Adapting without Access: Black-Box Bayesian Adaptation for Trustworthy Cross-Center Medical Image Segmentation

Xiaoxiang Han[1⋆], Yiman Liu[2,3⋆], Lixin Xu[4], Meng Wang[1], Xiang Xu[1], Yan Wang[3], Yuqi Zhang[2(✉)], and Qi Zhang[1(✉)]

[1] The SMART (Smart Medicine and AI-based Radiology Technology) Lab, School of Communication and Information Engineering, Shanghai University, Shanghai, China
[2] Department of Pediatric Cardiology, Shanghai Children's Medical Center, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[3] Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai, China
[4] Department of Ultrasound, The Wenzhou Third Clinical Institute Affiliated to Wenzhou Medical University, Wenzhou People's Hospital, Wenzhou, Zhejiang, China

Abstract. Cross-center deployment of medical image segmentation models faces dual challenges: domain shifts from varying imaging protocols and strict clinical constraints on privacy, security, and trustworthiness. Most existing domain adaptation methods require access to source-domain data or model parameters, which may cause privacy leakage and model attacks in real clinical settings. To bridge this gap, we propose a general, privacy-preserving, and trustworthy online adaptation framework. Under a strictly black-box setting (no source data, no parameter access), yet it continuously improves segmentation performance on the target domain and quantifies reliable uncertainty. We reformulate adaptation as approximate Bayesian inference over input-space perturbations: (1) we replace parameter updates with function-space posterior sampling via frequency-domain amplitude perturbations that preserve anatomical structures through phase spectrum invariance; (2) predictive uncertainty from posterior ensembles guides source-consistent adaptation by filtering unreliable samples; (3) semantic anchors and a memory-based empirical prior jointly prevent error accumulation. Experiments on multiple public and private cross-center datasets demonstrate consistent out-of-domain performance gains and well-calibrated uncertainty estimates under strict black-box settings.

Keywords: Domain Adaptation · Bayesian Inference · Medical Image Segmentation · Uncertainty Quantification.

⋆ Xiaoxiang Han and Yiman Liu contributed equally to this work.
Corresponding authors: Qi Zhang (zhangq@t.shu.edu.cn) and Yuqi Zhang (zhangyuqi@scmc.com.cn).
The code is available at: https://github.com/GtLinyer/AWA.

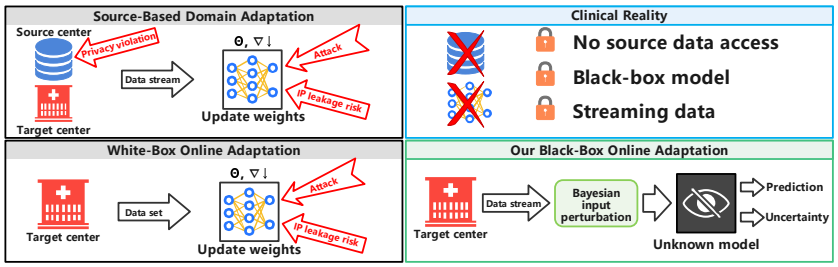

Fig. 1. Limitations of conventional source-based domain adaptation and white-box online adaptation under realistic clinical constraints, and our proposed black-box solution tailored to real-world deployment.

1 Introduction

Medical image segmentation is essential in clinical practice. Although deep learning models perform well on single-center datasets [26], their performance often degrades in multi-center, out-of-distribution (OOD) settings due to domain shifts arising from variations in imaging devices, protocols, and patient populations, which limits clinical applicability [5]. Furthermore, real-world deployment requires not only accuracy but also trustworthiness: models should quantify reliable uncertainty when encountering OOD inputs, noise, or artifacts [9,6].

Domain adaptation methods have improved robustness to distribution shift, yet their assumptions frequently conflict with real clinical constraints. Unsupervised domain adaptation typically requires access to source data, which may be prohibited by privacy regulations [2,8]. Source-free domain adaptation removes this requirement, but many approaches rely on multiple passes over the full target dataset and thus do not naturally support streaming clinical data [3,11]. Online [18]/test-time adaptation (TTA) [12] updates model parameters on incoming test samples via entropy minimization [20,27], batch normalization updates [23,7], or self-training [14,17]. Despite empirical success, these methods generally require white-box access (parameters/gradients) and online weight updates, which can be infeasible in practice due to intellectual property protection, restricted execution environments, and security concerns. Paranjape *et al.* proposed a black-box adaptation solution, but the approach relies on specific foundation models and prompts [15]. In clinical settings, deployed models are often unknown (model-agnostic) to protect security and intellectual property (IP) [1].

