indicating sustainable continual adaptation under repeated domain shifts.

| $\mathcal{L}_{hard}$ | | | $\mathcal{L}_{soft}$ | Gate $r$ | Average $DSC\uparrow$ |
|---|---|---|---|---|---|
| $\mathcal{L}_{hard}^{src}$ | $\mathcal{L}_{hard}^{fm}$ | Uncertainty $\Phi$ | | | |
| - | - | - | - | - | $64.68 \pm 0.00$ |
| ✓ | - | - | - | - | $65.89 \pm 0.10$ |
| - | ✓ | - | - | - | $65.80 \pm 0.22$ |
| ✓ | ✓ | - | - | - | $71.07 \pm 0.23$ |
| ✓ | ✓ | ✓ | - | - | $71.85 \pm 0.13$ |
| ✓ | ✓ | ✓ | ✓ | - | $72.70 \pm 0.11$ |
| ✓ | ✓ | ✓ | ✓ | ✓ | $\mathbf{73.59 \pm 0.05}$ |

Table 4: Ablation study on OD/OC segmentation. Results are reported as mean ± std over three random seeds, averaged over five domain sequences.

## 4.4 Ablation Study

**Effectiveness of Core Components.** We conduct an ablation study on OD/OC segmentation (Table 4) to assess the contribution of each component. We evaluate hard distillation from the source model $\mathcal{L}_{hard}^{src}$, hard distillation from the SAM branch $\mathcal{L}_{hard}^{fm}$, pixel-wise uncertainty weighting, hard class-balanced soft distillation $\mathcal{L}_{soft}$, and the consensus-aware gate $r$. Hard distillation from the source model provides a moderate gain, and adding the SAM branch intro-

| Metrics | TTCS | MedCLIP-SAM | TENT | CoTTA | SAR | PALM | SURGEON | ReDi-FM | ReDi-FM (FM=ViT-B) |
|---|---|---|---|---|---|---|---|---|---|
| Average $DSC$ ↑ | 28.22 | 42.43 | 69.09 | 47.83 | 68.93 | 70.79 | 68.86 | 74.84 | 74.94 |
| Latency (ms) ↓ | 4409.65 | 2709.88 | 23.78 | 527.85 | 50.52 | 129.86 | 187.20 | 345.54 | 144.11 |

Table 5: Efficiency comparison on polyp segmentation. Latency is measured in ms per test image under the same CTTA setting.

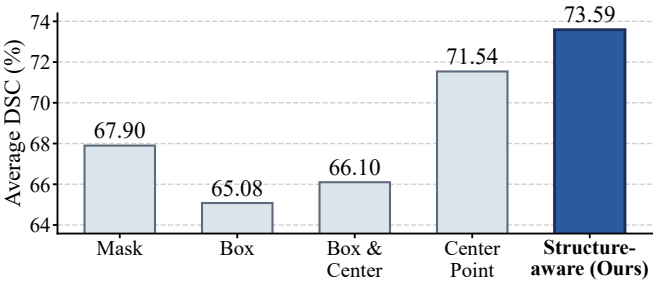


Figure 3: Comparison of different prompting strategies on OD/OC segmentation (average over five domain sequences).

| Strategies | w/o CB | Vanilla CB | **Hard CB (Ours)** |
|---|---|---|---|
| Average $DSC$ ↑ | 72.17 ± 0.32 | 72.41 ± 0.35 | **73.59 ± 0.05** |

Table 6: Ablation of class balancing (CB) strategies on OD/OC segmentation (average over five domain sequences).

duces external knowledge for further improvement. Pixel-wise uncertainty weighting enhances robustness against noisy pseudo-labels. Incorporating class-balanced soft distillation enables the model to absorb richer knowledge from the foundation model, achieving further improvement. Finally, the consensus-aware gate $r$ stabilizes distillation under teacher disagreement by down-weighting unreliable source supervision, leading to the best overall performance. Overall, each component incrementally improves performance.

**Effectiveness of Prompting Strategies.** We investigate different prompting strategies for guiding the foundation-model branch, as shown in Figure 3. Mask and Box prompts are less stable under domain shift, as they directly inherit errors from source pseudo-labels and may amplify them during adaptation. Using a Center point improves localization, while adding a Box on top brings only marginal gains. Our structure-aware prompt, which combines an informative hard point with the foreground center, achieves the best overall performance.

**Effectiveness of Class-Balancing Strategies.** We evaluate different class-balancing strategies for adaptation in Table 6. Removing class balancing leads to suboptimal performance, while conventional class-balanced weighting can be brittle under noisy pseudo-labels. In contrast, our hard class balancing provides a more stable weighting scheme and yields the best result.

