RIR: RELATIVE INTENSITY RELATIONS FOR SINGLE-SOURCE GENERALIZATION UNDER UNSEEN PATHOLOGY SHIFTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical domain generalization typically addresses shifts in acquisition conditions, institutions, or patient cohorts, yet costly annotation and fragmented clinical datasets may leave deployment conditions beyond the pathology coverage available during source training. We study brain tumor segmentation under such external shifts and ask which source-derived evidence remains useful when absolute intensity, morphology, and ordinal intensity position may change across tumor categories. We propose Relative Intensity Relation (RIR), a source-only framework that distinguishes absolute intensity from image-internal contrast organization. RIR derives within-volume intensity rank and local intensity ordering from each input, reducing direct dependence on absolute intensity scale while encoding image-internal relations. Because these relations are recomputed from each input, RIR does not transfer a fixed source intensity or morphology template across cohorts. Source segmentation supervision grounds these rank-derived cues into tumor consistency, tumor-background separation, and boundary evidence, which is integrated with the original image representation through a bounded residual correction. The framework further selects its deployment checkpoint and decision threshold using held-out source validation only, without target-domain feedback. Across five source-supervision sizes, RIR achieves the highest mean two-cohort external macro Dice among the compared methods and the highest mean sensitivity on both BraTS-Africa cohorts at all five sizes. The broader comparison also shows substantial degradation beyond the source glioma category across multiple DG strategies. A representation-level diagnostic further shows smaller source-to-target displacement of source-defined tumor scores in the pre-grounding RIR space than in absolute FLAIR. Together, these results support the utility of the complete RIR framework and provide evidence that rank-derived relational structure remains useful under pathology-associated external shifts.

1 INTRODUCTION

Deep learning-based medical image segmentation models often degrade under domain shifts caused by differences in imaging centers, scanners, acquisition protocols, and patient populations (Zhang et al., 2020; Guan & Liu, 2021; Gil-Terrón et al., 2024). Domain generalization (DG) aims to improve robustness to unseen target domains without access to target-domain data during training. In medical imaging, costly expert annotation, privacy restrictions, and fragmented datasets make it difficult to assemble multiple densely annotated source domains, motivating single-source domain generalization (SSDG) (Liu et al., 2022; Ouyang et al., 2022; Jiao et al., 2024). These constraints can also leave available source data incomplete with respect to medical conditions and imaging protocols (Dey et al., 2025; Zhu et al., 2026). We therefore study brain tumor segmentation under incomplete source coverage in the SSDG setting.

Incomplete source coverage is particularly relevant when tumor categories encountered at deployment are absent from training data. Different tumor populations can exhibit distinct morphology and spatial organization, introducing changes in the segmentation target beyond acquisition or cohort variation (Elzayat et al., 2026). We consider a source glioma cohort, an unseen glioma cohort that retains the broad source pathology category, and BraTS-Africa Other Neoplasms as a hetero-

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.

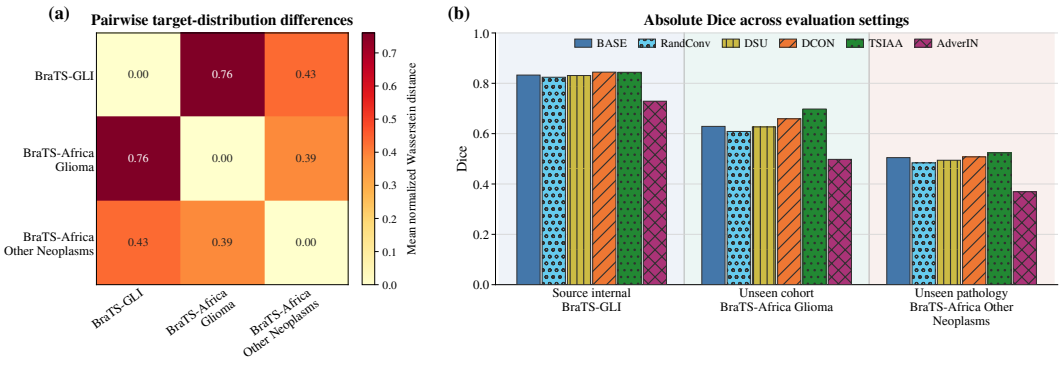


Figure 1: **Pathology-category shift under matched limited source supervision.** (a) Whole-tumor distribution differences across the source cohort, an unseen glioma cohort, and a cohort beyond the source glioma category, summarized over tumor burden, multiplicity, shape irregularity, and spatial location. (b) Broader Dice comparison across source-internal, unseen-cohort, and beyond-source-pathology settings. All methods use matched source supervision without target-domain adaptation.

geneous diagnostic-category cohort beyond the source glioma category. Throughout this paper, we use unseen pathology operationally to denote evaluation beyond the source glioma category rather than a causally isolated pathology shift. Figure 1(a) shows that these cohorts differ in tumor burden, lesion multiplicity, shape irregularity, and spatial location.

To examine whether vulnerability to this setting is specific to a particular baseline or DG strategy, Fig. 1(b) compares the baseline with a broader set of recent medical image segmentation DG methods under matched limited source supervision. All evaluated methods exhibit larger Dice degradation beyond the source glioma category than on the unseen glioma cohort, indicating that the observed degradation is not confined to a single source-side strategy. This breadth check motivates treating pathology coverage as an explicit axis of SSDG evaluation. The controlled source-size study in Sec. 4 then asks whether the complete RIR framework remains useful as labeled source supervision changes.

Existing medical SSDG methods mainly improve robustness through source-side appearance diversification, invariant representation learning, or structural priors (Su et al., 2023; Yi et al., 2024; Liu et al., 2025b; 2022). These approaches reduce reliance on domain-specific appearance, but do not directly identify which source-derived evidence remains useful when pathology coverage changes. We instead focus on relational evidence that can be recomputed within each input and complement the original image representation. We do not assume that absolute intensity, morphology, or ordinal intensity position is preserved across tumor categories. Our hypothesis is that local within-tumor consistency and tumor–background separation can remain useful segmentation evidence under such shifts.

This hypothesis is motivated by magnetic resonance imaging (MRI) and medical image interpretation. MRI signal intensity depends on tissue properties and acquisition settings, while image contrast is defined relative to surrounding tissue (Alzola-Aldamizetxebarria et al., 2022). Medical image perception studies further suggest that experts use global image information to guide local search (Sheridan & Reingold, 2017). These observations motivate a representation that captures image-internal intensity relations while retaining the original image context.

We propose **Relative Intensity Relation (RIR)** to operationalize this idea. RIR derives within-volume rank and local intensity ordering from each input and uses source segmentation supervision to ground these cues into local and regional tumor consistency and tumor–background separation. Because the relational cues are computed from each input rather than transferred as a fixed intensity or shape template, the representation does not rely on source-pathology morphology being preserved at deployment. The resulting relational evidence is integrated with the original image representation as a complementary correction rather than a replacement.

RIR also determines its deployment configuration from source data alone. It retains checkpoints with near-optimal source-validation Dice and prioritizes tumor sensitivity among these candidates. This

choice is supported by expert-centered evaluation showing that Dice alone has only weak agreement with perceived brain tumor segmentation quality, while sensitivity provides complementary information with more consistent expert agreement (Hoebel et al., 2023). The decision threshold is then selected using held-out source validation, requiring no target-domain feedback. Together, relational representation, source-supervised grounding, image-anchored integration, and source-only deployment selection form the complete RIR framework.

Our contributions are threefold.

- **Rank-derived relational representation beyond absolute intensity.** RIR constructs within-volume rank and local intensity ordering from each input as a relational basis that reduces direct dependence on source-specific absolute intensity magnitude. These relations are recomputed within each input rather than transferred as a fixed source intensity or shape template.
- **Source-supervised grounding into pathology-relevant evidence.** Because generic rank-derived cues do not identify tumor-relevant relations by themselves, RIR uses source annotations to learn local and regional tumor consistency and tumor–background separation. The resulting relational evidence complements the original image representation.
- **Source-only deployment selection as part of the complete framework.** Because target-domain validation is unavailable in SSDG, RIR selects among near-optimal source-validation checkpoints using tumor sensitivity and determines the decision threshold from held-out source validation. This Select stage is an explicit component of the proposed framework and requires no target-domain feedback.

2 Related Work

2.1 Single-Source Domain Generalization for Medical Image Segmentation

Medical SSDG methods improve robustness from a single source through source-side diversification, invariant representation learning, and pretrained priors. Recent approaches combine image- and feature-level perturbations with representation constraints (Wang et al., 2026; Huang et al., 2025), address structural and feature biases (Ren et al., 2025), or incorporate vision foundation models to broaden source-side knowledge (Liu et al., 2025a). Limited source supervision has also received increasing attention. Ye et al. (2025) study domain generalization with few source-domain annotations using both labeled and unlabeled source data.

A related line of work broadens pathology coverage itself. Multi-population models combine knowledge across different tumor types (Elzayat et al., 2026), while open-world segmentation frameworks use large-scale heterogeneous pretraining to generalize beyond the observed categories (Zhang et al., 2025). These approaches demonstrate the value of broader pathology coverage, but they do not remove the setting in which the pathology encountered at deployment remains unseen. We study this setting from the perspective of source-derived evidence that may remain useful across pathology shifts.

2.2 Structural and Relational Evidence for Robust Segmentation

Structural information has been used to improve robustness through structure-aware augmentation, anatomical priors, and ordinal constraints. Jiang et al. (2025) disentangle structure from appearance for robust augmentation, while Liu et al. (2020) incorporate anatomical shape priors into segmentation. Zhang et al. (2026) preserve local intensity ordering while generating adversarial intensity variations. Unlike Zhang et al. (2026), which use intensity ordering as a structural constraint during source-side intensity augmentation, RIR uses ordinal intensity organization itself as the basis for constructing the relational representation.

Relative representations provide another way to reduce reliance on absolute features. Cao & Chen (2025) encode samples relative to source-derived anchors and define relations through geometry in a learned feature space. RIR differs in where these relations are constructed. It derives relational cues directly from within-image ordinal intensity organization before relational feature learning, rather than from sample-to-anchor geometry after conventional feature extraction.

2.3 Model Selection under Domain Shift

Model selection remains challenging when target-domain feedback is unavailable. Source-domain validation may not reliably predict generalization under distribution shift (Efthymiadis et al., 2025), and out-of-distribution performance can vary substantially across the training trajectory in single-source learning (Cho et al., 2025). These findings motivate treating deployment selection as part of the DG pipeline rather than as a purely post hoc evaluation choice. Accordingly, RIR determines both the deployment checkpoint and decision threshold using source validation only, without target-domain feedback.