Therefore, an important yet under-explored scenario is **online cross-center segmentation under strict black-box access** (Fig. 1), where models achieve continuous performance improvement while providing reliable uncertainty estimation. A safer and more trustworthy clinical adaptation setting should therefore satisfy three requirements: (1) no access to source-domain data, (2) no visibility into model internals (input-output access only, model-agnostic), and (3) support for uncertainty quantification. Under these constraints, source-based distribution alignment and gradient-based parameter updates are infeasible, invalidating the

core assumptions of traditional adaptation methods. The core challenge therefore shifts from *"how to update network parameters"* to *"how to reliably change the model's functional behavior without touching its internals, and how to know when the model is unreliable."* This leads to three closely related questions in the black-box setting: (1) How can we achieve sustained online improvement? (2) How can we align target samples with the source distribution? (3) How can we obtain reliable uncertainty estimates?

To address this setting, we reformulate online adaptation as approximating Bayesian inference in function space rather than optimization in parameter space. Concretely, we treat the deployed segmenter as a fixed black-box predictor that can only be queried with inputs and returns prediction logits. Adaptation is achieved by learning a sampleable distribution over input-space perturbations: for each incoming target sample, a probabilistic perturbation network amortizes an approximate posterior over a latent variable, from which we generate constrained frequency-domain amplitude noise [29] to obtain a source-consistent input. To encourage source-consistent behavior without access to source data, we combine two complementary regularizers: (1) a memory bank that stores latent codes from low-uncertainty cases and forms an empirical prior to guide posterior updates and mitigate drift; and (2) semantic anchors constrain the perturbation to preserve anatomical content and prevent degenerate transformations. Multiple posterior samples yield an ensemble of perturbed inputs whose aggregated predictions improve robustness, while the predictive dispersion naturally provides an uncertainty estimate. This framework enables continuous, privacy-preserving, and model-agnostic online adaptation for cross-center medical image segmentation under strict black-box access.

2 Methodology

2.1 Problem Statement and Revisiting Online Adaptation

Let $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$ denote a pre-trained medical image segmentation model deployed in a target clinical center, where $\mathcal{X} \subset \mathbb{R}^{H\times W}$ is the image space and $\mathcal{Y} \subset \mathbb{R}^{H\times W\times C}$ denotes the pixel-wise logits. Conventional online or test-time adaptation methods address domain shift by optimizing model parameters θ using target samples, typically by minimizing entropy, enforcing consistency, or performing self-training. Formally, such methods solve: $\min_{\theta} \mathbb{E}_{x_t\sim\mathcal{D}_t}\left[\mathcal{L}\big(\mathcal{F}_\theta(x_t)\big)\right]$ [25], where $\mathcal{D}_t$ denotes the target-domain data distribution, and the optimization requires gradients of the objective with respect to θ.

However, in realistic clinical deployments, this optimization paradigm becomes infeasible. In particular, we consider a **strict black-box setting**, where: (1) no access to source-domain data is available; (2) model parameters θ of $\mathcal{F}$ are fixed and inaccessible; for any input $x \in \mathcal{X}$, the model returns logits $y = \mathcal{F}(x)$. Target-domain samples $\{x_t\}_{t=1}^{\infty}$ arrive sequentially in an **online** manner. The objective is to improve **segmentation performance** under domain shift, while simultaneously providing reliable **uncertainty estimates**, without modifying $\mathcal{F}$ or accessing any source information.