**FM-only vs. full framework.** To isolate the role of reciprocal distillation, we report ReDi-FM$^{\text{FM}}$ that uses only the FM branch outputs (Table 7). ReDi-FM$^{\text{FM}}$ is not consistently strong across benchmarks and can be inferior to the source model, indicating that frozen FM outputs alone are insufficient under domain shifts. In contrast, ReDi-FM yields consistent improvements, suggesting that the gains mainly come from transferring complementary knowledge into the target model via reciprocal distillation, rather than relying on FM outputs alone.

| Methods | Average $DSC$ ↑ | | |
|---|---|---|---|
| | OD/OC | Prostate | Polyp |
| Source Only | 64.68 | 45.77 | 72.41 |
| ReDi-FM$^{\text{FM}}$ | 57.70 | 52.58 | 72.18 |
| **ReDi-FM (Ours)** | **73.59** | **61.81** | **74.84** |

Table 7: FM-only vs. full framework (Average $DSC$). ReDi-FM$^{\text{FM}}$ denotes using only the FM branch outputs.

**Efficiency–Accuracy Trade-off.** Table 5 reports the accuracy–latency trade-off on polyp segmentation under the same CTTA setting. Lightweight CTTA baselines (e.g., TENT/SAR) are fast but less accurate, while VFM-based zero-shot methods are much slower and still underperform. ReDi-FM attains the best accuracy with moderate latency. With a lighter FM backbone (ViT-B), latency drops with negligible accuracy change, indicating the gain mainly comes from our distillation framework.

## 5 Conclusion

We propose ReDi-FM, a reciprocal distillation framework designed to address the challenges of CTTA in medical image segmentation. To the best of our knowledge, this is the first work to tackle CTTA by leveraging an off-the-shelf frozen vision foundation model as a heterogeneous external teacher. This alleviates the limitations of previous approaches that relied primarily on source-driven self-supervised adaptation. By introducing external knowledge from the foundation model, ReDi-FM pushes beyond the knowledge upper bound of source-model-driven adaptation on unlabeled target streams. Our framework integrates structure-aware prompting, which enables the foundation model to generate task-adapted supervision under the guidance of the source model, with complementary distillation strategies that reliably transfer both generic and target-adapted knowledge to the target model. Concretely, ReDi-FM distills complementary signals into the target model via uncertainty-driven hard distillation and hard class-balanced soft distillation, and uses a consensus-aware gate to coordinate the two teachers under disagreement. Extensive experiments on three medical segmentation benchmarks demonstrate that ReDi-FM consistently outperforms strong CTTA baselines and zero-shot foundation-model approaches in both adaptation accuracy and long-term stability.

## Acknowledgements

This research was supported by NSFC-FDCT under its Joint Scientific Research Project Fund (Grant No. 0004/2025/AFJ), National Natural Science Foundation of China (Grant No. 62571173), and Zhejiang Provincial Natural Science Foundation of China (Grant Nos. LD25F020005 and LQN25F030009).

## References

[Bernal *et al.*, 2015] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando 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.

[Bloch *et al.*, 2015] N Bloch, A Madabhushi, H Huisman, J Freymann, J Kirby, M Grauer, A Enquobahrie, C Jaffe, L Clarke, and K Farahani. Nci-isbi 2013 challenge: Automated segmentation of prostate structures. The Cancer Imaging Archive, 2015.

[Chen *et al.*, 2024a] Haotian Chen, Yonghui Xu, Yanyu Xu, Yixin Zhang, and Lizhen Cui. Test-time medical image segmentation using clip-guided sam adaptation. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1866–1873. IEEE, 2024.

[Chen *et al.*, 2024b] Ziyang Chen, Yongsheng Pan, Yiwen Ye, Mengkang Lu, and Yong Xia. Each test image deserves a specific prompt: Continual test-time adaptation for 2d medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11184–11193, 2024.

[Fang *et al.*, 2022] Huihui Fang, Fei Li, Huazhu Fu, Junde Wu, Xiulan Zhang, and Yanwu Xu. Dataset and evaluation algorithm design for goals challenge. In *International Workshop on Ophthalmic Medical Image Analysis*, pages 135–142. Springer, 2022.

[Fumero *et al.*, 2011] Francisco Fumero, Silvia Alayón, José L Sanchez, Jose Sigut, and M Gonzalez-Hernandez. Rim-one: An open retinal image database for optic nerve evaluation. In *2011 24th international symposium on computer-based medical systems (CBMS)*, pages 1–6. IEEE, 2011.

[Guan and Liu, 2021] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021.

[Jha *et al.*, 2019] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*, pages 451–462. Springer, 2019.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[Koleilat *et al.*, 2024] Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Medclip-sam: Bridging text and image towards universal medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 643–653. Springer, 2024.

[Koleilat *et al.*, 2025] Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Medclip-samv2: Towards universal text-driven medical image segmentation. *Medical Image Analysis*, page 103749, 2025.

[Konwer *et al.*, 2025] Aishik Konwer, Zhijian Yang, Erhan Bas, Cao Xiao, Prateek Prasanna, Parminder Bhatia, and Taha Kass-Hout. Enhancing sam with efficient prompting and preference optimization for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20990–21000, 2025.

[Lemaître *et al.*, 2015] Guillaume Lemaître, Robert Martí, Jordi Freixenet, Joan C Vilanova, Paul M Walker, and Fabrice Meriaudeau. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: a review. *Computers in biology and medicine*, 60:8–31, 2015.

