SPDA: Efficient Online Test-Time Adaptation for Promptable Medical Segmentation

Huixuan Xu[1,2], Hu Han[1,2], Shiguang Shan[1,2], and Xilin Chen[1,2]

[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
{xuhuixuan24z,hanhu,sgshan,xlchen}@ict.ac.cn

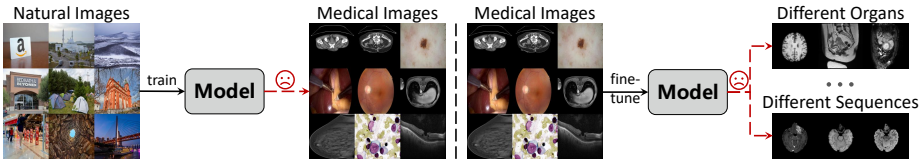
Abstract. Promptable foundation segmentation models exhibit impressive capabilities across diverse tasks, yet their performance deteriorates substantially on medical images due to domain and sub-domain shifts. Online test-time adaptation (OTTA) helps alleviate these shifts, but existing methods remain limited in effectiveness and efficiency. To this end, we introduce Semantic-Preserving Dual-Perturbation Adaptation (SPDA), a resource-efficient OTTA framework tailored for promptable medical segmentation. SPDA constructs diverse and robust learning signals by enforcing feature and prediction consistency across two semantic-preserving perturbations: (i) a domain-adaptive low-frequency radial-spectrum perturbation that simulates task-irrelevant intra-sub-domain variations without out-of-domain distortions, and (ii) a prompt undersampling perturbation that reduces prompt density while preserving target semantics. These consistency constraints drive the optimization of a lightweight post-encoder adapter, facilitating efficient adaptation with minimal computational overhead. Extensive experiments on varied 2D/3D medical data demonstrate that SPDA consistently outperforms state-of-the-art OTTA baselines under box and point prompts, while using substantially fewer resources. Ablation studies validate each component and indicate the suitability for real-time clinical deployment.

Keywords: Medical Segmentation · Promptable Foundation Model · Online Test-Time Adaptation

1 Introduction

Medical segmentation underpins essential clinical workflows, including diagnosis, treatment planning, and surgical navigation [2, 37, 48]. Promptable foundation models like SAM [23] and SAM2 [39] have demonstrated remarkable capabilities in segmenting natural images and videos using interactive prompts. Domain-adapted variants, such as MedSAM [28], MedSAM2 [30], and Medical-SAM2 [60], mitigate the natural-to-medical domain shift (Fig. 1a) via fine-tuning on large-scale medical data. Nevertheless, their performance degrades on unseen medical

This work was supported in part by the National Natural Science Foundation of China under Grants 62576358 and 62176249.



(a) Domain shift **(b)** Sub-domain shift

Fig. 1: Challenges in transferring general promptable segmentation models to medical image segmentation. (a) Models trained on natural images degrade on medical images due to substantial domain shift between natural and medical domains. (b) Even after fine-tuning on large-scale medical datasets, models are still sensitive to sub-domain shifts arising from data heterogeneity (*e.g.*, different organs, different MR sequences).

data due to sub-domain shifts (Fig. 1b) stemming from data heterogeneity across imaging modalities, target organs, acquisition protocols, and clinical centers.

Test-time adaptation (TTA) [7, 8, 12, 38, 51–53], which refines models on unlabeled test data in a privacy-preserving manner, offers a solution to mitigate domain and sub-domain shifts without source data access or extensive retraining. Specifically, online test-time adaptation (OTTA) [7, 8, 53] is suited for adapting promptable foundation segmentation models to clinical settings where data arrives sequentially. OTTA adapts the model to each incoming data batch, accumulating domain-specific knowledge while capturing instance-specific features [24].

Despite the rapid progress, existing OTTA methods face significant challenges in real-world clinical deployment. (i) Acquiring sufficient and reliable learning signals at test time remains challenging. To address this, alignment-based approaches [7] match source and target feature statistics but require source statistics access, which raises privacy concerns [25] and prevents direct learning from valuable target domains [35]. Consistency-based methods instead enforce agreement across semantic-preserving perturbations, which encourages the model to learn invariance to task-irrelevant information and adapt to the current sub-domain. Input-perturbation methods [8] use data augmentations to promote invariance. However, intra-sub-domain variations in medical data predominantly manifest within low-frequency components, whereas conventional augmentations mainly modify high-frequency details (Fig. 3). This mismatch can introduce out-of-domain distortions, violating semantic invariance and causing misaligned gradients [45, 55]. Inference-perturbation methods, such as student-teacher frameworks [53], generate weak signals due to limited perturbation diversity [1, 27] and amplify confirmation bias [43, 47], risking model collapse. (ii) Efficiency is another major concern in resource-constrained settings such as hospital edge devices. Many methods are impractical due to the substantial computational and memory overhead from full-model optimization and backpropagation costs [8].

To address the aforementioned challenges, we propose Semantic-Preserving Dual-Perturbation Adaptation (SPDA), a novel OTTA method for efficiently adapting promptable foundation models to medical segmentation. First, we construct diverse and robust consistency-based learning signals via two semantic-preserving perturbations. Driven by the observation that medical intra-sub-domain variations predominantly manifest in low-frequency spectra, we design a

domain-adaptive low-frequency radial-spectrum perturbation to simulate task-irrelevant variations without out-of-domain distortions. Additionally, we introduce a prompt under-sampling perturbation to yield diverse yet semantically consistent views during inference. These dual perturbations provide strong supervision via feature and prediction consistency losses. Second, we introduce a lightweight learnable adapter after the image encoder, keeping all parameters of the promptable foundation segmentation model frozen. This design minimizes trainable parameters and computational cost by avoiding backpropagation through the heavy encoder, enhancing the efficiency of the OTTA process.

To summarize, our main contributions are as follows: (1) We develop a highly efficient online test-time adaptation framework that effectively enhances the transferability of promptable foundation models across diverse medical segmentation tasks. (2) We introduce novel techniques to overcome key challenges of online test-time adaptation, specifically a domain-adaptive low-frequency radial-spectrum perturbation and a prompt under-sampling perturbation to construct reliable learning signals, alongside a lightweight post-encoder adapter to minimize computational overhead. (3) We extensively evaluate our approach on a wide range of 2D and 3D medical data covering multiple imaging modalities, demonstrating consistent state-of-the-art performance under the stringent resource constraints of real-world clinical environments.

