RELIABILITY-GATED TEST-TIME ADAPTATION FOR CROSS-DOMAIN 2D MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical image segmentation models often encounter acquisition conditions that differ from their source data. Test-time adaptation (TTA) can update a model without target labels, but image-wide entropy and consistency objectives allow ambiguous regions and unstable boundaries to contribute directly to the gradient. We study RELIABILITY-GATED TTA, a spatially routed objective that assigns different roles to target pixels according to their predictive stability. At each step, predictive entropy, disagreement between aligned intensity views, and disagreement between their boundary responses form a detached reliability gate. Selected pixels supervise entropy and consistency, their local neighborhood supports boundary consistency, and the complementary region is anchored to the source posterior from the second update onward. Optimization is restricted to normalization-affine parameters while normalization statistics and stochastic layers remain fixed. Across three independently trained MSCANet checkpoints, the routed update decreases DDTI Dice for every seed (-0.00159 ± 0.00089, mean paired change $\pm$ sample standard deviation) and produces only small positive changes on STU ($+0.00025 \pm 0.00019$). A seed-9 spatial audit further shows that the gate places errors disproportionately in its excluded half (mean enrichment 1.91 on DDTI and 1.99 on STU), while adding almost no separation within the target boundary band. These results separate a useful spatial diagnostic from an update objective that does not deliver a consistent, material segmentation improvement across the evaluated domains.

1 INTRODUCTION

Medical image segmentation has become a standard interface between image acquisition and downstream measurement. Encoder–decoder models such as U-Net made dense prediction practical with a simple skip-connected design (Ronneberger et al., 2015), and self-configuring training systems such as nnU-Net showed that careful preprocessing and optimization can be as important as an architectural change (Isensee et al., 2021). These systems perform well on in-domain benchmarks, yet performance can degrade when scanners, acquisition protocols, resolution, field of view, or annotation practice differ between development and deployment data.

The central difficulty is deciding which parts of a source prediction provide credible supervision when the target domain is unlabeled. Test-time training and adaptation address this setting by learning from information available at inference time (Sun et al., 2020; Wang et al., 2021; Zhang et al., 2022; Niu et al., 2022; Wang et al., 2022). For segmentation, this permits a trained checkpoint to adjust without target annotations. Most objectives, however, summarize the image or average over all pixels. Ambiguous regions, acquisition artifacts, and unstable lesion boundaries can therefore influence the same update as reliable interiors.

Confidence alone does not show whether a spatial prediction is stable under plausible appearance changes or whether its contour is reliable. Medical TTA has demonstrated adaptation under acquisition changes (Karani et al., 2021), while calibrated entropy and uncertainty-guided updates regulate when a model should learn from target data (Yang et al., 2024; Tan et al., 2025). These approaches leave a finer-grained question unresolved: can spatial reliability determine where an update receives supervision while a restricted parameter subspace determines how far the model can move?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We address this question with a spatially selective adaptation procedure. It estimates a pixelwise uncertainty score from three complementary signals: predictive entropy captures posterior ambiguity, disagreement between aligned weak and strong intensity views captures local instability, and disagreement between their boundary responses captures contour instability. A detached reliability gate converts these signals into an update mask. Entropy and consistency act on selected low-uncertainty pixels, and boundary consistency acts on a local dilation of that region. After the first update, a frozen source posterior constrains the complement. The default update changes normalization-affine parameters in evaluation mode, leaving convolutional weights, normalization statistics, and stochastic layers fixed.

Our contributions are:

- We formulate a pixel-routed TTA objective for cross-domain 2D medical segmentation, so target regions contribute according to their estimated reliability rather than through a uniform image-wide average.
- We construct a detached quantile gate from predictive ambiguity, aligned-view instability, and boundary instability, and couple reliable interiors, boundary neighborhoods, and complementary regions to different objectives.
- We couple the intervention to a matched diagnostic protocol that holds uncertainty readouts and stochastic perturbations fixed, logs the gate used by each successful update, and measures segmentation quality, selective risk, and spatial error enrichment on separate axes.
- We expose a concrete selector–update failure mode: the recorded gate concentrates whole-image errors outside its retained region, yet provides little boundary-specific separation; across three source checkpoints, the routed update decreases DDTI Dice and yields only tiny STU changes.

The protocol keeps three effects distinct: changes in the segmentation mask, changes in case-level risk ordering, and changes in the spatial support of the update. This decomposition shows that a spatially informative selector is not, by itself, evidence that its associated unsupervised update is beneficial.

2 Related Work

2.1 Medical image segmentation under domain shift

Strong 2D segmentation systems establish a useful source model but do not remove domain shift. U-Net remains a common architectural reference (Ronneberger et al., 2015), while nnU-Net illustrates how preprocessing and optimization choices shape segmentation performance (Isensee et al., 2021). Domain generalization benchmarks such as DomainBed and WILDS make the source–target mismatch explicit across environments (Gulrajani & Lopez-Paz, 2021; Koh et al., 2021), and flat-minima training provides one route to improve out-of-domain behavior without target labels (Cha et al., 2021). Biomedical test-time adaptation further shows that a source-trained segmenter can respond to changes in image appearance (Karani et al., 2021). Our setting asks which spatial evidence should control that response when target annotations are unavailable.

2.2 Test-time adaptation and selective updating

Test-time training introduces an auxiliary self-supervised objective at inference time (Sun et al., 2020); Tent adapts normalization parameters by minimizing prediction entropy on each target batch (Wang et al., 2021). MEMO, CoTTA, and EATA extend this setting through augmentation-based objectives, continual updates, anti-forgetting mechanisms, and selective sample filtering (Zhang et al., 2022; Wang et al., 2022; Niu et al., 2022). Their selection rules are primarily sample-level or derived from aggregated confidence. A segmentation mask instead contains stable interiors, ambiguous boundaries, and imaging artifacts within the same sample.

Medical single-image adaptation provides the closest protocol comparison. InTEnt integrates predictions produced with different normalization statistics (Dong et al., 2024), while SicTTA carries information through a continual single-image stream (Wu et al., 2026). Uncertainty-calibrated adaptation and calibrated entropy objectives regulate updates at the model or sample level (Tan et al.,

2025; Yang et al., 2024). We instead use uncertainty as a spatial routing variable, keep its evaluation readout separate from the intervention, and reserve a distinct neighborhood for boundary consistency.

2.3 Uncertainty, consistency, and boundary-aware segmentation

Predictive entropy and ensembles are established tools for representing uncertainty (Guo et al., 2017; Lakshminarayanan et al., 2017; Gal & Ghahramani, 2016), while selective prediction evaluates whether a model can abstain on risky examples (Geifman & El-Yaniv, 2019; 2017). Consistency learning uses perturbation agreement as a training signal (Tarvainen & Valpola, 2017; Sohn et al., 2020), and AugMix demonstrates that augmentation consistency can improve robustness and uncertainty jointly (Hendrycks et al., 2020). These ideas motivate our aligned-view disagreement cue, but they do not specify how a segmentation TTA update should be spatially gated.