[Li *et al.*, 2024] Jingjing Li, Zhiqi Yu, Zhekai Du, Lei Zhu, and Heng Tao Shen. A comprehensive survey on source-free domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5743–5762, 2024.

[Litjens *et al.*, 2014] Geert Litjens, Robert Toth, Wendy Van De Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram Van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014.

[Liu *et al.*, 2020] Quande Liu, Qi Dou, and Pheng-Ann Heng. Shape-aware meta-learning for generalizing prostate mri segmentation to unseen domains. In *International conference on medical image computing and computer-assisted intervention*, pages 475–485. Springer, 2020.

[Liu *et al.*, 2024a] Jiaming Liu, Ran Xu, Senqiao Yang, Renrui Zhang, Qizhe Zhang, Zehui Chen, Yandong Guo, and Shanghang Zhang. Continual-mae: Adaptive distribution masked autoencoders for continual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28653–28663, 2024.

[Liu *et al.*, 2024b] Jiaming Liu, Senqiao Yang, Peidong Jia, Renrui Zhang, Ming Lu, Yandong Guo, Wei Xue, and Shanghang Zhang. Vida: Homeostatic visual domain adapter for continual test time adaptation. In *International Conference on Learning Representations*, volume 2024, pages 48396–48417, 2024.

[Ma *et al.*, 2024] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024.

[Ma *et al.*, 2025a] Ke Ma, Jiaqi Tang, Bin Guo, Fan Dang, Sicong Liu, Zhui Zhu, Lei Wu, Cheng Fang, Ying-Cong Chen, Zhiwen Yu, et al. Surgeon: Memory-adaptive fully test-time adaptation via dynamic activation sparsity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 30514–30523, 2025.

[Ma *et al.*, 2025b] Qinghe Ma, Jian Zhang, Zekun Li, Lei Qi, Qian Yu, and Yinghuan Shi. Steady progress beats stagnation: Mutual aid of foundation and conventional models in mixed domain semi-supervised medical image segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5175–5185, 2025.

[Maharana *et al.*, 2025] Sarthak Kumar Maharana, Baoming Zhang, and Yunhui Guo. Palm: Pushing adaptive learning rate mechanisms for continual test-time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19378–19386, 2025.

[Ngoc Lan *et al.*, 2021] Phan Ngoc Lan, Nguyen Sy An, Dao Viet Hang, Dao Van Long, Tran Quang Trung, Nguyen Thi Thuy, and Dinh Viet Sang. Neounet: Towards accurate colon polyp segmentation and neoplasm detection. In *International Symposium on Visual Computing*, pages 15–28. Springer, 2021.

[Niu *et al.*, 2023] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*, 2023.

[Orlando *et al.*, 2020] José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel Van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis*, 59:101570, 2020.

[Sankaranarayanan *et al.*, 2018] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3752–3761, 2018.

[Silva *et al.*, 2014] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9(2):283–293, 2014.

[Sivaswamy *et al.*, 2014] Jayanthi Sivaswamy, SR Krishnadas, Gopal Datt Joshi, Madhulika Jain, and A Ujjwaft Syed Tabish. Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation. In *2014 IEEE 11th international symposium on biomedical imaging (ISBI)*, pages 53–56. IEEE, 2014.

[Tang *et al.*, 2024] Song Tang, Wenxin Su, Mao Ye, and Xiatian Zhu. Source-free domain adaptation with frozen multimodal foundation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23711–23720, 2024.

[Wang *et al.*, 2020] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

[Wang *et al.*, 2022] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7201–7211, 2022.

[Wang *et al.*, 2023a] Jinke Wang, Xiang Li, and Yuanzhi Cheng. Towards an extended efficientnet-based u-net framework for joint optic disc and cup segmentation in the fundus image. *Biomedical Signal Processing and Control*, 85:104906, 2023.

[Wang *et al.*, 2023b] Yan Wang, Jian Cheng, Yixin Chen, Shuai Shao, Lanyun Zhu, Zhenzhou Wu, Tao Liu, and Haogang Zhu. Fvp: Fourier visual prompting for source-free unsupervised domain adaptation of medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(12):3738–3751, 2023.

[Xie *et al.*, 2021] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[Yang *et al.*, 2024] Danni Yang, Jiayi Ji, Yiwei Ma, Tianyu Guo, Haowei Wang, Xiaoshuai Sun, and Rongrong Ji. Sam as the guide: Mastering pseudo-label refinement in semi-supervised referring expression segmentation. *arXiv preprint arXiv:2406.01451*, 2024.

[Zhang *et al.*, 2010] Zhuo Zhang, Feng Shou Yin, Jiang Liu, Wing Kee Wong, Ngan Meng Tan, Beng Hai Lee, Jun Cheng, and Tien Yin Wong. Origa-light: An online retinal fundus image database for glaucoma analysis and research. In *2010 Annual international conference of the IEEE engineering in medicine and biology*, pages 3065–3068. IEEE, 2010.

[Zhang *et al.*, 2023] Jian Zhang, Lei Qi, Yinghuan Shi, and Yang Gao. Domainadaptor: A novel approach to test-time adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18971–18981, 2023.