2 Related Work

2.1 Promptable foundation models for medical segmentation

Promptable foundation segmentation models, exemplified by SAM [23], have significantly advanced segmentation in 2D natural images, demonstrating strong zero-shot capabilities. To address the suboptimal performance of SAM in medical image analysis [20, 32], researchers have fine-tuned it for 2D medical images, yielding domain-adapted models such as MedSAM [28] and SAM-Med2D [9]. A primary challenge, however, is that most real-world clinical data are inherently volumetric. Efforts to extend promptable segmentation to 3D include training full 3D encoders from scratch (*e.g.*, SAM-Med3D [46]) or integrating 3D adapters into the original SAM architecture (e.g., Med-SA [54], MA-SAM [3]). Other works also design promptable 3D medical models [13, 14, 22]. Despite these advancements, a unified model capable of seamlessly processing 2D images, 3D volumes, and medical videos remained absent. The introduction of SAM2 [39], a state-of-the-art promptable video segmentation model, presented a promising direction. This led to models like Medical SAM2 [60] and MedSAM2 [30], which were fine-tuned on large-scale medical image (2D and 3D) and video datasets. Nevertheless, these methods incur substantial retraining costs, and their performance can often degrade in real-world clinical settings due to severe data heterogeneity [53]. These gaps motivate the development of efficient online test-time adaptation (OTTA) for promptable medical foundation models, which aims to preserve their general object segmentation capabilities while improving robustness to domain and sub-domain shifts under resource-constrained scenarios.

2.2 Test-time adaptation

Test-time adaptation (TTA) refers to refining pre-trained models at test time using unlabeled target-domain data to mitigate performance degradation due to domain and sub-domain shifts, without accessing source-domain data [24]. Early research on TTA focused on image classification [4,5,25,34,36,41,59] and video classification [19,26,57,58], primarily evaluating methods on corrupted benchmarks [15,16,40,56] which are designed to simulate domain shifts. More recently, TTA has been extended to medical imaging, addressing more realistic and clinically relevant domain and sub-domain shifts across scanners, institutions, protocols, and modalities. Beyond medical classification-oriented TTA [31,49], medical segmentation-oriented TTA has also attracted increasing attention. These methods can be grouped by their adaptation strategy: (i) test-time domain adaptation (TTDA) performs a one-shot adaptation on the entire unlabeled target dataset prior to inference (*e.g.*, UPL-SFDA [52], DeTTA [51]), capturing dataset-level knowledge; (ii) test-time batch adaptation (TTBA) adapts on a per-batch basis and immediately infers on that batch (*e.g.*, InTEnt [12]), improving responsiveness. However, TTDA cannot handle sequential data arrival and may overlook instance-specific features, while TTBA discards accumulated domain knowledge across batches. Online test-time adaptation (OTTA) [50] addresses these limitations by continually updating parameters across the input stream by aligning the source and target feature statistics (*e.g.*, VPTTA [7]) or enforcing consistency across outputs from perturbations (*e.g.*, GraTa [8], SAM-TTA [53]). Despite recent advances, several challenges persist in clinically realistic settings, especially for promptable foundation segmentation models (*e.g.*, SAM-family variants and their medical derivatives): (i) alignment-based methods access the source-domain statistics, raising privacy concerns and failing to learn target-specific knowledge; (ii) existing consistency-based methods often rely on non-semantic-preserving perturbations that can lead to misleading supervision; (iii) many methods fail to meet the requirement for resource efficiency in clinical scenarios. Motivated by these gaps, we study efficient and consistency-based OTTA for promptable medical segmentation. In contrast to existing OTTA methods, our approach employs semantic-preserving dual perturbations and a lightweight adapter to obtain robust and misinformation-resistant consistency-based learning signals while substantially reducing computational and memory overhead.

3 Method

We study efficient online test-time adaptation (OTTA) for promptable medical segmentation. Given a visual input $\mathbf{x} \in \mathbb{R}^{D\times C\times H\times W}$ ($D = 1$ for 2D, and $D > 1$ for 3D/video) with prompts $\mathbf{p}$ (*e.g.*, points or boxes), a promptable model f_θ predicts a binary mask $\mathbf{y} \in \{0,1\}^{D\times H\times W}$. The unlabeled target-domain data $\mathcal{D}_\mathcal{T} = \{(\mathbf{x}^{(t)}, \mathbf{p}^{(t)})\}_{t=1}^{T}$ arrives sequentially in mini-batches $\{\mathcal{B}_i\}_{i=1}^{I} (\mathcal{B}_i \subset \mathcal{D}_\mathcal{T})$. For $\mathcal{B}_i$, we adapt and infer online:

$$\mathcal{Y}_i = f_{\theta^{(i)}}(\mathcal{B}_i), \quad \theta^{(i)} = \mathcal{A}\Big(\theta^{(i-1)}; \mathcal{B}_i\Big), \tag{1}$$

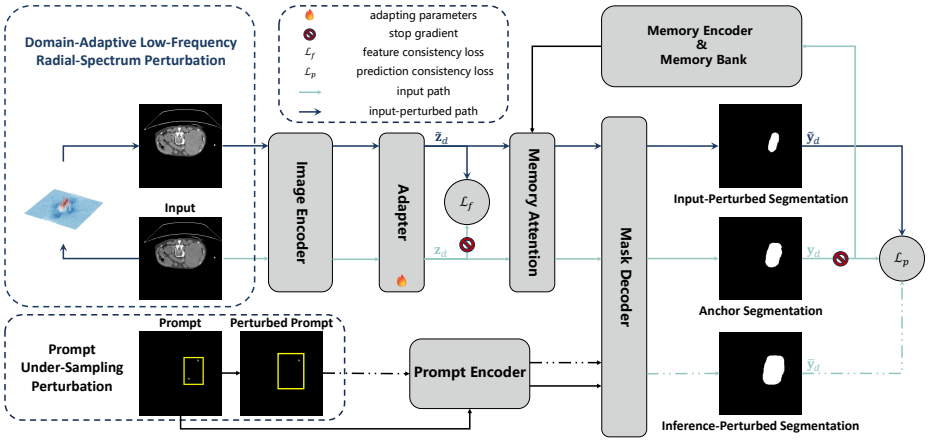


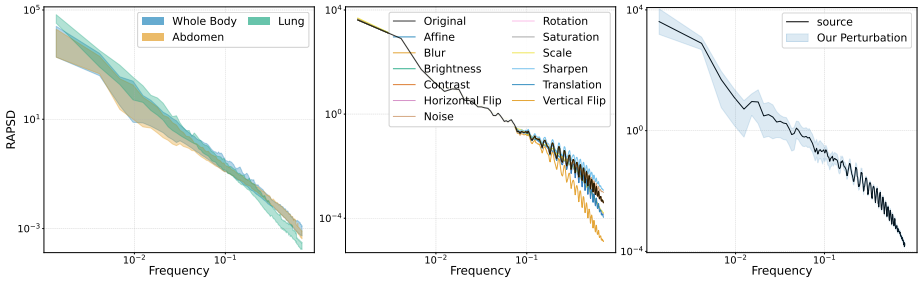
Fig. 2: Overview of our SPDA method, illustrated on the SAM2 architecture. Our method optimizes a lightweight, post-encoder adapter using a dual-consistency objective derived from two perturbations: (1) a Domain-Adaptive Low-frequency Radial-Spectrum Perturbation on the input, and (2) a Prompt Under-Sampling Perturbation applied during promptable inference. Specifically, we enforce feature consistency ($\mathcal{L}_f$) between the features $\mathbf{z}_d$ and $\widetilde{\mathbf{z}}_d$ derived from the original and perturbed inputs, alongside prediction consistency ($\mathcal{L}_p$) that aligns the anchor prediction ($\mathbf{y}_d$) with predictions from the input-perturbed ($\widetilde{\mathbf{y}}_d$) and inference-perturbed ($\bar{\mathbf{y}}_d$) views. The combined loss only updates the adapter.