3 Methods

3.1 RIR Framework Overview

We study single-source domain generalization for medical image segmentation using a labeled source dataset $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$. Model optimization uses source-training data, while validation, checkpoint selection, and decision-threshold selection use a held-out source split. Target domains remain inaccessible throughout.

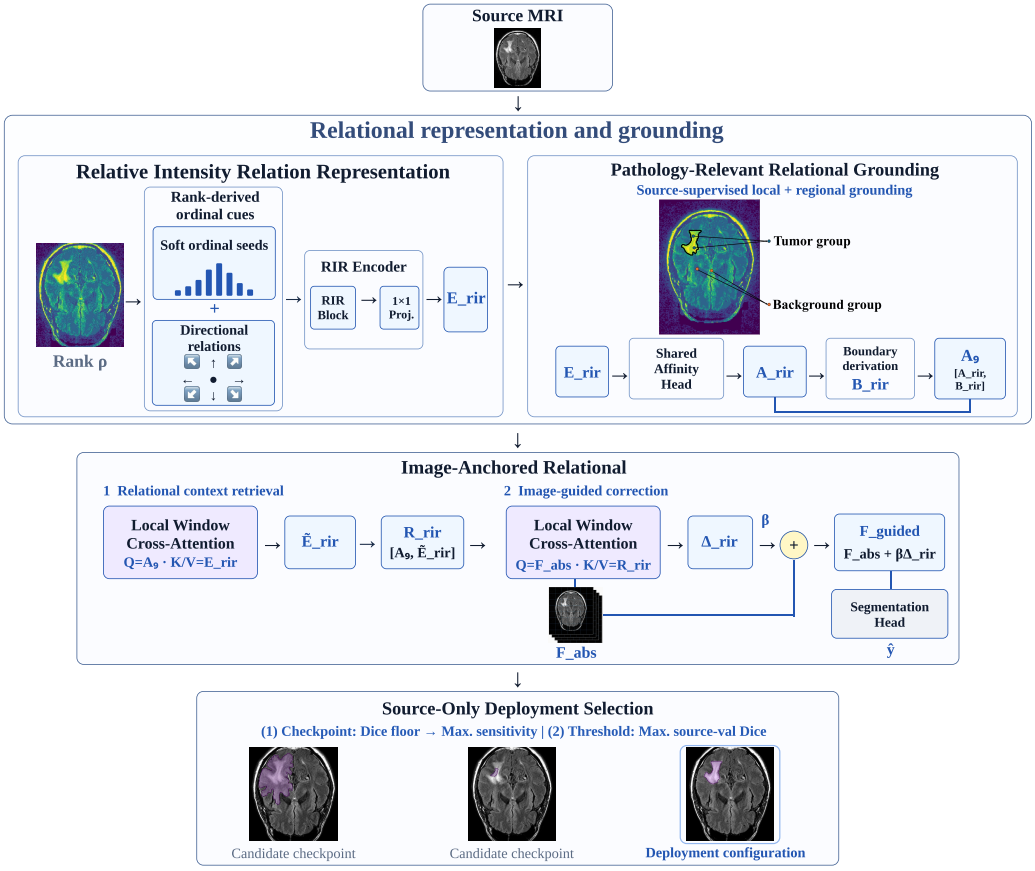


Figure 2: Overview of the proposed RIR framework.

As shown in Fig. 2, the proposed framework follows four stages, *Represent*, *Ground*, *Integrate*, and *Select*. *Represent* encodes within-image rank-derived ordinal cues into the relational representation E_{rir}. *Ground* uses source annotations during training to learn tumor-relevant consistency and separation relations at local and regional scales. The resulting directional affinities A_{rir} are further used to derive boundary evidence B_{rir}, forming the grounded structural query $A_9 = [A_{\text{rir}}, B_{\text{rir}}]$.

Integrate contextualizes E_{rir} using A_9 and uses the original image feature F_{abs} to retrieve the resulting relational evidence. This produces an image-conditioned relational correction controlled by a bounded scalar gate while retaining F_{abs} as the anchor representation. Finally, *Select* chooses a deployment checkpoint from the source-validation trajectory using a Dice-quality constraint followed by sensitivity-based ranking, and then selects the decision threshold using held-out source validation. These four stages together define the complete RIR framework evaluated in Sec. 4.

3.2 Relative Intensity Relation Representation

Because absolute intensity scales can vary across imaging domains, we construct the RIR representation from within-volume ordinal organization rather than absolute intensity values, as illustrated in Fig. 2. For a normalized image volume x, each spatial location p is assigned a volume-relative rank $\rho(p) \in [0, 1]$, indicating the quantile position of $x(p)$ within the same volume. In our 2D implementation, ρ is computed once from the full input volume and then indexed slice-wise, while directional relation learning and segmentation operate on individual slices. This computation uses only image intensities and requires no annotation. The final model instantiates only the stride-1 RIR pathway.

From ρ, we construct two rank-derived ordinal cues. A soft ordinal seed $S(p)$ represents the ordinal position of p, while directional relations compare its rank with those of neighboring locations. For a neighboring direction $d \in \mathcal{D}$,

$$r_d(p) = \sigma\left(\frac{\rho(p) - \rho(p+d)}{\tau_r}\right), \tag{1}$$

where $\tau_r = 0.1$ is a fixed temperature controlling the sharpness of the ordinal comparison. The resulting directional responses are concatenated with the soft ordinal seeds to form Z_{ord}. This construction describes each location using both its within-volume ordinal position and its relative rank relationships with surrounding locations, thereby reducing direct reliance on source-specific absolute intensity scale.

The RIR Encoder maps these ordinal cues to the learned relational representation

$$E_{\mathrm{rir}} = f_{\mathrm{RIR}}(Z_{\mathrm{ord}}).$$

Whereas F_{abs} retains image-derived segmentation information, E_{rir} encodes ordinal position and local rank ordering. These cues provide a generic relational basis and are not assumed to be invariant across pathologies. No separate direct target is imposed on E_{rir}. Source supervision optimizes the RIR pathway end-to-end through the relational grounding objectives in Sec. 3.3 and the segmentation objective, allowing the representation to learn tumor consistency and tumor-background separation. Additional construction details are provided in Appendix B.

3.3 Pathology-Relevant Relational Grounding

Although E_{rir} encodes rank-derived ordinal organization, it does not by itself indicate which relations are relevant to tumor segmentation. We therefore use source segmentation annotations during training to ground the generic relational representation into tumor-relevant consistency and separation relations. Neighboring pairs are categorized as tumor-tumor (TT), tumor-background (TB), or background-background (BB). TT and TB pairs are assigned high- and low-affinity targets, respectively. For BB pairs, rather than enforcing uniformly high affinity across heterogeneous background regions, we use an adaptive soft target that combines decoder-space similarity with rank-seed consistency,

$$s_d(p) = c_d(p)\, a_{\mathrm{dec},d}(p) + \big(1 - c_d(p)\big) a_{\mathrm{rank},d}(p), \tag{2}$$

where $a_{\mathrm{dec},d}(p)$ measures neighboring decoder-feature similarity, $a_{\mathrm{rank},d}(p)$ measures overlap between eight-channel soft ordinal seed distributions, and $c_d(p)$ weights these cues by current segmentation confidence. Thus, $a_{\mathrm{rank},d}$ uses fixed rank-derived seeds rather than the learned embedding E_{rir}. Ambiguous or invalid pairs are excluded from supervision. Decoder- and prediction-derived quantities and the resulting BB target are detached to prevent gradients through target construction. No additional relational input is required at inference. Exact definitions of $a_{\mathrm{dec},d}$, $a_{\mathrm{rank},d}$, and c_d are provided in Appendix C.

Local and regional affinity. Given these pairwise targets, the Shared Affinity Head predicts directional affinity from neighboring relational embeddings. For each direction d,

$$A_d(p) = \sigma(g_\theta\left(|E_{\text{rir}}(p) - E_{\text{rir}}(p+d)|\right)), \tag{3}$$

where E_{rir} denotes the learned affinity embedding and the predictor parameters are shared across directions. The directional predictions are optimized with

$$\mathcal{L}_{\text{dir}} = \mathcal{L}_{\text{TT}} + \mathcal{L}_{\text{TB}} + 0.5\,\mathcal{L}_{\text{BB}},$$

and collected as

$$A_{\text{rir}} = [A_d]_{d\in\mathcal{D}},$$

representing grounded consistency and separation relations between immediate neighbors.

To extend relational grounding beyond the fixed local neighborhood, we additionally sample regional pairs during training. Positive pairs connect locations within the same tumor component, whereas negative pairs connect tumor and background locations across different spatial ranges. The same affinity predictor is applied to these sampled pairs through $\mathcal{L}_{\text{region}}$, encouraging regional tumor consistency and tumor-background separation. Exact sampling details are provided in Appendix C.

Boundary evidence. To complement affinity-based consistency with explicit separation evidence, source annotations supervise boundary-positive locations adjacent to tumor-background transitions and boundary-negative confident tumor-interior locations without such transitions. Both the structural boundary representation and its supervision are derived from the same directional disaffinities $1 - A_d$. For relational integration, B_{rir} averages the two largest valid directional disaffinities at each location, providing a stable summary of local boundary structure. Boundary supervision instead uses the single largest valid directional disaffinity to capture the strongest local separation cue. The complete grounding objective is

$$\mathcal{L}_{\text{affinity}} = \mathcal{L}_{\text{dir}} + 0.5\,(\mathcal{L}_{\text{bnd-BCE}} + \mathcal{L}_{\text{bnd-Dice}}) + 0.5\,\mathcal{L}_{\text{region}}. \tag{4}$$

The directional affinities and boundary evidence are combined as

$$A_9 = [A_{\text{rir}}, B_{\text{rir}}],$$

which forms the grounded structural query passed with E_{rir} to relational integration. Source annotations supervise relational grounding only during training, while relational inference requires no annotation.

3.4 Image-Anchored Relational Integration

We first retrieve relational context through local cross-attention, using the structural query A_9 and relational representation E_{rir} as $Q = A_9$ and $K = V = E_{\text{rir}}$. The resulting context C_{att} is added to E_{rir} through a gated residual update,

$$\widetilde{E}_{\text{rir}} = E_{\text{rir}} + \gamma C_{\text{att}}, \qquad \gamma = 0.2\tanh(\hat{\gamma}). \tag{5}$$

The contextualized representation is then combined with the grounded structural query as

$$R_{\text{rir}} = [A_9, \widetilde{E}_{\text{rir}}],$$

retaining both explicit affinity and boundary evidence from A_9 and contextualized relational information from E_{rir}.