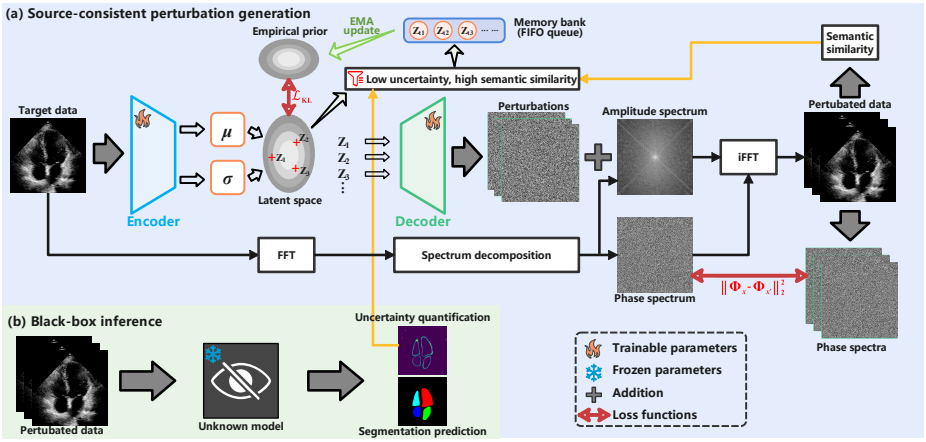

Fig. 2. The proposed framework: (a) source-consistent perturbation via amplitude-spectrum modulation with phase preservation and memory-guided posterior; (b) uncertainty-aware inference through posterior ensemble predictions.

Function-Space View: In white-box adaptation, domain shift is typically addressed by updating model parameters θ. Under black-box constraints, θ is not accessible. Instead, we reformulate adaptation as modifying the effective input-output function via input-space transformations. Let $g_\phi : \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$ be a stochastic perturbation function parameterized by ϕ, where $z \in \mathcal{Z}$ is a latent variable. The adapted predictor is then defined as the composition:

$$\mathcal{F}_\phi(x, z) = \mathcal{F}\big(g_\phi(x, z)\big). \tag{1}$$

Rather than optimizing a single deterministic perturbation, we model a distribution over perturbations, enabling Bayesian inference and uncertainty quantification. The core idea is to learn a posterior distribution over perturbation functions that align target samples with the source-domain behavior of $\mathcal{F}$.

2.2 Bayesian Perturbation Modeling for Domain Alignment

Latent Variable Model: For each target sample x, we introduce a latent variable z that controls the perturbation:

$$z \sim p(z),\ x' = g_\phi(x, z), \tag{2}$$

where $p(z)$ is a prior over perturbations. The prediction is $y' = \mathcal{F}(x')$. We aim to infer a posterior distribution $p(z|x, \mathcal{F})$ that induces perturbations yielding source-consistent predictions under $\mathcal{F}$.

Perturbations on Frequency-Domain Amplitude: Cross-center domain shifts in medical imaging are dominated by low-level appearance variations (e.g., intensity, contrast, scanner characteristics), while high-level anatomy remains stable [16]. To exploit this property, perturbations are applied in the frequency domain. Let $\mathcal{F}_{\text{FFT}}(\cdot)$ denote the Fourier transform. For an input image x,

$\mathcal{F}_{\text{FFT}}(x) = A_x \odot e^{i\mathbf{\Phi}_x}$, where A_x and $\mathbf{\Phi}_x$ are amplitude and phase components, respectively. The decoder maps z to an amplitude perturbation $N(z)$, yielding $A_{x'} = A_x + N(z)$, $\mathbf{\Phi}_{x'} = \mathbf{\Phi}_x$. The adapted image is then obtained by inverse FFT: $x' = \mathcal{F}_{\text{FFT}}^{-1}(A_{x'} \odot e^{i\mathbf{\Phi}_x})$. This design guarantees structural preservation while enabling flexible appearance adaptation.

2.3 Uncertainty-Regularized Posterior Inference

Under strict black-box constraints, reliable online adaptation requires preventing drift caused by self-reinforcing errors while maintaining sufficient plasticity to accommodate domain shifts. To this end, we introduce a semantics-gated Bayesian updating mechanism driven by uncertainty, which jointly governs posterior regularization and memory-based prior estimation.

Uncertainty Estimation: Given multiple posterior samples $\{z_k\}_{k=1}^K \sim q_\phi(z|x, y)$, we obtain an ensemble of predictions:

$$\left\{y_k = \mathcal{F}\big(g_\phi(x, z_k)\big)\right\}_{k=1}^K. \tag{3}$$

Uncertainty is quantified as predictive dispersion: $\mathcal{U}(x) = \text{Var}_k(y_k)$, which captures ambiguity in identifying source-consistent input transformations under black-box access. It reflects functional sensitivity to admissible perturbations and directly characterizes the reliability of the adaptation outcome.

