Memory-Supported Synergistic Adaptation for Training-Free Test-Time Medical Image Segmentation

Lingrui Li[1], Nan Pu[2*], Dong Zhao[3*], Wenjing Li[2], Andrew P French[1], Zhun Zhong[2], and Xin Chen[1]

[1] School of Computer Science, University of Nottingham, UK
[2] School of Computer Science and Information Engineering, Hefei University of Technology, China
[3] Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore

Abstract. Test-time adaptation (TTA) aims to mitigate distribution shifts by adapting models with unlabeled target data at inference time. While TTA with vision-language models (VLMs) has shown promising results in classification, extending it to medical image segmentation remains challenging. In this setting, the adaptation gains from optimizing on VLM-generated predictions are often outweighed by the degradation to the VLM's strong pretrained features caused by noisy, update-driven learning, resulting in limited and unstable improvements. We therefore propose **Memory-Supported Synergistic Adaptation (MSSA)**, a novel **training-free** TTA framework for medical image segmentation. Without updating model parameters, MSSA dynamically selects reliable image–text predictions to construct an online memory, uses them as text-guided semantic priors, and couples them with cross-image structural alignment for robust adaptation. Specifically, MSSA consists of (i) a noise-aware memory construction module that filters and stabilizes cross-modal predictions, and (ii) a relevance-driven prototype alignment module that aligns the target sample with structurally consistent memory samples and their reliable predictions to improve adaptation. Extensive experiments on multiple medical segmentation benchmarks demonstrate that MSSA consistently improves VLM-based segmentation models and outperforms existing fine-tuning-based TTA methods by a clear margin, with gains of up to 12.2% DSC and 11.7% mIoU. Project Page: https://lingrayy.github.io/MSSA/.

Keywords: Training-Free Test-Time Adaptation · Medical Image Segmentation · Prototype Matching

* Corresponding author.

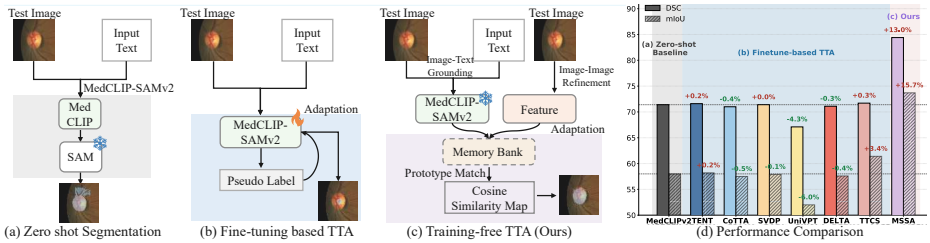

Fig. 1: Comparison of Test-Time Adaptation (TTA) paradigms. (a) Zero-shot VLM segmentation [13] relies on image–text alignment and often suffers from unstable localization under domain shifts. (b) Fine-tuning-based TTA [4] updates model parameters using noisy image–text pseudo-labels, leading to performance drift. (c) Our MSSA achieves training-free adaptation through synergistic image–text grounding and image–image prototype refinement. (d) Quantitative comparison showing that MSSA achieves consistent improvements while fine-tuning-based TTA methods [4, 17, 29, 30, 33, 38] are prone to noise amplification. Improvements are measured relative to Zero Shot in (a).

1 Introduction

Recent advances in large-scale Vision–Language Models (VLMs) have significantly reshaped the landscape of medical image analysis [11, 25, 35]. These VLMs demonstrate strong zero-shot segmentation capabilities by aligning visual and textual embeddings through attention mechanisms, enabling flexible and prompt-driven segmentation [12, 13] (see Fig. 1(a)). Despite their impressive generalization ability, we observe that foundation models remain vulnerable to domain shifts in real-world medical scenarios. When the downstream target data exhibit clear discrepancies from the pre-training distribution, performance degradation becomes inevitable. This domain gap limits the practical deployment of VLM-based segmentation systems.

Test-Time Adaptation (TTA) [14, 29, 30, 33] aims to mitigate distribution shifts by adapting models using unlabeled target data on-the-fly. In segmentation tasks, most TTA methods [29] primarily focus on Vision Foundation Models (VFMs), while effective adaptation of Vision–Language Models (VLMs) remains largely underexplored. TTCS [4] represents an early attempt to extend TTA to medical VLM-based segmentation (see Fig. 1(b)); however, the performance gains are marginal. Moreover, our empirical study shows that when representative training-based TTA methods designed for VFMs [30, 33, 38] are applied to this VLM-based segmentation framework, the improvements remain limited and can even become negative (see Fig. 1(d)). We attribute this phenomenon to the nature of pretrained VLMs. Unlike VFMs, VLMs already learn an aligned image–text representation space during pretraining. Directly fine-tuning VLMs with noisy image–text pseudo-labels may disrupt this alignment, leading to unstable optimization and limited performance improvements.

To overcome these limitations, we propose a novel training-free TTA framework for medical image segmentation, termed Memory-Supported Synergistic

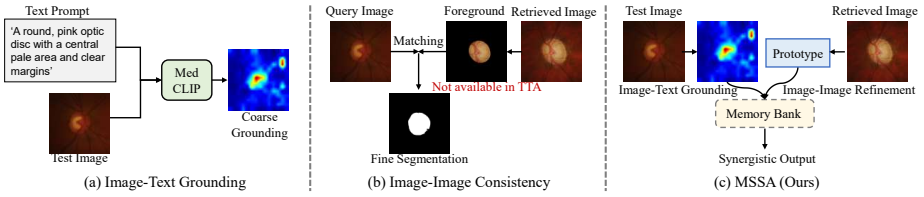

Fig. 2: Motivation of MSSA. (a) Image–text grounding provides semantic priors for localization but becomes coarse under domain shifts. (b) Image–image consistency enforces structural alignment, yet lacks semantic anchoring to reliably localize foreground regions in the target domain.(c) MSSA synergistically combines both, achieving robust and accurate segmentation.