Region overlap can hide contour errors, so boundary-sensitive objectives and surface distances are important complements to Dice. Boundary loss focuses optimization on contour geometry (Kervadec et al., 2019), and Hausdorff distance losses directly target outlier surface errors (Karimi & Salcudean, 2020). We therefore pair overlap with HD95 and ASD. Boundary disagreement supplies an adaptation signal; ground-truth surface metrics remain the independent measure of contour error.

3 Reliability-Gated Test-Time Adaptation

3.1 Problem formulation

Let f_θ be a source-trained 2D segmentation model that maps an image $x \in \mathbb{R}^{C\times H\times W}$ to logits $z = f_\theta(x) \in \mathbb{R}^{K_{\text{out}}\times H\times W}$. For a binary head, $K_{\text{out}} = 1$ and the foreground probability is $p_{\text{fg}} = \sigma(z)$; before computing reliability, we expand it to the two-class posterior $p = (1 - p_{\text{fg}}, p_{\text{fg}})$, so the class count in Eq. 1 is $K = 2$. For multiclass segmentation, p is the softmax over the K_{out} output channels and $K = K_{\text{out}}$. At test time, target images are observed without masks. The aim is to improve a prediction after a small number of updates using only x, while limiting each update to a deliberately small parameter subspace and spatial support.

We partition the parameters as $\theta = (\theta_{\text{frozen}}, \phi)$, where ϕ contains normalization-affine parameters and, optionally, named lightweight adapters. The default protocol updates only ϕ. This choice isolates the adaptation mechanism from backbone retraining and reduces the dimensionality of the update state.

3.2 Overview and reliability cues

The pipeline computes predictions for the original image and two spatially aligned intensity views. The views change brightness, contrast, noise, or blur but do not move pixels, so disagreement can be interpreted at the same spatial location. Let p, p_w, and p_s denote the corresponding posteriors. Figure 1 summarizes how these views produce the reliability gate and the gated update. We define three normalized uncertainty maps:

Predictive ambiguity. The normalized entropy map is

$$u_{\text{ent}}(i) = -\frac{1}{\log K}\sum_{k=1}^{K} p_k(i)\log p_k(i). \tag{1}$$

Thus, the binary case uses the two-class posterior defined above rather than a single foreground probability.

Augmentation instability. We measure the total variation between the weak and strong views,

$$u_{\text{aug}}(i) = \frac{1}{2}\sum_{k=1}^{K} |p_{w,k}(i) - p_{s,k}(i)|. \tag{2}$$

This cue is high when a small appearance change changes the local prediction.

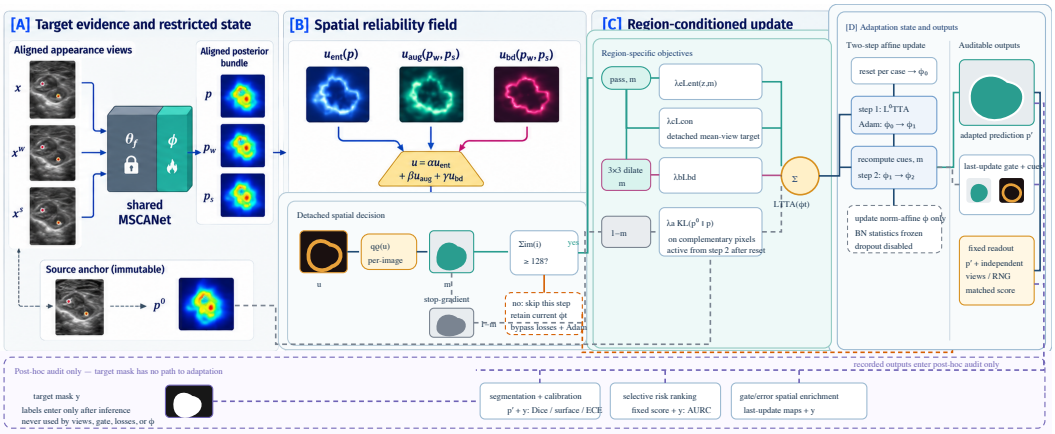

Figure 1: Reliability-Gated TTA and its evidence boundary. (A) The target image and two spatially aligned appearance views share one MSCANet with a frozen backbone θ_f and trainable normalization-affine state ϕ; a separate immutable source pass produces p^0. (B) Predictive entropy, view disagreement, and boundary disagreement form the spatial score u. A per-image quantile and stop-gradient yield the detached gate m; an update is skipped if fewer than 128 pixels are selected. (C) Reliable pixels support entropy and consistency, a 3×3 dilation of m supports boundary consistency, and the complementary region is anchored to p^0. (D) Each case resets to the source affine state and performs two Adam steps while BatchNorm statistics and stochastic layers remain fixed; the adapted prediction, last-update maps, and fixed readout are recorded without feedback into optimization. The lower violet rail is post-hoc evaluation only: target masks never enter view generation, gating, losses, or parameter updates. All image, probability, and uncertainty maps are schematic rather than experimental observations.

Boundary instability. Let $\mathcal{B}(p)$ be the Sobel magnitude of the foreground probability (or the maximum non-background probability for multiclass masks). We use

$$u_{\mathrm{bd}}(i) = \mathrm{scale}\left(|\mathcal{B}(p_w)(i) - \mathcal{B}(p_s)(i)|\right), \tag{3}$$

where scale uses per-image 5th and 95th percentiles and clips to $[0, 1]$. The robust scaling prevents a few high-gradient pixels from dominating the gate.

The combined uncertainty score is converted into a detached reliability gate. For the default ratio-based rule, let $q_\rho(u)$ denote the per-image ρ-quantile of the spatial scores, where ρ is the retained fraction. The gate is then

$$u(i) = \alpha u_{\mathrm{ent}}(i) + \beta u_{\mathrm{aug}}(i) + \gamma u_{\mathrm{bd}}(i), \tag{4}$$

$$m(i) = \mathbf{1}[u(i) \le \mathrm{sg}(q_\rho(u))], \tag{5}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient and the nonnegative cue weights are normalized to sum to one. The default configuration uses $\rho = 0.5$, so the gate retains the lowest-uncertainty half of the pixels and is insensitive to the absolute scale of u. When ratio selection is disabled, we use the alternative rule $m(i) = \mathbf{1}[u(i) \le \tau]$ with $\tau = 0.45$; a soft sigmoid version is available for controlled ablations. The minimum selected-pixel count is 128. Target masks do not enter this selection rule or any hyperparameter decision.