The subsequent image-guided correction conditions this relational evidence on the original image representation. Specifically, F_{abs} serves as the query and R_{rir} as both key and value, producing an image-conditioned relational correction Δ_{rir}. The correction is applied through a second gated residual update,

$$F_{\text{guided}} = F_{\text{abs}} + \beta\Delta_{\text{rir}}, \qquad \beta = 0.2\tanh(\hat{\beta}). \tag{6}$$

Thus, the relational evidence neither replaces nor is directly concatenated with the original image representation. F_{abs} remains the residual anchor, while the relational branch contributes an image-conditioned correction controlled by a bounded scalar gate before the segmentation head. Exact local-attention and projection details are provided in Appendix D.

The segmentation branch is optimized with equally weighted binary cross-entropy and soft Dice losses, and the complete framework jointly combines the segmentation and relational grounding objectives,

$$\mathcal{L}_{\text{seg}} = \frac{1}{2}\mathcal{L}_{\text{BCE}} + \frac{1}{2}\mathcal{L}_{\text{Dice}}, \qquad \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{affinity}}. \tag{7}$$

Additional optimization and gradient-flow details are provided in Appendix E.

3.5 SOURCE-ONLY DEPLOYMENT SELECTION

The deployment checkpoint is selected solely from held-out source-validation statistics. Let $D_e^{0.5}$ denote the source-validation Dice of checkpoint e evaluated at a fixed decision threshold of 0.5. We first retain checkpoints whose Dice remains within a predefined quality range of the best source-validation performance,

$$\mathcal{C}_{\text{sel}} = \left\{ e \;\middle|\; D_e^{0.5} \geq 0.95 \max_j D_j^{0.5} \right\}.$$

The fixed 0.95 factor is a predefined design choice that imposes a source-Dice quality constraint before sensitivity-based ranking. It is determined independently of external-cohort performance and prevents sensitivity from being optimized at the expense of substantial source-validation segmentation quality. Among the retained checkpoints, we select the checkpoint with the highest source sensitivity at the same threshold. Ties are resolved by higher source Dice and then by the earlier epoch.

Checkpoint selection and decision-threshold selection are performed sequentially. After fixing the checkpoint, we select the decision threshold using only the held-out source-validation set. The threshold is searched from 0.10 to 0.90 in increments of 0.05 and chosen to maximize mean patient-level Dice, with ties resolved in favor of the lower threshold. The threshold search does not affect checkpoint ranking.

Neither stage uses external-cohort statistics, target-domain observations, or target-domain performance. The selected checkpoint and source-derived decision threshold together define the final deployment configuration. This source-only selection procedure is treated as the *Select* component of the complete RIR framework rather than as a post-hoc target-domain evaluation step. Additional selection details are provided in Appendix F.

4 EXPERIMENTS

We evaluate the complete RIR framework under external shifts using a broader DG comparison, a controlled source-size study, and a Select-stage ablation with varying labeled source supervision.

4.1 EXPERIMENTAL SETUP

Data and evaluation protocol. The source cohort contains 147 glioma patients, divided into 102 training, 22 validation, and 23 source-internal test patients. We use five nested training subsets with $N \in \{10, 20, 31, 51, 102\}$ while keeping the validation and internal-test sets fixed. All inputs are FLAIR images, and the task is 2D whole-tumor binary segmentation.

External evaluation uses all 95 prepared eligible BraTS-Africa Glioma patients and all 51 prepared eligible BraTS-Africa Other Neoplasms patients. No additional preprocessing failures were recorded in the final prepared-data pipeline. BraTS-Africa Glioma retains the broad source pathology category, whereas BraTS-Africa Other Neoplasms is a heterogeneous diagnostic-category cohort beyond the source glioma category.

Compared methods. Figure 1 provides a broader comparison with recent medical segmentation DG methods, including DCON, TSIAA, and AdverIN. The repeated source-size study uses BASE, RandConv, DSU, and RIR under the same nested source subsets, three random seeds, shared U-Net backbone, fixed data splits, and target-blind external evaluation. RandConv provides an input-level appearance perturbation control, while DSU perturbs feature statistics. RIR is evaluated as the complete proposed framework, including the source-only deployment selection in Sec. 3.5. No external-cohort information is used for checkpoint or decision-threshold selection. Appendix H.1 additionally compares the same trained RIR trajectories with and without the proposed Select stage to isolate its deployment-level contribution.

Metrics and repeated runs. Dice and sensitivity are computed per patient and averaged within each cohort. The primary endpoint is two-cohort macro Dice, with sensitivity as a secondary cohort-specific diagnostic. Results are mean $\pm$ population standard deviation over three runs. Exact seeds, statistical conventions, and Select-ablation aggregation are provided in Appendix H.

Table 1: **External Dice across source supervision sizes.** Values are mean ± population standard deviation over three runs. The macro averages BraTS-Africa Glioma and BraTS-Africa Other Neoplasms within each matched run. Bold indicates the highest unrounded mean.

Method	$N = 10$	$N = 20$	$N = 31$	$N = 51$	$N = 102$
Two-cohort external macro Dice					
BASE	0.470 ± 0.037	0.481 ± 0.044	0.416 ± 0.058	0.547 ± 0.011	0.570 ± 0.007
RandConv	0.361 ± 0.004	0.410 ± 0.021	0.452 ± 0.045	0.478 ± 0.052	0.525 ± 0.024
DSU	0.466 ± 0.017	0.475 ± 0.011	0.471 ± 0.020	0.562 ± 0.006	0.573 ± 0.015
RIR	$\mathbf{0.548} \pm 0.006$	$\mathbf{0.499} \pm 0.027$	$\mathbf{0.517} \pm 0.046$	$\mathbf{0.565} \pm 0.013$	$\mathbf{0.589} \pm 0.009$
BraTS-Africa Glioma Dice					
BASE	0.533 ± 0.032	0.543 ± 0.047	0.481 ± 0.062	0.614 ± 0.012	0.637 ± 0.008
RandConv	0.434 ± 0.005	0.474 ± 0.024	0.518 ± 0.050	0.543 ± 0.053	0.591 ± 0.022
DSU	0.529 ± 0.014	0.536 ± 0.008	0.546 ± 0.019	0.630 ± 0.007	0.640 ± 0.016
RIR	$\mathbf{0.607} \pm 0.007$	$\mathbf{0.562} \pm 0.024$	$\mathbf{0.579} \pm 0.038$	$\mathbf{0.633} \pm 0.015$	$\mathbf{0.649} \pm 0.006$
BraTS-Africa Other Neoplasms Dice					
BASE	0.407 ± 0.042	0.420 ± 0.041	0.351 ± 0.054	0.481 ± 0.012	0.504 ± 0.008
RandConv	0.288 ± 0.009	0.346 ± 0.019	0.385 ± 0.041	0.412 ± 0.051	0.458 ± 0.026
DSU	0.404 ± 0.019	0.414 ± 0.016	0.396 ± 0.021	0.493 ± 0.008	0.507 ± 0.014
RIR	$\mathbf{0.490} \pm 0.007$	$\mathbf{0.437} \pm 0.031$	$\mathbf{0.455} \pm 0.054$	$\mathbf{0.496} \pm 0.010$	$\mathbf{0.528} \pm 0.012$

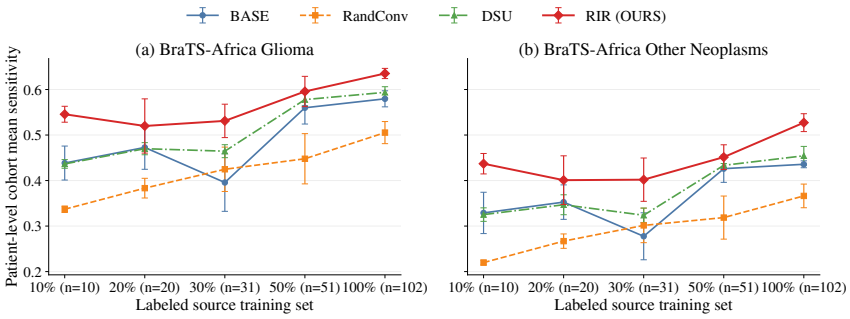


Figure 3: **Sensitivity across external cohorts and source sizes.** (a) BraTS-Africa Glioma. (b) BraTS-Africa Other Neoplasms. Mean ± population standard deviation over three runs.

4.2 Utility across Source Supervision Sizes

Table 1 shows that the complete RIR framework achieves the highest mean two-cohort external macro Dice among the four controlled-study methods at all five source sizes. The Select-stage ablation in Appendix H.1 further shows that RIR without Select exceeds BASE in mean macro Dice at four of five source sizes. Adding the proposed Select stage increases mean macro Dice at all five sizes, with gains of 0.001–0.056 over the same RIR trajectories without Select, thereby isolating the deployment contribution from the jointly optimized *Represent–Ground–Integrate* pathway.

4.3 Sensitivity under External Shifts

Figure 3 shows that the complete RIR framework achieves the highest mean sensitivity on both external cohorts at every source size. Sensitivity complements Dice and has shown more consistent agreement with perceived brain tumor segmentation quality (Hoebel et al., 2023). Appendix H.1 shows that Select adds 0.012–0.089 mean macro sensitivity while mean macro Dice also increases at every source size. No external-cohort sensitivity is used during selection.

4.4 Qualitative Segmentation under External Shifts

Figure 4 illustrates prediction differences on selected external cases. RIR recovers more of the annotated tumor and increases the displayed Dice from 0.472 to 0.654 on BraTS-Africa Glioma and from 0.252 to 0.521 on BraTS-Africa Other Neoplasms. These selected examples complement the cohort-level results in Table 1 and Fig. 3 and are not intended to represent the full cohort distributions.

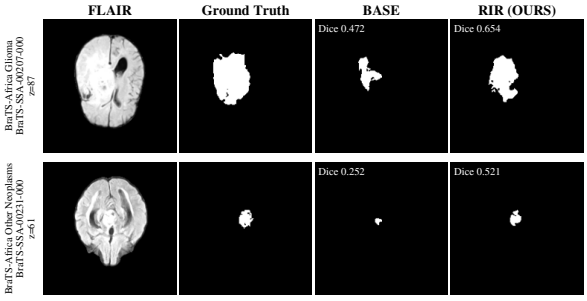


Figure 4: **Selected external segmentation examples.** Rows show BraTS-Africa Glioma and Other Neoplasms. Columns show FLAIR, ground truth, BASE, and RIR.