Functionally-Gated Memory Bank Construction: Low adaptation uncertainty alone is insufficient to guarantee semantic correctness, especially during early stages of online adaptation. We therefore introduce a functionally-gated memory bank to construct a robust empirical prior over latent perturbations. Let $y = \mathcal{F}(x)$ and $y' = \mathcal{F}(x')$ denote the predictions of the original and perturbed inputs, respectively. We admit a latent code z into the memory bank only if it satisfies: $\mathcal{U}_{\text{adapt}}(x) < \tau$, and $\text{sim}(y, y') > \delta$, where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity. The uncertainty threshold τ is defined as a running percentile to ensure scale invariance. This functional gating mechanism ensures that stored perturbations not only yield reliable predictions but also preserve the model's intrinsic semantic decision structure, effectively mitigating confirmation bias and long-term drift during online adaptation.

Online Empirical Prior with Temporal Smoothing: The memory bank is implemented as a fixed-length FIFO queue with expiration, allowing outdated latent codes to be gradually discarded as the target distribution evolves. From the current memory contents, we estimate an empirical Gaussian prior $p_{\mathcal{M}}(z) = \mathcal{N}(\mu_{\mathcal{M}}, \Sigma_{\mathcal{M}})$. To ensure temporal stability, the prior parameters are updated using exponential moving averages (EMA): $\mu_{\mathcal{M}}^{(t)} = (1 - \alpha)\mu_{\mathcal{M}}^{(t-1)} + \alpha\overline{z}_t$, where $\overline{z}_t$ denotes the mean of newly admitted latent codes at step t. This smoothed update balances stability and adaptability, preventing abrupt prior shifts caused by transient or noisy samples. Posterior inference is then regularized via:

$$\mathcal{L}_{\text{prior}} = \text{KL}\big(q_\phi(z \mid x, y) \,\|\, p_{\mathcal{M}}(z)\big), \tag{4}$$

which encourages consistency with historically reliable adaptations while preserving flexibility for ongoing domain alignment.

Table 1. Quantitative comparison on echocardiography datasets. A and B denote Private-Echo-A and Private-Echo-B, respectively. $*$, †, and ‡ indicate significant improvements of our methods over Source, TENT, and TEGDA, respectively (paired t-test, $p < 0.05$).

Methods	CardiacUDA→A		CardiacUDA→B		A→CardiacUDA		A→B	
	Avg.DSC	Avg.95HD	Avg.DSC	Avg.95HD	Avg.DSC	Avg.95HD	Avg.DSC	Avg.95HD
Source	$67.43_{3.15}$	$38.16_{8.42}$	$81.54_{1.83}$	$25.72_{4.12}$	$71.28_{3.08}$	$36.52_{7.95}$	$74.18_{2.96}$	$31.84_{6.91}$
Ours	$70.05_{2.94}{}^{*}$	$35.21_{7.48}{}^{*}$	$83.39_{1.66}{}^{*}$	$21.75_{3.97}{}^{*}$	$74.88_{2.72}{}^{*}$	$32.94_{6.80}{}^{*}$	$77.25_{2.68}{}^{*}$	$29.38_{6.20}{}^{*}$
TENT [19]	$69.12_{3.02}$	$36.24_{7.81}$	$82.89_{1.75}$	$22.17_{4.02}$	$73.64_{2.91}$	$34.88_{7.21}$	$76.41_{2.81}$	$30.12_{6.54}$
SAR [22]	$70.43_{2.88}$	$34.92_{7.31}$	$83.57_{1.79}$	$21.84_{3.83}$	$74.21_{2.79}$	$33.57_{6.98}$	$76.93_{2.77}$	$29.74_{6.33}$
CoTTA [21]	$71.16_{2.74}$	$34.28_{6.94}$	$83.90_{1.67}$	$21.55_{3.83}$	$74.86_{2.63}$	$33.12_{6.74}$	$77.68_{2.63}$	$29.21_{6.18}$
InTEnt [4]	$70.22_{2.90}$	$34.91_{7.35}$	$82.88_{1.72}$	$22.24_{3.96}$	$74.09_{2.85}$	$33.81_{7.03}$	$76.85_{2.72}$	$29.89_{6.41}$
TEGDA [28]	$72.84_{2.61}$	$32.83_{6.42}$	$85.28_{1.59}$	$20.92_{3.60}$	$76.38_{2.52}$	$31.47_{6.32}$	$79.86_{2.46}$	$27.68_{5.87}$
Ours+TENT	$72.11_{2.67}{}^{\dagger}$	$33.63_{6.58}{}^{\dagger}$	$84.61_{1.66}{}^{\dagger}$	$21.23_{3.72}{}^{\dagger}$	$75.92_{2.58}{}^{\dagger}$	$32.26_{6.41}{}^{\dagger}$	$78.94_{2.54}{}^{\dagger}$	$28.63_{6.02}{}^{\dagger}$
Ours+TEGDA	$\mathbf{76.82}_{2.58}{}^{\ddagger}$	$\mathbf{29.56}_{6.07}{}^{\ddagger}$	$\mathbf{87.45}_{1.46}{}^{\ddagger}$	$\mathbf{20.69}_{3.53}{}^{\ddagger}$	$\mathbf{79.24}_{2.41}{}^{\ddagger}$	$\mathbf{29.92}_{6.09}{}^{\ddagger}$	$\mathbf{82.97}_{2.19}{}^{\ddagger}$	$\mathbf{25.92}_{5.33}{}^{\ddagger}$

