LEARNING STRUCTURAL DEVIATION FROM NORMAL ANATOMY FOR PANCREATIC TUMOR SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Pancreatic tumor segmentation on CT remains challenging because CT-only predictions may fail to reliably distinguish the pancreatic tumor from normal tissue. We investigate whether complementary evidence can be obtained from how a patient's anatomy deviates from plausible normal states. To this end, we introduce a Deviation-Aware Residual Adaptation (DARA) model, which consists of Starchain Anatomy Graph (SAG) and Bounded Residual Reconciliation (BRR). SAG learns a multi-prototype normal anatomy universe from normal reference data and matches each patient's anatomy to class-compatible normal prototypes. SAG measures structural anatomical deviation between the patient and the normal reference and projects structural anatomical deviation into image space as a deviation field that complements the CT-only prediction. BRR trains a refiner using CT-only and SAG-aware out-of-fold (OOF) predictions and applies its bounded residual correction to the CT-only tumor logit. Effectiveness of the proposed DARA model has been validated by using three widely-used datasets (PanTSMini, MSD07, and PANORAMA) with five metrics (DSC, IoU, NSD, HD95, Recall). Experimental results show that the proposed model outperforms all competing methods in DSC, IoU, and NSD. In particular, 3.96%, 3.86%, 0.66% DSC gains can be achieved compared with the second-best method in each dataset respectively. These results suggest that learning what normal anatomy should look like provides a complementary cue to recognizing what is abnormal.

1 INTRODUCTION

Pancreatic tumor segmentation on CT remains challenging because pancreatic tumors often have a weak and heterogeneous appearance. Lesions can be small, low-contrast, and poorly defined, while their locations and relationships with surrounding anatomy vary substantially across patients. (Moglia et al., 2025; Li et al., 2024; Dong et al., 2025). Consequently, lesion appearance may provide insufficient evidence for segmentation in difficult cases. This motivates a complementary question: instead of asking only what a tumor looks like, can we also characterize abnormality by how a patient's anatomy deviates from plausible normal anatomy?

Both anatomy-aware segmentation and normal-reference modeling provide useful signals for answering this question. Anatomical priors can improve localization accuracy and structural reasoning capability (Dong et al., 2025; Jeon et al., 2025), while healthy-counterfactual and normality-modeling approaches detect abnormalities through comparison with learned normal states (Liu et al., 2025; Hinge et al., 2025). We connect these two perspectives by considering normality directly in a structural anatomical space. Specifically, we compare the observed anatomy of a patient with compatible normal anatomical configurations and transform the resulting structural deviation into spatial evidence for tumor segmentation.

Specifically, we propose a DARA model, which consists of SAG and BRR. SAG represents patient anatomy with a hierarchical star–chain graph, learns a normal anatomy universe, and derives structural deviations from compatible normal states. Structural patient-to-normal deviation is then projected into the CT volume as a deviation field. Since anatomical deviation is informative but not necessarily tumor-specific, BRR reconciles the CT-only and SAG-aware predictions through a bounded residual correction.

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

We evaluate the proposed DARA model on three widely-used datasets (PanTSMini, MSD07, and PANORAMA), which cover different data distributions and anatomical contexts. Experimental results show that DARA consistently outperforms the compared methods in tumor segmentation. Across all three datasets, it achieves the best DSC, IoU, and NSD, demonstrating the effectiveness of structural patient-to-normal deviation as complementary evidence beyond CT-only predictions. Our contributions are summarized as follows.

- We formulate pancreatic tumor segmentation from the perspective of structural patient-to-normal deviation, providing a complementary view beyond conventional CT-only and anatomy-guided predictions.
- We introduce SAG, which represents patient anatomy with a hierarchical star–chain graph, learns a multi-prototype normal anatomy universe, and converts structural patient-to-normal deviation into spatial evidence for tumor segmentation.
- We introduce BRR, which reconciles CT-only and SAG-aware predictions through a bounded residual correction to obtain the final tumor segmentation.

2 Related Work

2.1 Anatomy-Aware Pancreatic Tumor Segmentation

Medical image segmentation has benefited from powerful visual representations. Convolutional encoder–decoder frameworks, such as nnU-Net, remain strong baselines for volumetric segmentation, while Transformer, large-kernel convolution, and state-space architectures further improve long-range dependency modeling (Isensee et al., 2021; Hatamizadeh et al., 2022; Roy et al., 2023; Xing et al., 2024). Recent pretrained and self-supervised 3D models, together with generative tumor synthesis, aim to further improve robustness and generalization across datasets and medical-imaging tasks (Gao et al., 2026; Wu et al., 2026; Chen et al., 2024; Zhang et al., 2025). In pancreatic tumor segmentation, anatomical priors and positional context can additionally improve tumor localization (Dong et al., 2025).

Anatomical knowledge in medical image segmentation has evolved from fixed spatial constraints toward explicit shape, topology, adaptive atlases, and structured relational modeling. Shape- and topology-aware priors can improve the anatomical plausibility of segmentation outputs (Wyburd et al., 2024; You et al., 2025). MultiMorph and AtlasMorph construct population-specific or conditional deformable templates to improve the adaptability of anatomical references (Abulnaga et al., 2025; Rakic et al., 2026). Graph-based methods provide another mechanism for representing spatial and semantic relationships among anatomical entities (Li et al., 2026). These studies reveal that existing anatomical priors mainly constrain segmentation rather than explicitly characterizing patient-to-normal anatomical differences. Therefore, we model such differences as complementary evidence for tumor segmentation.

