SOURCE-FREE TEST-TIME ADAPTATION FOR BRAIN TUMOR SEGMENTATION BY JOINTLY HANDLING COVARIATE AND LABEL SHIFTS

Anonymous authors
Paper under double-blind review

ABSTRACT

In real deployment scenarios, covariate and label shifts typically co-occur, yet most Test-Time Adaptation (TTA) methods account only for the covariate shift. This limitation arises primarily because covariate shift adaptation assumes a stationary label distribution, whereas label shift correction presumes a fixed class-conditional input distribution. Consequently, neither approach is directly applicable when both shifts are simultaneously present. Moreover, the label shift is inferred from model predictions, while the covariate shift distorts these predictions, making accurate label shift correction feasible only after the covariate shift has been effectively mitigated. We further observe that the label shift is particularly detrimental in brain tumor segmentation, where the evaluation regions are hierarchically nested and any deviation in a class propagates to every region containing that class, thereby amplifying the shift in the predictions toward the source-domain label distribution. We term this effect *hierarchical label shift*. Based on the above analysis, we propose **H**ierarchical-**R**estoring **A**daptation and **C**orrection (HRAC), a source-free TTA framework that jointly handles the covariate shift and the hierarchical label shift in an *adapt-then-correct* manner by explicitly structuring adaptation and correction around the nested evaluation regions. Our HRAC comprises two synergistic components: (i) **S**tructure-**C**onsistency **A**daptation (SCA), which mitigates the covariate shift by restoring the spatial structure it corrupts, with consistency losses imposed across the nested regions; and (ii) **H**ierarchical **L**abel **S**hift **C**orrection (HLSC), which corrects the hierarchical label shift by applying a bounded logit bias to the violated region ratios. To the best of our knowledge, HRAC is the first source-free method to jointly mitigate the covariate shift and correct the hierarchical label shift without estimating a target prior. Extensive experiments on three BraTS benchmarks show that HRAC outperforms seven state-of-the-art adaptation baselines in mean Dice, mean HD95, and mean IoU, demonstrating its effectiveness and strong generalization in brain tumor segmentation.

1 INTRODUCTION

Semantic segmentation enables precise delineation of anatomical structures, including organs and pathological lesions, and serves as a foundational component for critical clinical decision-making, such as diagnosis and treatment response assessment in medical image analysis (Long et al., 2015; Ronneberger et al., 2015; Isensee et al., 2021). However, when a model trained on source-domain data is deployed to a target domain, distribution shift, especially across different imaging centers and scanner platforms, often leads to significant performance degradation (Guo et al., 2026; Zech et al., 2018). Conventional domain adaptation approaches typically rely on either access to the original source-domain data or offline collection of the complete target-domain dataset (Ganin et al., 2016; Tsai et al., 2018; Tranheden et al., 2021; Hoyer et al., 2023; Liu et al., 2021; Liang et al., 2020), prerequisites that are frequently unattainable due to stringent privacy regulations and the inherently sequential, case-by-case nature of clinical data acquisition. In contrast, test-time adaptation (TTA) has recently emerged as a promising paradigm for mitigating distribution shift without requiring prior access to source data or complete target data (Wang et al., 2021; Liang et al., 2025).

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.

Existing TTA methods address distribution shift by leveraging diverse optimization signals, such as entropy minimization (Wang et al., 2021), anti-forgetting–guided sample selection (Niu et al., 2022), and flatness-aware regularization (Niu et al., 2023). While most of these signals were originally developed on natural images, their medical segmentation counterparts have introduced domain-specific enhancements, including evaluation-guided adaptation (Zhou et al., 2025), style-modulated feature alignment (Wang et al., 2025), and hypothesis-driven safety mechanisms (Jhawar & Wang, 2026). However, none of these approaches explicitly incorporate class composition into their objective functions. Instead, they implicitly assume a stationary class prior. Consequently, they are fundamentally ill-equipped to handle label shift, a scenario where the marginal label distribution changes while the conditional input distribution given each class remains invariant. Under label shift, such methods treat the model's biased predictions as reliable supervision signals, rendering them incapable of correcting compositional deviations in the label space and potentially exacerbating them.

This gap is particularly pronounced in brain tumor segmentation. The tumor biology of adult and pediatric patients differs systematically (Mackay et al., 2017; Louis et al., 2021; Ostrom et al., 2021), leading to markedly distinct sub-region compositions between adult glioma cohorts (Baid et al., 2021; Adewole et al., 2026) and pediatric brain tumor cohorts (Kazerooni et al., 2024), as illustrated in Fig. 1(a). Segmentation performance in this domain is evaluated across three hierarchically nested regions: the whole tumor (WT) $\supseteq$ the tumor core (TC) $\supseteq$ the enhancing tumor (ET) (Baid et al., 2021). This nested anatomical–functional relationship implies that shifts in the prior distributions of mutually exclusive tissue classes propagate across hierarchical levels, inducing a systematic change in regional composition, which we term *hierarchical label shift*. Thus, any deviation in the proportion of a foreground tissue class propagates upward to all encompassing regions. Given that foreground tissues constitute only a small fraction of each volumetric scan (Fig. 1(b)), even minor changes in class proportions yield substantial relative changes in region volumes.

The classical approach to addressing label shift is prior-ratio correction, which adjusts the model's predicted posterior probabilities by weighting them with the ratio of target-domain to source-domain class priors (Saerens et al., 2002; Lipton et al., 2018). However, source domain data is unavailable at test time. Therefore, the target-class priors must be estimated solely from the model's own predictions, typically via bias rectification of the classifier, ensemble-based calibration, or empirical frequency counting of predicted classes (Liu et al., 2022; Ma et al., 2022; Weerasinghe et al., 2026). The rarer a foreground class is, the fewer predictions it receives, and the noisier its prior estimate becomes. Critically, estimation error for any foreground class propagates to all regions containing that class and is substantially amplified for rare classes.

This label shift seldom occurs in isolation during deployment. Correcting it typically relies on estimating the underlying label distribution from the model's predictions. However, covariate shift distorts those predictions, thereby compromising the reliability of such estimates. Consequently, accurate estimation of the label composition is feasible only after covariate shift has been adequately mitigated. Existing approaches that aim to address both shifts simultaneously (Jang & Chung, 2025; Park et al., 2023) effectively circumvent their interdependence rather than resolving it. Covariate-shift adaptation presumes a stationary label distribution, whereas label-shift correction assumes that model predictions remain unbiased and undistorted. Performing adaptation first entrenches an inaccurate label composition, while performing correction first yields erroneous estimates due to reliance on distorted predictions. Thus, either sequential strategy fails in the presence of concurrent shifts, regardless of execution order. The core challenge lies in the mutual dependence of the two procedures: both rely on the same set of predictions, yet each type of shift corrupts precisely the signal required by the other.

To this end, we propose **H**ierarchical-**R**estoring **A**daptation and **C**orrection (HRAC), a source-free TTA framework that jointly handles both covariate and label shifts in an *adapt-then-correct* manner, without requiring estimation of the target prior distribution. Our HRAC comprises two synergistic components. First, **S**tructure-**C**onsistency **A**daptation (SCA) is introduced to mitigate covariate shift by restoring the spatial structural integrity degraded by domain shift. This is achieved through two novel consistency losses imposed across nested image regions, neither relying on entropy minimization nor pseudo-labeling. Second, **H**ierarchical **L**abel **S**hift **C**orrection (HLSC) is deployed *after* adaptation concludes: it compares region-wise composition ratios against pre-specified, loose one-sided bounds and applies a bounded logit bias only to those ratios violating the bounds, requiring neither source data, target labels, nor knowledge of the target prior. Consequently, SCA removes the structural impediments that hinder effective label shift correction, while HLSC rectifies com-

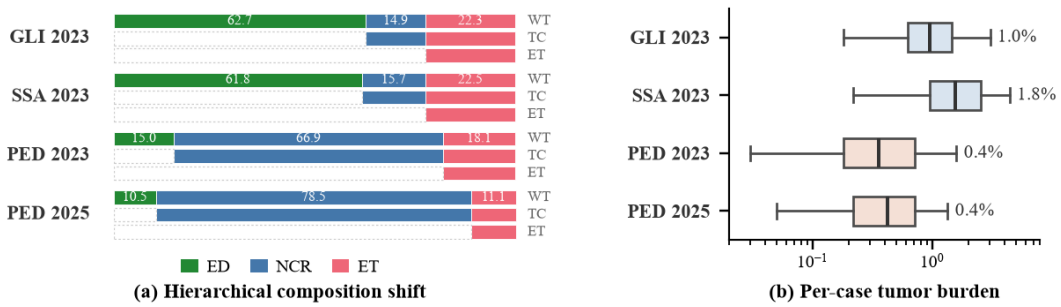


Figure 1: **Label distribution across the BraTS series**. (a) Ground-truth composition per dataset, where the tumor core (TC) and enhancing tumor (ET) bars nest inside the whole tumor (WT) bar, which is split into edema (ED), necrotic or non-enhancing core (NCR), and ET. The number inside each segment gives its share in percent. (b) Per-case tumor burden, measured as the percentage of voxels in the foreground and shown on a log scale, with boxes spanning the interquartile range and tints separating the adult glioma datasets from the pediatric brain tumor datasets.

position deviations that adaptation alone cannot resolve. By orchestrating these two mechanisms, HRAC jointly tackles both shifts within the source-free TTA setting, thereby closing a critical gap left unaddressed by existing methods. In summary, our main contributions are as follows:

- We formulate the co-occurrence of the covariate shift and hierarchical label shift as a prediction-composition confounding problem for brain tumor segmentation, where covariate-induced distortion obscures genuine composition change. To address this challenge, we propose to develop a source-free TTA framework by separating prediction restoration from composition correction for jointly handling the covariate and hierarchical label shifts.
- We propose HRAC with two components, SCA and HLSC, in an *adapt-then-correct* manner to structure around the nested evaluation regions. SCA mitigates covariate shift using structural self-consistency, while HLSC corrects hierarchical label shift with bounded logit biases on violated region ratios, without estimating the target prior. To the best of our knowledge, HRAC is the first source-free TTA method to jointly address both shifts.
- Extensive experiments across three BraTS datasets show that HRAC consistently outperforms seven adaptation baselines in mean Dice, mean HD95, and mean IoU. The gains grow with the magnitude of the cross-dataset distribution shift, demonstrating its effectiveness and cross-population generalization in brain tumor segmentation.