Semantic Anchoring via Observation Likelihood: Under black-box constraints, two types of observations are available. **(1)** a structural observation is derived from the phase spectrum. Since the phase is an explicit invariant of the generative process, we model its consistency via: $p(\mathbf{\Phi}_{x'} \mid \mathbf{\Phi}_x) = \mathcal{N}(\mathbf{\Phi}_x; \mathbf{\Phi}_{x'}, \sigma_\Phi^2 I)$, leading to a phase consistency loss $\|\mathbf{\Phi}_x - \mathbf{\Phi}_{x'}\|_2^2$. **(2)** Prediction consistency is modeled as a heteroscedastic likelihood whose negative log-likelihood is proportional to predictive dispersion: $-\log p(y \mid z, x) \propto \mathrm{Var}_k\big(\mathcal{F}(g_\phi(x, z_k))\big)$, which favors perturbations yielding stable predictions under the black-box model.

Variational Objective: We maximize the evidence lower bound (ELBO):

$$\mathbb{E}_{q(z|x)}\big[\log p(y \mid z, x) + \log p(\mathbf{\Phi}_{x'} \mid \mathbf{\Phi}_x)\big] - \mathrm{KL}\big(q_\phi(z \mid x, y) \,\|\, p_{\mathcal{M}}(z)\big). \quad (5)$$

Importantly, optimization does not require gradients or parameter updates of $\mathcal{F}$. The deployed model is queried only through forward inference $\mathcal{F}(g_\phi(x, z_k))$.

Inference: Multiple samples from the learned posterior generate an ensemble of adapted predictions. The final segmentation is obtained via aggregation based on uncertainty, while uncertainty map is reported alongside predictions.

3 Experiments and Results

3.1 Experimental Setup

Datasets: We evaluated the proposed black-box online adaptation framework on 2D echocardiography and 3D CT segmentation tasks using both public and private clinical datasets. For echocardiography, we used CardiacUDA [24] (public, 1932 adult images) together with two private datasets: Private-Echo-A (pediatric, 1014 images) and Private-Echo-B (adult, 157 images), which exhibit substantial variations in anatomy, demographics, imaging devices, and acquisition protocols. For abdominal CT, experiments were conducted on public BTCV [10] and WORD [13] datasets, covering diverse scanner vendors and protocols.

Table 2. Quantitative comparison on CT datasets. Abbreviations: Spl = Spleen, R.kid = Right Kidney, L.kid = Left Kidney, Gall = Gallbladder, Eso = Esophagus, Sto = Stomach, Pan = Pancreas. (*,†,‡paired t-test, $p < 0.05$)