5 Discussion

RIR examines whether source-derived relational evidence remains useful when deployment extends beyond source pathology coverage. Across all five source sizes, the complete framework achieves the highest mean two-cohort external macro Dice among the four controlled methods and the highest mean sensitivity on both external cohorts. The Select-stage ablation in Appendix H.1 further shows that source-only deployment selection provides additional Dice and sensitivity gains over the same learned RIR pathway. Together, these results support the utility of the complete framework while separating the deployment contribution from the jointly optimized relational pathway.

Absolute external Dice remains modest across methods, while Fig. 1 shows substantial degradation beyond the source glioma category across several DG strategies. We therefore interpret the results as evidence that source-derived relational structure remains useful under a challenging external shift rather than as uniformly large gains or clinical readiness. Appendix G provides complementary representation-level evidence without establishing pathology invariance or causal transferability.

More broadly, the findings motivate pathology coverage as an additional robustness dimension in medical DG. BraTS-Africa Other Neoplasms is heterogeneous and differs from the source in factors beyond tumor category, so the study does not causally isolate pathology identity. Limitations include FLAIR-only 2D segmentation and one fixed nested source-subset sequence. Future work should examine multimodal 3D settings and broader source sampling.

6 Conclusion

We studied whether source-derived relational evidence remains useful for single-source brain tumor segmentation beyond source pathology coverage. RIR combines rank-derived representation, source-supervised grounding, image-anchored integration, and source-only deployment selection without target-domain feedback. Across five source-supervision sizes, the complete framework shows utility in Dice and sensitivity, while qualitative and representation-level analyses provide complementary evidence for useful relational structure under the evaluated external shifts. The broader comparison further highlights pathology coverage as an important robustness dimension in medical DG. Overall, our findings demonstrate the potential of RIR as a source-only framework for robust brain tumor segmentation under pathology-associated external shifts.

AI USE STATEMENT

Generative AI tools were used only for limited language editing and stylistic refinement of the manuscript, including improvements to grammar, clarity, and readability. They were not used to generate the scientific claims, methodology, experimental results, or conclusions presented in this work. All AI-assisted edits were reviewed by the authors to ensure technical accuracy and consistency with the intended scientific content. We take responsibility for the final content of this work, including any text revised with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The source data split, source-supervision sizes, evaluation protocol, and compared methods are described in Sec. 4. Detailed definitions of the RIR representation, relational grounding, relational integration, training objective and gradient flow, and source-only deployment selection are provided in Appendices B–F. The target-distribution characterization and representation-level analysis are documented in Appendices A and G, respectively, while qualitative segmentation examples are provided in Fig. 4. Exact random seeds, patient-level aggregation rules, the population standard-deviation convention, and matched-seed comparison procedures are specified in Appendix H. The same appendix also reports the Select-stage ablation, including the conventional deployment rule used for RIR without Select and its comparison with the complete RIR framework. Together, these sections document the model construction, deployment procedure, evaluation protocol, and statistical reporting used in this study.

REFERENCES

Saioa Alzola-Aldamizetxebarria, Laura Fernández-Méndez, Daniel Padro, Jesús Ruíz-Cabello, and Pedro Ramos-Cabrer. A comprehensive introduction to magnetic resonance imaging relaxometry and contrast agents. *ACS omega*, 7(42):36905–36917, 2022.

Meng Cao and Songcan Chen. A simple yet lightweight module for enhancing domain generalization through relative representation. *Pattern Recognition*, pp. 112423, 2025.

Dong Kyu Cho, Inwoo Hwang, and Sanghack Lee. Peer pressure: Model-to-model regularization for single source domain generalization. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15360–15370. IEEE, 2025.

Neel Dey, Benjamin Billot, Hallee Wong, Clinton Wang, Mengwei Ren, Ellen Grant, Adrian Dalca, and Polina Golland. Learning general-purpose biomedical volume representations using randomized synthesis. In *International Conference on Learning Representations*, volume 2025, pp. 32033–32064, 2025.

Nikos Efthymiadis, Giorgos Tolias, and Ondřej Chum. Crafting distribution shifts for validation and training in single source domain generalization. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1883–1892. IEEE, 2025.

Ahmed Elzayat, Nourhan Hanafy, Mariem Magdy, Hazem Zakaria, Mina Tayeh, Abdulkhalek Al-Fakih, Yeong Hyeon Gu, Mohammed A Al-Masni, and Meena M Makary. Distilling population specific expertise into a unified model for generalizable brain tumor segmentation. *Scientific Reports*, 16(1):12969, 2026.

F Javier Gil-Terrón, Pablo Ferri, Víctor Montosa-i Micó, María Gómez Mahiques, Carles Lopez-Mateu, Pau Martí, Juan M García-Gómez, and Elies Fuster-Garcia. Exploring the trade-off between generalist and specialized models: A center-based comparative analysis for glioblastoma segmentation. *International Journal of Medical Informatics*, 191:105604, 2024.

Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021.

Katharina V Hoebel, Christopher P Bridge, Sara Ahmed, Oluwatosin Akintola, Caroline Chung, Raymond Y Huang, Jason M Johnson, Albert Kim, K Ina Ly, Ken Chang, et al. Expert-centered evaluation of deep learning algorithms for brain tumor segmentation. *Radiology: Artificial Intelligence*, 6(1):e220231, 2023.

Yunzhi Huang, Luyi Han, and Haoran Dou. Generative feature style augmentation for domain generalization in medical image segmentation. *Pattern Recognition*, 162:111416, 2025.

Jia-Xuan Jiang, Yuee Li, and Zhong Wang. Structure-aware single-source generalization with pixel-level disentanglement for joint optic disc and cup segmentation. *Biomedical Signal Processing and Control*, 99:106801, 2025.

Rushi Jiao, Yichi Zhang, Le Ding, Bingsen Xue, Jicong Zhang, Rong Cai, and Cheng Jin. Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. *Computers in Biology and Medicine*, 169:107840, 2024.

Chuang Liu, Yichao Cao, YingYing Zhang, Xiu Su, and Haogang Zhu. Perturbating, tuning, and collaborating: Harnessing vision foundation models for single domain generalization on medical imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5370–5378, 2025a.

Quande Liu, Qi Dou, and Pheng-Ann Heng. Shape-aware meta-learning for generalizing prostate mri segmentation to unseen domains. In *International conference on medical image computing and computer-assisted intervention*, pp. 475–485. Springer, 2020.

Quande Liu, Cheng Chen, Qi Dou, and Pheng-Ann Heng. Single-domain generalization in medical image segmentation via test-time adaptation from shape dictionary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 1756–1764, 2022.

Wentao Liu, Zhiwei Ni, Xuhui Zhu, Qian Chen, Liping Ni, and Pingfan Xia. Spectrum intervention based invariant causal representation learning for single-domain generalizable medical image segmentation. *Medical Image Analysis*, pp. 103741, 2025b.

Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai, and Daniel Rueckert. Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(4):1095–1106, 2022.

Bijuan Ren, Yanfeng Li, Jia Sun, Houjin Chen, and Luyifu Chen. Anatomically-robust and feature-unbiased domain generalization for medical segmentation. *Expert Systems with Applications*, pp. 129752, 2025.

Heather Sheridan and Eyal M Reingold. The holistic processing account of visual expertise in medical image perception: A review. *Frontiers in psychology*, 8:1620, 2017.

Zixian Su, Kai Yao, Xi Yang, Kaizhu Huang, Qiufeng Wang, and Jie Sun. Rethinking data augmentation for single-source domain generalization in medical image segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2366–2374, 2023.

Ruofan Wang, Jintao Guo, Jian Zhang, Lei Qi, Qian Yu, and Yinghuan Shi. A hybrid dual-augmentation constraint framework for single-source domain generalization in medical image segmentation. *Pattern Recognition*, 170:112082, 2026.

Zanting Ye, Ke Wang, Wenbing Lv, Qianjin Feng, and Lijun Lu. Fsda-dg: Improving cross-domain generalizability of medical image segmentation with few source domain annotations. *Medical Image Analysis*, 105:103704, 2025.

Jingjun Yi, Qi Bi, Hao Zheng, Haolan Zhan, Wei Ji, Yawen Huang, Shaoxin Li, Yuexiang Li, Yefeng Zheng, and Feiyue Huang. Hallucinated style distillation for single domain generalization in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 438–448. Springer, 2024.

Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Bradford J Wood, Holger Roth, Andriy Myronenko, Daguang Xu, et al. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE transactions on medical imaging*, 39(7):2531–2540, 2020.

Siqi Zhang, Qizhe Zhang, Shanghang Zhang, Xiaohong Liu, Jingkun Yue, Ming Lu, Huihuan Xu, Jiaxin Yao, Xiaobao Wei, Jiajun Cao, et al. A generalist foundation model and database for open-world medical image segmentation. *Nature Biomedical Engineering*, pp. 1–16, 2025.

Zheyuan Zhang, Bin Wang, Lanhong Yao, Elif Keles, Debesh Jha, Matthew Antalek, Gorkem Durak, Alpay Medetalibeyoglu, Concetto Spampinato, Baris Turkbey, et al. Adverin: Monotonic adversarial intensity attack for domain generalization in medical image segmentation. *Medical Image Analysis*, pp. 103848, 2026.

Ziquan Zhu, Hanruo Zhu, Si-Yuan Lu, Xiang Li, Yanda Meng, Yunxiao Zhang, Gaojie Jin, Lu Yin, Lijie Hu, Di Wang, et al. Dual-kernel adapter: Expanding spatial horizons for data-constrained medical image analysis. In *International Conference on Learning Representations*, volume 2026, pp. 116057–116088, 2026.

A TARGET DISTRIBUTION CHARACTERIZATION

To complement the target-distribution summary in Fig. 1, we characterize the three cohorts used in that analysis with descriptor-level distributions and pairwise distances. The analysis considers tumor burden, lesion multiplicity, surface irregularity, and centroid displacement, which capture complementary aspects of whole-tumor extent, lesion organization, shape, and spatial location. Tumor burden is measured as the percentage of the valid field of view occupied by the whole-tumor mask, while lesion multiplicity is defined as the number of three-dimensional connected tumor components per scan.

For each descriptor k, values from all three cohorts are pooled before normalization. Let P_k denote the pooled descriptor values, with pooled median m_k and pooled interquartile range s_k. Each descriptor value is normalized as

$$x_k^{\text{norm}} = \frac{x_k - m_k}{s_k}.$$

If the pooled interquartile range is zero, the implementation falls back to a scaled median absolute deviation and then to the pooled standard deviation. Pairwise cohort differences for descriptor k are then computed as

$$D_k(i,j) = W_1\left(X_{i,k}^{\text{norm}}, X_{j,k}^{\text{norm}}\right),$$

where W_1 denotes the one-dimensional Wasserstein distance. The compact target-distribution summary shown in Fig. 1(a) is the equal-weight mean across the four descriptors,

$$D(i,j) = \frac{1}{4}\sum_{k=1}^{4} D_k(i,j).$$

This aggregation is used as a descriptive summary rather than as a learned or clinically weighted distance.