3.3 Gated Adaptation Objective

The objective pairs adaptation losses on selected pixels with source-posterior anchoring on the complementary region. We use the weak and strong predictions as a detached consistency target $\bar{z} = \frac{1}{2}(z_w + z_s)$ and optimize

$$\mathcal{L}_{\mathrm{TTA}} = \lambda_e \mathcal{L}_{\mathrm{ent}}(z, m) + \lambda_c \mathcal{L}_{\mathrm{con}}(z, \bar{z}, m) + \lambda_b \mathcal{L}_{\mathrm{bd}}(z, \bar{z}, m) + \lambda_a \mathcal{L}_{\mathrm{anchor}}(z, z^0, 1 - m), \tag{6}$$

where $z^0 = f_{\theta_0}(x)$ is computed by an immutable source anchor. The first term is entropy averaged over m; the second is a probability-space mean-squared consistency loss; and the third is an L_1

Table 1: Case-wise reliability-gated adaptation protocol.

Input: target image x, source model f_{θ_0}, update subset ϕ, steps S.
1. Compute $z^0 = f_{\theta_0}(x)$ and z, z_w, z_s for the current model and aligned views.
2. Estimate $u_{\mathrm{ent}}, u_{\mathrm{aug}}, u_{\mathrm{bd}}$ and form m using Eq. 5.
3. If $\sum_i m(i)$ is below the minimum, skip the update. Otherwise minimize Eq. 6 over ϕ only.
4. Repeat for S steps, then evaluate $f_\theta(x)$.
Output: adapted logits and the score, cue maps, and detached gate from the last successful update.

distance between Sobel boundary responses. Entropy and consistency are averaged over selected pixels, whereas the boundary loss applies the locally dilated gate used in the implementation. The anchor term is a pixelwise KL divergence from the source posterior, averaged over $1 - m$. At source initialization the anchor term and its parameter gradient are zero. It can constrain a case-wise update only from the second optimization step onward; consequently, anchor ablations use at least two matched steps.

The hard gate is detached before backpropagation. This prevents the optimizer from changing the reliability decision through a shortcut in the loss. When the selected region contains fewer than a minimum number of pixels, the code skips the update and returns the current model prediction. This guard is active when thresholding or a small keep ratio produces too few selected pixels; with a fixed 50% keep ratio it is normally inactive at the evaluated resolution.

3.4 Parameter-efficient update and protocols

The default update subspace contains the scale and bias of BatchNorm, InstanceNorm, GroupNorm, or LayerNorm modules. The model remains in evaluation mode during this update: BatchNorm running statistics are frozen and stochastic layers such as dropout remain disabled. A user backbone can instead expose a small adapter by including "adapter" or "tta_adapter" in its parameter name and selecting the corresponding mode. Adam updates the selected parameters with a fixed learning rate and step count. Cost measurements therefore describe the specified adaptation subspace rather than backbone retraining or BatchNorm statistics adaptation.

We evaluate two adaptation protocols. In the case-wise protocol, the adaptive model and optimizer are reset before every target image. In the continual protocol, the state is carried through a fixed target stream order. The former measures per-case recovery without forgetting; the latter measures accumulated drift and is not interchangeable with independent test images. Both protocols are implemented in the supplied runner. Main results use the case-wise protocol unless stated otherwise.

3.5 Algorithm and computational scope

Table 1 summarizes one case-wise update. With S adaptation steps and P_ϕ trainable parameters, the persistent additions are the frozen anchor and the optimizer state for P_ϕ; activation memory depends on the implementation. The adaptation path uses one anchor evaluation, three current-model evaluations per step, and one post-update evaluation, for $3S + 2$ forward passes. A fixed composite uncertainty readout adds two aligned-view evaluations. The exact MSCANet parameter count, peak memory, and synchronized latency are computed from the same checkpoint used for evaluation.

4 Experiments

4.1 Datasets and task audit

MSCANet (Pan et al., 2026) is trained on the TN3K thyroid-ultrasound training split (Gong et al., 2021; 2023), and the checkpoint is selected on the source validation split. The cross-domain evaluation includes DDTI (Pedraza et al., 2015); the locally held collection labeled STU enters only if its foreground masks denote the same anatomical target as TN3K. Every target collection is audited for mask semantics, empty-mask frequency, duplicate images, patient-level separation, preprocessing, and licence terms before inclusion. TG3K is not treated as an ordinary target domain unless

this audit confirms label equivalence; the standard TG3K task uses thyroid-gland masks (Chi et al., 2023), whereas TN3K uses nodule masks. Target annotations are never passed to the adaptation computation; they are accessed only for post-hoc metric calculation and visualization.

We evaluate three independently trained source checkpoints with training seeds 7–9 and a 100-epoch budget. Their SHA-256 digests identify checkpoint identity; the evaluator's deterministic random seed is a separate quantity and is not used to infer the training seed. All three checkpoints load strictly into the same MSCANet constructor with no missing or unexpected keys, and all protocols within a checkpoint use the same target cases, preprocessing, and fixed uncertainty readout. The retained training histories have unequal lengths (84, 96, and 69 entries), so the exact checkpoint-selection trajectory remains a provenance limitation even though the effective training configuration is matched.

4.2 Compared protocols and matched randomization

We compare frozen source inference, AdaBN, entropy minimization, and Reliability-Gated TTA. Entropy minimization and Reliability-Gated TTA update the same normalization-affine parameter subset for the same number of steps and learning rate. Both models remain in evaluation mode, so BatchNorm running statistics and dropout are fixed. AdaBN is the separate control that updates BatchNorm statistics without gradient-based parameter adaptation. In the primary case-wise protocol, the model and optimizer return to the source state before every image. A fixed-order continual stream provides a separate test of accumulated drift.

Stochastic corruptions and TTA views use independent generators. Each generator is seeded by a stable hash of the run seed, case identifier, stage, corruption, and severity. Excluding the protocol name ensures that competing methods receive the same corrupted image and uncertainty-readout views. Reliability-Gated TTA uses two update steps, equal cue weights, and a 50% keep ratio. The two-step setting is necessary for the source-anchor gradient to become non-zero in a case-wise reset. Hyperparameters are selected without target-test masks.

4.3 Metrics and uncertainty readouts

Segmentation quality is measured with Dice, IoU, HD95, and average surface distance (ASD). HD95 and ASD use pixels on the resized evaluation grid; each aggregate is accompanied by the number of cases with a finite value. Calibration is summarized by expected calibration error (ECE) and Brier score. For selective prediction, case risk is $1 - \mathrm{Dice}$ and coverage is obtained by accepting cases in ascending uncertainty order. We report the area under this risk–coverage curve (AURC), where lower is better. Non-finite surface distances are reported with their finite-case count and are never replaced by zero.