2 Related Work

2.1 Test-Time Adaptation for Brain Tumor Segmentation

Source data are typically unavailable at deployment in brain tumor segmentation. Test-time adaptation addresses this constraint by updating the deployed model in a single pass over the test stream (Liang et al., 2025). Recent methods introduce evaluation guidance that dynamically steers adaptation (Zhou et al., 2025), style modulation that aligns cross-domain appearance (Wang et al., 2025), and hypothesis-driven safeguards that stabilize the model updates (Jhawar & Wang, 2026) to tune the deployed model itself. Cache-driven spatial adaptation also restores source weights and ensembles multi-head predictions (Li et al., 2024), and external priors from a pre-trained segmentation model further inform the adaptation (Wu et al., 2026). These designs improve the generalization of the adapted model, yet the gains are obtained without making use of the hierarchical structure that the nested evaluation regions provide.

2.2 Distribution Shift Handling

Covariate shift adaptation. Covariate shift changes the class-conditional densities $p(\mathbf{x} \mid y)$. TTA methods use entropy minimization (Wang et al., 2021), anti-forgetting selection (Niu et al., 2022), and flatness-aware stabilization (Niu et al., 2023), signals that act on per-voxel confidences or on feature statistics, while generative domain adaptation restyles the source images toward the target appearance (Zhu et al., 2017; Hoffman et al., 2018; Vu et al., 2019; Yang & Soatto, 2020; Hu et al., 2024). However, both paradigms treat each voxel independently and thus fail to explicitly model inter-voxel dependencies, leaving the underlying spatial structure unaccounted for.

Label shift correction. Label shift changes the class marginal $p(y)$. Prior-ratio correction reweights outputs by the source–target prior ratio, estimated from the source by EM (Saerens et al., 2002) or from confusion matrices (Lipton et al., 2018), and is unavailable once the source is withheld. Output-distribution correction takes the same route, reweighting the predicted composition as in Undoing_UDA (Liu et al., 2022) or applying post-hoc logit correction as in BLDA (Li et al., 2025), online test-time correction estimates the target prior from the running predictions (Jang & Chung, 2025; Ma et al., 2022; Weerasinghe et al., 2026), and training-time remedies reshape the composition instead, by re-weighting (Cui et al., 2019; Lin et al., 2017; Cao et al., 2019), re-sampling (He & Garcia, 2009), or logit adjustment (Menon et al., 2021). Together, these corrections rebalance the predicted class proportions toward the target ones. Nevertheless, none of them handles the covariate shift and the label shift at once.

Although existing methods have made this progress in mitigating the label shift and the covariate shift, they have treated the two shifts separately, on the covariate side by the signals above and on the label side by output-distribution and prior-based correction, and the adapters that address both shifts together rely on auxiliary tasks or source-domain information (Park et al., 2023). Therefore, *how to leverage hierarchical signals as auxiliary information to jointly address both shifts in the source-free setting remains an open problem.*

3 Methodology

3.1 Preliminary

Problem definition. In TTA for brain tumor segmentation, a segmentation network f_θ is pre-trained on a labeled source domain $\mathcal{D}_S = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{N_s}$, with $\mathbf{y}_i^s$ the voxel-wise label of the volume $\mathbf{x}_i^s$, and then adapted online to an unlabeled target stream $\mathcal{D}_T = \{\mathbf{x}_i^t\}_{i=1}^{N_t}$ with data distribution p_T. The network outputs per-voxel logits $\mathbf{z} = f_\theta(\mathbf{x}^t)$.

The classical label correction. The classical remedy for the label component adjusts the logits by the ratio of the target and source priors (Saerens et al., 2002; Lipton et al., 2018), where $\boldsymbol{\pi}_S$ and $\boldsymbol{\pi}_T$ denote the class marginals of the source and target domains and $\pi_S(c)$, $\pi_T(c)$ the marginals of class c. To analyze it and the correction of Section 3.4, we introduce a constant class bias $\mathbf{b} \in \mathbb{R}^4$ applied to every voxel, which gives the biased predictor $p_{v,c}(\mathbf{b}) = \mathrm{softmax}(z_v + \mathbf{b})_c$ and the target risk by:

$$\mathcal{L}_T(\mathbf{b}) = \mathbb{E}_{\mathbf{x}\sim p_T}\, \mathbb{E}_{y\sim p_T(y|\mathbf{x})}\big[-\log p_{\mathbf{x},y}(\mathbf{b})\big], \tag{1}$$

where $\hat{\pi}_c(\mathbf{b}) = \mathbb{E}_{\mathbf{x}\sim p_T} \frac{1}{|\mathcal{V}|} \sum_v p_{v,c}(\mathbf{b})$ is the predicted marginal and $\mathbf{g} \triangleq \hat{\boldsymbol{\pi}}(\mathbf{0}) - \boldsymbol{\pi}_T$ is the marginal error. Setting $b_c^* = \log(\pi_T(c)/\pi_S(c))$ reproduces the target posterior under pure label shift and attains the infimum of $\mathcal{L}_T$ over constant biases.

Definition 1. ***Hierarchical label shift.*** *Let $\boldsymbol{\pi}_S$ and $\boldsymbol{\pi}_T$ be two class marginals with $\boldsymbol{\pi}_S \neq \boldsymbol{\pi}_T$, and let $\Delta_c \triangleq \pi_T(c) - \pi_S(c)$. The deviation induced on region R can be expressed as:*

$$\Delta m_R = \sum_{c\in R} \Delta_c, \qquad R \in \{\mathrm{ET}, \mathrm{TC}, \mathrm{WT}\}, \tag{2}$$

where m_R is the region marginal. We term hierarchical label shift the case in which a predictor trained under $\boldsymbol{\pi}_S$ is evaluated by region-level scores while its predictions remain biased toward $\boldsymbol{\pi}_S$ at every level of the nesting, so that the deviation of each class in equation 2 enters every region containing it.

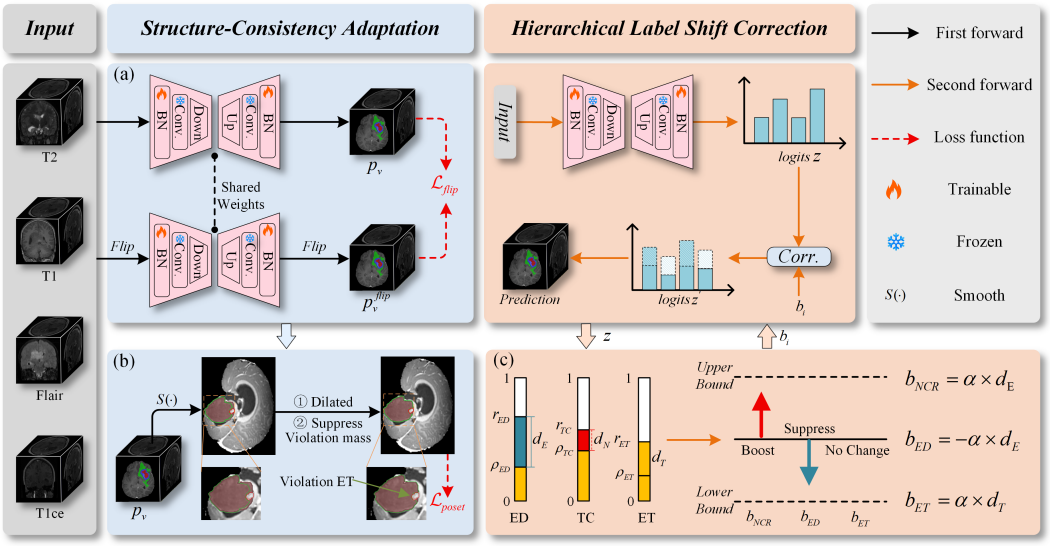


Figure 2: The pipeline of the proposed HRAC. (a) *Flip consistency*. The flipped volume serves as a teacher for the online branch. (b) *Partial-order consistency*. Penalize predictions that fall outside the delineated region. (c) *Bounded correction*. It combines the three intra-foreground ratios with the one-sided structural prior and applies a bounded bias to the final prediction.

Adaptation before correction. Any label-shift signal has to be read from the predictions, whereas the covariate component distorts exactly these predictions, so the covariate-induced distortion obscures the genuine composition change that the correction has to measure. This prediction-composition confounding fixes the order of operations, adaptation before correction, while the fragility of prior estimation rules out the classical route altogether, so the correction must be driven by something other than an estimated π_T. Therefore, Section 3.4 drives the correction with structural deficits alone.

3.2 Overview

To address the prediction-composition confounding formulated above, we propose **H**ierarchical-**R**estoring **A**daptation and **C**orrection (HRAC), a source-free TTA framework that handles the covariate shift and the hierarchical label shift in an *adapt-then-correct* manner by structuring adaptation and correction around the nested evaluation regions, as illustrated in Fig. 2. HRAC proceeds in two stages. The first stage (Section 3.3) restores the spatial structure that the covariate component corrupts, using structural consistency alone. The second stage (Section 3.4) then corrects the label shift, applying a bounded correction driven by the violated bounds. Algorithm 1 in Appendix B gives the full procedure. We next detail the two key components that drive our framework.

3.3 Structure-Consistency Adaptation

Generally, the covariate component manifests itself in the predictions as a loss of spatial structure, so that sub-regions are predicted outside their containers and the nesting that the evaluation protocol presupposes no longer holds. Based on the above analysis, we propose Structure-Consistency Adaptation (SCA) to restore the structure by driving adaptation with structural consistency alone. As illustrated in Figs. 2(a) and (b), it comprises two label-free consistency objectives, the flip consistency on the bilateral symmetry of the brain and the partial-order consistency across the nested evaluation regions, and updates only the BatchNorm affine parameters of f_θ. These parameters form a low-dimensional subspace, which stabilizes optimization under label-free objectives.