Adaptation (MSSA) (see Fig. 1(c)). Instead of updating model parameters with noisy image–text predictions, MSSA transforms them into a support memory that leverages image-to-image consistency to guide the segmentation of all subsequent test images. Our motivation stems from an intriguing observation illustrated in Fig. 2. Image–text grounding in VLMs provides strong semantic priors due to vision–language pretraining, yet it often produces structurally coarse predictions under domain shifts. In contrast, image-to-image similarity preserves structural consistency across samples but lacks explicit semantic anchoring to identify target regions. Motivated by this complementarity, MSSA integrates semantic grounding with structural consistency, enabling robust adaptation under domain shifts.

Specifically, MSSA maintains an online memory that collects reliable predictions produced by image–text grounding and uses them as structural references for subsequent test images. For each incoming query image, MSSA retrieves representative samples from the memory and performs prototype-level image-to-image matching to refine the segmentation prediction. In this paradigm, we identify two key challenges: how to identify reliable predictions from noisy image–text outputs, and how to effectively utilize the memory to guide adaptation without parameter updates. To address the first challenge, we introduce a noise-aware memory construction mechanism that stabilizes cross-modal predictions through multi-augmentation consistency and dual-criterion reliability scoring. To address the second challenge, we design a relevance-driven image-to-image adaptation strategy that performs similarity-aware prototype alignment, enabling the memory to provide structured guidance for each query image. Together, these designs enable MSSA to convert noisy image–text predictions into structured guidance for training-free TTA, leading to notable improvements over existing TTA methods (e.g., +12.2% DSC and +11.7% mIoU in Fig. 1(d)).

Our contributions are summarized as follows:

- We propose the first training-free TTA framework for medical image segmentation, enabling effective adaptation of VLMs at test time while mitigating error accumulation, without introducing sensitive hyperparameter tuning.

- We introduce a noise-aware memory construction mechanism that identifies reliable image–text predictions via augmentation consistency and reliability scoring, enabling stable memory formation for adaptation.
- We develop a relevance-driven image-to-image adaptation strategy that retrieves representative samples from memory and performs prototype alignment to guide segmentation without parameter updates.
- Extensive experiments on multiple medical datasets demonstrate that MSSA achieves state-of-the-art performance and exhibits superior generalization compared to existing TTA methods.

2 Related works

Foundation Models for Medical Image Segmentation. The advent of large-scale VLMs like CLIP [25] and segmentation models like SAM [11] has spurred their adaptation for medical imaging, aiming for versatile, zero-shot segmentation. Models like BiomedCLIP [35] have advanced visual-language alignment in the medical domain, demonstrating strong cross-modal retrieval capabilities. To inject semantic awareness into the promptable but semantic-agnostic SAM, a line of research [2, 12, 13], employs VLMs to generate semantic-aware prompts or attention maps for SAM automatically. However, these methods based on the image-text paradigm often suffer from inconsistent and unstable attention generation. To address the inherent instability, we propose to incorporate the methods based on the image-image paradigm that can provide stability and reliable segmentations into a unified design.

Test-Time Adaptation. TTA aims to adapt a source-pre-trained model to unlabeled test data in a source-free and online manner [30]. Mainstream TTA methods often rely on self-supervised signals, such as entropy minimization [15,21,29]) or consistency regularization with pseudo-labels [30, 40], and frequently update model parameters, e.g., via Batch Normalization layers [19]. While effective for classification, these strategies are notoriously vulnerable in dense prediction tasks like segmentation, where errors in pseudo-labels can accumulate and lead to catastrophic performance collapse [5]. This risk is exacerbated in a fully source-free medical setting, where no ground truth is available. Unlike finetuning-based approaches, which may distort pretrained features, our work explores a training-free adaptation strategy that leverages foundational models directly, circumventing the peril of irreversible error propagation.

Cross-Image Consistency. The principle of cross-image semantic consistency, where shared categories exhibit invariant feature representations, has been widely exploited to enhance segmentation robustness [32,34,39]. In supervised and few-shot learning, prototype-matching [3, 16, 20, 34] leverages labeled support sets to resolve local ambiguities. In semi-supervised segmentation, Querying Labeled for Unlabeled [39] utilizes this consistency to propagate knowledge from reliable anchors to unlabeled samples. These works collectively demonstrate that inter-image relationships can effectively mitigate individual instance errors. However, existing consistency-based frameworks fundamentally rely on ground-truth labels or source-domain priors, limiting their utility in TTA. To circumvent the

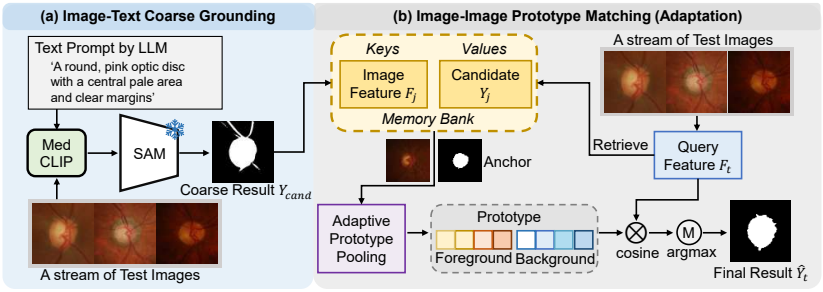

Fig. 3: The overall framework of MSSA operating in two stages. **(a)** Image-Text Coarse Grounding produces initial coarse masks from input images and text prompts, stabilized via our dual majority voting. **(b)** Noise-Aware Memory Construction filters these candidates by evaluating their semantic alignment and spatial smoothness, populating a dynamic memory bank with only the highest-quality samples. Image-Image Prototype Matching retrieves the most similar anchor example for a query image to compute a prototype and generate the final, accurate segmentation.

need for manual labels, we construct a high-quality memory bank autonomously curated from the outputs of image-text foundational models.

3 Methods

Task Definition. We address the problem of TTA for medical image segmentation. Given a stream of unlabeled test images $\{I^i\}_{i=1}^N$ from a target domain, our goal is to predict their corresponding segmentation masks $\{\hat{Y}_t^i\}_{i=1}^N$. The objective is to adapt to the target domain without access to any source data or ground-truth labels, relying solely on pre-trained source or foundation models and continuously arriving target samples.