Every protocol is scored with two fixed readouts. The confidence readout is the mean pixelwise normalized predictive entropy. The composite readout averages the same equally weighted entropy, aligned-view disagreement, and boundary disagreement used throughout the evaluation. The resulting 2×2 design crosses frozen or adapted predictions with entropy or composite ranking. Cue ablations alter the adaptation gate while the evaluation readouts remain fixed, so a change in ranking is not conflated with a change in the scorer.

Paired case-level differences in Dice, IoU, surface distance, calibration, and AURC use 95% percentile bootstrap intervals from 10,000 resamples. When several images come from the same patient, the patient is the resampling unit. Patient identifiers are unavailable in the current DDTI/STU copies, so the reported intervals resample images and may be too narrow if multiple images belong to one patient. We report each checkpoint separately and summarize the three source-training runs with their mean and sample standard deviation. The case bootstrap and the across-checkpoint standard deviation answer different questions and are not substituted for one another.

The matched readout design, stage-specific randomization, and post-hoc spatial audit are documented in Appendix B; they are controls for the result-bearing comparisons, not experimental results themselves.

Table 2: Cross-domain segmentation across three independently trained MSCANet checkpoints. Dice is reported as mean ± sample standard deviation across source-training seeds 7–9. Paired changes subtract source-only performance from the same checkpoint; the final column exposes their direction and magnitude instead of treating case bootstrap as checkpoint replication.

Domain	Protocol	Dice ↑	ΔDice vs. source ↑	Seed-wise ΔDice (7/8/9)
DDTI	Source-only	0.73505 ± 0.03396	–	–
	AdaBN	0.73763 ± 0.02737	+0.00259 ± 0.00665	-0.00263 / +0.01008 / +0.00031
	Entropy TTA	0.73310 ± 0.03494	-0.00195 ± 0.00098	-0.00118 / -0.00305 / -0.00160
	Reliability-TTA	0.73346 ± 0.03483	-0.00159 ± 0.00089	-0.00100 / -0.00261 / -0.00116
STU	Source-only	0.81312 ± 0.01293	–	–
	AdaBN	0.81394 ± 0.01393	+0.00082 ± 0.00104	+0.00039 / +0.00005 / +0.00201
	Entropy TTA	0.81368 ± 0.01286	+0.00055 ± 0.00036	+0.00015 / +0.00085 / +0.00065
	Reliability-TTA	0.81337 ± 0.01288	+0.00025 ± 0.00019	+0.00004 / +0.00041 / +0.00029

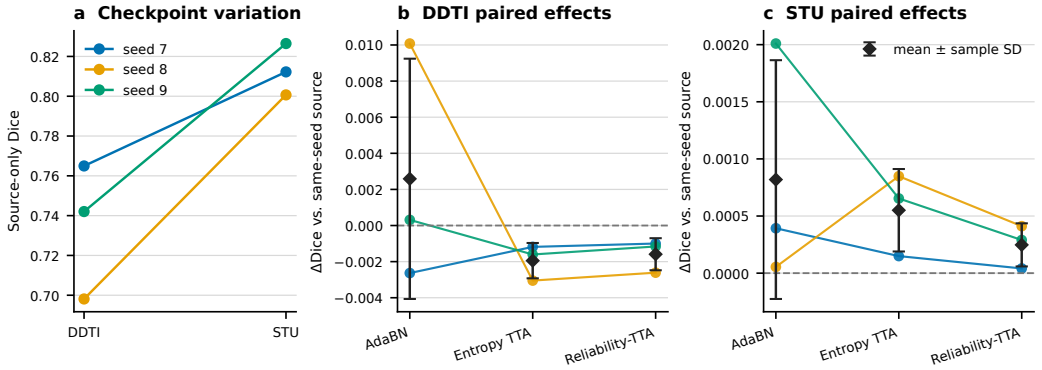

Figure 2: Checkpoint variability and paired segmentation effects. (a) Source-only Dice for source-training seeds 7–9. (b–c) Protocol-minus-source Dice for the same checkpoint on DDTI and STU. Colored points and connecting lines identify individual seeds; black diamonds and bars show the mean and sample standard deviation across the three checkpoints. The bars are not confidence intervals, and no significance claim is made from three runs.

4.4 Primary comparison

Checkpoint variation is much larger than any Reliability-TTA effect: source Dice ranges from 0.69813 to 0.76496 on DDTI and from 0.80067 to 0.82648 on STU. Relative to its same-checkpoint source model, Reliability-Gated TTA decreases DDTI Dice for all three seeds, with mean paired change -0.00159 ± 0.00089 (sample standard deviation). It increases STU Dice for all three seeds, but only by $+0.00025 \pm 0.00019$; the individual changes range from $+0.00004$ to $+0.00041$. AdaBN has the highest mean Dice on both domains, yet its DDTI changes are directionally inconsistent (-0.00263, $+0.01008$, and $+0.00031$). These results do not support a consistent cross-domain segmentation improvement from Reliability-Gated TTA.

4.5 Mechanism and anchor ablation

The following mechanism and anchor analyses use seed 9 only. Although the update does not improve DDTI Dice, its spatial score is strongly associated with pixel errors. The excluded 50% of pixels captures 95.7% of DDTI errors (95% CI $[95.1, 96.2]\%$) and 99.3% of STU errors ($[98.6, 99.9]\%$), corresponding to error enrichment of 1.91 ($[1.90, 1.92]$) and 1.99 ($[1.97, 2.00]$), respectively. In contrast, enrichment inside the target-derived boundary band is only 1.008 on DDTI and 1.002 on STU. The selector therefore separates easy interiors from error-prone pixels, but does not specifically rank errors within the boundary band and does not by itself establish that its routed losses provide a corrective gradient. The matched-readout intervals and spatial enrichment plot are retained in Appendix E rather than treated as multi-checkpoint evidence.

Removing the source anchor changes Dice from 0.74080 to 0.74089 on DDTI ($+0.00009$, 95% CI $[0.00005, 0.00014]$) and from 0.82677 to 0.82677 on STU (-0.00001, $[-0.00005, 0.00004]$). The DDTI effect is statistically resolved under case resampling but practically very small; STU is indistinguishable from zero. We therefore do not claim that anchoring materially drives the observed behavior in this configuration.

4.6 Single-checkpoint sensitivity and corruption stress tests

The completed incremental queue adds 22 sensitivity runs on the seed-9 checkpoint and six severity-3 corruption suites on 3,493 TN3K source-validation images. The former is one-checkpoint sensitivity, not replication; the latter is a source-distribution stress test, not external domain generalization. The full Dice values, paired deltas, and provenance checks are reported in Appendix D. AdaBN is positive for all six corruptions ($+0.00106$ to $+0.00489$), whereas Reliability-Gated TTA changes range from -0.00020 to $+0.00039$. The corrected CUDA latency and memory benchmark remains outstanding and contributes no number to the conclusions.