Flip consistency. The brain is approximately symmetric about the mid-sagittal plane, which makes the left-right flipped volume a label-free augmentation. Let p^{flip} denote the prediction on the flipped volume and then flipped back to the original geometry. We distill the online prediction toward it by:

$$\mathcal{L}_{\text{flip}} = \frac{1}{|\mathcal{V}|} \sum_{v} KL\big(p_v^{\text{flip}} \,\big\|\, p_v\big), \tag{3}$$

where $KL(\cdot \,\|\, \cdot)$ denotes the Kullback-Leibler divergence between the per-voxel categorical distributions. The two views describe the same anatomy, so the distillation penalizes any prediction whose spatial layout is inconsistent with its own mirror view, restoring the bilateral structure that the covariate component distorts.

Partial-order consistency. Because the nesting relation holds between regions, SCA enforces it indirectly through voxel-wise masks built from the per-voxel probabilities $p_{v,c}$ of the adapted model:

$$M_{\text{ET}} = \sigma\Big(\tfrac{p_{\text{ET}}-\tau}{T}\Big), \quad M_{\text{TC}} = \sigma\Big(\tfrac{p_{\text{TC}}-\tau}{T}\Big), \quad M_{\text{WT}} = \sigma\Big(\tfrac{p_{\text{WT}}-\tau}{T}\Big), \tag{4}$$

where $p_{\text{TC}} = p_{\text{NCR}} + p_{\text{ET}}$ and $p_{\text{WT}} = p_{\text{NCR}} + p_{\text{ED}} + p_{\text{ET}}$, $\sigma(\cdot)$ is the logistic function, $\tau = 0.5$ is the mask threshold, and the temperature T is annealed from 0.1 towards a hard-mask limit. Formally, let $\tilde{M}_i = \text{MaxPool}_3(M_i)$ denote a $3 \times 3 \times 3$ max-pooling dilation of the container mask, and define the violation rate as:

$$\mathcal{L}_{\text{poset}} = \frac{\sum_v M_{\text{ET},v}\big(1 - \tilde{M}_{\text{TC},v}\big) + \sum_v M_{\text{TC},v}\big(1 - \tilde{M}_{\text{WT},v}\big)}{\sum_v M_{\text{ET},v} + \sum_v M_{\text{TC},v} + \epsilon}, \tag{5}$$

where $\epsilon > 0$ is a numerical constant. The dilation absorbs the boundary slack of the soft masks, so that the penalty fires only on mass lying genuinely outside its container.

Adaptation objective. The overall training objective is defined as a weighted sum of two loss terms:

$$\mathcal{L}_{\text{SCA}} = \lambda_{\text{poset}}\, \mathcal{L}_{\text{poset}} + \lambda_{\text{cons}}\, \mathcal{L}_{\text{flip}}, \qquad \lambda_{\text{poset}} = 1, \;\; \lambda_{\text{cons}} = 10, \tag{6}$$

The choice of λ_{cons} is discussed in Appendix C.3. Each volume is adapted with Adam, updating only the BatchNorm affine parameters. This minimal capacity rescales and shifts the normalized features, so it absorbs the covariate component while keeping the learned weights fixed.

3.4 Hierarchical Label Shift Correction

Building on the adaptation by SCA, we further propose Hierarchical Label Shift Correction (HLSC) to correct the label shift by measuring the genuine composition change that the prediction-composition confounding obscures. As illustrated in Fig. 2(c), it comprises a bounded correction built on the structural ratios of the adapted predictions and the one-sided structural deficits against a fixed structural prior, and applies the correction to the final prediction only.

Structural ratios. With the per-voxel probabilities $p_{v,c}$ of the adapted model, we define the marginals $P_{\text{ED}} = \sum_v p_{v,\text{ED}}$, $P_{\text{ET}} = \sum_v p_{v,\text{ET}}$, $P_{\text{WT}} = \sum_v p_{v,\text{WT}}$, and $P_{\text{TC}} = \sum_v p_{v,\text{TC}}$, and the three intra-foreground ratios:

$$r_{\text{TC}} = \frac{P_{\text{TC}}}{P_{\text{WT}}}, \qquad r_{\text{ED}} = \frac{P_{\text{ED}}}{P_{\text{WT}}}, \qquad r_{\text{ET}} = \frac{P_{\text{ET}}}{P_{\text{TC}}}, \tag{7}$$

where r_{TC} is the share of the core within the whole tumor, r_{ED} the share of edema within the whole tumor, and r_{ET} the share of ET within the core. The ratios read the class composition at every level of the nesting independently of the foreground size, and are the quantities that the hierarchical label shift of Definition 1 moves.

One-sided structural deficits. We compare the ratios against a structural composition prior $\boldsymbol{\rho} = (\rho_{\text{TC}}, \rho_{\text{ED}}, \rho_{\text{ET}}) = (0.4, 0.6, 0.6)$, a fixed structural bound rather than an estimate of the composition of any population (Appendix C.2), and define the deficits as follows:

$$d_{\text{N}} = \frac{[\rho_{\text{TC}} - r_{\text{TC}}]_+}{\rho_{\text{TC}}}, \qquad d_{\text{E}} = \frac{[r_{\text{ED}} - \rho_{\text{ED}}]_+}{1 - \rho_{\text{ED}}}, \qquad d_{\text{T}} = \frac{[\rho_{\text{ET}} - r_{\text{ET}}]_+}{\rho_{\text{ET}}}, \tag{8}$$

where $[t]_+ = \max(t, 0)$, $\boldsymbol{\rho}$ is enforced one-sidedly, so that any ratio meeting its bound receives no correction. The bounds double as normalizers, bringing the three violations into the correction as relative deviations on a common scale.

Table 1: Comparison with the state-of-the-art TTA methods on the BraTS 2023 PED, BraTS 2025 PED, and BraTS 2023 SSA datasets. **Bold** and underlined values denote the best and second-best results in each column within each dataset.

Dataset	Method	Dice (%) ↑				HD95 (mm) ↓				IoU (%) ↑			
		WT	TC	ET	Mean	WT	TC	ET	Mean	WT	TC	ET	Mean
BraTS 2023 PED	w/o adaptation	72.11	32.88	47.44	50.81	29.70	117.72	118.09	88.50	62.13	25.42	40.15	42.57
	Tent	81.08	36.05	54.15	57.09	15.39	82.30	85.70	61.13	71.65	27.45	45.86	48.32
	EATA	80.62	39.45	52.54	57.54	12.05	64.29	73.51	49.95	70.89	29.92	44.37	48.39
	Undoing_UDA	73.60	36.67	45.02	51.76	22.50	76.07	91.63	63.40	63.60	27.97	36.24	42.60
	SAR	80.37	37.69	54.11	57.39	15.85	67.73	82.44	55.34	70.43	28.43	45.56	48.14
	TEGDA	**82.55**	39.10	**55.01**	58.89	8.90	63.48	81.67	51.35	**73.37**	29.71	**46.59**	49.89
	SmaRT	75.39	43.32	49.22	55.98	20.09	39.30	55.60	38.33	63.48	32.62	39.97	45.36
	BLDA	73.48	36.81	46.20	52.16	37.76	69.61	102.58	69.98	63.85	27.87	37.89	43.20
	HRAC (Ours)	82.19	**50.40**	51.87	**61.49**	**8.80**	**32.45**	**47.59**	**29.61**	72.89	**38.15**	42.88	**51.31**
BraTS 2025 PED	w/o adaptation	52.68	20.63	36.09	36.47	55.76	102.31	141.87	99.98	42.95	15.13	31.15	29.74
	Tent	66.12	18.88	**47.53**	44.18	**37.83**	129.62	122.98	96.81	**56.57**	13.58	**42.28**	37.48
	EATA	62.84	25.16	38.08	42.03	53.59	74.45	129.31	85.78	50.55	17.82	32.36	33.58
	Undoing_UDA	62.16	19.78	38.13	40.02	47.82	108.21	132.42	96.15	52.07	13.89	33.27	33.08
	SAR	62.46	24.74	32.64	39.95	52.99	71.67	151.34	92.00	50.45	17.53	27.11	31.70
	TEGDA	**67.75**	25.93	41.58	45.09	47.70	75.42	**121.24**	81.45	56.30	18.84	35.68	36.94
	SmaRT	57.19	19.13	31.76	36.03	44.56	89.81	148.37	94.25	45.37	13.63	26.22	28.41
	BLDA	50.45	18.70	23.94	31.03	59.96	66.42	152.91	93.10	41.04	12.65	19.17	24.29
	HRAC (Ours)	67.40	**45.78**	30.90	**48.03**	49.24	**45.25**	141.68	**78.72**	55.08	**33.74**	24.82	**37.88**
BraTS 2023 SSA	w/o adaptation	69.64	58.34	49.57	59.18	24.65	39.81	57.04	40.50	57.87	48.27	39.41	48.52
	Tent	87.66	72.95	66.42	75.68	14.45	17.69	27.04	19.73	79.92	63.41	55.45	66.26
	EATA	86.12	71.86	66.04	74.67	**10.86**	18.16	32.97	20.66	77.80	62.12	54.95	64.96
	Undoing_UDA	77.79	61.48	49.40	62.89	19.24	28.28	47.55	31.69	67.96	50.56	37.44	51.99
	SAR	85.71	72.20	65.33	74.41	11.37	**13.83**	27.46	17.55	77.68	62.44	54.25	64.79
	TEGDA	**87.72**	74.01	68.17	76.63	15.39	17.53	26.43	19.78	**80.26**	64.67	57.47	67.47
	SmaRT	76.74	70.67	66.29	71.23	29.49	23.66	19.59	24.25	65.20	59.85	52.94	59.33
	BLDA	78.32	65.90	61.52	68.58	18.90	42.32	45.53	35.58	68.96	56.61	50.04	58.54
	HRAC (Ours)	86.81	**74.87**	**70.28**	**77.32**	15.20	18.31	**14.55**	**16.02**	78.73	**65.39**	**59.30**	**67.81**

Additive logit correction. Based on the deficits, the per-volume additive logit correction can be computed as follows:

$$\mathbf{b} = \alpha \left(0,\, d_{\mathrm{N}},\, -d_{\mathrm{E}},\, d_{\mathrm{T}}\right), \qquad \mathbf{z}' = \mathbf{z} + \mathbf{b}, \qquad \hat{\mathbf{y}}_v = \arg\max_c \, \mathrm{softmax}(\mathbf{z}'_v)_c. \tag{9}$$

The correction re-aligns the predicted composition with the structural prior in a single gradient-free pass, firing only where a deficit is nonzero. The prior $\boldsymbol{\rho}$ and the scale α are discussed in Appendix C.2.