where $\mathcal{A}$ is an adaptation operator that accesses neither labels nor source data. The objective is to minimize the expected segmentation error over the sequence:

$$\min_{\mathcal{A}} \frac{1}{\sum_{i=1}^{I}|\mathcal{B}_i|} \sum_{i=1}^{I} \sum_{(\mathbf{x},\mathbf{p})\in\mathcal{B}_i} \ell\left(f_{\theta^{(i)}}(\mathbf{x},\mathbf{p}),\ \widehat{\mathbf{y}}\right), \tag{2}$$

where $\widehat{\mathbf{y}}$ is the latent ground-truth segmentation mask of the given input $(\mathbf{x},\mathbf{p})$ and ℓ represents a segmentation discrepancy.

To achieve this goal, our proposed Semantic-Preserving Dual-Perturbation Adaptation (SPDA) concentrates on two key challenges: generating robust learning signals and ensuring resource efficiency. To construct robust learning signals, we devise two semantic-preserving perturbations: a Domain-Adaptive Low-Frequency Radial-Spectrum Perturbation (detailed in Sec. 3.1) and a Prompt Under-Sampling Perturbation (detailed in Sec. 3.2). For resource efficiency, we introduce a lightweight adapter positioned after the image encoder, which is the only module to be updated during adaptation (detailed in Sec. 3.3). These components are integrated via a dual-consistency objective, enabling stable, source-free OTTA with minimal overhead (detailed in Sec. 3.4). An overview of our SPDA method is illustrated in Fig. 2.



(a) RAPSD on different sub-tasks. **(b)** RAPSD on augmentations. **(c)** RAPSD on our perturbation.

Fig. 3: Analysis of Radially Averaged Power Spectral Density (RAPSD) motivating our domain-adaptive perturbation. (a) Min-max bands of RAPSD across different CT sub-tasks. Fluctuations are concentrated in the low-frequency region, and the magnitude of variation differs between sub-tasks. (b) RAPSD plots for a CT Abdomen sample under various traditional data augmentations. These augmentations primarily affect high-frequency components. (c) RAPSD band for the same CT Abdomen sample under our domain-adaptive low-frequency radial-spectrum perturbation. Our perturbation effectively introduces diverse variations specifically in the low-frequency region.

3.1 Domain-adaptive low-frequency radial-spectrum perturbation

To generate diverse yet semantic-preserving consistency-based learning signals, we propose a domain-adaptive low-frequency radial-spectrum perturbation on the visual input. Our strategy is motivated by radially averaged power spectral density (RAPSD) analysis [44] (see Fig. 3): medical intra-sub-domain variations predominantly manifest in the low-frequency spectrum (Fig. 3a), whereas traditional augmentations primarily perturb high-frequency details [35] (Fig. 3b). Additionally, minor high-frequency differences can induce substantial semantic shifts in medical segmentation. Consequently, this mismatch of traditional augmentations can introduce out-of-domain distortions, violating semantic invariance and leading to misaligned gradients [45, 55]. To generate diverse and sub-domain-aligned perturbations while preserving semantics, our approach exclusively perturbs the low-frequency components, dynamically adapting to the characteristics of the target sub-domain. By enforcing consistency under our semantic-preserving domain-adaptive low-frequency radial-spectrum perturbation, models learn invariance to task-irrelevant intra-sub-domain variations while enhancing their sensitivity to task-discriminative features, thereby effectively mitigating the performance degradation caused by inter-sub-domain shifts.

Formally, for a given input $\mathbf{x}^{(t)}$, we compute its 2D Fourier transform $\mathbf{X}^{(t)}_{d,c} \in \mathbb{C}^{H\times W}$ and associated power $\mathbf{P}^{(t)}_{d,c}$ for each slice-channel pair (d, c). The frequency plane is then partitioned into K radial bands $\{\Omega_b\}_{b=1}^{K}$:

$$\Omega_b = \left\{ (i,j) \middle| \left(\frac{b-1}{K}\right)^{\gamma} \le \rho(i,j) < \left(\frac{b}{K}\right)^{\gamma} \right\}, \tag{3}$$

where $\rho(i, j)$ is the normalized distance from the center, K is the number of bands, and hyperparameter γ controls band resolution.

We calculate the average log-radial power $R^{(t)}_{d,c,b} = \frac{1}{|\Omega_b|}\sum_{(i,j)\in\Omega_b} \log P^{(t)}_{d,c}(i,j)$ for each band and intra-sample mean $\mu^{(t)}_{\text{intra},b}$ and variance $v^{(t)}_{\text{intra},b}$. To capture sub-domain characteristics, we update inter-sample statistics ($\mu^{(t)}_{\text{inter},b}$, $v^{(t)}_{\text{inter},b}$) using exponentially weighted moving average (EWMA) and variance (EWVar) with momentum hyperparameter η:

$$\delta^{(t)}_b = \mu^{(t)}_{\text{intra},b} - \mu^{(t-1)}_{\text{inter},b}, \tag{4}$$

$$\mu^{(t)}_{\text{inter},b} = \mu^{(t-1)}_{\text{inter},b} + \eta\delta^{(t)}_b, \tag{5}$$

$$v^{(t)}_{\text{inter},b} = (1-\eta)v^{(t-1)}_{\text{inter},b} + \eta(\delta^{(t)}_b)^2. \tag{6}$$

We compute the combined variance $v^{(t)}_b = \lambda_{\text{intra}} v^{(t)}_{\text{intra},b} + \lambda_{\text{inter}} v^{(t)}_{\text{inter},b}$, then sample a perturbation factor $\alpha^{(t)}_b$ from an adaptive log-normal distribution:

$$\epsilon^{(t)}_b \sim \mathcal{N}\left(0, v^{(t)}_b\right), \tag{7}$$

$$\alpha^{(t)}_b = \exp\left(\epsilon^{(t)}_b - \frac{1}{2}v^{(t)}_b\right). \tag{8}$$

Critically, this perturbation is applied only to the low-frequency bands ($b \le K_{\text{low}}$, where K_{low} is a hyperparameter defining the low-frequency threshold) by scaling the real part of the Fourier coefficients $\mathbf{A}^{(t)}_{d,c} = \Re(\mathbf{X}^{(t)}_{d,c})$:

$$\mathbf{A}^{(t)}_{d,c}(i,j) \leftarrow \alpha^{(t)}_b \mathbf{A}^{(t)}_{d,c}(i,j) \quad \forall (i,j) \in \Omega_b. \tag{9}$$

The perturbed input $\tilde{\mathbf{x}}^{(t)}$ is reconstructed by applying the inverse Fourier transform shift, followed by the 2D inverse Fourier transform ($\mathcal{F}_2^{-1}$).