5 Diagnostic Analysis

5.1 Separating adaptation from risk scoring

The two uncertainty readouts create a 2×2 diagnostic. Holding the prediction fixed while replacing entropy with the composite score isolates the reliability estimator. Holding the score fixed while replacing the source prediction with the adapted prediction isolates the update. Their interaction reveals whether adaptation changes the ordering of difficult cases beyond the effect of rescoring a frozen model. The matched-readout analysis is currently available for seed 9: composite-readout AURC increases under Reliability-Gated TTA on DDTI and has an inconclusive increase on STU. Across all three source checkpoints, Dice effects are domain-dependent instead: all DDTI changes are negative, whereas all STU changes are positive but very small. AdaBN has the highest mean Dice while its DDTI direction varies by checkpoint, illustrating why segmentation, selective ranking, and source-training variability must be reported separately.

5.2 Testing the spatial selection mechanism

Case-level AURC cannot establish that the gate excludes erroneous pixels. The spatial analysis compares segmentation error rates on m and $1 - m$, both over the full image and within a fixed-width target boundary band. We additionally report error capture,

$$\text{capture}(1 - m) = \frac{\sum_i \mathbf{1}[m(i) = 0]\, \mathbf{1}[\hat{y}(i) \neq y(i)]}{\sum_i \mathbf{1}[\hat{y}(i) \neq y(i)]}, \tag{7}$$

alongside the area fraction of $1 - m$. Error enrichment is the ratio between the error rate outside the gate and the image-wide error rate. A value above one indicates that excluded pixels contain more errors than expected from their area. Segmentation metrics independently determine whether the subsequent update turns this spatial information into a better mask. The observed combination is diagnostic: whole-image enrichment is near the maximum of two implied by a 50% gate, but boundary-band enrichment is near one. Most errors are therefore outside the retained region largely because confident background and lesion interiors are easy to exclude from the error set; within the difficult boundary band, the current score provides little further separation.

The seed-9 qualitative diagnostic uses the median case within data-derived adverse, neutral, and favorable terciles of adapted-minus-source Dice; the use of target error for this post-hoc stratification is disclosed. Each panel displays the image, target mask, source and adapted predictions, composite uncertainty, actual update gate, and both error maps on a common scale. The verified DDTI and STU plates appear in Appendix F. The selected examples show visually small prediction changes, consistent with the aggregate Dice effects; they are mechanism illustrations rather than evidence of benefit.

5.3 Mechanism interpretation

Together, the seed-9 diagnostics localize the current failure. The gate carries a strong image-wide error-ranking signal, but this signal is largely absent after conditioning on the target boundary band. Replication across checkpoints then shows that the routed update consistently decreases DDTI Dice while producing only tiny STU gains. These measurements do not support the complete selector–update mechanism: they identify both boundary ranking and conversion of the selected signal into a corrective gradient as current bottlenecks.

6 Limitations and Responsible Use

The study is restricted to 2D segmentation and intensity views that preserve pixel correspondence. It does not characterize 3D-volume adaptation, long clinical streams, multi-site deployment, geometric shifts, or changes in the foreground definition. A collection with gland masks, for example, cannot serve as a domain-shift test for a nodule segmenter. A 3D extension would require separate memory, latency, and volumetric evaluation.

The main Dice table contains three source checkpoints, 637 DDTI images per run, and 42 STU images per run. Three checkpoints expose substantial source-training variation but remain too few for a precise distributional or significance claim. Moreover, their retained training-history lengths differ, and the exact checkpoint-selection trajectory has not been fully reconstructed. The locally available datasets do not expose verified patient identifiers, so seed-9 case bootstrap intervals use images as independent units; hidden within-patient correlation would make these intervals optimistic. The matched-readout AURC, spatial gate, update-map, and anchor analyses are also seed-9 diagnostics rather than replicated checkpoint evidence. The completed sensitivity queue uses only the seed-9 checkpoint (with evaluator seed metadata retained in the records), and the corruption queue covers only severity 3 on the TN3K source-validation split. These results do not establish robustness across checkpoints, severities, or external domains. Corrected CUDA latency and peak-memory measurements remain outstanding.

Reliability is estimated from the model's predictions and can itself be miscalibrated. Source anchoring constrains drift after the first update but cannot supply a missing semantic correction or repair an overconfident source error. The default quantile gate allocates a fixed fraction of every image and does not implement whole-case rejection; the absolute-threshold variant is therefore evaluated separately. Gate hyperparameters must be fixed without target-test masks. Case-wise resets avoid accumulated drift but do not represent continual deployment, and a finite fixed-order stream cannot establish long-horizon stability.

The near-two whole-image error enrichment should not be read as successful adaptation. With a fixed 50% gate, excluding confident easy pixels can produce strong enrichment even when the difficult boundary pixels are not ordered well. Indeed, boundary-band enrichment is approximately one, and the subsequent update decreases DDTI Dice across all three checkpoints. The spatial analysis is therefore a mechanism diagnostic, not an accuracy claim.

The uncertainty score represents model behavior, not disease probability, and has not been validated for clinical triage or decision support. Dataset licences, de-identification, patient-level separation, protected metadata, and site-specific performance require source-level verification before public release. Redistribution of the MSCANet implementation and checkpoint is subject to the same audit.

7 Conclusion

We frame label-free test-time adaptation for 2D medical segmentation as spatial evidence selection and audit both the selector and its update. The detached three-cue gate concentrates whole-image errors outside the selected region but adds little separation within the target boundary band. Across three checkpoints, Reliability-TTA decreases DDTI Dice for every seed and yields only tiny STU gains; the current evidence supports a diagnostic use of spatial selection, not a consistent accuracy-improvement claim. Boundary-conditioned ranking, corruption coverage, continual deployment, and CUDA efficiency remain priorities for a revised update objective.

AI USE STATEMENT

Generative AI tools assisted with research-direction brainstorming, method and experiment-design feedback, literature-search organization, manuscript drafting and editing, code implementation and debugging, and scientific-figure prototyping. AI-assisted material was treated as a draft and checked against cited sources, code tests, and experiment records before inclusion. The authors take responsibility for the final study design, claims, results, text, and artifacts. Generative AI was not used to create target annotations or experimental observations. Target annotations are used only for post-hoc evaluation and do not enter adaptation.

ETHICS STATEMENT

The experiments use ultrasound image–mask pairs for research evaluation. The adaptation loop does not consume target annotations, and its outputs have not been validated for clinical decision making. The dataset copies available to this study do not by themselves document de-identification, patient-level separation, or redistribution terms. Public release of data or derived models is therefore contingent on verification against the original data sources.