Methods	BTCV → WORD										WORD → BTCV	
	Spl	R.kid	L.kid	Gall	Eso	Liver	Sto	Pan	Avg.DSC	Avg.95HD	Avg.DSC	Avg.95HD
Source	87.35	85.14	81.49	52.78	48.52	94.44	75.13	56.60	$72.68_{23.11}$	$19.79_{15.01}$	$67.56_{25.16}$	$25.52_{19.18}$
Ours	87.60	86.83	83.25	56.38	57.43	94.59	75.87	60.26	$75.28_{24.58}^{*}$	$11.03_{13.81}^{*}$	$69.72_{16.84}^{*}$	$16.84_{16.72}^{*}$
TENT [19]	87.04	85.17	84.13	50.37	51.91	95.12	75.92	58.48	$73.52_{21.03}$	$13.45_{18.13}$	$70.84_{14.92}$	$14.92_{14.87}$
InTEnt [4]	87.60	86.83	89.25	55.38	52.43	93.59	75.87	60.26	$75.15_{24.58}$	$11.69_{15.93}$	$71.63_{13.78}$	$13.78_{14.94}$
TEGDA [28]	88.28	89.19	87.94	60.11	62.00	93.53	77.97	65.93	$78.12_{18.72}$	$10.07_{16.39}$	$76.18_{11.62}$	$11.62_{12.08}$
Ours+TENT	87.62	87.07	83.51	60.30	58.52	94.58	**82.66**	64.55	$77.35_{19.81}^{\dagger}$	$10.76_{14.75}^{\dagger}$	$76.42_{10.94}^{\dagger}$	$10.94_{11.67}^{\dagger}$
Ours+TEGDA	**90.76**	**92.49**	**90.45**	**70.09**	**68.11**	**94.79**	82.41	**72.51**	$\mathbf{82.70}_{11.04}^{\ddagger}$	$\mathbf{9.43}_{12.11}^{\ddagger}$	$\mathbf{79.23}_{13.51}^{\ddagger}$	$\mathbf{9.87}_{10.14}^{\ddagger}$

Table 3. Ablation study of core components (BP: Bayesian perturbation; UP: uncertainty-aware posterior; MB: memory bank; SA: semantic anchoring; C: CardiacUDA; A: Private-Echo-A).

BP	UP	MB	SA	C→A	BTCV→WORD
✓	–	–	–	$67.82_{3.11}$	$72.41_{26.34}$
✓	✓	–	–	$69.05_{3.02}$	$73.68_{25.17}$
✓	✓	✓	–	$69.70_{2.98}$	$74.53_{24.89}$
✓	✓	✓	✓	$70.05_{2.94}$	$75.28_{24.58}$

Table 4. Ablation study on filtering strategies (Unc: Uncertainty; Sem: Semantic; BT: BTCV; WO: WORD; Drift: % of sequences with >5% DSC degradation over 50 consecutive samples).

Filtering Strategy	C→A	BT→WO	Drift
None	$68.21_{3.24}$	$73.42_{22.87}$	15.6%
Unc-only ($\mathcal{U} < \tau$)	$69.18_{3.05}$	$74.35_{21.93}$	8.2%
Sem-only (sim $> \delta$)	$68.95_{3.12}$	$74.01_{22.15}$	9.7%
Dual	$70.05_{2.94}$	$75.28_{24.58}$	3.1%

Implementation Details: Strict black-box online adaptation setting: the deployed segmentation model was frozen and only queried for logits. Target samples arrived sequentially and were used once. The g_ϕ is an eight-layer convolutional architecture with a latent space dimension of 10. Our framework was trained online with Adam, drawing multiple posterior samples per input. A fixed-size FIFO memory bank with EMA maintained an empirical prior over latents. All datasets shared identical hyperparameters. U-Net (2D) and V-Net (3D) backbones were used, implemented in PyTorch 2.4 on a NVIDIA RTX 3090.

3.2 Comparison with Existing Methods

We experimentally compared our method against existing approaches. While ours operates under strict black-box constraints, TENT [19], SAR [22], CoTTA [21], InTEnt [4], and TEGDA [28] assume white-box access and rely on parameter updates or gradient-based adaptation. Under these distinct settings, our approach delivered consistent performance gains across all cross-center benchmarks without compromising clinical privacy requirements, as shown in Tables 1 and 2, . More importantly, our Bayesian perturbation framework exhibited strong complementarity with white-box adaptation: integrating it with TENT and TEGDA yielded significant additional improvements, demonstrating that input-space adaptation operated orthogonally to parameter-space methods and provides a practical enhancement for privacy-sensitive deployments.

Table 5. Ablation study of frequency-domain perturbation (pert.: perturbation; freq: frequency. amp.: amplitude).

SA	C→A	BTCV→WORD
Baseline (no pert.)	$67.43_{3.15}$	$72.68_{23.11}$
Spatial noise	$69.14_{3.11}$	$73.98_{24.54}$
Phase pert.	$68.06_{3.08}$	$73.12_{24.41}$
Low-freq amp. only	$69.34_{3.12}$	$74.55_{24.69}$
Ours	$70.05_{2.94}$	$75.28_{24.58}$

Table 6. Uncertainty reliability and selective risk evaluation (Spe.: Spearman).