The variations introduced by our domain-adaptive perturbation (Fig. 3c) closely align with the authentic data diversity observed in real-world clinical settings (the Abdomen sub-task in Fig. 3a). This alignment ensures the generation of semantically consistent and domain-relevant views, thereby providing highly reliable consistency-based learning signals for OTTA.

3.2 Prompt under-sampling perturbation

We introduce a lightweight perturbation to provide diverse yet semantically consistent views during inference. This approach perturbs the prompting process, preserving semantic invariance by maintaining the original target semantics, as the under-sampled prompts continue to reference the same anatomical or pathological regions with reduced specificity.

The prompt under-sampling procedure varies depending on the prompt type to yield the under-sampled prompt $\tilde{\mathbf{p}}$ for the target object:

- Box Prompts: Given a box prompt $\mathbf{p}$, we generate $\tilde{\mathbf{p}}$ by isotropically scaling its area by a factor $\zeta > 1$, which creates a less specific prompt that still fully encompasses the same target object.
- Point Prompts: Given a set of $N \ge 2$ point prompts $\mathbf{p}$, the perturbed prompt subset $\tilde{\mathbf{p}} \subsetneq \mathbf{p}$ is generated by randomly sampling $N' < N$ points.

3.3 Resource-efficient adapter

Resource efficiency is critical for OTTA in clinical settings. The primary bottleneck in backpropagation-based OTTA is the backward pass of the large image encoder, which incurs substantial computational latency for gradient computation and memory overhead for storing intermediate activations [6]. While Parameter-Efficient Fine-Tuning (PEFT) strategies [17,18] reduce trainable parameters and gradient computation, modules inserted within the encoder still require a full backward pass, failing to address this bottleneck.

To overcome this, we introduce a lightweight adapter positioned solely after the image encoder. This adapter, implemented as a simple bottleneck convolution for each output feature map (supporting hierarchical encoders), is the only component updated. This post-encoder design completely bypasses backpropagation through the heavy image encoder, which yields significant efficiency gains: (i) it dramatically reduces computational latency by computing gradients only for the lightweight adapter, and (ii) it substantially saves memory by obviating the need to store the encoder's intermediate activations for the backward pass.

3.4 Overall Objective

During adaptation for an incoming batch, we update only the lightweight adapter (described in Sec. 3.3) using a dual-consistency objective. This objective provides robust and misinformation-free consistency-based learning signals by leveraging our semantic-preserving perturbations (described in Sec. 3.1 and Sec. 3.2).

Let $\mathbf{z}_d$ and $\widetilde{\mathbf{z}}_d$ be the set of features for the d-th slice extracted by the encoder and processed by our adapter from the original input $\mathbf{x}$ and the perturbed input $\widetilde{\mathbf{x}}$, respectively. We formally define the anchor prediction $\mathbf{y}_d$, the input-perturbed prediction $\widetilde{\mathbf{y}}_d$ (Sec. 3.1), and the inference-perturbed prediction $\bar{\mathbf{y}}_d$ (Sec. 3.2) as the outputs generated from the feature-prompt pairs $(\mathbf{z}_d, \mathbf{p}_d)$, $(\widetilde{\mathbf{z}}_d, \mathbf{p}_d)$, and $(\mathbf{z}_d, \widetilde{\mathbf{p}}_d)$, respectively. Our optimization objective consists of a feature consistency loss and a prediction consistency loss.

Feature consistency loss. We define the feature consistency loss to encourage the robustness of adapter-produced representations against the domain-adaptive input perturbation. Since our adapter supports hierarchical encoders, we enforce feature consistency by minimizing the cosine dissimilarity at all feature levels:

$$\mathcal{L}_f = \frac{1}{L} \sum_{l=1}^{L} \left(1 - \cos\left(\mathbf{z}_d^{(l)}, \widetilde{\mathbf{z}}_d^{(l)}\right)\right), \tag{10}$$

where $\mathbf{z}_d^{(l)}$ and $\widetilde{\mathbf{z}}_d^{(l)}$ are the features at the l-th level.

Prediction consistency loss. We impose a prediction consistency loss that aligns the anchor prediction with those derived from each perturbation, using the Dice loss $\mathcal{L}_{\text{dice}}$ [33] to encourage agreement:

$$\mathcal{L}_p = \frac{1}{2} \left(\mathcal{L}_{\text{dice}}\left(\mathbf{y}_d, \bar{\mathbf{y}}_d\right) + \mathcal{L}_{\text{dice}}\left(\mathbf{y}_d, \widetilde{\mathbf{y}}_d\right)\right), \tag{11}$$

Final loss. We form the final loss as a weighted sum of these terms:

$$\mathcal{L} = \lambda_f \mathcal{L}_f + \lambda_p \mathcal{L}_p \tag{12}$$

where λ_f and λ_p are hyperparameters that balance each term. By optimizing this combined objective $\mathcal{L}$ via backpropagation, gradients are computed and applied only to the adapter parameters, thereby achieving efficient and effective OTTA.

4 Experiments

4.1 Experimental setup

Datasets. We conduct extensive experiments on a large-scale dataset collected by the CVPR 2024 Segment Anything in Medical Images Challenge, covering 11 imaging modalities (CT, Dermoscopy, Endoscopy, Fundus, Mammography, Microscopy, MR, OCT, PET, Ultrasound, and XRay) [29]. Each modality comprises diverse sub-tasks, such as segmenting different anatomical regions or pathologies, which inevitably introduce significant sub-domain shifts. The dataset is curated and pre-processed following established methodologies [21, 28].

Evaluation metric. We report the Dice Similarity Coefficient (DSC) as the primary evaluation metric [11, 42]. The DSC quantitatively measures the spatial overlap between the model's prediction and the ground truth mask, providing a robust measure of overall segmentation accuracy.