Method Overview. As illustrated in Fig. 3, our training-free framework achieves the above objective through a synergistic, two-stage pipeline without any parameter updates or backpropagation: (1) Image-Text Coarse Grounding for producing high-quality initial predictions, (2) Adaptation: Noise-Aware Memory Construction for filtering and retaining reliable samples; and Image-Image Prototype Matching for robust and consistent segmentation.

3.1 Image-Text Coarse Grounding

This stage aims to produce initial, class-aware segmentations by leveraging the zero-shot capability of BiomedCLIP [35]. We build upon MedCLIP-SAMv2 [13] as the base pipeline, and propose several key improvements to further enhance its robustness, as shown in Fig 3 (a).

Base Pipeline. Given a test target image I, we utilize a descriptive text prompt T (generated via an LLM GPT-4 [1]) and the BiomedCLIP model to obtain a saliency map M_s through the M2IB module [31]. The coarse saliency map is

then post-processed into a binary mask M, which serves to generate the visual prompts (points and bounding boxes) for SAM to obtain a zero-shot segmentation prediction Y_{cand}, following [13]. Details are shown in the supplementary.

Gaussian Point Selection. A key limitation of the standard pipeline is the reliance on random point prompts within the coarse mask M, which can be sensitive to boundary inaccuracies. Instead, we propose a Gaussian Point Selection strategy. We compute the distance transform of the coarse mask and sample points with a probability weighted by a Gaussian kernel centered at the mask centroid (x_0, y_0):

$$GM(x, y) = \exp\left(-\frac{(x - x_0)^2 + (y - y_0)^2}{2\sigma^2}\right), \tag{1}$$

where σ is the scale parameter controlling the spatial spread of the Gaussian distribution. This prioritizes points near the object center, which are more stable and reliable, leading to more consistent refinements by SAM.

Stabilization via Dual Majority Voting. To counteract the instability of VLM-generated saliency maps, we employ a dual majority voting scheme.

For Saliency Maps: We apply photometric augmentations (e.g., CLAHE, Gamma Correction) to the input image, generate a saliency map for each variant, and aggregate them via majority voting to produce a consolidated, more stable saliency-based mask M.

For SAM Refinement: We apply geometric augmentations (e.g., flips, scales, rotations) to the image-prompt input pair of SAM and aggregate the corresponding SAM outputs via majority voting. This enforces geometric consistency and yields the candidate segmentation Y_{cand}.

3.2 Noise-Aware Memory Construction

The coarse candidate segmentations $\{Y_{cand}\}$ generated from the previous stage exhibit high potential but are unstable for direct use. To exploit their reliability while suppressing noise, we employ a selective scoring procedure to construct a high-quality **Memory Bank**, denoted as $\mathcal{B} = \{(F_j, Y_j)\}$.

It stores pairs of image features F_j and their corresponding segmented masks Y_j. To determine which candidates should be retained, we introduce a dual-faceted scoring mechanism that jointly evaluates two complementary aspects of segmentation quality: (1) semantic alignment score $\mathcal{Q}_{sem}$ with the target textual concept, and (2) spatial smoothness score $\mathcal{Q}_{smo}$ that reflects boundary consistency. The overall score $\mathcal{Q}_{total}$ for each candidate is defined as:

$$\mathcal{Q}_{total} = \mathcal{Q}_{sem} + \mathcal{Q}_{smo}, \tag{2}$$

Semantic Alignment Score ($\mathcal{Q}sem$). This score quantifies the congruence between the segmented region and the textual concept. Given the candidate mask Y_{cand}, we extract the foreground region from the original image I to obtain a

masked image I_{masked}. We then compute the BiomedCLIP image-text similarity between I_{masked} and the original text prompts T:

$$\mathcal{Q}_{sem} = \text{BiomedCLIP}(I_{masked}, T). \tag{3}$$

A high score indicates that the masked region is semantically well-aligned with the target class description.

Spatial Smoothness Score ($\mathcal{Q}_{smo}$). This score evaluates the structural integrity and boundary smoothness. The computation process is summarized in Algorithm 1, which integrates global shape regularity and local boundary stability through two key components: (1) Shape Regularity (R_s): This metric, derived from the area $A(C)$ and perimeter $P(C)$, favors shapes that are spatially consolidated. It ensures the mask resembles a well-formed anatomical structure rather than fragmented noise, without over-penalizing natural biological variations. (2) Boundary Smoothness (B_s): Calculated from the vertex density of the approximated polygon C_{approx}, this term focuses on local contour stability. Fewer high-curvature vertices imply a smoother contour, effectively reducing artifacts and noise-induced irregularities.

Algorithm 1 Spatial Smoothness Score Computation

Input: Binary mask Y_{cand}
Output: Spatial smoothness score $\mathcal{Q}_{smo}$
Step 1: Contour Extraction & Approximation
Extract the maximum external contour C from Y_{cand}
Approximate C with a polygon C_{approx}
Step 2: Shape Regularity
Compute area $A(C)$ and perimeter $P(C)$
Calculate shape regularity score R_s
Step 3: Boundary Smoothness
Compute vertex number N_v of C_{approx}
Calculate boundary smoothness score B_s
Step 4: Score Fusion
$\mathcal{Q}_{smo} \leftarrow R_s + B_s$
return $\mathcal{Q}_{smo}$

Adaptive Candidate Selection. Candidates with a total score $\mathcal{Q}_{total}$ exceeding a dynamic threshold τ_t are regarded as *promising* and admitted into the memory bank $\mathcal{B}$ under a FIFO (First-In-First-Out) update policy:

$$\mathcal{Q}_{total}^{t} \geq \tau_t. \tag{4}$$

Instead of relying on a fixed, dataset-specific threshold which may be suboptimal under varying domain shifts, we introduce an adaptive thresholding mechanism that automatically calibrates the selection criterion according to the evolving score distribution of the target stream. Specifically, we maintain a historical score buffer

$$\mathcal{H}_t = \{\mathcal{Q}_{total}^{i}\}_{i=1}^{t}, \tag{5}$$

which records the total scores of all processed samples up to time t. After a short warm-up period of N_{warm} samples ($N_{warm} = 10$), the threshold is updated:

$$\tau_t = \begin{cases} \tau_{\text{init}}, & t < N_{warm}, \\ \max(\tau_{t-1}, \text{Percentile}(\mathcal{H}_t, P)), & t \geq N_{warm}, \end{cases} \tag{6}$$

where P denotes a high percentile (i.e., $P = 80$).