4 Experiments

4.1 Experimental Setup

Datasets. Following the standard TTA protocol, we evaluate our method on four BraTS datasets. The source domain is the BraTS 2023 adult glioma (GLI, 1251 labeled cases) training set, and the unlabeled targets are the BraTS 2023 pediatric (PED, 99 cases), BraTS 2025 pediatric (PED, 257 cases), and BraTS 2023 Sub-Saharan Africa (SSA, 60 cases) cohorts, which differ from the source in population and acquisition. Each case has four standard MRI modalities (T1, T1ce, T2, Flair).

Implementation Details. The backbone is a 3D U-Net with feature widths $(32, 64, 128, 256)$ and dropout 0.3. We compare HRAC against the unadapted source model and seven test-time adaptation methods, including Tent (Wang et al., 2021), EATA (Niu et al., 2022), SAR (Niu et al., 2023), Undoing_UDA (Liu et al., 2022), TEGDA (Zhou et al., 2025), SmaRT (Wang et al., 2025), and BLDA (Li et al., 2025), all re-run by us on the same pipeline. We report Dice, HD95, and IoU for WT, TC, and ET, with the Mean column averaging performance across the three regions. Details

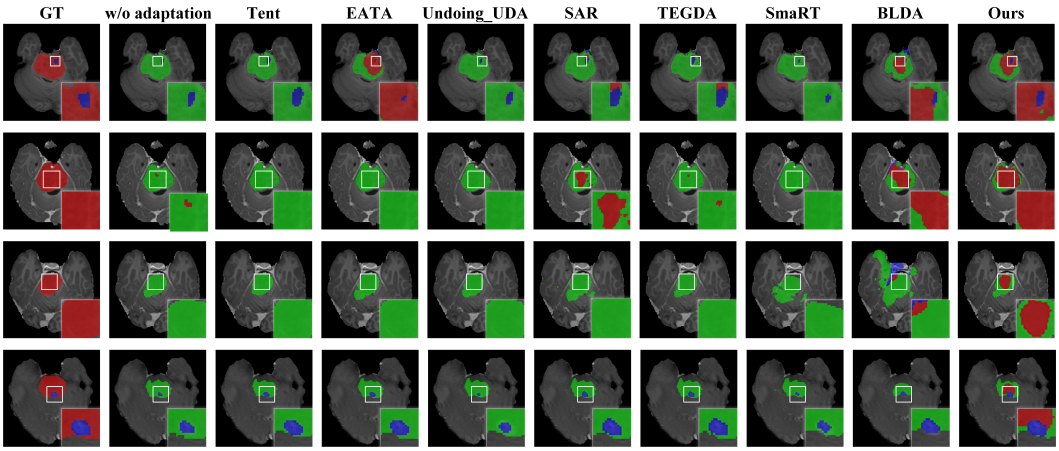


Figure 3: Qualitative comparison on four target volumes from the BraTS 2023 PED dataset. The columns show the ground truth, the unadapted source model, the seven baselines, and HRAC, with the necrotic core in red, edema in green, and the enhancing tumor in blue. In all cases, baseline methods misclassify the lesion as edema, whereas HRAC successfully identifies and corrects it.

Table 2: Component ablation of the proposed HRAC on the BraTS 2023 PED dataset. **Bold** and underlined values denote the best and second-best results in each column, respectively.

SCA	HLSC	Dice (%) ↑				HD95 (mm) ↓				IoU (%) ↑			
		WT	TC	ET	Mean	WT	TC	ET	Mean	WT	TC	ET	Mean
		72.11	32.88	47.44	50.81	29.70	117.72	118.09	88.50	62.13	25.42	40.15	42.57
✓		**83.05**	42.62	**55.63**	60.43	9.04	48.60	78.07	45.24	**73.77**	32.31	**46.99**	51.02
	✓	67.65	41.39	46.56	51.87	40.01	56.29	108.04	68.11	57.74	31.31	39.20	42.75
✓	✓	82.19	**50.40**	51.87	**61.49**	**8.80**	**32.45**	**47.59**	**29.61**	72.89	**38.15**	42.88	**51.31**

of source training, the BraTS 2025 PED class mapping, and all hyperparameters are provided in Appendix B.2. No hyperparameter is tuned using target labels.

4.2 Comparison with State-of-the-Art Methods

Quantitative evaluation results. Table 1 shows that our HRAC achieves the highest mean Dice on all three target datasets. Compared with the best baseline, *i.e.*, TEGDA (Zhou et al., 2025), the gains are 0.69, 2.60, and 2.94 points on BraTS 2023 SSA, BraTS 2023 PED, and BraTS 2025 PED, respectively. The improvements are most pronounced for TC, where HRAC raises Dice by 0.86, 7.08, and 19.85 points over the best baseline on the three datasets. In contrast, WT performance remains close to the best baseline. This pattern indicates that the overall gain mainly comes from correcting the under-predicted tumor core rather than uniformly enlarging the predicted tumor region. We also observe that HRAC achieves the lowest mean HD95 on all three datasets, reducing it from 17.55 to 16.02 mm on BraTS 2023 SSA, from 38.33 to 29.61 mm on BraTS 2023 PED, and from 81.45 to 78.72 mm on BraTS 2025 PED. Mean IoU shows the same overall trend, with HRAC attaining the best result on all three datasets. Together, these results demonstrate that HRAC improves both regional overlap and boundary accuracy across different target shifts.

Qualitative evaluation results. Figure 3 compares HRAC with the baselines on four cases. All methods localize the lesion, while the main difference lies in the predicted class composition. Rows 1 to 3 are core-dominant cases where the ground truth contains little or no edema. The unadapted model and most baselines fill the lesion with edema and leave the necrotic core unpredicted, reflecting the under-prediction of NCR and the over-prediction of ED caused by the prediction-composition confounding of covariate shift and hierarchical label shift. A few baselines recover part of the core unstably with loose or over-extended masks. In contrast, HRAC recovers a necrotic core consistent with the ground truth with only a thin edema rim and preserves the enhancing component in Row 1. Row 4 is enhancing-dominant and HRAC reproduces the enhancing and surrounding necrotic

Table 3: Plug-in effect of the HLSC module on three TTA backbones on the BraTS 2023 PED dataset. Bold marks the better value within each backbone pair.

Method	Dice (%) ↑				HD95 (mm) ↓				IoU (%) ↑			
	WT	TC	ET	Mean	WT	TC	ET	Mean	WT	TC	ET	Mean
Tent	**81.08**	36.05	**54.15**	57.09	**15.39**	82.30	85.70	61.13	**71.65**	27.45	**45.86**	**48.32**
Tent + HLSC	79.43	**40.38**	51.56	**57.12**	19.45	**59.94**	**83.17**	**54.19**	69.53	**30.42**	43.11	47.69
SAR	**80.37**	37.69	**54.11**	57.39	15.85	67.73	82.44	55.34	**70.43**	28.43	**45.56**	48.14
SAR + HLSC	78.66	**46.52**	51.34	**58.84**	**13.85**	**33.75**	**52.51**	**33.37**	68.45	**34.91**	42.61	**48.66**
EATA	**80.62**	39.45	**52.54**	57.54	**12.05**	64.29	73.51	49.95	**70.89**	29.92	**44.37**	48.39
EATA + HLSC	79.43	**46.54**	50.60	**58.86**	18.43	**44.97**	**59.06**	**40.82**	69.53	**35.19**	41.87	**48.86**

Table 4: Internal ablation of the structure-consistency objectives within SCA on the BraTS 2023 PED dataset. Poset denotes the partial-order consistency loss and Flip the flip-consistency loss.

Poset	Flip	Dice (%) ↑				HD95 (mm) ↓				IoU (%) ↑			
		WT	TC	ET	Mean	WT	TC	ET	Mean	WT	TC	ET	Mean
		72.11	32.88	47.44	50.81	29.70	117.72	118.09	88.50	62.13	25.42	40.15	42.57
✓		82.73	41.53	53.84	59.37	9.33	48.97	81.93	46.74	73.47	31.51	45.13	50.04
	✓	**83.07**	41.38	54.52	59.66	**8.95**	48.96	81.73	46.55	**73.78**	31.33	46.24	50.45
✓	✓	83.05	**42.62**	**55.63**	**60.43**	9.04	**48.60**	**78.07**	**45.24**	73.77	**32.31**	**46.99**	**51.02**

components while the baselines still predict the periphery as edema. HRAC also keeps the mask compact, whereas BLDA and SmaRT over-extend into the normal tissue in Rows 3 and 4 and BLDA leaves a detached component. Overall, HRAC yields the composition closest to the ground truth, validating the adapt-then-correct framework.

4.3 Ablation Experiments

Component ablation. To verify the effectiveness of SCA and HLSC, we conduct a combined ablation study on the BraTS 2023 PED dataset. As shown in Table 2, the mean Dice is 50.81 without adaptation. Introducing SCA raises it to 60.43, while using HLSC alone improves the TC Dice from 32.88 to 41.39. When the two are combined, our HRAC achieves the best mean Dice (61.49) and mean HD95 (29.61), with the largest single improvement on TC Dice (+7.78). In addition, we apply HLSC as a plug-and-play module to several classic TTA methods to verify its generality. As shown in Table 3, HLSC shows an overall upward trend in performance across backbones.

SCA internal ablation. Table 4 evaluates the contribution of the two structure-consistency objectives in SCA. Both objectives provide substantial improvements over the unadapted model, with Poset and Flip achieving mean Dice scores of 59.37 and 59.66, respectively. Their combination further improves mean Dice to 60.43, together with the best TC Dice and mean HD95, demonstrating that the two objectives provide complementary benefits. Specifically, Poset consistency directly constrains the nested regional relationships, while Flip consistency improves prediction stability through bilateral symmetry. These results effectively support our design of combining complementary structural signals for covariate-shift adaptation. The effect of the flip-consistency weight is further analyzed in Appendix C.3.