Baseline. We evaluate our proposed SPDA approach against four recent state-of-the-art test-time adaptation methods for medical segmentation (including both OTTA and TTBA methods) by applying them to adapt the promptable foundation segmentation models:

- **InTEnt** [12]: InTEnt is a TTBA method that generates an ensemble of models by interpolating the normalization layer statistics. The final segmentation prediction is then constructed by weighting the ensemble outputs based on their respective prediction entropy.
- **VPTTA** [7]: VPTTA is an OTTA method that optimizes a trainable low-frequency prompt. The learning objective is to align the feature map statistics of the target domain with those of the source domain.
- **GraTa** [8]: GraTa is an OTTA method that optimizes model parameters by implicitly aligning the gradients from an entropy minimization objective and a consistency-based objective.
- **SAM-TTA** [53]: SAM-TTA is an OTTA method that employs a student-teacher framework. The method mainly enforces dual-scale prediction consistency to optimize the model and a learnable input transformation.

We make some necessary modifications to the evaluated baseline methods to ensure compatibility with promptable segmentation models. All further implementation details are provided in the supplementary materials.

Implementation details. For all experiments, we set the batch size to 1 and adapt the model for a single iteration per batch, which aligns with the protocols of the evaluated OTTA baselines [7,8,53]. Following the MedSAM2 protocol, the initial prompts inherently contain realistic noise (*e.g.*, up to a 10% box offset). To mitigate catastrophic forgetting during continuous adaptation, we stochastically restore 1% of the adapter parameters to their initial pre-adaptation weights during the adaptation of each sample. We employ the AdamW optimizer for our SPDA and the recommended optimizer for each baseline method. The optimal learning rate for each method (if applicable) on the whole dataset is determined via a grid search on its sampled subset (20 samples per sub-task). For SPDA, we set the remaining hyperparameters as follows: the number of radial frequency bands $K = 24$, the number of perturbed low-frequency bands $K_{\text{low}} = 8$, the band resolution controller $\gamma = 1.7$, the EWMA momentum $\eta = 0.1$, the variance aggregation weights $\lambda_{\text{intra}} = 0.3$ and $\lambda_{\text{inter}} = 0.7$, the box scaling factor $\zeta = 1.005$, the number of under-sampled points $N' = 1$, and the consistency loss weights $\lambda_f = 0.1$ and $\lambda_p = 1.0$. Notably, this unified configuration consistently yields competitive performance across all evaluated medical modalities. Sensitivity analysis provided in the supplementary materials demonstrates that our method exhibits adequate robustness to hyperparameter variations.

4.2 Experimental results

SPDA and the other TTA methods are evaluated on the SAM2 architecture, initialized with pre-trained parameters of either SAM2 or MedSAM2. We also report the baseline performance of the original pretrained model without any TTA method as a comparison reference (Source-Only).

Performance with general foundation model (SAM2). We first assess the model's ability to adapt from natural to medical domains using SAM2 pretrained weights. We present results for both box prompts and 3-click point prompts. For 3D samples, prompts are provided for the key slice (defined as the slice with the largest mask area) and up to four adjacent slices, with the remaining slices segmented using SAM2's object tracking capability. As shown in Tab. 1, SPDA consistently outperforms the source-only model and baselines across modalities. It achieves average DSC improvements of 7.07% and 9.96% for box and point prompts, respectively. This demonstrates SPDA's effectiveness in mitigating the natural-to-medical domain shift without source data access.

Performance with medical foundation model (MedSAM2). We adopt a more challenging and realistic segmentation setup where prompts are provided exclusively for the key slice in 3D samples. The results, presented in Tab. 2, confirm the superiority of our method. SPDA achieves state-of-the-art average performance, improving upon the strongest baseline (GraTa) by 1.79% for box prompts and 2.06% for point prompts. This highlights SPDA's robustness and effectiveness in refining models that are already fine-tuned for the medical domain but still struggle with sub-domain shifts from data heterogeneity.

Table 1: Quantitative comparison of our SPDA against SOTA methods using the general SAM2 foundation model. Evaluation metric is DSC (%). The best results are highlighted in **bold** and the second-best results are underlined.

Prompt	Method	DSC (%) ↑											
		CT	Derm.	Endo.	Fundus	Mammo.	Micro.	MR	OCT	PET	US	XRay	Avg. DSC
box	Source-Only	47.04	86.39	84.59	82.02	66.36	74.56	56.82	62.86	17.36	86.38	73.87	67.11
	InTEnt [12]	49.72	88.33	85.96	82.59	66.28	74.42	57.53	59.80	19.25	84.62	73.92	67.49
	VPTTA [7]	52.36	89.63	87.98	82.73	66.84	78.37	69.75	80.11	22.82	88.99	77.81	72.49
	GraTa [8]	49.90	89.74	87.93	82.49	67.62	77.92	69.58	81.92	22.91	90.11	80.49	72.78
	SAM-TTA [53]	50.59	89.75	87.94	82.62	66.98	80.14	67.05	80.56	17.01	90.16	80.93	72.16
	SPDA (ours)	**54.39**	**90.84**	**88.72**	**83.33**	**69.19**	**80.25**	**70.27**	**82.18**	**24.65**	**90.77**	**81.41**	**74.18**
points	Source-Only	36.65	63.42	65.78	55.86	31.76	67.55	43.90	68.14	10.22	65.35	57.68	51.48
	InTEnt [12]	34.26	63.33	66.33	54.55	29.80	66.48	44.57	71.37	9.95	66.78	58.08	51.41
	VPTTA [7]	37.56	69.33	75.63	60.95	36.14	67.83	45.04	71.75	17.79	70.42	59.25	55.61
	GraTa [8]	36.47	63.83	73.04	60.19	34.18	68.16	46.20	68.70	11.68	60.42	58.76	52.88
	SAM-TTA [53]	35.74	68.90	81.31	58.59	32.16	67.74	48.85	71.43	15.64	73.95	60.93	55.93
	SPDA (ours)	**47.13**	**78.69**	**82.94**	**62.39**	**38.91**	**70.53**	**55.39**	**73.77**	**23.18**	**75.05**	**67.85**	**61.44**

Table 2: Quantitative comparison of our SPDA against SOTA methods using the domain-adapted MedSAM2 foundation model. Evaluation metric is DSC (%). The best results are highlighted in **bold** and the second-best results are underlined.