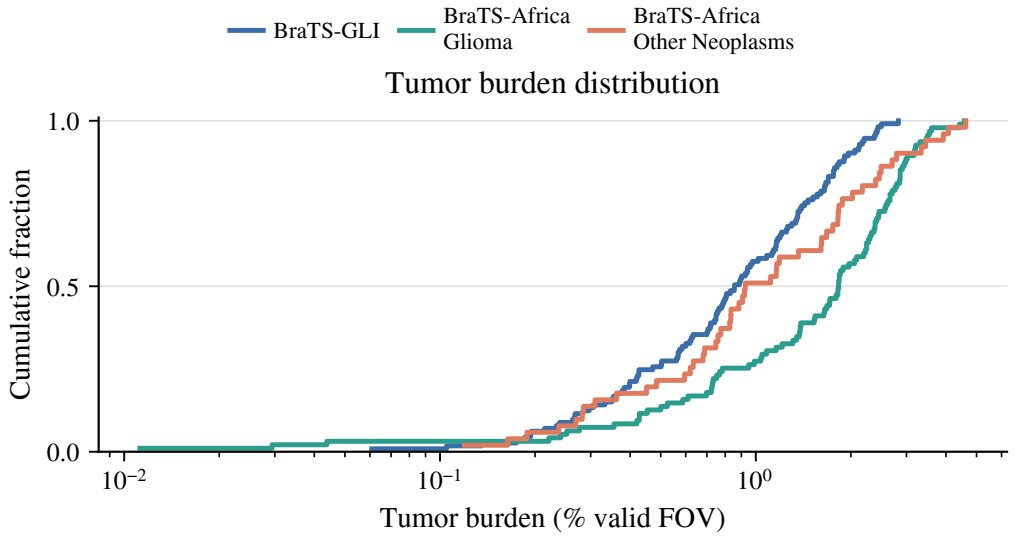


Figure A.1: **Tumor burden distributions across evaluation cohorts.** Empirical cumulative distributions of whole-tumor burden measured as a percentage of the valid field of view.

Figure A.1 shows overlapping but distinct tumor-burden distributions across BraTS-GLI, BraTS-Africa Glioma, and BraTS-Africa Other Neoplasms. The differences are therefore not described by a single directional change shared by both external cohorts. Instead, the three cohorts occupy different ranges and cumulative profiles of tumor extent.

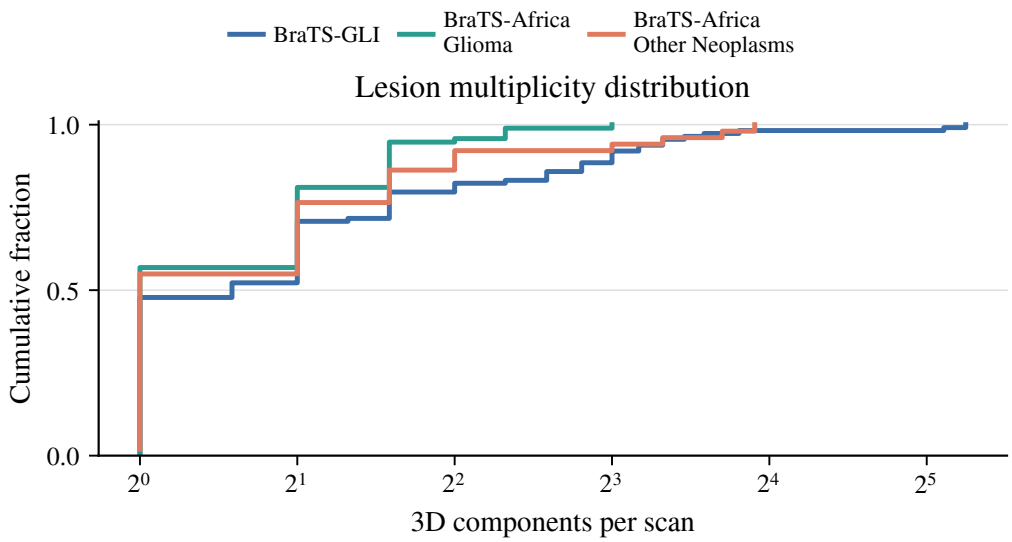


Figure A.2: **Lesion multiplicity distributions across evaluation cohorts.** Empirical cumulative distributions of the number of three-dimensional tumor components per scan.

Lesion multiplicity reveals a complementary form of target-structure variation, as shown in Fig. A.2. All three cohorts contain substantial numbers of low-multiplicity cases, while their cumulative profiles and upper tails differ. Thus, the external cohorts do not depart from the source through a single common structural change and instead exhibit distinct lesion-organization patterns.

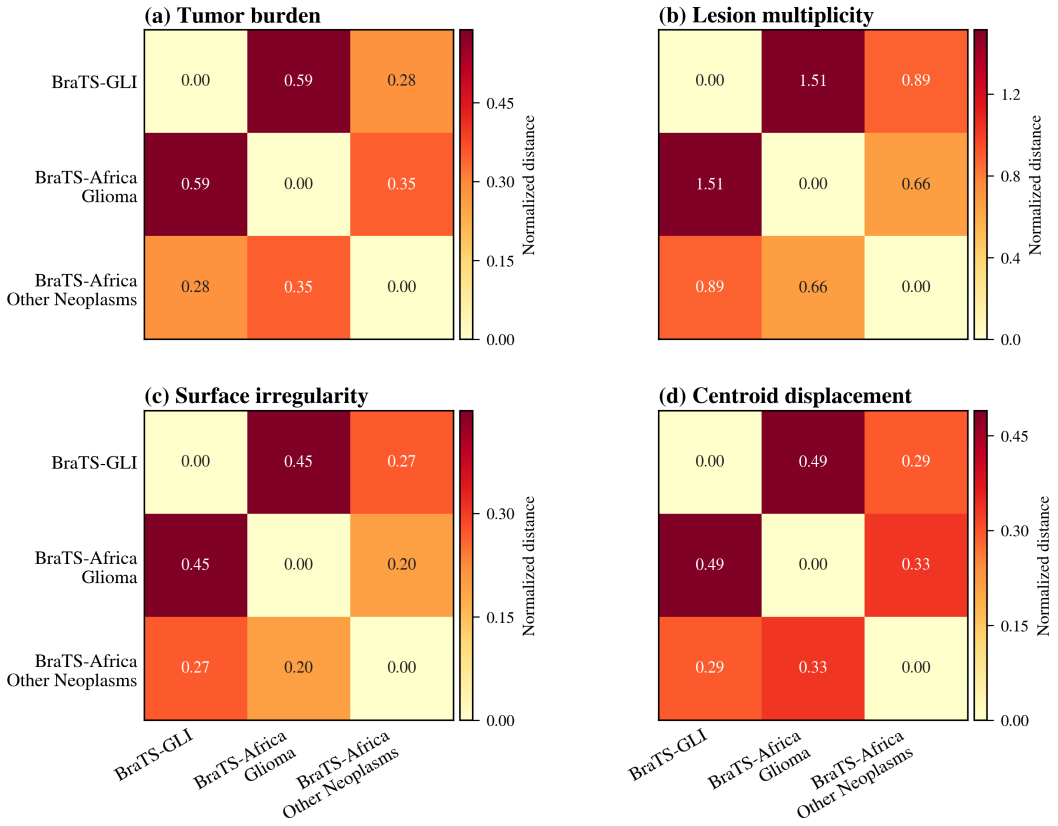


Figure A.3: **Descriptor-specific pairwise target-distribution distances.** Normalized Wasserstein distances are shown separately for (a) tumor burden, (b) lesion multiplicity, (c) surface irregularity, and (d) centroid displacement. Each panel uses an independent color scale to emphasize within-descriptor pairwise differences. Color intensity should therefore be interpreted within each panel rather than compared across panels.

Figure A.3 decomposes the aggregate target-distribution summary in Fig. 1(a) into its four descriptor-specific components. The pairwise distances show that cohort differences arise along multiple aspects of the segmentation target rather than through tumor burden or lesion multiplicity alone. Different cohort pairs also exhibit different relative differences across descriptors, indicating that target-distribution shift is heterogeneous across structural axes. Because each panel uses an independent color scale, color intensity should be compared only among cohort pairs within the same descriptor.

Taken together, these analyses characterize the evaluation cohorts as occupying distinct regions of a multi-dimensional target-structure space spanning tumor extent, multiplicity, shape, and spatial organization. The equal-weight distance reported in the main text is intended only as a compact descriptive summary of this multi-axis heterogeneity. These cohort-level differences motivate treating pathology coverage as an explicit evaluation dimension without assuming that absolute morphology, lesion burden, spatial location, or ordinal intensity position is preserved across tumor categories.

B Relative Intensity Relation Construction

This section provides the implementation details of the Relative Intensity Relation (RIR) representation introduced in Sec. 3.2. RIR is constructed from the within-volume ordinal organization of each normalized input image and consists of two complementary rank-derived cues, namely soft ordinal position and local directional rank relations.

Volume-relative intensity rank. For each normalized image x, we assign every spatial location p a volume-relative intensity rank $\rho(p) \in [0, 1]$. The rank is computed only from finite normalized voxels within the valid field of view of the same volume. We construct a quantile lookup table using 257 uniformly spaced cumulative-probability points

$$u_j = \frac{j}{256}, \qquad j = 0, \ldots, 256,$$

and their corresponding empirical quantiles. At runtime, the rank of $x(p)$ is obtained by locating its surrounding quantile entries using `searchsorted`, linearly interpolating the corresponding cumulative probabilities, and clamping the result to $[0, 1]$. Thus, $\rho(p)$ represents the quantile position of $x(p)$ within its own volume rather than its original numerical intensity. This transformation preserves the ordering of intensities within the volume while discarding their absolute scale.

Soft ordinal and directional relation cues. The volume-relative rank is first converted into an eight-channel soft ordinal seed $S(p) \in \mathbb{R}^8$. We use eight uniformly distributed ordinal centers

$$\mu_k = \frac{k + 0.5}{8}, \qquad k = 0, \ldots, 7,$$

with a fixed Gaussian width

$$\sigma_s = 0.125.$$

For each valid location, the ordinal logits are

$$\ell_k(p) = -\frac{(\rho(p) - \mu_k)^2}{2\sigma_s^2},$$

and the soft ordinal seed is obtained by channel-wise softmax,

$$S_k(p) = M(p) \frac{\exp(\ell_k(p))}{\sum_{j=0}^{7} \exp(\ell_j(p))}, \tag{B.1}$$

where $M(p) \in \{0, 1\}$ denotes the valid-field-of-view mask. Hence, locations outside the valid field of view contribute zero-valued ordinal seed channels.