5 Conclusion

In this paper, we propose a source-free TTA framework, called HRAC. Our HRAC first adapts with SCA to restore the spatial structure corrupted by covariate shift through consistency losses imposed across the nested evaluation regions, and then corrects the hierarchical label shift with HLSC by applying a bounded additive correction driven by the violated structural bounds without estimating a target prior. Our HRAC is thus the first source-free method that jointly handles the two shifts underlying this confounding problem. Across three BraTS datasets, HRAC outperforms seven TTA baselines in mean Dice, mean HD95, and mean IoU under distribution shifts of varying magnitude.

AI USE STATEMENT

In this work, we used generative AI tools in an assistive capacity only, and no task was carried out by AI in full. They assisted with part of the implementation of the proposed method and the experimental pipeline, and with part of the work on the mathematical proofs. Additionally, we used them to polish the writing of the manuscript and to search for and identify related literature. We have reviewed all AI-assisted work. All AI-assisted code was inspected and tested by the authors before any reported result was produced. All AI-assisted proof steps were verified line by line before inclusion in the appendix. Every reference suggested by these tools was checked against the original publication before being cited. All AI-assisted text was revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All experiments rely exclusively on publicly released BraTS datasets, so every evaluation dataset can be obtained from its official distributor. The source code and the trained model weights are being prepared for public release and will be made available upon acceptance of the paper.

REFERENCES

Maruf Adewole, Jeffrey D Rudie, Anu Gbadamosi, Dong Zhang, Confidence Raymond, Oluyemisi Toyobo, Olubukola Omidiji, Rachel Akinola, Mohammad Abba Suwaid, Feyisayo Daji, et al. Brain tumor segmentation in Sub-Saharan Africa patient population: The BraTS-Africa challenge. *Neuro-Oncology Advances*, 8(1), Jan 2026. doi: 10.1093/noajnl/vdag082. URL `http://dx.doi.org/10.1093/noajnl/vdag082`.

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9260–9269. IEEE, 2019.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Brian Guo, Darui Lu, Gregory Szumel, Rongze Gui, Pengyu Li, Tingyu Wang, Nicholas Konz, and Maciej A Mazurowski. The impact of scanner domain shift on deep learning performance in medical imaging: an experimental study. *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–11, 2026.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, pp. 1989–1998. Pmlr, 2018.

Lukas Hoyer, Dengxin Dai, Qin Wang, Yuhua Chen, and Luc Van Gool. Improving semi-supervised and domain-adaptive semantic segmentation with self-supervised depth estimation. *International Journal of Computer Vision*, 131(8):2070–2096, 2023.

Kai Hu, Jinhao Li, Yuan Zhang, Xiongjun Ye, and Xieping Gao. One-to-multiple: A progressive style transfer unsupervised domain-adaptive framework for kidney tumor segmentation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 24496–24522, 2024.

Fabian Isensee, Paul F Jaeger, Simon AA 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.

Minguk Jang and Hye Won Chung. Label distribution shift-aware prediction refinement for test-time adaptation. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=c7AAHdEYz5.

Kartik Jhawar and Lipo Wang. Hd-tta: Hypothesis-driven test-time adaptation for safer brain tumor segmentation. *arXiv preprint arXiv:2602.19454*, 2026.

Anahita Fathi Kazerooni, Nastaran Khalili, Xinyang Liu, Deep Gandhi, Zhifan Jiang, Syed Muhammed Anwar, Jake Albrecht, Maruf Adewole, Udunna Anazodo, Hannah Anderson, et al. The brain tumor segmentation in pediatrics (brats-peds) challenge: focus on pediatrics (cbtn-connect-dipgr-asnr-miccai brats-peds). *arXiv preprint arXiv:2404.15009*, 2024.

Wangkai Li, Rui Sun, Bohao Liao, Zhaoyang Li, and Tianzhu Zhang. Balanced learning for domain adaptive semantic segmentation. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25, pp. 35858–35883. JMLR.org, 2025.

Xiang Li, Huihui Fang, Changmiao Wang, Mingsi Liu, Lixin Duan, and Yanwu Xu. Cache-driven spatial test-time adaptation for cross-modality medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 146–156. Springer, 2024.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 6028–6039. PMLR, 2020.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64, 2025.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, pp. 3122–3130. PMLR, 2018.

Yahao Liu, Jinhong Deng, Jiale Tao, Tong Chu, Lixin Duan, and Wen Li. Undoing the damage of label shift for cross-domain semantic segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7032–7042. IEEE, 2022.

Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1215–1224, June 2021.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

David N Louis, Arie Perry, Pieter Wesseling, Daniel J Brat, Ian A Cree, Dominique Figarella-Branger, Cynthia Hawkins, HK Ng, Stefan M Pfister, Guido Reifenberger, et al. The 2021 who classification of tumors of the central nervous system: a summary. *Neuro-Oncology*, 23(8):1231–1251, 2021.

Wenao Ma, Cheng Chen, Shuang Zheng, Jing Qin, Huimao Zhang, and Qi Dou. Test-time adaptation with calibration of medical image classification nets for label distribution shift. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 313–323. Springer, 2022.

Alan Mackay, Anna Burford, Diana Carvalho, Elisa Izquierdo, Janat Fazal-Salom, Kathryn R Taylor, Lynn Bjerke, Matthew Clarke, Mara Vinci, Meera Nandhabalan, et al. Integrated molecular meta-analysis of 1,000 pediatric high-grade and diffuse intrinsic pontine glioma. *Cancer Cell*, 32(4): 520–537, 2017.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=37nvvqkCo5.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning*, pp. 16888–16905. PMLR, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=g2YraF75Tj.

Quinn T Ostrom, Gino Cioffi, Kristin Waite, Carol Kruchko, and Jill S Barnholtz-Sloan. Cbtrus statistical report: primary brain and other central nervous system tumors diagnosed in the united states in 2014–2018. *Neuro-Oncology*, 23(Supplement_3):iii1–iii105, 2021.

Sunghyun Park, Seunghan Yang, Jaegul Choo, and Sungrack Yun. Label shift adapter for test-time adaptation under covariate and label shifts. In *2023 IEEE/CVF International Conference on Computer Vision*, pp. 16375–16385. IEEE, 2023.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural Computation*, 14(1):21–41, 2002.

Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *2021 IEEE Winter Conference on Applications of Computer Vision*, pp. 1378–1388. IEEE, 2021.

Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7472–7481. IEEE, 2018.

Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2512–2521. IEEE, 2019.

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.

Yuanhan Wang, Yifei Chen, Shuo Jiang, Wenjing Yu, Mingxuan Liu, Beining Wu, Jinying Zong, Feiwei Qin, Changmiao Wang, and Qiyuan Tian. Smart: Style-modulated robust test-time adaptation for cross-domain brain tumor segmentation in mri. *arXiv preprint arXiv:2509.17925*, 2025.

Thilina Weerasinghe, Ruwan Tennakoon, WeiQin Chuah, and Alireza Bab-Hadiashar. Logit-adjusted test-time adaptation under partial class imbalance. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5977–5985. IEEE, 2026.

Jianghao Wu, Yicheng Wu, Yutong Xie, Wenjia Bai, You Zhang, Feilong Tang, Yulong Li, Imran Razzak, Daniel F Schmidt, and Yasmeen George. Sam-aware test-time adaptation for universal medical image segmentation. *IEEE Transactions on Image Processing*, 2026.

Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4084–4094. IEEE, 2020.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Medicine*, 15(11):e1002683, 2018.

Yubo Zhou, Jianghao Wu, Wenjun Liao, Shichuan Zhang, Shaoting Zhang, and Guotai Wang. 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.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2242–2251, 2017.

Appendix

Source-Free Test-Time Adaptation for Brain Tumor Segmentation by Jointly Handling Covariate and Label Shifts

The Supplementary Material of the paper is organized as:

- Appendix A: Insight Analysis.
- Appendix B: Algorithm Design and Implementation Details.
- Appendix C: Extended Experiments.
- Appendix D: Method Comparison.
- Appendix E: Generalization Discussion.

A Insight analysis

A.1 Why the Co-occurrence Is the Hard Case

Problem setting. Let the source joint distribution $p_S(\mathbf{x}, y)$ be known, and let only the marginal $p_T(\mathbf{x})$ be observed on the target domain. Write the density ratio as:

$$\rho(\mathbf{x}, y) = \frac{p_T(\mathbf{x}, y)}{p_S(\mathbf{x}, y)} = r(y)\,\kappa_y(\mathbf{x}), \qquad r(y) = \frac{p_T(y)}{p_S(y)}, \qquad \kappa_y(\mathbf{x}) = \frac{p_T(\mathbf{x} \mid y)}{p_S(\mathbf{x} \mid y)}. \tag{10}$$

The two standard assumptions correspond to degenerate forms of the density ratio. Label shift (LS) means $\kappa_y \equiv 1$, and covariate shift (CS) means $\rho(\mathbf{x}, y) = w(\mathbf{x})$, in which case $\kappa_y(\mathbf{x}) = w(\mathbf{x})/r(y)$. If $\kappa_y \not\equiv 1$ and $r(y)\,\kappa_y(\mathbf{x})$ cannot be reduced to a function of $\mathbf{x}$ alone, the shift is said to be in the state of *mechanism coexistence*, of which the co-occurrence of sensor-channel variation and scene variation is an instance. The following analysis shows that mechanism coexistence triggers three difficulties at once, namely the lack of identifiability, the ill-posedness of the inverse problem, and the amplification of decision errors.