Prompt	Method	DSC (%) ↑											
		CT	Derm.	Endo.	Fundus	Mammo.	Micro.	MR	OCT	PET	US	XRay	Avg. DSC
box	Source-Only	70.89	94.62	92.24	94.09	80.56	79.72	76.21	82.00	68.71	89.00	88.87	83.36
	InTEnt [12]	68.73	94.55	92.14	93.92	81.93	78.48	76.33	79.83	69.19	89.92	88.73	83.07
	VPTTA [7]	71.37	94.45	93.62	95.86	81.72	86.31	78.42	86.58	75.61	91.33	89.85	85.92
	GraTa [8]	71.99	94.57	93.06	94.83	81.18	87.72	79.89	87.17	76.97	91.63	89.89	86.26
	SAM-TTA [53]	69.80	93.42	93.02	94.68	82.14	87.63	78.02	85.59	75.45	93.75	89.14	85.69
	SPDA (ours)	**76.58**	**94.81**	**93.87**	**95.89**	**83.08**	**90.82**	**81.94**	**88.23**	**79.14**	**93.94**	**90.22**	**88.05**
points	Source-Only	68.79	86.44	88.46	80.84	63.67	81.80	72.89	74.36	73.57	83.69	83.74	78.02
	InTEnt [12]	67.86	87.12	86.21	81.27	62.63	81.29	71.06	76.73	72.39	85.74	82.85	77.74
	VPTTA [7]	67.47	90.21	86.88	**82.77**	65.76	81.73	71.11	76.65	72.11	85.45	85.74	78.72
	GraTa [8]	68.19	87.76	87.35	82.57	66.59	81.22	75.76	76.87	73.34	86.39	83.52	79.05
	SAM-TTA [53]	68.61	87.63	88.07	81.08	65.43	81.86	76.33	77.32	73.24	86.44	83.14	79.01
	SPDA (ours)	**70.45**	**91.03**	**89.05**	82.17	**66.85**	**82.08**	**77.40**	**79.16**	**74.58**	**87.51**	**91.91**	**81.11**

Table 3: Comparison between our method and baseline TTA methods in terms of average computational cost per sample (TFLOPs) and the number of parameters (M).

Method	Comput. Cost (TFLOPs)↓	#params (M)↓
InTEnt [12]	204.56	**38.97**
VPTTA [7]	233.2	38.98
GraTa [8]	497.18	38.99
SAM-TTA [53]	63.65	78.04
SPDA (ours)	**39.88**	39.16

Computational cost and parameter efficiency. We quantify the per-sample computational cost (TFLOPs) and parameter count (including model parameters, memory bank variables, and statistics) for each method, measured under the MedSAM2 setup. As detailed in Tab. 3, SPDA maintains a competitive parameter count (39.16M) that is comparable to most baselines and significantly more parameter-efficient than SAM-TTA (78.04M). Crucially, its computational cost (39.88 TFLOPs) is substantially lower than that of other baselines. This validates our design as a lightweight framework suitable for clinical deployment.

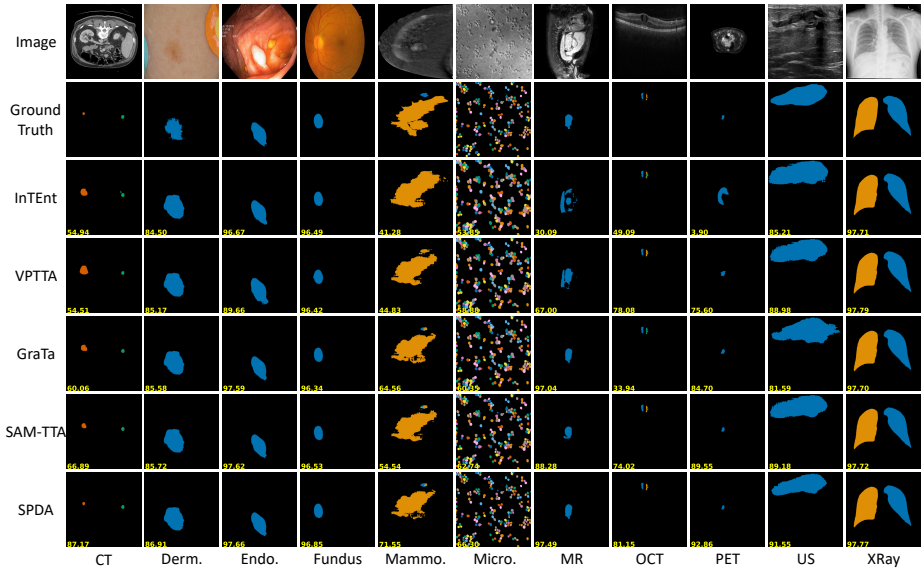


Fig. 4: Visualization of segmentation results when applying our method and the SOTA TTA methods to MedSAM2. Different colors denote distinct objects in a column and the number of each predicted mask indicates the DSC (%).

Table 4: Ablation study with different promptable foundation segmentation models.

Method	Avg. DSC (%)↑
SAM	65.48
SAM + SPDA	70.16
MedSAM	71.64
MedSAM + SPDA	72.82

Visualization and comparisons. We visualize the segmentation results of our method and the baselines when performing TTA on MedSAM2 in Fig. 4. Due to the capability of MedSAM2, most approaches produce reasonable results. However, our proposed SPDA consistently yields more precise and robust segmentation masks. Compared to other methods, the outputs of SPDA more closely align with the target structures, particularly in challenging cases with ambiguous edges or complex anatomies. These qualitative results confirm SPDA's superior capability in mitigating the subtle yet critical sub-domain shifts inherent in heterogeneous medical data, thereby further validating its effectiveness.

4.3 Ablation study

To validate the key components of SPDA, we conduct ablation studies on a subset comprising 20 randomly selected samples from each sub-task. Unless specified otherwise, we use MedSAM2 as the backbone and evaluate using box prompts. Hyperparameter selection analysis is provided in the supplementary materials.

Table 5: Comparison between our adapter and other optimization strategies.

Opt. strategy	Avg. DSC (%)↑	TFLOPs↓
LN Optimization	80.85	<u>77.61</u>
LoRA Optimization	**85.11**	78.90
Our Adapter	<u>85.04</u>	**39.88**

Table 6: Ablation studies for the key components of our SPDA.

Input Pert.		Prompt Pert.	Avg. DSC (%)↑
Trad.	Ours		
			80.07
	✓		83.91
		✓	81.45
✓		✓	83.78
	✓	✓	**85.04**

Effectiveness with different foundation segmentation models. To verify the generalizability of our method, we apply SPDA to other promptable foundation segmentation models, namely the original SAM [23] and MedSAM [28]. As shown in Tab. 4, SPDA consistently improves performance across different models, demonstrating its model-agnostic benefits.

Comparison with other optimization strategies. We compare our efficient adapter with alternative strategies, *i.e.*, layer normalization (LN) and LoRA optimization. As shown in Tab. 5, our adapter (85.04% DSC) performs comparably to the best alternative, LoRA (85.11% DSC). However, our computational cost (39.88 TFLOPs) is substantially lower, at roughly half that of LoRA (78.90 TFLOPs) and LN (77.61 TFLOPs). This result validates the superior efficiency of our post-encoder design, which avoids backpropagation through the heavy encoder while maintaining high adaptation accuracy.