Methods	CardiacUDA → A		BTCV → WORD	
	Spe. ρ	AUROC	Spe. ρ	AUROC
Entropy	0.21	0.62	0.18	0.59
TENT	0.34	0.71	0.29	0.68
TEGDA	0.41	0.76	0.36	0.74
Ours	**0.56**	**0.83**	**0.49**	**0.81**

3.3 Ablation Study

To validate the contribution of each component, we conducted comprehensive ablation studies on cross-center echocardiography and CT segmentation tasks. As shown in Table 3, the Bayesian perturbation (BP) module alone yielded modest improvements over the source model, while incorporating uncertainty-aware posterior inference (UP) provided consistent gains by guiding reliable adaptation. The memory bank (MB) further enhanced stability through empirical prior regularization, and semantic anchoring (SA) yielded the final performance boost by preserving anatomical fidelity during perturbation. BP defines a perturbation hypothesis space, whereas UP and SA suppress unreliable perturbations and stabilize adaptation. Crucially, Table 4 showed that our dual filtering strategy, which required both low adaptation uncertainty and high semantic similarity, reduced error accumulation compared to single-criterion filtering or no filtering. This approach cut catastrophic drift sequences by over 80% while preserving superior segmentation accuracy. Our frequency-domain ablation (Table 5) showed amplitude perturbations outperformed spatial noise or phase perturbations, while preserving phase spectra was critical for anatomical fidelity. Additionally, hyperparameter ablation yielded optimal settings. We used K=6 posterior samples, a memory bank of 128 with EMA rate α=0.05. Uncertainty filtering adopted the 30th percentile threshold τ, and semantic gating used cosine similarity δ=0.9. Amplitude perturbations were limited to the lowest 25% frequency band.

3.4 Uncertainty Reliability Evaluation

Beyond accuracy, we assessed the reliability and clinical utility of predictive uncertainty. Uncertainty was measured as the variance across posterior perturbation samples and was correlated with segmentation error (1−DSC) using Spearman's coefficient. As shown in Table 6, our method exhibited markedly stronger uncertainty–error correlation than the source and test-time adaptation baselines on both echocardiography and CT datasets, indicating that higher uncertainty reliably reflected poorer segmentation under domain shift. For risk-aware decision making, we further evaluated bad-case detection by computing AUROC with uncertainty as the confidence score. Our method achieved the highest AUROC, demonstrating effective failure identification. Overall, our uncertainty was aligned with errors and supported trustworthy black-box deployment.

3.5 Conclusion

We present a privacy-preserving black-box online adaptation framework for cross-center medical image segmentation. By reformulating adaptation as approximate Bayesian inference over input-space perturbations, our method achieves continuous performance improvement and reliable uncertainty estimation without accessing source data or model parameters. Experiments across echocardiography and CT benchmarks demonstrate consistent gains and strong uncertainty reliability under strict black-box constraints. Limitations include increased inference cost from posterior sampling and reliance on handcrafted frequency-domain priors. Future work will explore more efficient sampling strategies and extend the framework to video and multi-modal clinical data.

Acknowledgments. This work was supported by the National Natural Science Foundation of China [No. 62571309], the Natural Science Foundation of Shanghai Municipality [No. 25ZR1401135], the Fundamental Research Funds for the Central Universities [No. YG2025QNA07], the Pudong New Area Science and Technology Development Fund [No. PKJ2025-Y04], and Shanghai Technical Service Center of Science and Engineering Computing, Shanghai University.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bassily, R., Thakkar, O., Guha Thakurta, A.: Model-agnostic private learning. Advances in neural information processing systems **31** (2018)
2. Cai, Z., Xin, J., You, C., Shi, P., Dong, S., Dvornek, N.C., Zheng, N., Duncan, J.S.: Style mixup enhanced disentanglement learning for unsupervised domain adaptation in medical image segmentation. Medical Image Analysis **101**, 103440 (2025)
3. Chen, Y., Luo, X., Chen, R., Li, Y., Zhang, H., Lyu, H., Song, H., Li, K.: Source-free active domain adaptation via influential-points-guided progressive teacher for medical image segmentation. IEEE Transactions on Medical Imaging (2025)
4. Dong, H., Konz, N., Gu, H., Mazurowski, M.A.: Medical image segmentation with intent: Integrated entropy weighting for single image test-time adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5046–5055 (2024)
5. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Transactions on Biomedical Engineering **69**(3), 1173–1185 (2021)
6. Huang, Q., Gao, S., Li, G., Wang, W.: Fuzzy knowledge perception with fault tolerance and brain-inspired reasoning for diagnosis of ultrasound images. IEEE Transactions on Fuzzy Systems (2026)
7. Kang, J., Kim, N., Ok, J., Kwak, S.: Membn: Robust test-time adaptation via batch norm with statistics memory. In: European Conference on Computer Vision. pp. 467–483. Springer (2024)
8. Kollias, D., Arsenos, A., Kollias, S.: Domain adaptation explainability & fairness in ai for medical image analysis: Diagnosis of covid-19 based on 3-d chest ct-scans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4907–4914 (2024)