Lack of identifiability. The observation $(p_S, p_T(\mathbf{x}))$ determines the feasible parameter set:

$$\mathcal{F} = \Big\{(r, \kappa) : \int r(y)\,\kappa_y(\mathbf{x})\,p_S(\mathbf{x}, y)\,dy = p_T(\mathbf{x}) \ \text{ for all } \mathbf{x}\Big\}, \tag{11}$$

which is an infinite-dimensional family. For any positive integrable function r, let $u(\mathbf{x}) = p_T(\mathbf{x})/p_S(\mathbf{x})$ and $w_r(\mathbf{x}) = u(\mathbf{x})/\mathbb{E}_S[r \mid \mathbf{X}]$. The pair (w_r, r) then matches the observed marginals exactly and is normalized automatically, while the induced target prior $p_T^{(r)}(y)$ varies with r. The same observation therefore corresponds to a family of different posteriors $p_T(y \mid \mathbf{x})$ and different optimal decisions, and $p_T(y \mid \mathbf{x})$ is not uniquely determined by the observation. This is not a matter of statistical efficiency but of structural non-identifiability, since the unlabeled target marginal does not contain the information required to identify $p_T(y)$. The feasible solutions can be compressed to a single point only by external structure, such as sensor channels that are known or can be calibrated, invariant representations, a small number of labeled target samples, multi-environment invariance constraints, and moment constraints.

Ill-posedness of the inverse problem. Even within the identifiable LS subclass, where $p_T = Ar$ with $A = [\pi_1 f_1, \ldots, \pi_K f_K]$, $\pi_y = p_S(y)$ and $f_y = p_S(\cdot \mid y)$, the estimation error satisfies:

$$\|\delta r\| \le \frac{\|\delta p_T\|}{\sigma_{\min}(A)}, \tag{12}$$

and in the two-class case $\|\delta p_T\|/|\delta r_1| = \pi_1 \|f_1 - f_2\|$, so the smallest singular value is of the same order as the distance between the class-conditional distributions. The more severely the class-conditional distributions overlap, the larger the condition number $\kappa(A) = \sigma_{\max}/\sigma_{\min}$ and the more pronounced the error amplification. At the same time, the effective sample size of importance weighting, $\mathrm{ESS} = n/(1 + \mathrm{CV}^2(\rho))$, degrades as the variation of ρ grows, which makes reweighting-based correction unreliable under finite samples. In other words, class overlap makes even the identifiable subproblem ill-posed.

Amplification of decision errors. Under the condition that the denominator is uniformly bounded below, the target posterior is Lipschitz continuous in (r, κ):

$$p_T(y \mid \mathbf{x}) = \frac{r(y)\,\kappa_y(\mathbf{x})\,p_S(y \mid \mathbf{x})}{\sum_{y'} r(y')\,\kappa_{y'}(\mathbf{x})\,p_S(y' \mid \mathbf{x})}, \tag{13}$$

so a parameter error ϵ induces only an $O(\epsilon)$ perturbation of the posterior. The 0-1 risk, however, is not linearly stable. With the margin $m(\mathbf{x}) = p_T(y^\star \mid \mathbf{x}) - \max_{y \neq y^\star} p_T(y \mid \mathbf{x})$, the decision flips only on $\{m(\mathbf{x}) \leq C\epsilon\}$, and the excess risk satisfies:

$$R_T(\hat{f}) - R_T(f^\star) \;\leq\; P_T\big(m(\mathbf{X}) \leq C\epsilon\big), \tag{14}$$

which is $O(\epsilon^\alpha)$ under the Tsybakov-type margin condition $P_T(m \leq t) \leq t^\alpha$. Intensified overlap piles the margin mass near 0 (α decreases), and the parameter error is markedly amplified into decision risk. Under unbounded losses such as the log loss, when the posterior lacks a uniform positive lower bound, the excess risk $\mathbb{E}_T[\mathrm{KL}]$ may fail to converge quadratically in ϵ.

Summary. Mechanism coexistence constitutes the hard core of the joint shift for the following reasons. First, the two standard invariance assumptions cannot strictly coexist, since classes that are inseparable in features are forced to share prior ratios and strict coexistence degenerates into trivial reweighting by overlap-connected components, so the two standard correction routes of LS and CS fail at the same time. Second, even when one falls back on estimating a single mechanism, class overlap renders the inverse problem ill-posed. Third, the parameter error is ultimately amplified nonlinearly at the decision level. The difficulty of the joint shift is therefore structural in nature. Without external structure, the target posterior cannot be uniquely determined from the observation, and any method that relies only on the unlabeled target marginal is fundamentally limited by non-identifiability.

A.2 Why HRAC Works

The difficulty of the joint shift is structural in nature and cannot be resolved at the statistical level. In the density ratio $\rho(\mathbf{x}, y) = r(y)\,\kappa_y(\mathbf{x})$ of Appendix A.1, mechanism coexistence means that neither $\kappa_y \equiv 1$ nor $\rho(\mathbf{x}, y) = w(\mathbf{x})$ holds, so both standard correction routes fail at once. Worse, the two components are mutually confounded in the predictions, since the label-shift signal has to be read from the predictions, and the covariate component distorts exactly these predictions. HRAC addresses this confounding at three levels.

Identifiability. The unlabeled target marginal determines the pair (r, κ) only up to a family of feasible solutions, so the target composition is not recoverable by estimation from target data alone. HRAC therefore does not estimate $\hat{\pi}_T$ at all. It replaces estimation with structure. The nesting $\mathrm{WT} \supseteq \mathrm{TC} \supseteq \mathrm{ET}$ and the bilateral symmetry of the brain are invariants that hold across populations, so composition is bounded within a coarse anatomical range, since the core is nested within the enclosing regions. The structural prior of Section 3.4 is accordingly a fixed structural bound, not an estimate of the composition of any population, and the pipeline requires no source data. In short, the method buys identifiability with prior structure instead of solving an underdetermined inverse problem.

Stability. Classical prior-ratio correction is ill-conditioned. For the rare foreground classes, an arbitrarily small absolute error in the estimated prior yields an arbitrarily large bias error, and the estimated ratios may even prescribe an ascent direction of $\mathcal{L}_T$ at the uncorrected point. HLSC removes this amplification source through a direction condition that requires no estimate. Only ratios that violate their bounds receive a one-sided deficit, the bounds double as normalizers and thus express the violations as relative deviations on a common scale, and the correction is a single gradient-free pass, firing only where a deficit is nonzero, with the magnitude capped by α. The correction therefore does nothing when no structural bound is violated, and its effect is decoupled from the accuracy of any prior estimate.

Decision risk. Small parameter errors are amplified into decision errors wherever the margin is small. SCA addresses the margin directly. The flip consistency distills each prediction toward its own mirror view, the partial-order consistency penalizes mass lying genuinely outside its container, and together they stabilize the prediction and move its mass away from the region boundaries, while the bounded bias keeps the injected perturbation capped. Consistent with this account, the gains

Algorithm 1 HRAC

Require: source-pretrained f_θ, unlabeled volume $\mathbf{x}$, structural prior $\boldsymbol{\rho}$, scales α, λ_{poset}, λ_{cons}, steps N

1: freeze f_θ, enable gradients on BatchNorm affine parameters only
2: **for** $n = 1$ **to** N **do**
3: $\quad \mathbf{z} \leftarrow f_\theta(\mathbf{x}), \quad \mathbf{p} \leftarrow \mathrm{softmax}(\mathbf{z}), \quad M_{\text{ET}}, M_{\text{TC}}, M_{\text{WT}} \leftarrow$ soft masks equation 4
4: $\quad p^{\text{flip}} \leftarrow f_\theta(\mathrm{flip}_W(\mathbf{x}))$ **(no grad)**, flipped back to the original geometry
5: $\quad \mathcal{L} \leftarrow \lambda_{\text{poset}}\mathcal{L}_{\text{poset}}(M_{\text{ET}}, M_{\text{TC}}, M_{\text{WT}}) + \lambda_{\text{cons}}\mathcal{L}_{\text{flip}}(\mathbf{p}, p^{\text{flip}})$ equation 5–equation 3
6: $\quad$ update BatchNorm affine parameters by one Adam step on $\mathcal{L}$ (clip $\|\nabla\|_2 \le 1$), anneal mask temperature
7: **end for**
8: $\mathbf{z} \leftarrow f_\theta(\mathbf{x}), \quad r_{\text{TC}}, r_{\text{ED}}, r_{\text{ET}} \leftarrow$ structural ratios equation 7
9: $d_{\text{N}}, d_{\text{E}}, d_{\text{T}} \leftarrow$ one-sided deficits equation 8, $\quad \mathbf{b} \leftarrow \alpha\,(0, d_{\text{N}}, -d_{\text{E}}, d_{\text{T}})$
10: **return** $\hat{\mathbf{y}} = \arg\max_c \mathrm{softmax}\big(f_\theta(\mathbf{x}) + \mathbf{b}\big)_c$

concentrate on the rarest region, with the largest single improvement on TC Dice ($+7.78$), they come from correcting the under-predicted tumor core, and they grow with the magnitude of the cross-dataset composition shift.

In short, HRAC is effective not because it estimates the shift more accurately, but because it sidesteps the non-identifiability of the joint shift. Structure-consistency adaptation absorbs the covariate component and restores the invariant spatial organization from which the label-shift signal is read, while HLSC acts on the structural ratios as bounded, one-sided directions instead of estimated priors. This adapt-then-correct decomposition is what makes both shifts jointly correctable in the source-free setting.

B Algorithm Design and Implementation Details

B.1 Algorithm Pseudocode

Algorithm 1 restates the two components described in Sections 3.3 and 3.4 as one procedure, with SCA (lines 3–7) followed by HLSC (lines 8–9).

B.2 Implementation Details

Source training. The source model is a 3D U-Net with feature widths $(32, 64, 128, 256)$ and dropout 0.3, trained on the BraTS 2023 GLI training set to minimize a Dice loss of weight 0.6 plus a cross-entropy loss of weight 0.4 for up to 300 epochs, with Adam at learning rate 5×10^{-4}, weight decay 10^{-4}, batch size 2, an $80/20$ training-validation split, and a fixed random seed. All experiments are run on a single NVIDIA A800 GPU.

Test-time adaptation. Only the BatchNorm affine parameters are updated, which amounts to 0.03% of the 21.37M parameters of the network. SCA adapts each volume for $N = 2$ steps with Adam at learning rate 3×10^{-4} and gradient-norm clipping at 1.0, using $\lambda_{\text{poset}} = 1$, $\lambda_{\text{cons}} = 10$, mask threshold $\tau = 0.5$, and a mask temperature annealed from 0.1 towards 0.02 by a factor of 0.95 per step. HLSC then applies one gradient-free correction per volume with $\alpha = 4.0$ and $\boldsymbol{\rho} = (0.4, 0.6, 0.6)$. The affine parameters are carried across the target stream and are never reset between volumes.