Analysis of component contributions. We analyze the contributions of our domain-adaptive low-frequency radial-spectrum perturbation (Input Pert.) and prompt under-sampling perturbation (Prompt Pert.), while comparing against traditional augmentations (Trad.) [10]. In Tab. 6, the baseline without adaptation achieves 80.07% DSC. Introducing our Input Pert. alone boosts performance to 83.91%, whereas Prompt Pert. yields a smaller gain (81.45%). Notably, replacing our Input Pert. with traditional augmentations (83.78%) results in slightly lower performance, confirming that traditional methods fail to adequately target the relevant sub-domain variations. Finally, the full SPDA model, combining both our semantic-preserving input perturbation and the prompt perturbation, achieves the highest accuracy (85.04%), demonstrating that the two components are complementary and essential for optimal performance.

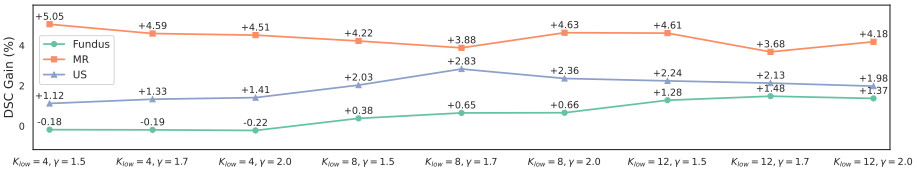


Fig. 5: Modality-specific sensitivity of the frequency hyperparameters K_{low} and γ.

Modality-specific sensitivity to hyperparameters. Although our unified global hyperparameters establish a robust baseline across the dataset, relying on a single configuration may not fully exploit the potential of modalities with distinct spectral properties (*e.g.*, Fundus). As illustrated in Fig. 5, we analyze the frequency hyperparameter sensitivity of individual modalities by conducting a small-scale grid search ($K_{\text{low}} \in \{4, 8, 12\}$, $\gamma \in \{1.5, 1.7, 2.0\}$) on three representative modalities (US, MR, and Fundus). For example, MR exhibits a strong preference for a narrower low-frequency band ($K_{\text{low}} = 4$, $\gamma = 1.5$), achieving a substantial +5.05% DSC gain. Conversely, Fundus images benefit from a broader perturbed spectrum ($K_{\text{low}} = 12$, $\gamma = 1.7$) to better accommodate their unique frequency characteristics, yielding an additional +1.48% DSC improvement. These findings indicate that while a global configuration suffices for general deployment, a lightweight modality-specific grid search serves as an effective strategy to address diverse spectral distributions and maximize adaptation potential for heterogeneous clinical data.

5 Discussion

In this paper, we presented Semantic-Preserving Dual-Perturbation Adaptation (SPDA), an efficient and effective online test-time adaptation (OTTA) method for promptable medical segmentation. It optimizes a lightweight post-encoder adapter by enforcing dual consistency across two semantic-preserving perturbations: a domain-adaptive low-frequency radial-spectrum perturbation targeting sub-domain shifts, and a prompt under-sampling perturbation providing diverse inference views. The resulting framework achieves strong performance with high resource efficiency, demonstrating its suitability for clinical deployment.

One limitation is that SPDA is designed for architectures with distinct image encoders, prompt encoders, and mask decoders. Its applicability to more integrated, end-to-end models remains unexplored. Future work could investigate extending the framework to be more architecture-agnostic, enhancing its versatility across different foundation model designs.

Acknowledgements

The AI-driven experiments, simulations and model training were supported in part by the robotic AI-Scientist platform of Chinese Academy of Sciences.

References

1. Baek, K., Choi, Y., Uh, Y., Yoo, J., Shim, H.: Rethinking the truly unsupervised image-to-image translation. In: ICCV. pp. 14154–14163 (2021)
2. Cao, K., Xia, Y., Yao, J., Han, X., Lambert, L., Zhang, T., Tang, W., Jin, G., Jiang, H., Fang, X., et al.: Large-scale pancreatic cancer detection via non-contrast CT and deep learning. Nat. Med. **29**(12), 3033–3043 (2023)
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. Med. Image Anal. **98**, 103310 (2024)
4. Chen, D., Wang, D., Darrell, T., Ebrahimi, S.: Contrastive test-time adaptation. In: CVPR. pp. 295–305 (2022)
5. Chen, G., Niu, S., Chen, D., Zhang, S., Li, C., Li, Y., Tan, M.: Cross-device collaborative test-time adaptation. NeurIPS **37**, 122917–122951 (2024)
6. Chen, T., Xu, B., Zhang, C., Guestrin, C.: Training deep nets with sublinear memory cost. arXiv preprint arXiv:1604.06174 (2016)
7. Chen, Z., Pan, Y., Ye, Y., Lu, M., Xia, Y.: Each test image deserves a specific prompt: Continual test-time adaptation for 2D medical image segmentation. In: CVPR. pp. 11184–11193 (2024)
8. Chen, Z., Ye, Y., Pan, Y., Xia, Y.: Gradient alignment improves test-time adaptation for medical image segmentation. In: AAAI. vol. 39, pp. 2429–2437 (2025)
9. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., Sun, H., He, J., Zhang, S., Zhu, M., Qiao, Y.: SAM-Med2D. arXiv:2308.16184 (2023)
10. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: RandAugment: Practical automated data augmentation with a reduced search space. NeurIPS **33**, 18613–18624 (2020)
11. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
12. Dong, H., Konz, N., Gu, H., Mazurowski, M.A.: Medical image segmentation with intent: Integrated entropy weighting for single image test-time adaptation. In: CVPRW. pp. 5046–5055 (2024)
13. Du, Y., Bai, F., Huang, T., Zhao, B.: SegVol: Universal and interactive volumetric medical image segmentation. NeurIPS **37**, 110746–110783 (2024)
14. He, Y., Guo, P., Tang, Y., Myronenko, A., Nath, V., Xu, Z., Yang, D., Zhao, C., Simon, B., Belue, M., et al.: VISTA3D: A unified segmentation foundation model for 3D medical imaging. In: CVPR. pp. 20863–20873 (2025)
15. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: ICCV. pp. 8340–8349 (2021)
16. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: ICLR (2019)
17. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for NLP. In: ICML (2019)
18. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022)
19. Huang, Y., Yang, X., Zhang, J., Xu, C.: Relative alignment network for source-free multimodal video domain adaptation. In: ACM MM. pp. 1652–1660 (2022)
20. 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? Med. Image Anal. **92**, 103061 (2024)

21. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
22. Isensee, F., Rokuss, M., Krämer, L., Dinkelacker, S., Ravindran, A., Stritzke, F., Hamm, B., Wald, T., Langenberg, M., Ulrich, C., et al.: nnInteractive: Redefining 3D promptable segmentation. arXiv preprint arXiv:2503.08373 (2025)
23. 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: ICCV. pp. 4015–4026 (2023)
24. Liang, J., He, R., Tan, T.: A comprehensive survey on test-time adaptation under distribution shifts. IJCV **133**(1), 31–64 (2025)
25. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: ICML. pp. 6028–6039 (2020)
26. Lin, W., Mirza, M.J., Kozinski, M., Possegger, H., Kuehne, H., Bischof, H.: Video test-time adaptation for action recognition. In: CVPR. pp. 22952–22961 (2023)
27. Luo, X., Chen, J., Song, T., Wang, G.: Semi-supervised medical image segmentation through dual-task consistency. In: AAAI. vol. 35, pp. 8801–8809 (2021)
28. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nat. Commun. **15**(1), 654 (2024)
29. Ma, J., Li, F., Kim, S., Asakereh, R., Le, B.H., Nguyen-Vu, D.K., Pfefferle, A., Wei, M., Gao, R., Lyu, D., et al.: Efficient MedSAMs: Segment Anything in medical images on laptop. arXiv preprint arXiv:2412.16085 (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. Ma, W., Chen, C., Zheng, S., Qin, J., Zhang, H., Dou, Q.: Test-time adaptation with calibration of medical image classification nets for label distribution shift. In: MICCAI. pp. 313–323. Springer (2022)
32. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment Anything Model for medical image analysis: an experimental study. Med. Image Anal. **89**, 102918 (2023)
33. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3DV. pp. 565–571. IEEE (2016)
34. Mirza, M.J., Micorek, J., Possegger, H., Bischof, H.: The norm must go on: Dynamic unsupervised domain adaptation by normalization. In: CVPR (2022)
35. Niu, S., Chen, G., Zhao, P., Wang, T., Wu, P., Shen, Z.: Self-bootstrapping for versatile test-time adaptation. In: ICML. vol. 267, pp. 46611–46628. PMLR (2025)
36. Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., Tan, M.: Towards stable test-time adaptation in dynamic wild world. In: ICLR (2023)
37. Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C.P., Heidenreich, P.A., Harrington, R.A., Liang, D.H., Ashley, E.A., et al.: Video-based AI for beat-to-beat assessment of cardiac function. Nature **580**(7802), 252–256 (2020)
38. Park, H., Hwang, J., Mun, S., Park, S., Ok, J.: MedBN: Robust test-time adaptation against malicious test samples. In: CVPR. pp. 5997–6007 (2024)
39. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: SAM 2: Segment Anything in images and videos. In: ICLR (2025)

40. Schiappa, M.C., Biyani, N., Kamtam, P., Vyas, S., Palangi, H., Vineet, V., Rawat, Y.S.: A large-scale robustness analysis of video action recognition models. In: CVPR. pp. 14698–14708 (2023)
41. Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., Hardt, M.: Test-time training with self-supervision for generalization under distribution shifts. In: ICML. pp. 9229–9248. PMLR (2020)
42. Taha, A.A., Hanbury, A.: Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC medical imaging **15**(1), 29 (2015)
43. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. NeurIPS **30** (2017)
44. van der Schaaf, A., van Hateren, J.H.: Modelling the power spectra of natural images: statistics and information. Vision Research **36**(17), 2759–2770 (1996)
45. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: ICLR (2021)
46. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., Fu, B., Zhang, S., He, J.: SAM-Med3D: A vision foundation model for general-purpose segmentation on volumetric medical images. IEEE Trans. Neural Netw. Learn. Syst. **36**(10), 17599–17612 (2025)
47. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: CVPR. pp. 7201–7211 (2022)
48. Wang, Y.R., Yang, K., Wen, Y., Wang, P., Hu, Y., Lai, Y., Wang, Y., Zhao, K., Tang, S., Zhang, A., et al.: Screening and diagnosis of cardiovascular disease using artificial intelligence-enabled cardiac magnetic resonance imaging. Nat. Med. **30**(5), 1471–1480 (2024)
49. Wang, Z., Ye, M., Zhu, X., Peng, L., Tian, L., Zhu, Y.: MetaTeacher: Coordinating multi-model domain adaptation for medical image classification. NeurIPS **35**, 20823–20837 (2022)
50. Wang, Z., Luo, Y., Zheng, L., Chen, Z., Wang, S., Huang, Z.: In search of lost online test-time adaptation: A survey. IJCV **133**(3), 1106–1139 (2025)
51. Wen, R., Yuan, H., Ni, D., Xiao, W., Wu, Y.: From denoising training to test-time adaptation: Enhancing domain generalization for medical image segmentation. In: WACV. pp. 464–474 (2024)
52. Wu, J., Wang, G., Gu, R., Lu, T., Chen, Y., Zhu, W., Vercauteren, T., Ourselin, S., Zhang, S.: UPL-SFDA: Uncertainty-aware pseudo label guided source-free domain adaptation for medical image segmentation. IEEE Trans. Med. Imaging **42**(12), 3932–3943 (2023)
53. Wu, J., Wu, Y., Xie, Y., Bai, W., Zhang, Y., Tang, F., Li, Y., George, Y., Razzak, I.: SAM-aware test-time adaptation for universal medical image segmentation. arXiv preprint arXiv:2506.05221 (2025)
54. Wu, J., Wang, Z., Hong, M., Ji, W., Fu, H., Xu, Y., Xu, M., Jin, Y.: Medical SAM adapter: Adapting Segment Anything Model for medical image segmentation. Med. Image Anal. **102**, 103547 (2025)
55. Yang, L., Qi, L., Feng, L., Zhang, W., Shi, Y.: Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In: CVPR. pp. 7236–7246 (2023)
56. Yi, C., Yang, S., Li, H., Tan, Y.P., Kot, A.C.: Benchmarking the robustness of spatial-temporal models against corruptions. In: NeurIPS (2021)
57. Yi, C., Yang, S., Wang, Y., Li, H., Tan, Y.P., Kot, A.C.: Temporal coherent test-time optimization for robust video classification. In: ICLR (2023)
58. Zeng, R., Deng, Q., Xu, H., Niu, S., Chen, J.: Exploring motion cues for video test-time adaptation. In: ACM MM. pp. 1840–1850 (2023)

59. Zhang, M., Levine, S., Finn, C.: MEMO: Test time robustness via adaptation and augmentation. NeurIPS **35**, 38629–38642 (2022)
60. Zhu, J., Hamdi, A., Qi, Y., Jin, Y., Wu, J.: Medical SAM 2: Segment medical images as video via Segment Anything Model 2. arXiv preprint arXiv:2408.00874 (2024)