To explicitly represent local ordinal relationships, we additionally compare the rank at p with those of its eight spatial neighbors. We use the direction set

$$\mathcal{D} = \{\mathrm{N}, \mathrm{S}, \mathrm{W}, \mathrm{E}, \mathrm{NW}, \mathrm{NE}, \mathrm{SW}, \mathrm{SE}\}.$$

For each valid neighboring pair $(p, p+d)$, the directional ordinal relation is computed as

$$r_d(p) = \sigma\left(\frac{\rho(p) - \rho(p+d)}{0.1}\right), \qquad d \in \mathcal{D}, \tag{B.2}$$

where $\sigma(\cdot)$ denotes the logistic sigmoid and the fixed temperature 0.1 controls the sharpness of the directional comparison.

For each direction, we additionally maintain a pair-validity mask $M_d(p)$, which is one only when both endpoints belong to the valid field of view. When $p+d$ is invalid, the corresponding directional relation is assigned the neutral value 0.5. The associated rank-affinity quantity used by the grounding target is set to zero, and the pair-validity mask excludes the pair from supervised losses and structural aggregation. Thus, the neutral value serves only as a numerically well-defined placeholder and is not interpreted as a supervised relation.

The eight directional responses are stacked as

$$R_{\mathrm{ord}}(p) = [r_d(p)]_{d\in\mathcal{D}} \in \mathbb{R}^8,$$

and concatenated with the eight soft ordinal seed channels to obtain

$$Z_{\mathrm{ord}} = \mathrm{Concat}\,(S, R_{\mathrm{ord}}) \in \mathbb{R}^{B\times 16\times H\times W}. \tag{B.3}$$

Accordingly, Z_{ord} jointly represents the ordinal position of each location within the volume and its relative ordering with respect to neighboring locations.

RIR encoder. Only the stride-1 RIR pathway is instantiated in the final model. The active RIR block operates at 16 channels throughout. It first applies a 1×1 convolution from 16 to 16 channels, followed by a depthwise 3×3 convolution and a pointwise 1×1 convolution. The resulting features are processed by eight-group Group Normalization, GELU activation, and dropout with probability 0.05. A second depthwise–pointwise convolutional stage followed by Group Normalization and GELU is applied with a residual connection across the block.

The resulting block output is mapped to the learned relational embedding by a final 1×1 projection,

$$E_{\mathrm{rir}} = f_{\mathrm{RIR}}\,(Z_{\mathrm{ord}}) = \mathrm{Proj}_{1\times 1}\,(\mathrm{RIRBlock}\,(Z_{\mathrm{ord}}))\,, \tag{B.4}$$

where the final projection maps 16 channels to 16 channels while preserving spatial resolution. The resulting representation therefore satisfies

$$E_{\mathrm{rir}} \in \mathbb{R}^{B\times 16\times H\times W}.$$

No sigmoid, clipping, or explicit range normalization is applied to the output of the RIR Encoder. Unlike the directional relations in Eq. B.2, which are bounded ordinal cues, E_{rir} is treated as an unconstrained latent relational representation. The rank-derived cues are not assumed to be invariant across tumor categories. Instead, they provide an input-derived relational basis from which the learned encoder can construct task-adaptive representations.

No separate direct supervision target is imposed on E_{rir}. During end-to-end training, gradients from the tumor-relevant affinity, boundary, and sampled region-affinity objectives propagate through the grounding pathway into the RIR Encoder. Gradients from the subsequent relational integration and segmentation objective also update the RIR pathway. Thus, the ordinal cues are constructed directly from each input without annotation, while the learnable RIR parameters are optimized exclusively through source-domain supervision.

C Pathology-Relevant Relational Grounding Details

This section provides the implementation details of the pathology-relevant relational grounding described in Sec. 3.3, including local affinity supervision, the adaptive BB target, regional pair sampling, and boundary supervision.

Pair categorization and directional targets. For grounding supervision, a source-annotation location is treated as tumor when

$$y(p) \geq 0.75,$$

and as background when

$$y(p) \leq 0.$$

Locations between these thresholds are treated as ambiguous and excluded from pairwise supervision. A directional pair is further required to satisfy the directional validity mask $M_d(p)$.

Valid neighboring pairs are categorized as tumor–tumor (TT), tumor–background (TB), or background–background (BB). TT pairs receive target affinity one and TB pairs receive target affinity zero. BB pairs instead receive the adaptive soft target described below.

The Shared Affinity Head operates directly on absolute differences between the learned 16-channel relational embeddings,

$$|E_{\mathrm{rir}}(p) - E_{\mathrm{rir}}(q)|\,.$$

No additional pairwise embedding projection is applied before this difference. The shared predictor consists of a 1×1 projection from 16 to 16 channels, GELU activation, a 1×1 projection from 16 channels to one channel, and a sigmoid output. The same predictor parameters are shared across all local directions and sampled regional pairs, so both local and regional grounding operate in the same learned affinity space.

For numerical stability, probability-space binary cross-entropy is evaluated in FP32 after clipping predicted affinities to

$$[10^{-6}, 1 - 10^{-6}].$$

TT, TB, and BB losses are reduced separately over their valid members. If a group contains no valid pair, its loss is represented by a differentiable zero. The directional objective is the sum of the three group losses,

$$\mathcal{L}_{\mathrm{dir}} = \mathcal{L}_{\mathrm{TT}} + \mathcal{L}_{\mathrm{TB}} + 0.5\,\mathcal{L}_{\mathrm{BB}},$$

rather than an average across pair categories.

Adaptive BB soft target. Background regions can contain heterogeneous structures and are therefore not forced to have uniformly high affinity. For a valid BB pair $(p, p+d)$, the decoder-space similarity is obtained from the current forward pass as

$$a_{\mathrm{dec},d}(p) = M_d(p)\,\mathrm{clip}_{[0,1]}\left(\frac{1 + \cos(F(p), F(p+d))}{2}\right), \tag{C.1}$$

where F denotes the current 32-channel U-Net image feature used by the model.

The rank-seed affinity is computed directly from the eight-channel soft ordinal seeds defined in Eq. B.1,

$$a_{\mathrm{rank},d}(p) = M_d(p) \sum_{k=0}^{7} \sqrt{S_k(p) S_k(p+d)}. \tag{C.2}$$

This quantity measures the overlap between the ordinal membership distributions of the two locations. It is computed from the fixed rank-derived seeds and does not depend on the learned embedding E_{rir}.

Let

$$\pi(p) = \sigma(z(p))$$

denote the current segmentation probability. Prediction confidence for the pair is computed as

$$c_d(p) = M_d(p) \min\left(2|\pi(p) - 0.5|, 2|\pi(p+d) - 0.5|\right). \tag{C.3}$$

The BB supervision target combines decoder-space similarity and rank-seed affinity according to the current prediction confidence,

$$s_d(p) = c_d(p) a_{\mathrm{dec},d}(p) + (1 - c_d(p))\, a_{\mathrm{rank},d}(p). \tag{C.4}$$

The rank-seed affinity is masked by directional validity and is zero for invalid pairs.

The decoder feature and segmentation probability used in Eqs. C.1–C.4 are obtained from the current model in the same forward pass. They do not come from an exponential-moving-average model or a separate teacher network. All quantities used to construct the BB target are evaluated without gradient propagation, and s_d is detached before being used as supervision. Therefore, the BB loss updates the Shared Affinity Head and RIR pathway through the predicted affinity but does not back-propagate through the image feature or segmentation probability used to construct the target.

Regional affinity sampling. The local directional objective supervises only immediate neighboring locations. To provide longer-range relational grounding, we additionally sample regional pairs during training using connected components of the source tumor annotation.

Positive regional pairs connect locations belonging to the same tumor component. Positive partners are sampled from three component-relative distance strata, P_{near}, P_{mid}, and P_{far}, using 0.25 and 0.50 of the corresponding component bounding-box diagonal as the near/mid distance breakpoints, with longer-distance pairs assigned to the far stratum.

Negative regional pairs use a tumor location as the anchor and a background location as its partner. They are divided according to pixel distance from the tumor region into

$$N_{\text{boundary}} \quad d \le 2, \qquad N_{\text{mid}} \quad 2 < d \le 8, \qquad N_{\text{far}} \quad d > 8.$$

At most 64 tumor anchors are sampled from each slice, with at most 16 anchors drawn from any single tumor component. For each anchor, one partner is sampled from every available positive or negative distance stratum.

Sampling is deterministic for a fixed training run. The sampling seed is derived from the training seed together with the zero-based epoch index, batch index, and batch-element index, ensuring reproducible pair construction without fixing the same regional pairs throughout training.

The same Shared Affinity Head used for directional affinity prediction is applied to the sampled regional pairs. Positive pairs receive target one and negative pairs target zero. Each non-empty regional group is reduced by its own mean binary cross-entropy, and $\mathcal{L}_{\text{region}}$ is the unweighted mean over the available group losses. These sampled pairs extend the local supervision to regional tumor consistency and tumor-background separation without requiring spatial continuity of the tumor as a whole. The regional term enters the grounding objective with coefficient 0.5.

Boundary representation and supervision. Boundary evidence is derived from the same directional affinities used to represent local consistency and separation. Let

$$D_d(p) = (1 - A_d(p))\, M_d(p).$$

For the structural query used in relational integration, $B_{\text{rir}}(p)$ is the mean of the two largest valid directional disaffinities at p. The Top-2 aggregation reduces dependence on a single orientation while preserving strong local separation evidence.

Boundary supervision intentionally uses a different aggregation. The boundary prediction is

$$\widetilde{B}(p) = \max_{d \in \mathcal{D}} D_d(p). \tag{C.5}$$

A location is boundary-positive when at least one valid direction forms a tumor–background pair. Boundary-negative supervision is assigned only to confident tumor-interior locations that have at least one valid neighbor but no tumor–background direction. Background and ambiguous locations are not used as boundary negatives.

For boundary BCE, positive and negative samples are reduced separately and the available group means are then averaged. The boundary Dice term is evaluated over the full batch and spatial domain only when both positive and negative boundary supervision are present. Otherwise, the Dice contribution is set to a differentiable zero. This avoids unstable boundary Dice evaluation in batches containing only one supervision class.