This percentile-based relative thresholding removes dependence on absolute score magnitudes and enables the model to adaptively align its selection strictness with the intrinsic difficulty of the target domain.

Importantly, we enforce a *monotonically non-decreasing* update rule to prevent degradation of the memory bank quality over time. Without this constraint, temporary fluctuations in score distribution could lower the threshold and admit noisy candidates, leading to error accumulation in subsequent prototype matching. The monotonic design ensures that once high-quality samples are identified, the selection criterion will not relax, thereby maintaining a progressively refined and high-fidelity memory. In practice, the adaptive percentile-based update rapidly supersedes the initialization after the warm-up phase, rendering the framework largely insensitive to the specific choice of τ_{init}.

Each selected candidate Y_j is stored in $\mathcal{B}$ as the value, where the key is the corresponding DINOv2 [22] image feature F_j, as in Fig 3 (b).

Remark: This curation process ensures that $\mathcal{B}$ contains only semantically precise and spatially clean examples, forming a robust foundation for the subsequent prototype matching stage. Noise-Aware Memory Construction (NMC) enables the possibility of harnessing the high ceiling of the image-text modality while mitigating its inconsistency.

3.3 Robust Segmentation via Image-Image Prototype Matching

With a curated Memory Bank $\mathcal{B}$, we are able to segment testing samples using a robust, non-parametric, image-image prototype matching approach, as in Fig 3 (b). We treat the current test sample I as the query and extract its DINOv2 [22] feature F_t. If $\mathcal{B}$ is not empty, select the anchor pair (F_a, Y_a) that has the highest feature similarity to the test feature F_t; otherwise, the current initial segmentation Y_{cand} serves as anchor.

Anchor-Query Matching. For a test query image I, we first identify its most relevant anchor example from the bank. We compute the feature similarity between the test feature F_t and all anchor features F_j in $\mathcal{B}$:

$$s_j = \frac{F_t \cdot F_j}{\|F_t\|\|F_j\|}, \quad (F_j, Y_j) \in \mathcal{B}. \tag{7}$$

We select the anchor pair (F_a, Y_a) with the highest similarity s_j to guide the segmentation of I.

Non-Parametric Prototype Extraction. We adapt the ALPNet [24] into our framework. Using the selected anchor mask Y_a, we extract features from the corresponding anchor feature map F_a for both foreground ($c = 1$) and background

($c = 0$) classes. Specifically, we compute a set of local prototypes by applying average pooling with a sliding window over the regions defined by Y_a:

$$P_c^{(m,n)} = \frac{1}{|\Omega^{(m,n)}|} \sum_{u,v \in \Omega^{(m,n)}} F_a(u,v) \cdot f_\theta[Y_a(u,v) = c]. \tag{8}$$

where $\Omega^{(m,n)}$ is the local window centered at (m,n), and f_θ is the indicator function. This yields prototype sets $\mathcal{P}_1$ and $\mathcal{P}_0$ for foreground and background, respectively.

Similarity Computation & Mask Decoding. The test segmentation is generated by comparing its image feature F_t against the extracted prototypes. For each prototype $P_l^c \in \mathcal{P}_c$, we compute a cosine similarity map:

$$M_l^c(u,v) = \frac{F_t(u,v) \cdot P_l^c}{\|F_t(u,v)\| \|P_l^c\|}. \tag{9}$$

These maps are aggregated into a class-wise similarity map by taking the element-wise maximum for each class: $\tilde{M}^c = \max_l M_l^c$. The final segmentation logits L are obtained by normalizing the foreground and background similarity maps and predicted test mask $\hat{Y}_t$ is derived from L:

$$L = \text{softmax}([\tilde{M}^1; \tilde{M}^0]), \hat{Y}_t = \arg\max L. \tag{10}$$

The detailed algorithm is in the supplementary.

Remark: Crucially, the entire process is training-free. We do not fine-tune any models or parameters. Instead, we rely on the powerful representations of the foundation model and the similarity matching provided by our NMC. This design enables robust image-to-image matching without any ground-truth supervision.

4 Experiments

4.1 Datasets and Evaluation Metrics

Following TTCS [4], we evaluated our method under TTA setting on two medical segmentation tasks: optic disc (OD) segmentation in fundus images and lung segmentation in chest X-rays. Optic Disc & Cup: We collected five public fundus image datasets, treated as distinct domains. The domains include: A (RIM-ONE-r3) [8] with 159 images, B (REFUGE) [23] with 400 images, C (ORIGA) [37] with 650 images, D (REFUGE-Valid) [23] with 800 images, and E (Drishti-GS) [28] with 101 images. All images were cropped to a 1024×1024 region of interest centered on OD. Chest X-ray Lung Segmentation: To evaluate adaptation across diverse anatomical and pathological conditions, we utilized two benchmark chest X-ray datasets. First, the MC & SZ dataset [10] served as a cross-site benchmark, comprising 138 images from a US tuberculosis program and 566 from Shenzhen Hospital, China. To assess scalability and robustness, we employed the COVID-QU-Ex database [6, 27], specifically selecting the 957 test

images covering normal, lung opacity, viral pneumonia, and COVID-19 cases, as the target domain. This diverse selection tested the framework's stability against both institutional domain shifts and significant pathological variations. Following previous works [4, 18], we employed Dice Score Coefficient (DSC) and Mean Intersection over Union (mIoU) for evaluation.

4.2 Implementation Details

Following the protocols established in [4, 13], we conducted our segmentation TTA experiments using a batch size of 1 to ensure evaluation consistency. We employed SAM [11], BioMedCLIP [12], and DINOv2 [22] as our foundational models. All images were resized to 1024×1024 to meet SAM's input requirements. For data augmentation, we applied CLAHE and gamma correction [26] to generate attention maps, while random flipping, rotation, and scaling were used to create four augmented variants for SAM. The memory bank is set to store up to 15 samples. For the adaptive thresholding mechanism, we set the parameters to $P = 80$ and $N_{warm} = 10$, with the hyperparameter σ empirically determined as 0.25.