REPRODUCIBILITY STATEMENT

The accompanying artifact contains the method implementation, evaluator, configuration templates, and plotting script. Appendix A records the reset policy and target-mask exclusion; Appendix B records the view, randomization, and update conventions. The backbone insertion point is isolated so that the MSCANet implementation can be audited separately from the adaptation code.

REFERENCES

Junho Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. SWAD: Domain generalization by seeking flat minima. In *Advances in Neural Information Processing Systems*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/bcb41ccdc4363c6848a1d760f26c28a0-Abstract.html.

Jianning Chi, Zelan Li, Zhiyi Sun, Xiaosheng Yu, and Huan Wang. Hybrid transformer UNet for thyroid segmentation from ultrasound scans. *Computers in Biology and Medicine*, 153:106453, 2023. doi: 10.1016/j.compbiomed.2022.106453.

Haoyu Dong, Nicholas Konz, Hanxue Gu, and Maciej A. Mazurowski. 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 Workshops*, pp. 5046–5055, 2024. doi: 10.1109/CVPRW63382.2024.00511.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, 2016. URL https://proceedings.mlr.press/v48/gal16.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Haifan Gong, Guanqi Chen, Ranran Wang, Xiang Xie, Mingzhi Mao, Yizhou Yu, Fei Chen, and Guanbin Li. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 257–261, 2021. doi: 10.1109/ISBI48211.2021.9434087.

Haifan Gong, Jiaxin Chen, Guanqi Chen, Haofeng Li, Guanbin Li, and Fei Chen. Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. *Computers in Biology and Medicine*, 155:106389, 2023. doi: 10.1016/j.compbiomed.2022.106389.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1gmrxHFvB.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. doi: 10.1038/s41592-020-01008-z.

Neerav Karani, Ertunc Erdil, Krishna Chaitanya, and Ender Konukoglu. Test-time adaptable neural networks for robust medical image segmentation. *Medical Image Analysis*, 68:101907, 2021. doi: 10.1016/j.media.2020.101907.

Davood Karimi and Septimiu E. Salcudean. Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. *IEEE Transactions on Medical Imaging*, 39 (2):499–513, 2020. doi: 10.1109/TMI.2019.2930068.

Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. In *Proceedings of the Second Conference on Medical Imaging with Deep Learning*, volume 102 of *Proceedings of Machine Learning Research*, pp. 285–296, 2019. URL https://proceedings.mlr.press/v102/kervadec19a.html.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proceedings of the 39th International Conference on Machine Learning*, 2022. URL https://proceedings.mlr.press/v162/niu22a.html.

Zhongjie Pan, Ge Tang, Peishi Jiang, Jingchao Zhang, Longchun Dong, Jiafei Liu, Yi Sun, and Jun Tian. MSCANet: A lightweight and robust model for colorectal tumor segmentation in medical imaging. In *Advanced Intelligent Computing Technology and Applications*, Lecture Notes in Computer Science, pp. 111–122. Springer Nature Singapore, 2026. doi: 10.1007/978-981-92-3495-0_10.

Lina Pedraza, Carlos Vargas, Fabián Narváez, Oscar Durán, Emma Muñoz, and Eduardo Romero. An open access thyroid ultrasound image database. In *10th International Symposium on Medical Information Processing and Analysis*, volume 9287 of *Proceedings of SPIE*, pp. 92870W, 2015. doi: 10.1117/12.2073532.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/06964dce9addb1c5cb5d6e3d9838f733-Abstract.html.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9229–9248, 2020. URL https://proceedings.mlr.press/v119/sun20b.html.

Mingkui Tan, Guohao Chen, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Peilin Zhao, and Shuaicheng Niu. Uncertainty-calibrated test-time model adaptation without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6274–6289, 2025. doi: 10.1109/TPAMI.2025.3560696.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/68053af2923e00204c3ca7c6a3150cf7-Abstract.html.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7210, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Continual_Test-Time_Domain_Adaptation_CVPR_2022_paper.html.

Jianghao Wu, Xinya Liu, Guotai Wang, and Shaoting Zhang. Sictta: Single image continual test time adaptation for medical image segmentation. *Medical Image Analysis*, 108:103859, 2026. doi: 10.1016/j.media.2025.103859.

Hao Yang, Min Wang, Jinshen Jiang, and Yun Zhou. Towards test time adaptation via calibrated entropy minimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3736–3746, 2024. doi: 10.1145/3637528.3671672.

Marvin Zhang, Sergey Levine, and Chelsea Finn. MEMO: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems*, pp. 38629–38642, 2022. doi: 10.52202/068431-2799.

A REPRODUCIBILITY CHECKLIST

The implementation supplies a stable insertion point for MSCANet and writes the resolved model, checkpoint, TTA, uncertainty-readout, corruption, and random-seed metadata with every run. Per-case rows contain both uncertainty scores and all segmentation and calibration metrics. RELIABILITY-GATED TTA can additionally store the score, detached gate, and three cue maps used by the last successful update. Files without this map provenance are rejected by the spatial-analysis tool. No target mask is consumed by the adaptation loop.

B ADDITIONAL PROTOCOL DETAILS

The primary protocol resets the adaptive model and optimizer before each target image and applies two gradient steps. The model stays in evaluation mode during affine adaptation. AdaBN is implemented separately by enabling BatchNorm statistics only during its adaptation pass. A continual RELIABILITY-GATED TTA variant carries the state through a fixed target order and is analyzed separately.

Spatially misaligned geometric augmentations are excluded from the reliability estimate. Weak and strong views alter intensity, noise, and blur while preserving pixel correspondence. Corruptions, adaptation views, and evaluation readout views use independent deterministic generators. The evaluation readout has its own fixed configuration and is not changed by cue ablations.

C CONFIGURATION AND EVIDENCE RECORD

Each completed run record identifies the exact MSCANet constructor, checkpoint hash, source and target splits, foreground semantics, preprocessing, optimiser, random seeds, adaptation configuration, fixed uncertainty readout, and software environment. Efficiency records use synchronized post-warm-up timing and peak allocated device memory. Dataset and model licences are documented separately from the executable configuration.

The main Dice comparison uses source-training seeds 7–9. The accompanying multi-checkpoint record lists the checkpoint SHA-256, exact per-case IDs, and each protocol's aggregate and paired changes. The matched-readout bootstrap, spatial gate audit, qualitative update maps, and anchor ablation use the seed-9 checkpoint only; they diagnose one checkpoint and are not presented as checkpoint-replication evidence.

D INCREMENTAL SENSITIVITY AND CORRUPTION RESULTS

The following artifacts were generated from 46 corrected JSON files after checking schema, checkpoint and configuration hashes, fixed readout metadata, case counts, unique case IDs, JSON–CSV alignment, case-level Dice means, and the remote SHA-256 manifest. The sensitivity rows use the seed-9 checkpoint only; the evaluator seed field is preserved in the provenance record and is not treated as an independent source-training replication.