The complete grounding objective is therefore

$$\mathcal{L}_{\text{affinity}} = \mathcal{L}_{\text{dir}} + 0.5\,(\mathcal{L}_{\text{bnd-BCE}} + \mathcal{L}_{\text{bnd-Dice}}) + 0.5\,\mathcal{L}_{\text{region}},$$

as defined in Sec. 3.3.

D RELATIONAL INTEGRATION DETAILS

This section provides the architectural details of the two-stage relational integration described in Sec. 3.4.

Relational context retrieval. The first cross-attention stage uses the nine-channel structural query

$$A_9 = [A_{\mathrm{rir}}, B_{\mathrm{rir}}]$$

as the query and E_{rir} as both key and value. The query is projected from 9 to 16 channels, whereas key and value remain in a 16-dimensional embedding space. Query, key, and value projections are bias-free.

Attention is evaluated within non-overlapping 8×8 local windows using four heads. The total embedding dimension is 16, yielding a per-head dimension of 4. Non-affine Layer Normalization is applied within the attention module. The attention output is projected from 16 to 16 channels and followed by single-group Group Normalization.

The resulting context C_{att} is incorporated through

$$\widetilde{E}_{\mathrm{rir}} = E_{\mathrm{rir}} + \gamma C_{\mathrm{att}}, \qquad \gamma = 0.2 \tanh(\hat{\gamma}),$$

and the contextualized representation is concatenated with the structural query,

$$R_{\mathrm{rir}} = [A_9, \widetilde{E}_{\mathrm{rir}}] \in \mathbb{R}^{B \times 25 \times H \times W}.$$

Image-guided correction. The second cross-attention stage uses the original image feature as its query and R_{rir} as key and value. The image query is projected from 32 to 32 channels and R_{rir} is projected from 25 to 32 channels. The query, key, and value projections are bias-free.

This stage also uses 8×8 local windows and four attention heads, but operates with a 32-dimensional embedding, yielding a per-head dimension of 8. The output projection maps 32 channels to 32 channels and is followed by single-group Group Normalization. The resulting correction is denoted by Δ_{rir}, and

$$F_{\mathrm{guided}} = F_{\mathrm{abs}} + \beta \Delta_{\mathrm{rir}}, \qquad \beta = 0.2 \tanh(\hat{\beta}).$$

Both $\hat{\gamma}$ and $\hat{\beta}$ are initialized to zero. Accordingly, the bounded learnable coefficients γ and β initially equal zero, so both residual updates begin from an identity state.

Image-feature integration. The U-Net image representation is 32-dimensional and is used directly as F_{abs}. The 32-channel relational correction is added to this feature, and F_{guided} is passed to the segmentation head. Thus, the relational branch does not replace the original image-derived representation. F_{abs} remains the residual anchor throughout relational integration.

Warm-up behavior. During the first ten training epochs, the structure-conditioned contextualization stage is bypassed. It is activated from epoch 11 onward. During the bypass phase,

$$\widetilde{E}_{\mathrm{rir}} = E_{\mathrm{rir}},$$

and

$$R_{\mathrm{rir}} = [A_9, E_{\mathrm{rir}}].$$

The second image-guided correction stage remains active throughout training, and the segmentation, local affinity, boundary, and regional affinity objectives are evaluated during both the warm-up and post-warm-up phases.

Because the first contextualizer is not executed during the first ten epochs, its attention parameters and $\hat{\gamma}$ receive no gradient and do not acquire optimizer state during this phase. They are also unaffected by AdamW weight decay while the contextualizer is bypassed.

E TRAINING OBJECTIVE AND GRADIENT FLOW

Segmentation objective. Let z denote the segmentation logits and

$$P = \sigma(z)$$

the corresponding probability map. The segmentation loss is

$$\mathcal{L}_{\text{seg}} = 0.5\,\mathcal{L}_{\text{BCE}} + 0.5\,\mathcal{L}_{\text{Dice}}.$$

$\mathcal{L}_{\text{BCE}}$ is the unweighted mean binary-cross-entropy-with-logits loss. Soft Dice is computed independently for each sample,

$$\mathcal{L}_{\text{Dice}}^{(i)} = 1 - \frac{2\sum_p P_i(p)y_i(p) + 10^{-6}}{\sum_p P_i(p) + \sum_p y_i(p) + 10^{-6}}, \tag{E.1}$$

and then averaged over the batch.

The complete objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{affinity}}.$$

There is no additional outer weighting coefficient applied to $\mathcal{L}_{\text{affinity}}$.

Gradient flow. The local TT/TB/BB affinity losses, boundary supervision, and sampled region-affinity objective update the Shared Affinity Head and propagate through the learned E_{rir} embedding into the active stride-1 RIR Encoder. The relational integration pathway additionally allows the segmentation objective to update the RIR representation.

In contrast, the decoder- and prediction-derived quantities used to construct the adaptive BB supervision target are detached. Consequently, the BB target does not introduce a gradient path through the image feature or segmentation prediction from which the target is constructed. The target is generated from the current forward pass and does not involve a separate teacher or exponential-moving-average model.

At initialization, $\beta = 0$ because $\hat{\beta} = 0$. Therefore, the image-guided attention body initially receives no segmentation gradient through the residual correction term, while the raw learnable parameter $\hat{\beta}$ can receive gradient. Once β moves away from zero, segmentation gradients propagate through the image-guided relational correction normally.

F SOURCE-ONLY DEPLOYMENT SELECTION DETAILS

This section provides the exact source-only selection protocol corresponding to Sec. 3.5. The selection procedure is an explicit component of the complete RIR framework and uses only held-out source-validation information.

Checkpoint selection. Validation statistics are computed at a fixed decision threshold of 0.5. Dice, precision, and sensitivity are first computed per patient volume and then averaged across patients without volume weighting.

For checkpoint e, let $D_e^{0.5}$ denote its mean source-validation Dice at threshold 0.5. Candidate checkpoints satisfy

$$\mathcal{C}_{\text{sel}} = \left\{ e \;\middle|\; D_e^{0.5} \geq 0.95 \max_j D_j^{0.5} \right\}.$$

The fixed 0.95 factor is a predefined design choice that imposes a source-Dice quality constraint before sensitivity-based ranking. It is defined independently of external-cohort performance. Within this candidate set, the checkpoint with the highest source-validation sensitivity at threshold 0.5 is selected. Ties are resolved first by higher source-validation Dice and then by the earlier epoch. Source precision is retained as a diagnostic but does not participate in checkpoint ranking.

The selector receives only the checkpoint epoch and source-validation Dice, precision, and sensitivity. External-cohort statistics, target-domain observations, target-domain performance, target-validation signals, and oracle quantities are not accepted as selection inputs.

Decision-threshold selection. After the deployment checkpoint has been selected, its parameters are fixed. Candidate decision thresholds are evaluated on the held-out source-validation split over

$$\{0.10, 0.15, \ldots, 0.90\}.$$

For each threshold, patient-level Dice is computed and averaged without patient-volume weighting. The threshold with the highest mean source-validation Dice is selected. If multiple thresholds achieve the same Dice, the lower threshold is selected.

Checkpoint selection and decision-threshold selection are performed sequentially. The threshold sweep does not affect checkpoint ranking, and no external or target-domain information is used in either stage. The selected checkpoint and source-validation-derived decision threshold together constitute the final deployment configuration.

G Representation-Level Analysis under External Shifts

We examine whether the learned RIR representation retains source-defined tumor separation structure under the two external shifts. This post-hoc diagnostic uses the frozen source-selected model from seed 2026. Source statistics and the Source-derived projection direction are estimated from the fixed source-internal test cohort. These quantities are used only for the representation analysis and are not used for training, deployment selection, reported segmentation performance, or inference on the external cohorts.

Voxel support and sampling. For every cohort, valid voxels are restricted to the intersection of the explicit valid-field-of-view mask and the valid mask of the stride-1 RIR pathway. Within this support, tumor voxels satisfy $y(v) > 0.5$ and background voxels satisfy $y(v) \leq 0.5$. For each case, up to 4096 tumor voxels and 4096 background voxels are selected using deterministic uniform-priority reservoir sampling with seed 2026. Absolute FLAIR and RIR are evaluated at exactly the same sampled spatial locations.

Absolute FLAIR score. For a sampled voxel v, the absolute-space score is

$$q_{\mathrm{Abs}}(v) = s_{\mathrm{Abs}} I_{\mathrm{FLAIR}}(v),$$

where I_{FLAIR} denotes the prepared model-input FLAIR value rather than raw scanner-unit intensity. The orientation sign s_{Abs} is fixed once from the Source cohort so that larger scores correspond to the Source tumor direction,

$$s_{\mathrm{Abs}} = \begin{cases} +1, & \bar{I}_T^S \geq \bar{I}_B^S, \\ -1, & \bar{I}_T^S < \bar{I}_B^S. \end{cases}$$

The Source tumor and background means are 3.9839 and -0.0332, respectively, giving $s_{\mathrm{Abs}} = +1$. This sign is not refit on either external cohort.

Pre-grounding RIR score. We use the 16-dimensional stride-1 feature

$$E_{\mathrm{rir}} = \texttt{RIRPyramidOutput.features["stride1"]},$$

measured before affinity and boundary prediction in the forward pathway. Each voxel feature is first normalized as

$$\widetilde{\mathbf{f}}(v) = \frac{\mathbf{f}(v)}{\|\mathbf{f}(v)\|_2 + \epsilon}.$$

Using Source voxels only, we compute the tumor and background feature means

$$\boldsymbol{\mu}_T^S = \frac{1}{|V_T^S|} \sum_{v \in V_T^S} \widetilde{\mathbf{f}}(v),$$

and

$$\boldsymbol{\mu}_B^S = \frac{1}{|V_B^S|} \sum_{v \in V_B^S} \widetilde{\mathbf{f}}(v).$$

The Source-derived tumor direction is

$$\mathbf{d} = \frac{\boldsymbol{\mu}_T^S - \boldsymbol{\mu}_B^S}{\left\|\boldsymbol{\mu}_T^S - \boldsymbol{\mu}_B^S\right\|_2},$$

and the RIR voxel score is

$$q_{\mathrm{RIR}}(v) = \widetilde{\mathbf{f}}(v)^\top \mathbf{d}.$$

The same Source-derived direction is applied unchanged to BraTS-Africa Glioma and BraTS-Africa Other Neoplasms. No external-cohort projection fitting is performed. The term pre-grounding refers to the location at which the feature is measured in the forward pathway. The RIR Encoder itself is learned through source supervision as described in Sec. 3.2 and Appendix B.