9. Lambert, B., Forbes, F., Doyle, S., Dehaene, H., Dojat, M.: Trustworthy clinical ai solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis. Artif. Intell. Medicine **150**, 102830 (2024)
10. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge. vol. 5, p. 12 (2015)
11. Li, J., Yu, Z., Du, Z., Zhu, L., Shen, H.T.: A comprehensive survey on source-free domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(8), 5743–5762 (2024)
12. Liang, J., He, R., Tan, T.: A comprehensive survey on test-time adaptation under distribution shifts. International Journal of Computer Vision **133**(1), 31–64 (2025)
13. Luo, X., Liao, W., Xiao, J., Chen, J., Song, T., Zhang, X., Li, K., Metaxas, D.N., Wang, G., Zhang, S.: Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. Medical Image Analysis **82**, 102642 (2022)
14. Ma, J.: Improved self-training for test-time adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23701–23710 (2024)
15. Paranjape, J.N., Sikder, S., Vedula, S.S., Patel, V.M.: Black-box adaptation for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 454–464. Springer (2024)
16. Qiao, Q., Wang, W., Qu, M., Su, K., Jiang, B., Guo, Q.: Medical image segmentation via single-source domain generalization with random amplitude spectrum synthesis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 435–445. Springer (2024)
17. Su, Y., Xu, X., Jia, K.: Towards real-world test-time adaptation: Tri-net self-training with balanced normalization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 15126–15135 (2024)
18. Volpi, R., De Jorge, P., Larlus, D., Csurka, G.: On the road to online adaptation for semantic image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19184–19195 (2022)
19. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020)
20. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: 9th International Conference on Learning Representations, ICLR 2021 (2021)
21. 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)
22. Wang, W., Zhong, Z., Wang, W., Chen, X., Ling, C., Wang, B., Sebe, N.: Dynamically instance-guided adaptation: A backward-free approach for test-time domain adaptive semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24090–24099 (2023)
23. Wu, Y., Chi, Z., Wang, Y., Plataniotis, K.N., Feng, S.: Test-time domain adaptation by learning domain-aware batch normalization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 15961–15969 (2024)
24. Yang, J., Ding, X., Zheng, Z., Xu, X., Li, X.: Graphecho: Graph-driven unsupervised domain adaptation for echocardiogram video segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11878–11887 (2023)

25. Yuan, L., Xie, B., Li, S.: Robust test-time adaptation in dynamic scenarios. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15922–15932 (2023)
26. Zeng, X., Li, G., Li, J., Li, G., Huang, Q.: Self-supervised saliency-guided dual-attention fusion for benign-malignant thyroid nodule diagnosis in ultrasound. Neurocomputing p. 133641 (2026)
27. Zhou, X., Tian, Z., Zhang, B., Zhang, Y., Cheung, K.C., See, S., Yang, H., Zhou, Y., Zhang, N.L.: Test-time adaptation on noisy data via model-pruning-based filtering and flatness-aware entropy minimization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 10852–10860 (2025)
28. Zhou, Y., Wu, J., Liao, W., Zhang, S., Zhang, S., Wang, G.: Tegda: Test-time evaluation-guided dynamic adaptation for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 628–637. Springer (2025)
29. Zhou, Z., Qi, L., Shi, Y.: Generalizable medical image segmentation via random amplitude mixup and domain-specific image restoration. In: European Conference on Computer Vision. pp. 420–436. Springer (2022)