4.3 Comparison with State-of-the-art Methods

We conducted a comprehensive evaluation against existing SOTA methods, which we categorized into two primary groups based on their adaptation strategy: (1) **Training-based TTA methods**: These approaches adapt model parameters using test-time data. This group included TENT [29], EATA [21], CoTTA [30], DELTA [38], SVDP [33], and TTCS [4]; and (2) **Training-free Zero-shot Segmentation methods**: These methods leverage the inherent capabilities of large-scale foundation models without performing any parameter updates or online adaptation. Representative baselines included SaLIP [2] and MedCLIPv2 [13].

Following the experimental protocol established in TTCS [4], we adopted its prompt generation strategy for baseline methods not inherently compatible with SAM to ensure a fair comparison. In this framework, SAM served as the universal segmentation backbone across all evaluated approaches. Initial predictions generated by SAM were utilized as baseline outputs. Subsequently, each method was fine-tuned on these predictions using Low-Rank Adaptation (LoRA) [9], adhering to the specific optimization objectives and adaptation procedures defined in the original implementations.

Quantitative Results on Optic Disc/Cup Segmentation. As shown in Table 1, our method consistently outperforms all competitors across all five domains. Notably, we achieve an average DSC of 83.9% and IoU of 73.1%, surpassing the SOTA training-based TTA method, TTCS [4], by 12.2% in DSC and the strongest zero-shot method, MedCLIPSAMv2 [13], by 12.5%. This demonstrates that our framework successfully exploits the powerful zero-shot capabilities of large foundation models while effectively leveraging test-time target knowledge through training-free adaptation.

Table 1: Quantitative comparison of TTA methods built upon the VLM-based segmentation model MedCLIPv2 on the optic disc dataset. Subscripts denote the relative performance change compared with MedCLIPv2.

Technique	Method	Domain A		Domain B		Domain C		Domain D		Domain E		Average	
		DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU
VLM	SaLIP [2]	42.2	35.3	30.5	21.3	27.3	19.8	14.2	5.2	35.6	26.6	30.0	21.6
	SAMAug [7]	52.2	49.3	73.0	66.6	53.5	37.4	69.2	63.3	61.7	54.2	61.9	54.2
	MedCLIPv1 [12]	38.6	30.2	51.9	41.2	59.9	50.2	44.1	33.2	54.2	45.1	49.7	40.0
	MedCLIPv2 [13]	59.8	44.5	82.2	71.0	64.9	49.8	71.1	57.2	79.0	67.2	71.4	58.0
Training-based TTA	TENT [29]	59.8	44.6	82.1	70.8	64.9	49.8	71.4	57.5	79.6	68.2	71.6 (+0.3%)	58.2 (+0.3%)
	OCL [36]	59.6	44.3	82.0	70.8	64.8	49.7	70.9	56.9	79.6	68.0	71.4 (+0.0%)	57.9 (−0.2%)
	EATA [21]	58.6	43.4	82.2	71.0	64.7	49.6	71.1	57.2	78.6	66.8	71.0 (−0.6%)	57.6 (−0.7%)
	CoTTA [30]	59.3	44.0	82.0	70.7	65.0	49.9	71.3	57.4	77.4	65.7	71.0 (−0.6%)	57.5 (−0.9%)
	SVDP [33]	59.4	44.0	82.3	71.1	64.9	49.8	71.2	57.2	79.0	67.4	71.4 (+0.0%)	57.9 (−0.2%)
	UniVPT [17]	63.5	48.3	74.3	60.0	63.7	48.1	66.4	50.7	67.8	52.9	67.1 (−6.0%)	52.0 (−10.3%)
	DELTA [38]	59.1	43.7	82.1	70.8	64.8	49.7	71.2	57.2	78.2	66.8	71.1 (−0.4%)	57.6 (−0.7%)
	TTCS [4]	65.3	52.4	86.7	**79.1**	65.2	54.1	64.7	54.0	76.4	67.4	71.7 (+0.4%)	61.4 (+5.9%)
Training-free TTA	**MSSA(Ours)**	**76.6**	**63.1**	**87.0**	77.1	**84.6**	**73.8**	**81.5**	**69.5**	**89.9**	**82.1**	**83.9** (+17.5%)	**73.1** (+26.0%)

Table 2: Quantitative comparison of TTA methods built upon the MedCLIPv2 VLM segmentation baseline on different lung datasets. Subscripts denote relative performance changes with respect to MedCLIPv2.

Technique	Method	COVID		MC		SZ		Average	
		DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU
VLM	SaLIP [2]	63.1	56.4	62.9	45.9	66.1	49.4	64.1	53.9
	SAMAug [7]	47.0	40.3	65.3	53.2	53.7	47.1	55.3	46.8
	MedCLIPv1 [12]	32.6	29.0	35.5	33.6	25.2	21.9	31.1	28.2
	MedCLIPv2 [13]	54.6	40.7	62.1	54.6	60.2	50.5	59.0	48.6
Training-based TTA	TENT [29]	56.3	40.0	63.5	54.3	62.4	46.1	60.7 (+2.9%)	46.8 (−3.7%)
	OCL [36]	56.4	40.1	64.1	54.8	62.5	46.1	61.0 (+3.4%)	47.0 (−3.3%)
	EATA [21]	55.8	39.5	63.1	54.0	62.1	45.7	60.3 (+2.2%)	46.4 (−4.5%)
	CoTTA [30]	56.0	39.8	63.9	54.9	62.4	45.9	60.8 (+3.1%)	46.9 (−3.5%)
	SVDP [33]	56.3	40.1	64.0	54.9	62.4	46.0	60.9 (+3.2%)	47.0 (−3.3%)
	UniVPT [17]	20.4	11.9	29.1	17.7	13.3	7.5	20.9 (−64.6%)	12.4 (−74.5%)
	DELTA [38]	56.1	39.9	63.8	54.7	62.3	45.9	60.7 (+2.9%)	46.8 (−3.7%)
	TTCS [4]	57.8	44.6	65.2	43.1	65.2	43.1	62.8 (+6.4%)	43.6 (−10.3%)
Training-free TTA	**MSSA(Ours)**	**73.2**	**58.7**	**79.2**	**69.4**	**82.9**	**72.1**	**78.4** (+32.9%)	**66.7** (+37.2%)

Quantitative Results on Lung Segmentation. The superiority of our approach is further validated on different chest X-ray datasets (Table 2). We observe a significant performance gain, especially on the challenging COVID-QU-Ex dataset, where we improve the DSC by over 15% to TTCS. This highlights our method's robustness to substantial domain shifts and pathological variations, a scenario where pure TTA methods often fail due to noise propagation and where zero-shot methods lack specificity.