Label mapping. The BraTS 2025 PED release uses a five-class label set, which we map onto the four-class scheme of the source model before evaluation, taking the official enhancing tumor label to ET, the official non-enhancing tumor and cystic component labels to the necrotic and non-enhancing class, and the official edema label to ED. After the mapping the nested regions are formed exactly as for the other cohorts.

Table 5: Comparison with existing test-time adaptation methods on the reverse transfer direction, with BraTS 2023 PED as the source dataset and GLI as the target dataset. **Bold** and underlined values denote the best and second-best results in each column.

Method	Dice (%) ↑				HD95 (mm) ↓				IoU (%) ↑			
	WT	TC	ET	Mean	WT	TC	ET	Mean	WT	TC	ET	Mean
w/o adaptation	65.15	44.33	43.83	51.10	35.10	41.94	64.93	47.32	52.86	31.76	32.49	39.04
Tent	13.40	11.35	12.27	12.34	263.15	261.37	274.13	266.22	10.24	7.97	9.61	9.27
EATA	**72.21**	40.38	**52.46**	55.02	27.50	42.00	38.78	36.09	**59.95**	27.99	**39.91**	42.62
Undoing_UDA	60.96	43.31	40.55	48.27	23.32	**33.53**	38.80	31.88	48.75	31.66	28.82	36.41
SAR	72.17	40.03	51.91	54.70	27.50	42.49	40.69	36.89	59.92	27.73	39.55	42.40
TEGDA	3.16	2.52	4.80	3.49	198.15	201.87	204.77	201.60	2.42	1.76	4.09	2.76
SmaRT	42.68	19.89	50.83	37.80	56.73	72.35	55.38	61.49	31.75	12.44	38.53	27.57
BLDA	57.66	43.43	39.27	46.79	35.93	41.03	71.35	49.44	46.20	31.43	29.29	35.64
HRAC (Ours)	71.17	**44.92**	50.27	**55.45**	**22.78**	36.16	**31.39**	**30.11**	59.03	**31.86**	37.87	**42.92**

C Extended Experiments

C.1 Evaluation on the Reverse Transfer Direction

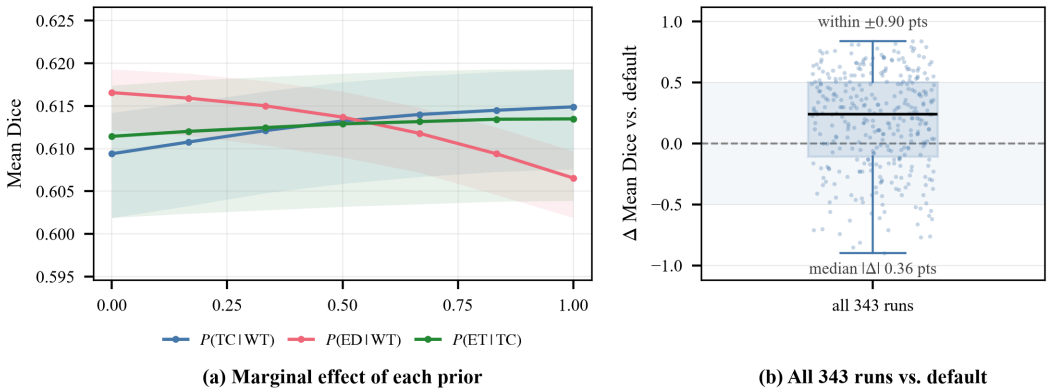
The main evaluation trains the source model on BraTS 2023 adult glioma (GLI) and adapts it to the pediatric and Sub-Saharan Africa targets. To examine whether this choice drives the results, we reverse the direction. The source model is now trained on the BraTS 2023 PED training set (99 cases) and evaluated on GLI (1251 cases), so the target is the large adult population, and the compositional shift reverses as well, since the adult target contains more edema and less necrotic core than the pediatric source, whereas the main transfers move the other way. All baselines are re-run under the same protocol, the hyperparameters of every method keep their main-evaluation values, and no hyperparameter is tuned on target labels. Table 5 reports the results.

Three observations follow. First, entropy-based adaptation collapses in this direction. Tent falls to 12.34 mean Dice with a mean HD95 of 266.22 mm, TEGDA falls to 3.49, and SmaRT falls to 37.80, all far below the unadapted source model at 51.10, which is consistent with the observation that objectives which sharpen predictions amplify the prevailing compositional bias. Second, adaptation can also hurt without collapsing. Undoing_UDA and BLDA reach 48.27 and 46.79 mean Dice, below the unadapted model, and on TC Dice every adaptation baseline stays below the unadapted value of 44.33, whereas HRAC is the only method that improves it, to 44.92. Third, HRAC attains the best mean in all three metrics, namely 55.45 Dice, 30.11 HD95, and 42.92 IoU. The margin in mean Dice is 0.43 points over EATA, while the margin in boundary quality is larger, with the mean HD95 1.77 mm below the best baseline value of 31.88 and the ET HD95 reduced from 38.78 to 31.39. On the regional Dice of WT and ET the strongest entries remain the covariate-shift baselines, and the advantage of HRAC concentrates in boundary quality and in the mean.

These results indicate that the benefit does not depend on the direction of the compositional change. The deficits of HLSC fire only where the predicted composition violates a one-sided structural bound, so the correction follows the violation observed on the target and not a ratio estimated against the source, and the bounded bias cannot push any class past the anatomical range in either direction. Reversing the transfer leaves the framework intact, which supports its use beyond the deployment scenario of the main evaluation.

C.2 The Structural Prior and the Correction Scale

Rationale for the Structural Prior Design. The vector $\boldsymbol{\rho}$ of Section 3.4 is not an estimate of the composition of any population, source or target, and this choice rests on two considerations. (1) Composition shifts across populations, so correction is necessary and the ratios carry a diagnostic signal. (2) Composition is bounded within a coarse anatomical range, since the core cannot vanish inside the whole tumor, edema cannot dominate it, and ET cannot vanish inside the core, so enforcing the bounds as floors and ceilings instead of exact targets is safe and requires no source or target statistics.



(a) Marginal effect of each prior **(b) All 343 runs vs. default**

Figure 4: **Sensitivity of HRAC to the target composition priors.** (a) Mean Dice when sweeping each prior over its grid (r_{TC} and $r_{\mathrm{ED}} \in [0.2, 0.8]$, $r_{\mathrm{ET}} \in [0.4, 1.0]$). Each line averages over all configurations of the other two priors, and the band spans their min–max range. (b) Deviation of each of the 343 grid configurations from the deployed configuration. All runs stay within ± 0.90 points.

Choosing the correction scale. The scale α is fixed and not calibrated, since the quantity that determines the effect of the correction, the marginal error $\mathbf{g}$, is not observable source-free. The fixed $\alpha = 4.0$ (Section 3.4) is a boundedness choice whose role is safety. We set $\alpha = 4.0 \approx \log 50$, where 50 is the largest source-to-target class-prior ratio we assume across the BraTS cohorts. Empirically, the full pipeline remains beneficial across the evaluated shifts (Tables 1 and 2).

Sensitivity to the structural prior. The prior sweep of Figure 4 is a direct validation of the rationale above, since the structural prior need only be coarsely specified. The three bounds are swept over a $7 \times 7 \times 7$ grid centered on the literature-median composition, and the deployed configuration $(0.4, 0.6, 0.6)$ is itself a grid point. Over the whole grid Mean Dice varies by 1.74 points (from 60.19 to 61.93), and every configuration stays within 0.90 points of the deployed one.

C.3 Flip-Consistency Weight Sweep

The flip-consistency term of Section 3.3 is normalized per voxel, and its gradients are far smaller in magnitude than those of the partial-order term, so its weight λ_{cons} is calibrated on that scale and is not transferable from a volume-summed objective. The sweep in Table 6 shows the consequence. Below the per-voxel scale the flip term is effectively inactive, so the rows for $\lambda_{\mathrm{cons}} \in \{0, 0.1, 1\}$ are numerically indistinguishable, whereas $\lambda_{\mathrm{cons}} = 10$ offsets this gradient mismatch and lets the flip signal contribute measurable gains in every region, and the gains carry through the correction. We therefore adopt $\lambda_{\mathrm{cons}} = 10$ throughout.

D Method Comparison

D.1 Qualitative Results on SSA

Fig. 5 compares HRAC with the baselines on four cases from the SSA dataset. Similar to Fig. 3, all methods localize the lesion while the main difference lies in the predicted composition. The ground truth in Row 1 is core-dominant with an enhancing band and limited edema. The unadapted model produces large spurious edema regions across the brain and some baselines fill the lesion with edema or distort the core and enhancing ratio, while HRAC recovers the core, the enhancing band and the thin edema consistent with the ground truth. Row 2 contains a mixed three-class composition and the baselines give fragmented masks with over-predicted edema or enhancing components, while HRAC reproduces the mixed composition. Row 3 is enhancing-dominant with a necrotic rim, where the unadapted model again spreads spurious edema over a large brain region, Tent and EATA fill the lesion with edema and TEGDA over-predicts the enhancing component, while HRAC matches the

Table 6: Sensitivity of SCA and HRAC to the flip-consistency weight λ_{cons} on the BraTS 2023 PED dataset. With HLSC disabled the framework reduces to SCA alone, and $\lambda_{\text{cons}} = 0$ reduces SCA to the partial-order objective. **Bold** and underlined values denote the best and second-best results in each column, respectively.