Source standardization. Scores in each space are standardized using the pooled Source tumor and background score distribution,

$$z_k(v) = \frac{q_k(v) - \mu_{k,\mathrm{Source}}}{\sigma_{k,\mathrm{Source}}}, \qquad k \in \{\mathrm{Abs}, \mathrm{RIR}\}.$$

The Source mean is

$$\mu_{k,\mathrm{Source}} = \mathrm{mean}\left(Q_{k,T}^S \cup Q_{k,B}^S\right),$$

and the population standard deviation is

$$\sigma_{k,\mathrm{Source}} = \sqrt{\frac{1}{N} \sum_{q \in Q_{k,T}^S \cup Q_{k,B}^S} (q - \mu_{k,\mathrm{Source}})^2}.$$

The resulting constants are

$$\mu_{\mathrm{Abs,Source}} = 1.927128683, \qquad \sigma_{\mathrm{Abs,Source}} = 2.142091573,$$

and

$$\mu_{\mathrm{RIR,Source}} = 0.072576457, \qquad \sigma_{\mathrm{RIR,Source}} = 0.599524562.$$

The two spaces are standardized independently, so their numerical score units should not be compared directly.

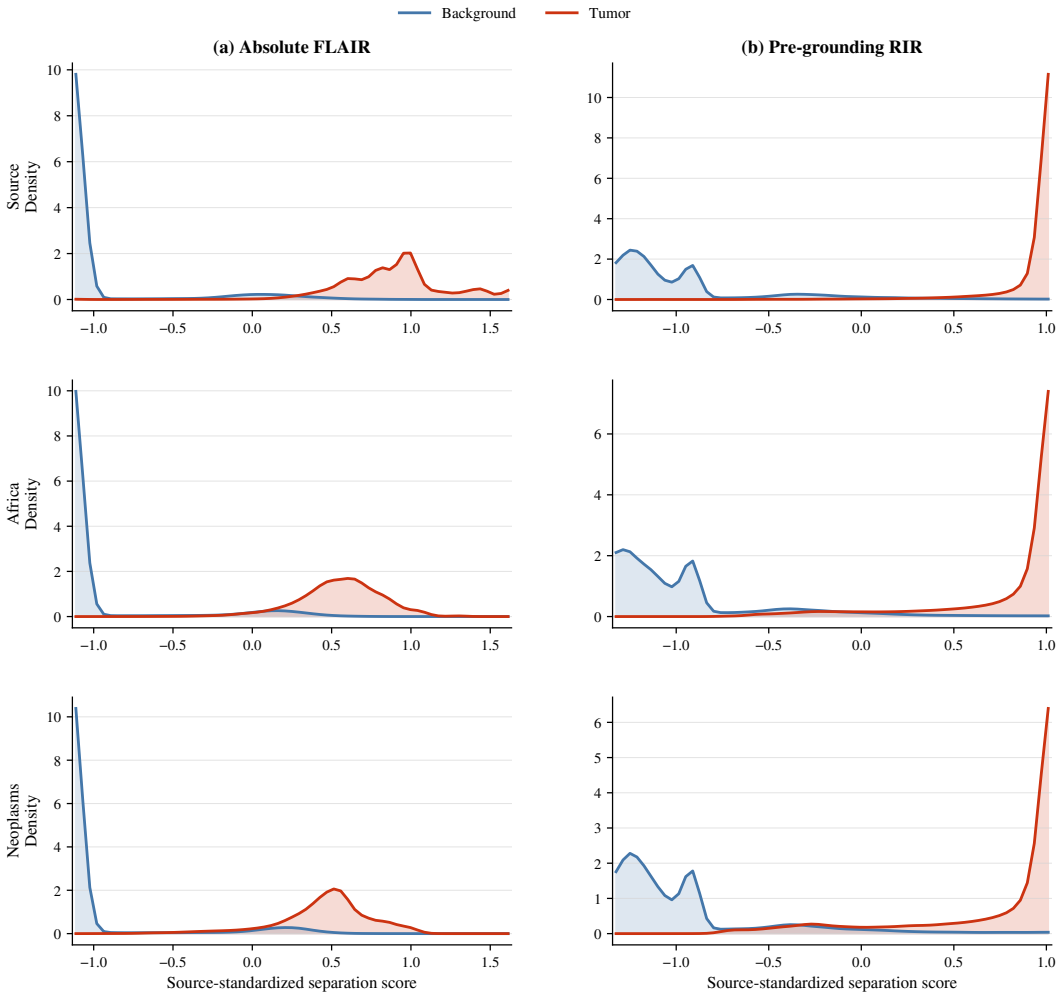


Figure G.1: **Source-standardized tumor and background score distributions across cohorts.** Rows show Source, BraTS-Africa Glioma, and BraTS-Africa Other Neoplasms. Column (a) uses the prepared absolute FLAIR score and column (b) uses the Source-derived projection of the pre-grounding RIR feature. Curves show density estimates over sampled valid tumor and background voxels. Scores are standardized using the pooled Source tumor and background mean and population standard deviation independently within each space. Larger values indicate greater alignment with the Source-defined tumor direction.

Figure G.1 shows that the background-score locations remain similar across cohorts in both spaces, while the external tumor-score distributions shift away from their Source locations. The Source tumor mean is 0.916 in the absolute space and 0.940 in the RIR space. For BraTS-Africa Glioma, the corresponding tumor means are 0.552 and 0.773, while for BraTS-Africa Other Neoplasms they are 0.457 and 0.677. Thus, the Source-to-target displacement of the tumor-score location is smaller in the RIR space for both external cohorts.

This analysis is a voxel-level distribution diagnostic rather than a case-level AUROC or effect-size analysis. It therefore does not establish superior AUROC preservation, pathology invariance, or causal transferability. Instead, it provides descriptive evidence that the Source-defined tumor-score location is retained more closely in the learned RIR representation under the evaluated external shifts.

H Run Statistics and Evaluation Convention

The controlled source-size study compares BASE, RandConv, DSU, and the complete RIR framework over three independent runs with random seeds 2026, 4513, and 6649. Dice and sensitivity are computed per patient and then averaged across patients within each cohort. For each run, the two-cohort external macro Dice is the unweighted mean of the BraTS-Africa Glioma and BraTS-Africa Other Neoplasms cohort means. Sensitivity in Fig. 3 is reported separately for the two external cohorts.

Let m_s denote a reported metric from run s and let $S = 3$. We report the across-run mean

$$\bar{m} = \frac{1}{S}\sum_{s=1}^{S} m_s$$

and the population standard deviation

$$\sigma_{\mathrm{pop}} = \sqrt{\frac{1}{S}\sum_{s=1}^{S}(m_s - \bar{m})^2}.$$

Accordingly, the standard deviation uses divisor S rather than $S - 1$. It summarizes the observed variability across the three runs and is reported descriptively rather than as an inferential estimate of uncertainty.

When Dice differences relative to BASE are reported, they are computed within matched seeds before averaging. For a method M, the per-run difference is

$$\Delta_s^M = m_s^M - m_s^{\mathrm{BASE}},$$

and the reported mean difference is

$$\overline{\Delta}^M = \frac{1}{S}\sum_{s=1}^{S}\Delta_s^M.$$

H.1 Select-Stage Ablation

We isolate the contribution of the proposed *Select* stage from the learned RIR pathway. The ablation reuses the same trained RIR trajectories and changes only the deployment rule, so no additional model training is performed. *RIR w/o Select* chooses the checkpoint with the highest source-validation Dice evaluated at threshold 0.5, resolves ties by the earlier epoch, and keeps the deployment threshold fixed at 0.5. *Full RIR* instead applies the source-only checkpoint and decision-threshold selection defined in Appendix F. Both variants therefore share identical *Represent*, *Ground*, and *Integrate* training procedures.

For compact comparison, the sensitivity reported in Table H.1 is the unweighted mean of the BraTS-Africa Glioma and BraTS-Africa Other Neoplasms cohort sensitivities within each matched run. Dice uses the same two-cohort external macro definition as Table 1.

Table H.1: **Ablation of source-only deployment selection across source supervision sizes.** Values are mean ± population standard deviation over three matched runs. Both Dice and sensitivity are unweighted two-cohort external macro scores within each run.

Method	$N = 10$	$N = 20$	$N = 31$	$N = 51$	$N = 102$
Two-cohort external macro Dice					
BASE	0.470 ± 0.037	0.481 ± 0.044	0.416 ± 0.058	0.547 ± 0.011	0.570 ± 0.007
RIR w/o Select	0.500 ± 0.071	0.493 ± 0.027	0.461 ± 0.058	0.564 ± 0.018	0.555 ± 0.033
Full RIR	0.548 ± 0.006	0.499 ± 0.027	0.517 ± 0.046	0.565 ± 0.013	0.589 ± 0.009
Two-cohort external macro sensitivity					
BASE	0.384 ± 0.041	0.413 ± 0.043	0.337 ± 0.057	0.493 ± 0.033	0.508 ± 0.013
RIR w/o Select	0.418 ± 0.074	0.420 ± 0.026	0.380 ± 0.056	0.512 ± 0.037	0.492 ± 0.032
Full RIR	0.491 ± 0.020	0.460 ± 0.056	0.466 ± 0.042	0.524 ± 0.030	0.581 ± 0.014

Removing Select separates the deployment contribution from the learned RIR pathway while leaving model training unchanged. RIR without Select exceeds BASE in both mean Dice and mean sensitivity at $N = 10$, 20, 31, and 51, although this pattern does not persist at $N = 102$. Thus, the learned relational pathway itself provides useful external segmentation behavior across most evaluated source-supervision sizes, but its benefit is not uniformly positive at every scale.

Adding Select to the same RIR training trajectories produces a positive mean change in both metrics at all five source sizes. Relative to RIR without Select, mean Dice increases by 0.001–0.056, while mean sensitivity increases by 0.012–0.089. The sensitivity gains are therefore accompanied by non-decreasing mean Dice across all evaluated source sizes rather than by an observed mean overlap trade-off. At $N = 102$, where RIR without Select falls below BASE in both reported metrics, Full RIR exceeds BASE after source-only deployment selection, illustrating the complementary role of the deployment stage.

This ablation directly characterizes the incremental contribution of Select because it modifies only the deployment rule after training. In contrast, *Represent*, *Ground*, and *Integrate* form a jointly optimized relational pathway rather than independent additive components. The rank-derived representation receives no standalone supervision, grounding makes these cues task-relevant through source-derived relational objectives, and integration determines how the grounded evidence contributes to the segmentation representation. We therefore interpret these stages as a coupled learned pathway and use the Select ablation to separate the deployment contribution from this pathway.

The repeated runs are used to characterize run-to-run variability. All controlled comparisons and the Select-stage ablation are descriptive, and no statistical-significance claim is made from the three-run estimates.