Qualitative Results. Fig. 4 and Fig. 5 provide visual comparisons. The baseline zero-shot methods (e.g., MedCLIPv2) often produce incomplete or semantically inaccurate segmentations. TTA methods (e.g., TTCS), while sometimes more stable, can introduce smoothing artifacts or propagate errors. In contrast, our method generates segmentation masks that are both semantically precise and anatomically coherent, closely aligning with the ground truth. This visually

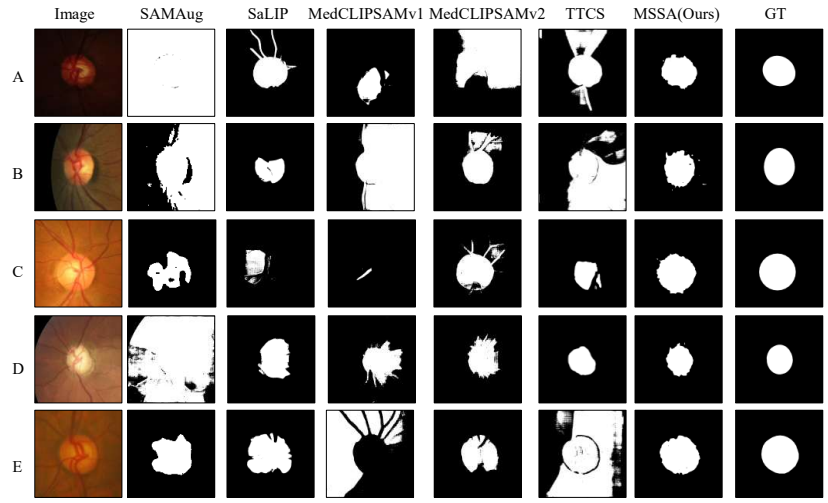

Fig. 4: Qualitative comparison of MSSA against state-of-the-art methods on optic disc datasets.

confirms the effectiveness of our selective memory bank in providing stable and accurate guidance.

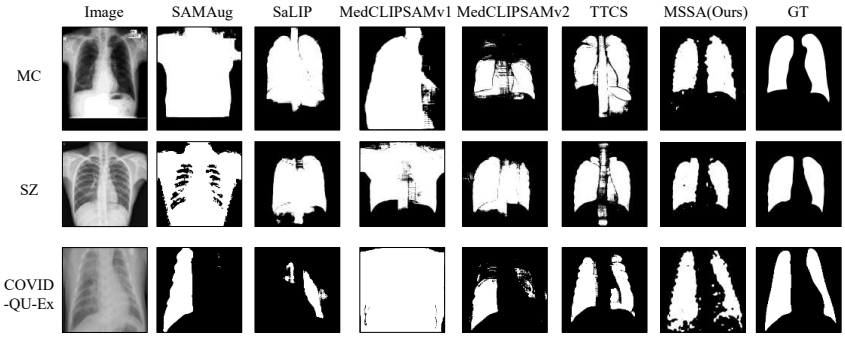

Fig. 5: Qualitative comparison of MSSA against other TTA methods on lung datasets.

4.4 Ablation Study

We performed extensive ablation studies to verify the contribution of each component in our framework.

Component-wise Analysis. The progression of results in Table 3 (rows 1 to 8) systematically validates our design choices. The experimental conclusions are:

Table 3: Ablation study of key MSSA components on Optic Disc segmentation task. We report mIoU for each target domain and DSC/mIoU averaged over all domains.

#	Components	Image-text			Image-Image		Domain (mIoU)					Avg	
		SAM-MV	ATN-MV	GP	NMC	Similar	A	B	C	D	E	DSC	mIoU
1	Baseline						44.5	71.0	49.8	57.2	67.2	71.4	58.0
2	+SAM-MV	✓					47.4	76.3	53.0	56.5	75.8	72.8	61.8
3	+ATN-MV	✓	✓				48.8	76.3	54.9	53.3	81.0	73.9	62.9
4	+GP	✓	✓	✓			51.2	76.5	57.1	57.7	81.0	75.7	64.7
5	+Bank				✓		53.0	67.6	74.1	66.4	73.9	79.2	67.0
6	+Select				✓	✓	58.3	70.7	74.1	18.6	76.5	71.2	59.7
7	All w/o Select	✓	✓	✓	✓		63.1	76.2	73.8	67.5	73.7	83.2	70.8
8	**Full Model**	✓	✓	✓	✓	✓	**63.1**	**77.1**	**73.8**	**69.5**	**82.1**	**83.9**	**73.1**

- Image-Text Branch Stabilization (Rows 2-4): The introduction of SAM Majority Voting (SAM-MV) and Attention Map Majority Voting (ATN-MV) progressively improves performance, confirming their necessity for stabilizing the initial VLM outputs. Gaussian Point (GP) Selection further provides a noticeable boost, underscoring the importance of robust prompt engineering for SAM.
- Image-Image Branch (Rows 5 & 6): To enable image matching in a zero-shot setting, this image-image variant is initialized with a single image-text sample as a warm-up support, and subsequently relies on its own predictions as the anchor. The results reveal a clear instability in this self-referential strategy, with performance degrading notably over adaptation, particularly causing a catastrophic failure on Domain D, even with NMC. This starkly illustrates the risk of error accumulation in a naive self-iterative paradigm.
- The Power of Selective Construction (Rows 4, 7 & 8): The critical step, our proposed Noise-Aware Memory Construction, leads to better performance across all domains. This proves that selecting high-quality candidates is paramount to enabling robust image-image prototype matching. Row 8 validated further improvement of selecting the most similar one as a anchor sample.