Table 3: Seed-9 checkpoint sensitivity. Each row is a single fixed-order or case-wise run on one checkpoint (checkpoint SHA-256 $408e230e \ldots 611e2ef2$); evaluator seed metadata is retained in the JSON record and is not interpreted as an independent training seed. Dice is the mean over all target cases. The primary 50% case-wise Reliability-TTA configuration is shown for context; this table is a sensitivity analysis, not multi-checkpoint replication.

Domain	Variant	n	Dice ↑	Δ vs. default
DDTI	default case-wise, keep ratio 50%	637	0.74089	–
	keep ratio 25%	637	0.74143	+0.00055
	keep ratio 75%	637	0.74078	-0.00011
	keep ratio 100%	637	0.74087	-0.00002
	one update step	637	0.74140	+0.00051
	without entropy loss	637	0.74211	+0.00122
	without consistency loss	637	0.74089	-0.00000
	without boundary loss	637	0.74087	-0.00002
	without entropy cue	637	0.74094	+0.00005
	without augmentation cue	637	0.74088	-0.00001
	without boundary cue	637	0.74088	-0.00000
	continual stream	637	0.74625	+0.00536
STU	default case-wise, keep ratio 50%	42	0.82677	–
	keep ratio 25%	42	0.82678	+0.00001
	keep ratio 75%	42	0.82676	-0.00001
	keep ratio 100%	42	0.82703	+0.00026
	one update step	42	0.82666	-0.00011
	without entropy loss	42	0.82644	-0.00033
	without consistency loss	42	0.82677	-0.00000
	without boundary loss	42	0.82677	+0.00000
	without entropy cue	42	0.82682	+0.00006
	without augmentation cue	42	0.82677	-0.00000
	without boundary cue	42	0.82678	+0.00001
	continual stream	42	0.82797	+0.00120

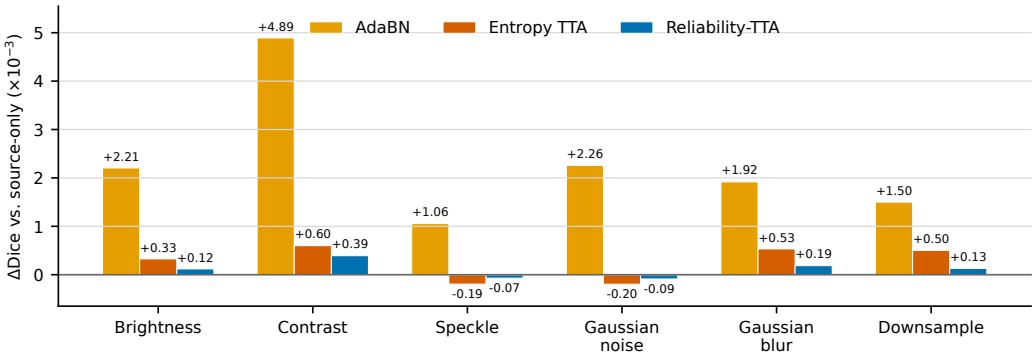

Figure 3: Verified severity-3 TN3K source-validation stress test. Bars show mean Dice relative to source-only inference under the same corruption for AdaBN, entropy-TTA, and Reliability-TTA. Each bar is computed from 3,493 cases on the seed-9 checkpoint; labels report the exact delta. This figure does not support claims about other severities or external target domains.

Table 4: TN3K source-validation stress test under severity-3 corruption (3,493 cases per row) using the seed-9 checkpoint. Values are mean Dice; Δ is relative to source-only inference under the same corruption. The corruption suite is a robustness stress test at one severity, not evidence for all corruption severities or external domains.

Corruption	Source-only		AdaBN		Entropy TTA		Reliability-TTA	
	Dice	Δ	Dice	Δ	Dice	Δ	Dice	Δ
Brightness	0.76741	–	0.76962	+0.00221	0.76774	+0.00033	0.76753	+0.00012
Contrast	0.73180	–	0.73669	+0.00489	0.73240	+0.00060	0.73219	+0.00039
Speckle noise	0.77975	–	0.78081	+0.00106	0.77956	-0.00019	0.77968	-0.00007
Gaussian noise	0.76462	–	0.76688	+0.00226	0.76443	-0.00020	0.76453	-0.00009
Gaussian blur	0.75093	–	0.75285	+0.00192	0.75146	+0.00053	0.75112	+0.00019
Downsampling	0.76700	–	0.76850	+0.00150	0.76751	+0.00050	0.76713	+0.00013

E SEED-9 MATCHED-READOUT AND SPATIAL DIAGNOSTICS

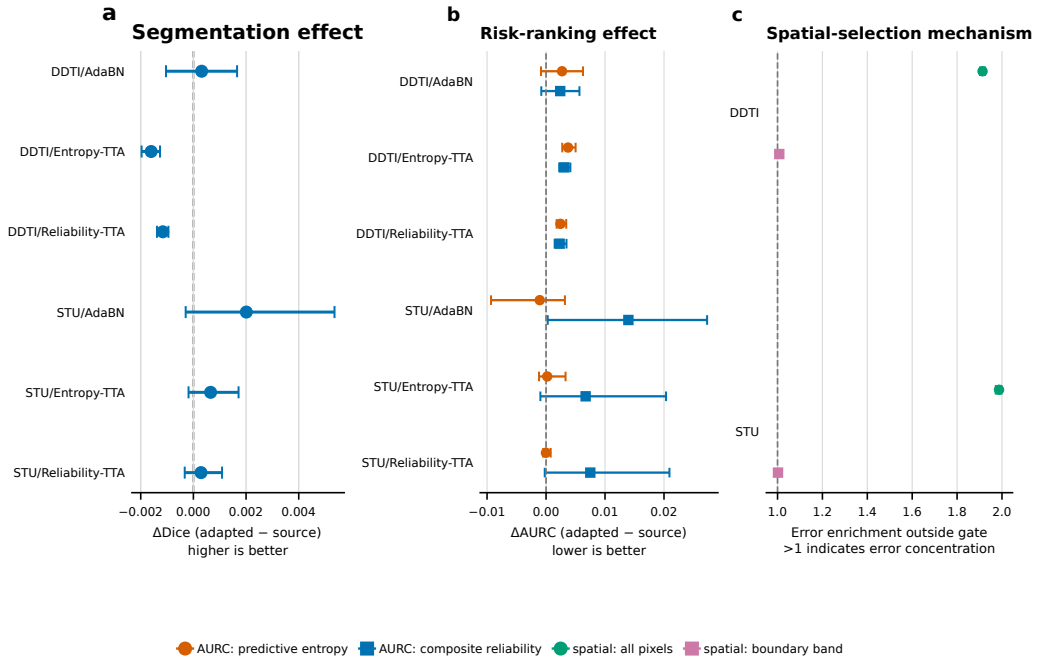

Figure 4: Seed-9 diagnostic results. (a) Case-paired Dice change relative to source-only inference. (b) Case-paired AURC change under predictive-entropy and composite-reliability readouts; lower is better. (c) Pixel- and target-boundary-error enrichment outside the actual Reliability-TTA update gate. Points and intervals use 10,000 case resamples within the seed-9 checkpoint; these intervals do not quantify source-training variability.