λ_{cons}	HLSC	Dice (%) ↑				HD95 (mm) ↓				IoU (%) ↑			
		WT	TC	ET	Mean	WT	TC	ET	Mean	WT	TC	ET	Mean
0		82.73	41.53	53.84	59.37	9.33	48.97	81.93	46.74	73.47	31.51	45.13	50.04
0.1		82.74	41.53	53.84	59.37	9.34	48.97	81.93	46.75	73.48	31.51	45.13	50.04
1		82.79	41.35	53.84	59.33	9.21	49.01	81.94	46.72	73.54	31.40	45.14	50.03
10		**83.05**	42.62	**55.63**	60.43	9.04	48.60	78.07	45.24	**73.77**	32.31	**46.99**	51.02
0	✓	81.94	48.21	52.37	60.84	8.94	33.30	50.74	30.99	72.67	36.43	43.24	50.78
0.1	✓	81.95	48.26	52.37	60.86	8.94	33.29	50.74	30.99	72.69	36.47	43.25	50.80
1	✓	82.02	48.41	52.32	60.92	11.91	33.12	54.34	33.12	72.76	36.58	43.23	50.86
10	✓	82.19	**50.40**	51.87	**61.49**	**8.80**	**32.45**	**47.59**	**29.61**	72.89	**38.15**	42.88	**51.31**

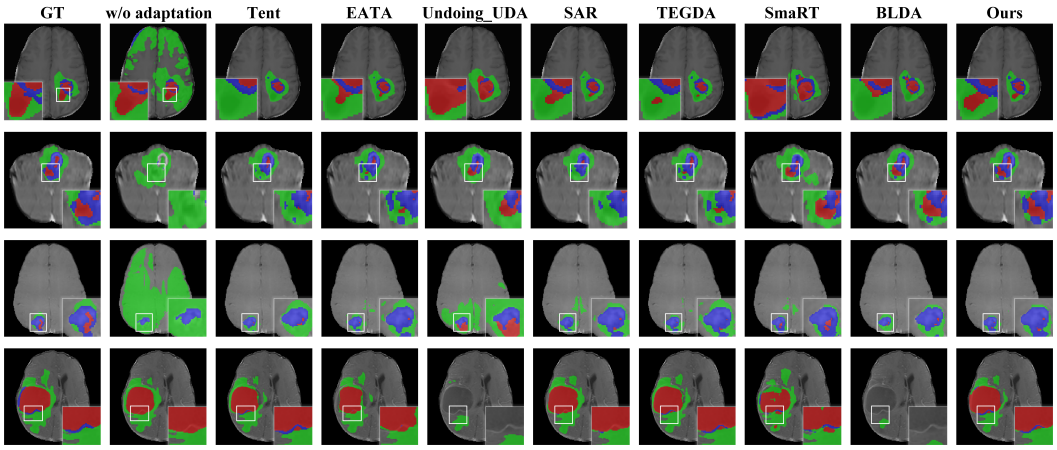


Figure 5: Qualitative comparison of HRAC and the baselines on four cases from the SSA dataset. The layout is the same as Fig. 3. Each row shows one case with the ground truth, the unadapted source model prediction, the predictions of the seven baselines and HRAC, and the white box indicates the zoomed region. HRAC produces the class composition closest to the ground truth, validating the cross-population generalization of the adapt-then-correct framework.

ground truth. Row 4 is core-dominant with an edema rim, where Undoing_UDA and BLDA fail to produce a complete foreground while HRAC segments the lesion completely and keeps the mask compact. Overall, HRAC again yields the composition closest to the ground truth on SSA, showing that the adapt-then-correct framework generalizes across populations beyond the adult-to-pediatric setting.

D.2 Comparison of Label-Shift Correction Strategies

Table 7 compares HLSC with representative label-shift correction strategies based on three properties relevant to source-free test-time correction, including access to source information, reliance on source priors, and explicit estimation of the target class prior.

Saerens EM (Saerens et al., 2002), BBSE (Lipton et al., 2018), and Undoing_UDA (Liu et al., 2022) explicitly estimate the target class prior and therefore require source-domain information or statistics to construct the correction. BLDA (Li et al., 2025) also relies on source-domain prior information, but does not explicitly estimate the target prior online. These methods consequently require some form of source-side information when performing label-shift correction.

Table 7: Comparison of HLSC with representative label-shift correction strategies. ✓ / × indicate whether a property is satisfied or not. Source-free denotes that the correction does not require access to source data or source-domain statistics at test time. Uses π_S indicates that the correction relies on source-domain class priors or related source statistics. Est. $\hat{\pi}_T$ indicates that the target class prior is explicitly estimated online.

Method	Source-free	Uses π_S	Est. $\hat{\pi}_T$
Saerens EM (Saerens et al., 2002)	×	✓	✓
BBSE (Lipton et al., 2018)	×	✓	✓
Undoing_UDA (Liu et al., 2022)	×	✓	✓
BLDA (Li et al., 2025)	×	✓	×
HLSC (Ours)	✓	×	×

In contrast, HLSC does not require source data, source statistics, or an explicit estimate of the target class prior. Instead, it derives the correction from structural ratios measured directly from the target prediction and uses these measurements to identify class-wise logit deficits or excesses. This design allows the correction to be performed entirely from the target volume during deployment.

The comparison highlights the specific distinction of HLSC. HLSC derives the logit correction signal from target-side structural constraints without relying on source priors or explicit estimation of the target class prior. This design enables a source-free correction strategy based solely on target-domain structural information.

D.3 Complexity Analysis

Table 8 summarizes the computational complexity of the compared methods in terms of the number of adapted parameters and per-case test-time cost. The reported time includes both adaptation and inference, while excluding data loading. HRAC adapts only the BatchNorm affine parameters, resulting in approximately 10K adapted parameters, which is comparable to Tent, SAR, and TEGDA.

HRAC performs two SCA adaptation steps ($N = 2$) for each target volume. The adapted affine parameters are continuously updated along the target stream and remain available for subsequent cases (Appendix B.2). HLSC then computes the structural ratios and performs the corresponding logit correction without gradient optimization. With these additional operations, HRAC achieves a per-case runtime of 1.080 s. This is slightly higher than Tent (0.639 s) and SAR (0.658 s), due to the extra adaptation and structural correction steps, but remains substantially lower than TEGDA (4.126 s).

The runtime comparison should also be interpreted in the context of the adaptation protocol adopted by each baseline. EATA achieves a per-case runtime of 0.140 s by selectively performing adaptation on only a subset of samples. Undoing_UDA and BLDA perform adaptation on the source side offline, and their source-side computation is therefore not included in the reported per-case time. SmaRT adapts a substantially larger parameter set of 580K and requires 0.922 s per case. In comparison, HRAC performs adaptation and structural correction with only 10K adapted parameters at a cost of 1.080 s per case.

Overall, HRAC introduces a moderate test-time overhead compared with lightweight BN-based adaptation methods, while remaining considerably less computationally expensive than TEGDA. The additional runtime is justified by the joint optimization of test-time adaptation and structural logit correction, which enables HRAC to address both feature shift and Hierarchical label shift with a small number of adapted parameters.

E Generalization Discussion

E.1 Theoretical Generalization

HRAC uses no tumor-specific property. The nesting $\mathrm{WT} \supseteq \mathrm{TC} \supseteq \mathrm{ET}$ is one instance of a containment-ordered label space. For a label set $\mathcal{Y}$ carrying a partial order $y_j \supseteq y_i$ (region i lies inside region j), equation 2 is defined without change as the sum of marginal deviations over

Table 8: Computational complexity comparison on the BraTS 2023 SSA dataset. Adapt (K) denotes the number of adapted parameters in thousands, and s/case denotes the per-case test-time cost, including adaptation and inference. Measurements use batch size 1, exclude data loading, discard the first 3 cases as warm-up, and are obtained on a single NVIDIA A800 GPU. For UDA baselines, source-side adaptation is performed offline and is therefore not included in the reported per-case time.

Method	Adapt (K)	s/case
w/o adaptation	0	0.143
Tent	10	0.639
EATA	10	0.140
Undoing_UDA	0	0.142
SAR	10	0.658
TEGDA	10	4.126
SmaRT	580	0.922
BLDA	0	0.142
HRAC (Ours)	10	1.080

any region. On a tree-structured label space, the ratios of equation 7 become node-to-parent ratios $r_i = p_i/p_{\mathrm{parent}(i)}$, the prior $\boldsymbol{\rho}$ assigns a one-sided upper bound to each ratio, and the deficits of equation 8 become per-node one-sided deficits. SCA and HLSC then operate as in the main text. Two components transfer as-is and one does not. The partial-order consistency term and the bounded correction depend only on containment and are domain-free, whereas flip consistency relies on the bilateral symmetry of the brain and must be replaced by a domain-appropriate equivariance prior (e.g., left-right symmetry of paired organs, temporal consistency in sequential data) in other domains.

E.2 Instantiation Conditions

HRAC instantiates on a new domain when three conditions hold. (i) The labels can be expressed, or restructured, as mutually exclusive classes whose marginals form a containment hierarchy. Purely semantic inclusion (e.g., a lane inside a drivable area) must first be cast as nested marginals of a common segmentation head. (ii) Geometric consistency between the compared views holds, so that the containment masks of equation 4 are not distorted by misregistration. (iii) $\boldsymbol{\rho}$ is a genuine safety bound rather than a target estimate. Any ratio meeting its bound is left uncorrected, so a conservative bound costs little and an aggressive one imports error. Violations of (i)–(iii) degrade to no correction and not to amplified error, since the bias is one-sided and bounded.

E.3 Applicability to Other Medical Tasks

Containment hierarchies are common in medical segmentation, as in kidney $\supseteq$ renal tumor, liver $\supseteq$ hepatic lesion, and organ $\supseteq$ substructure in general. Under a shift in scanner protocol (covariate) and disease prevalence (label), such cases can be handled by supplying the containment rules and a conservative lesion-to-organ bound, without estimating the target composition. This is a framework-level statement, and no cross-organ experiment is conducted in this work.

E.4 Extension to General Computer Vision

Two families of tasks expose analogous structure. In road scene segmentation, the class composition is nested (e.g., current lane within drivable area), and structural rules survive shifts such as clear-weather highway to snowy urban driving, where appearance and class proportions change jointly. In remote sensing, land $\supseteq$ urban area $\supseteq$ building footprints provide stable bounds under city or sensor transfer. In both cases HRAC would act as a lightweight, label-free correction once the labels are recast as nested marginals and a replacement equivariance prior is chosen for the consistency term. We stress that these are design implications of the tree abstraction in Section E.1 rather than empirical claims, and validating them on non-BraTS tasks is left to future work.