Fig. 6 also demonstrates the qualitative comparison of different paradigms on OD datasets, where relying solely on one paradigm leads to suboptimal results.

Relevant Anchor Sample Strategy. We also ablated the criteria for selecting the anchor sample for a given query in Table 4. Simply choosing a Random sample from the bank or selecting based solely on Semantic Alignment score like Q_{sem} improves but is suboptimal. Our final strategy, which selects the most *Similar* sample based on DINOv2 [22] feature similarity, achieves the best results. This indicates that feature-level similarity is a more reliable indicator for successful prototype transfer than the semantic score alone, as it captures visual and structural affinity.

Memory Bank Scoring Mechanism. We ablated the components of our selective scoring mechanism for memory bank Construction. Table 5 compares using only the semantic alignment score (Q_{sem}), only the spatial smoothness

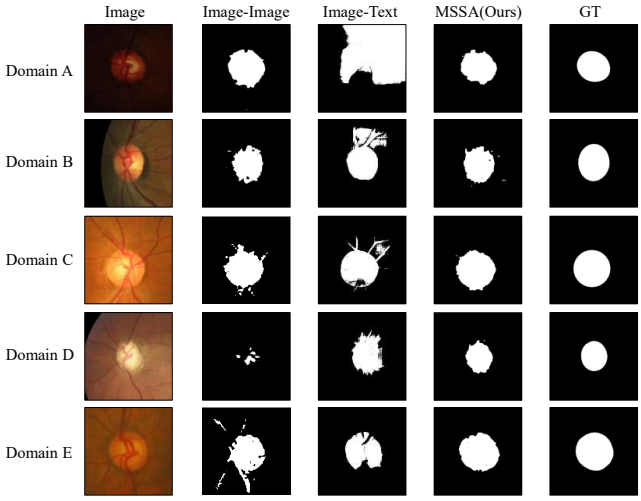

Fig. 6: Qualitative comparison of different paradigms on Optic Disc.

Table 4: Ablation studies of different anchor sample scoring mechanisms for optic disc segmentation.

Selection	**A**		**B**		**C**		**D**		**E**		**Avg**	
	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU
Random	76.6	63.1	86.3	76.2	83.9	73.8	79.4	67.5	89.9	73.7	83.2	70.8
Alignment	76.6	63.1	85.3	74.7	82.0	70.9	80.2	68.5	89.9	73.7	82.8	70.2
Similar	**76.6**	**63.1**	**87.0**	**76.7**	**84.6**	**73.8**	**81.5**	**69.5**	**89.9**	**82.1**	**83.9**	**73.1**

score (Q_{smo}), and our combined approach (Q_{total}). The results demonstrate that neither score alone is sufficient. Relying solely on *Semantic Alignment* can select masks that are semantically correct but spatially fragmented, limiting their effectiveness as anchor examples. Conversely, using only *Spatial Smoothness* may select well-shaped masks that are semantically incorrect, providing misleading guidance. Our dual-criteria strategy, requiring excellence in both semantic and spatial quality, achieved superior performance across all datasets. This confirms that these criteria are complementary: the semantic score verifies the accuracy of the segmented class, while the smoothness score ensures the structural integrity and reliability of the boundaries. Together, they provide a necessary filter for curating a high-quality, stable memory bank.

4.5 Limitation

Table 6 compares average per-image inference latency across different TTA methods on a single NVIDIA RTX A6000 GPU. While MSSA exhibits higher latency due to its multi-augmentation voting and memory retrieval, this is a deliberate trade-off to achieve notable gains in segmentation accuracy and stability.This trade-off is well-justified for non-real-time medical tasks (e.g., diagnostic plan-

Table 5: Ablation studies of selection score strategy on lung datasets.

Score	LungXray		MC		SZ		Avg	
	DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU
Q_{sem}	69.9	55.5	70.4	55.6	68.9	53.6	69.7	54.9
Q_{smo}	70.2	55.5	70.4	55.6	69.0	53.6	69.9	54.9
Q_{total}	**73.2**	**58.7**	**79.2**	**69.4**	**82.9**	**72.1**	**78.4**	**66.7**

Table 6: Comparison of inference time per image (seconds) on Drishti-GS.

Method	Performance(↑)	Time(s)(↓)
OCL [36]	61.0%	1.7
DELTA [38]	60.7%	1.7
SVDP [33]	60.9%	2.3
TTCS [4]	62.8%	5.9
MSSA(Ours)	78.4%	12.1

ning) where accuracy is critical to avoid misdiagnosis. Currently, MSSA is most effective for organs with regular geometries (e.g., lung, optic disc); its performance on highly intricate structures, like micro-vascular networks, remains a subject for future optimization.

5 Conclusion

In this paper, we propose MSSA, a novel training-free framework for test-time adaptation in medical image segmentation. By integrating three cohesive modules, stabilized candidate generation, dynamic noise-aware memory construction, and proto-type-based refinement, MSSA enables consistent and accurate adaptation without any source data or ground-truth labels. Conceptually simple and effective, MSSA is readily applicable to real-world medical scenarios without labeled data and with prevalent domain shifts, offering a new perspective for test-time adaptation.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (No. 62572166 & No. 62402157) and the Fundamental Research Funds for the Central Universities (No. JZ2025HGTB0219).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Aleem, S., Wang, F., Maniparambil, M., Arazo, E., Dietlmeier, J., Curran, K., Connor, N.E., Little, S.: Test-time adaptation with salip: A cascade of sam and clip for zero-shot medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5184–5193 (2024)
3. Ayzenberg, L., Giryes, R., Greenspan, H.: Protosam: One-shot medical image segmentation with foundational models. arXiv preprint arXiv:2407.07042 (2024)
4. Chen, H., Xu, Y., Xu, Y., Zhang, Y., Cui, L.: Test-time medical image segmentation using clip-guided sam adaptation. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1866–1873. IEEE (2024)

5. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11184–11193 (2024)
6. Chowdhury, M.E., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M.A., Mahbub, Z.B., Islam, K.R., Khan, M.S., Iqbal, A., Al Emadi, N., et al.: Can ai help in screening viral and covid-19 pneumonia? Ieee Access **8**, 132665–132676 (2020)
7. Dai, H., Ma, C., Yan, Z., Liu, Z., Shi, E., Li, Y., Shu, P., Wei, X., Zhao, L., Wu, Z., et al.: Samaug: Point prompt augmentation for segment anything model. arXiv preprint arXiv:2307.01187 (2023)
8. Fumero, F., Alayón, S., Sanchez, J.L., Sigut, J., Gonzalez-Hernandez, M.: Rim-one: An open retinal image database for optic nerve evaluation. In: 2011 24th International Symposium on Computer-based Medical Systems (CBMS). pp. 1–6. IEEE (2011)
9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
10. Jaeger, S., Karargyris, A., Candemir, S., Folio, L., Siegelman, J., Callaghan, F., Xue, Z., Palaniappan, K., Singh, R.K., Antani, S., et al.: Automatic tuberculosis screening using chest radiographs. IEEE transactions on medical imaging **33**(2), 233–245 (2013)
11. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
12. Koleilat, T., Asgariandehkordi, H., Rivaz, H., Xiao, Y.: Medclip-sam: Bridging text and image towards universal medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 643–653. Springer (2024)
13. Koleilat, T., Asgariandehkordi, H., Rivaz, H., Xiao, Y.: Medclip-samv2: Towards universal text-driven medical image segmentation. Medical Image Analysis p. 103749 (2025)
14. Li, L., Zhou, Y., Pu, N., Chen, X., Zhong, Z.: Multi-scale global-instance prompt tuning for continual test-time adaptation in medical image segmentation. In: 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 2395–2402. IEEE (2025)
15. Li, L., Zhou, Y., Yang, G.: Robust source-free domain adaptation for fundus image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7840–7849 (2024)
16. Liu, Y., Zhu, M., Li, H., Chen, H., Wang, X., Shen, C.: Matcher: Segment anything with one shot using all-purpose feature matching. arXiv preprint arXiv:2305.13310 (2023)
17. Ma, X., Wang, Y., Liu, H., Guo, T., Wang, Y.: When visual prompt tuning meets source-free domain adaptive semantic segmentation. Advances in Neural Information Processing Systems **36**, 6690–6702 (2023)
18. Mao, X., Xing, X., Meng, F., Liu, J., Bai, F., Nie, Q., Meng, M.: One polyp identifies all: One-shot polyp segmentation with sam via cascaded priors and iterative prompt evolution. arXiv preprint arXiv:2507.16337 (2025)
19. Mirza, M.J., Micorek, J., Possegger, H., Bischof, H.: The norm must go on: Dynamic unsupervised domain adaptation by normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14765–14775 (2022)

20. Nie, J., Xing, Y., Zhang, G., Yan, P., Xiao, A., Tan, Y.P., Kot, A.C., Lu, S.: Cross-domain few-shot segmentation via iterative support-query correspondence mining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3380–3390 (2024)
21. Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., Tan, M.: Efficient test-time model adaptation without forgetting. In: International conference on machine learning. pp. 16888–16905. PMLR (2022)
22. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
23. Orlando, J.I., Fu, H., Breda, J.B., van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.A., Kim, J., Lee, J., et al.: Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical Image Analysis **59**, 101570 (2020)
24. Ouyang, C., Biffi, C., Chen, C., Kart, T., Qiu, H., Rueckert, D.: Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In: European conference on computer vision. pp. 762–780. Springer (2020)
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
26. Rahman, S., Rahman, M.M., Abdullah-Al-Wadud, M., Al-Quaderi, G.D., Shoyaib, M.: An adaptive gamma correction for image enhancement. EURASIP Journal on Image and Video Processing **2016**(1), 35 (2016)
27. Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Kashem, S.B.A., Islam, M.T., Al Maadeed, S., Zughaier, S.M., Khan, M.S., et al.: Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. Computers in biology and medicine **132**, 104319 (2021)
28. Sivaswamy, J., Krishnadas, S., Chakravarty, A., Joshi, G., Tabish, A.S., et al.: A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis. JSM Biomedical Imaging Data Papers **2**(1), 1004 (2015)
29. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020)
30. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7201–7211 (2022)
31. Wang, Y., Rudner, T.G., Wilson, A.G.: Visual explanations of image-text representations via multi-modal information bottleneck attribution. Advances in Neural Information Processing Systems **36**, 16009–16027 (2023)
32. Wu, L., Fang, L., He, X., He, M., Ma, J., Zhong, Z.: Querying labeled for unlabeled: Cross-image semantic consistency guided semi-supervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(7), 8827–8844 (2023)
33. Yang, S., Wu, J., Liu, J., Li, X., Zhang, Q., Pan, M., Gan, Y., Chen, Z., Zhang, S.: Exploring sparse visual prompt for domain adaptive dense prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 16334–16342 (2024)
34. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Ma, X., Dong, H., Gao, P., Li, H.: Personalize segment anything model with one shot. arXiv preprint arXiv:2305.03048 (2023)

35. Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., et al.: Large-scale domain-specific pretraining for biomedical vision-language processing. arXiv preprint arXiv:2303.00915 **2**(3), 6 (2023)
36. Zhang, Y., Sun, Y., Zheng, S., Shui, Z., Zhu, C., Yang, L.: Test-time training for semantic segmentation with output contrastive loss. arXiv preprint arXiv:2311.07877 (2023)
37. Zhang, Z., Yin, F.S., Liu, J., Wong, W.K., Tan, N.M., Lee, B.H., Cheng, J., Wong, T.Y.: 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. pp. 3065–3068. IEEE (2010)
38. Zhao, B., Chen, C., Xia, S.T.: Delta: degradation-free fully test-time adaptation. arXiv preprint arXiv:2301.13018 (2023)
39. Zhao, D., Wang, S., Zang, Q., Jiao, L., Sebe, N., Zhong, Z.: Stable neighbor denoising for source-free domain adaptive segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23416–23427 (2024)
40. Zhao, D., Zang, Q., Pu, N., Wang, S., Sebe, N., Zhong, Z.: Secov2: Semantic connectivity-driven pseudo-labeling for robust cross-domain semantic segmentation. IEEE TPAMI **47**(11), 10378–10395 (2025). https://doi.org/10.1109/TPAMI.2025.3596943