F VERIFIED UPDATE-MAP AUDIT

F.1 DDTI

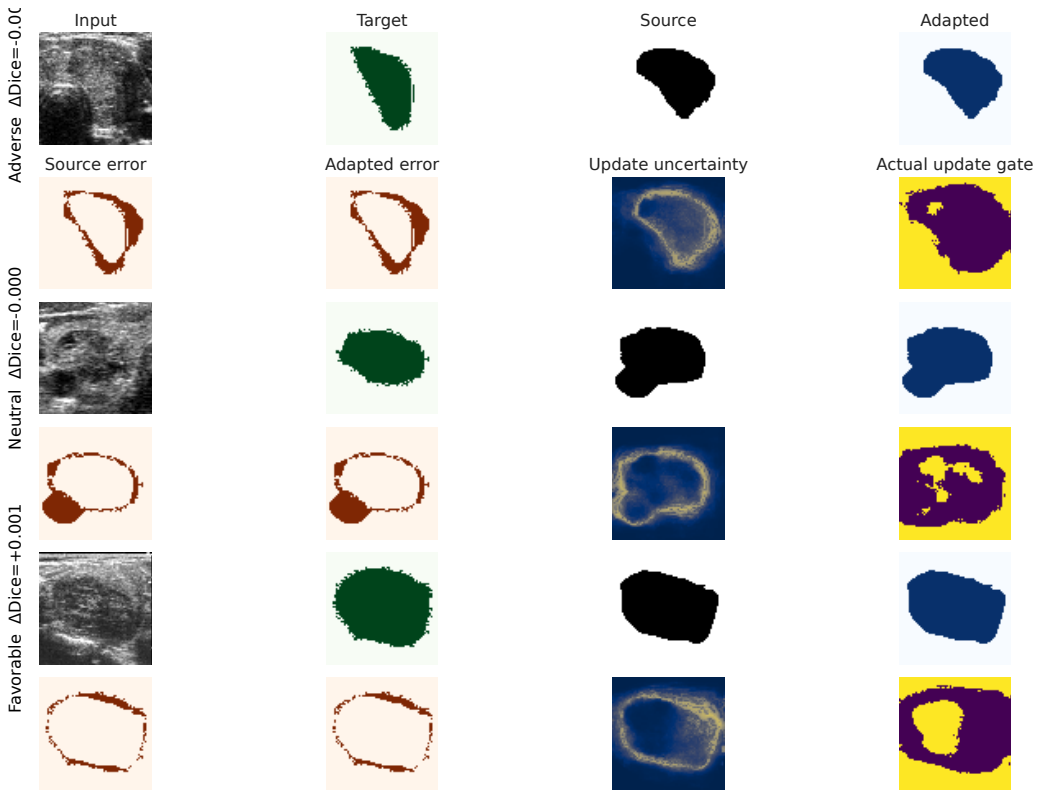

Figure 5: Verified DDTI qualitative audit using the actual gate and maps from the last successful Reliability-TTA update. Rows are selected post hoc from adverse, neutral, and favorable adapted-minus-source Dice strata. The gate is an inference-time artifact; target masks and error maps are shown only for post-hoc analysis.

F.2 STU

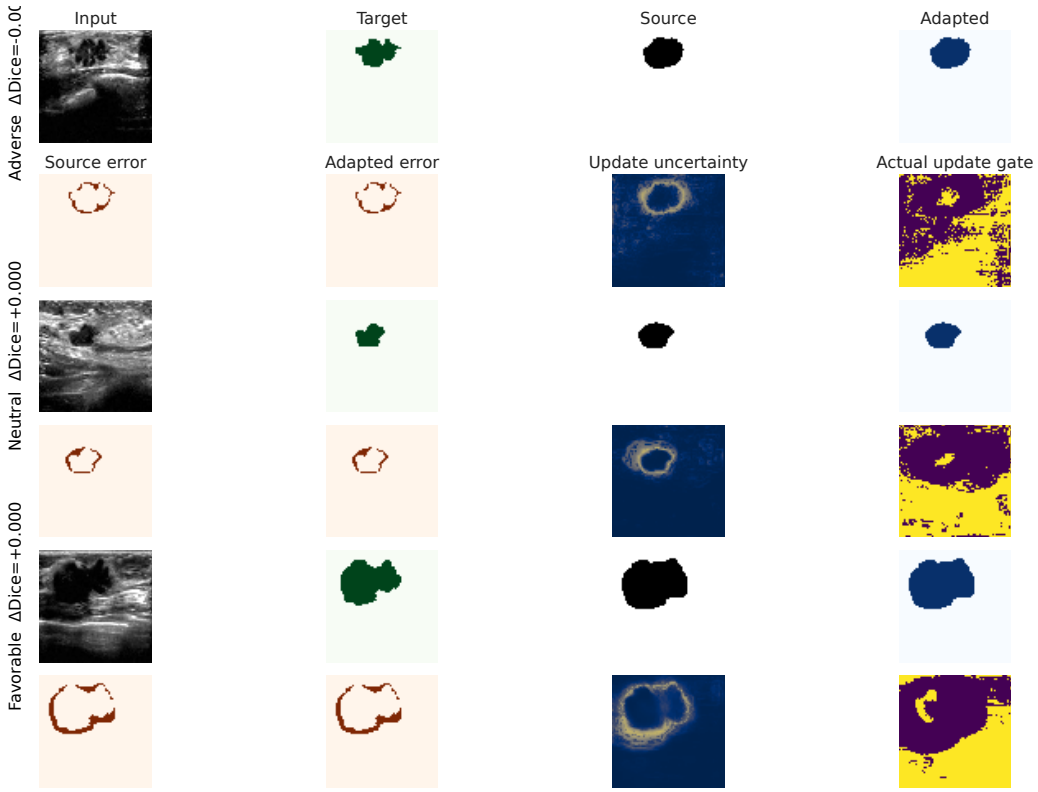

Figure 6: Verified STU qualitative audit using the actual gate and maps from the last successful Reliability-TTA update. Rows are selected post hoc from adverse, neutral, and favorable adapted-minus-source Dice strata. The gate is an inference-time artifact; target masks and error maps are shown only for post-hoc analysis.