2.2 Normal-Reference Modeling

Normal-reference modeling characterizes abnormalities by comparing patient observations with expected normal states, and has been used to healthy counterfactuals, individualized references, and feature-space normality modeling. Diffusion-based counterfactual methods and UNA generate disease-free counterparts or recover normal anatomy from pathological images for anomaly localization (Fontanella et al., 2025; Liu et al., 2025), while Normal Twin PET and MatchGen further emphasize normal references that are matched to the individual patient (Hinge et al., 2025; Ma et al., 2025). Normality can also be represented through anatomy-specific feature distributions or multiple representative normal states (Cao et al., 2025; Schulthess & Konukoglu, 2025). Across these approaches, abnormality is defined relative to a normal reference that can be patient-matched, anatomy-specific, or multi-state. These studies show the value of normal references for characterizing abnormality, but their use in anatomical comparison for pancreatic tumor segmentation remains limited. Thus, we address this by using deviations from normal anatomy as complementary evidence.

3 METHOD

3.1 OVERVIEW

As illustrated in Fig. 1, the proposed DARA model consists of two main components: SAG and BRR. SAG includes anatomical grounding, normal anatomy universe, and structural deviation. SAG uses the input CT volume to produce the deviation field D_{SAG}. We then obtain two predictions: the CT-only prediction is produced from the CT volume alone; the SAG-aware prediction is produced from the CT volume together with D_{SAG}. BRR uses the CT-only and SAG-aware predictions to estimate a bounded residual correction, which is then applied to the CT-only tumor logit to produce the final tumor prediction.

3.2 ANATOMICAL GROUNDING (AG)

AG takes the CT volume I as input and builds a patient-specific anatomical representation for patient-to-normal comparison. Starting from I, an anatomy predictor segments the organs and then produces organ masks. We construct two types of anatomical nodes from these masks. First, each predicted organ, including the pancreas, is represented as one global organ node. Second, for the pancreas, we further sort its voxels along the longest physical axis and divide them into twelve ordered regions with similar voxel counts, each represented as a pancreatic chain node. Therefore, the organ node and the twelve pancreatic chain nodes coexist as distinct nodes. We use i as a common index for both types of nodes. The anatomical class label c_i specifies whether node i corresponds to a particular organ node or to a particular pancreatic region, and M_i^P denotes its voxel mask in the CT volume. Together, the organ nodes and pancreatic chain nodes define the star–chain anatomical structure. Consecutive pancreatic chain nodes are assigned fixed anatomical relations, while additional relations connect the pancreas and its chain nodes to surrounding anatomical structures. Missing anatomical structures are marked as invalid and excluded from subsequent patient-to-normal comparison.

For each valid node i, we extract an 11-D node state n_i^P and a 9-D shape descriptor s_i^P. We then build the patient relation matrix A^P. For two valid nodes i and j, let μ_i and μ_j be their normalized center positions. Their center distance and fixed anatomical relation are used to compute A^P:

$$A_{ij}^P = \mathrm{Softmax}_{j\neq i}(\exp\left(-\frac{\|\mu_i-\mu_j\|_2^2}{\sigma^2}\right) + R_{ij}^{\mathrm{anat}}), \qquad A_{ii}^P = 0. \tag{1}$$

Here, σ controls the distance term and is set to 0.25, while R_{ij}^{anat} encodes the fixed anatomical relations described above. The softmax function is applied over all valid non-self nodes.

Finally, AG summarizes the patient anatomy and outputs the node states $N^P = [n_1^P, \ldots, n_N^P]^\top$, shape descriptors $S^P = [s_1^P, \ldots, s_N^P]^\top$, anatomical relations $A^P = [A_{ij}^P]_{N\times N}$, and voxel masks $M^P = \{M_i^P\}_{i=1}^N$. Details of the dataset-specific anatomy predictors and node construction are provided in Appendix A.2.

3.3 NORMAL ANATOMY UNIVERSE (NAU) AND STRUCTURAL DEVIATION

NAU: We learn the NAU from a normal reference set $\mathcal{D}_{\mathrm{ref}}$. We first apply AG to each CT in $\mathcal{D}_{\mathrm{ref}}$ and encode its node states using the Node Attribute Encoder E_G. NAU learns node prototypes U, a relation matrix A^U, and shape prototypes S^U:

$$(U, A^U, S^U) = \mathrm{NAU}(\mathcal{D}_{\mathrm{ref}}). \tag{2}$$

Here, $U = \{u_k\}_{k=1}^M$; u_k is the k-th node prototype and M is the total number of prototypes. We use four prototypes for each anatomical class. Notably, the NAU is trained independently with normal reference set and has been frozen in AG.The complete NAU training objective and implementation details are provided in Appendix A.3.

For the current patient I, we encode N^P with the frozen E_G to obtain patient embeddings H^P. Then, each patient node is matched only with normal prototypes from the same anatomical region. Let ℓ_k denote the anatomical class of prototype u_k. For node i, the matching scores over its candidate

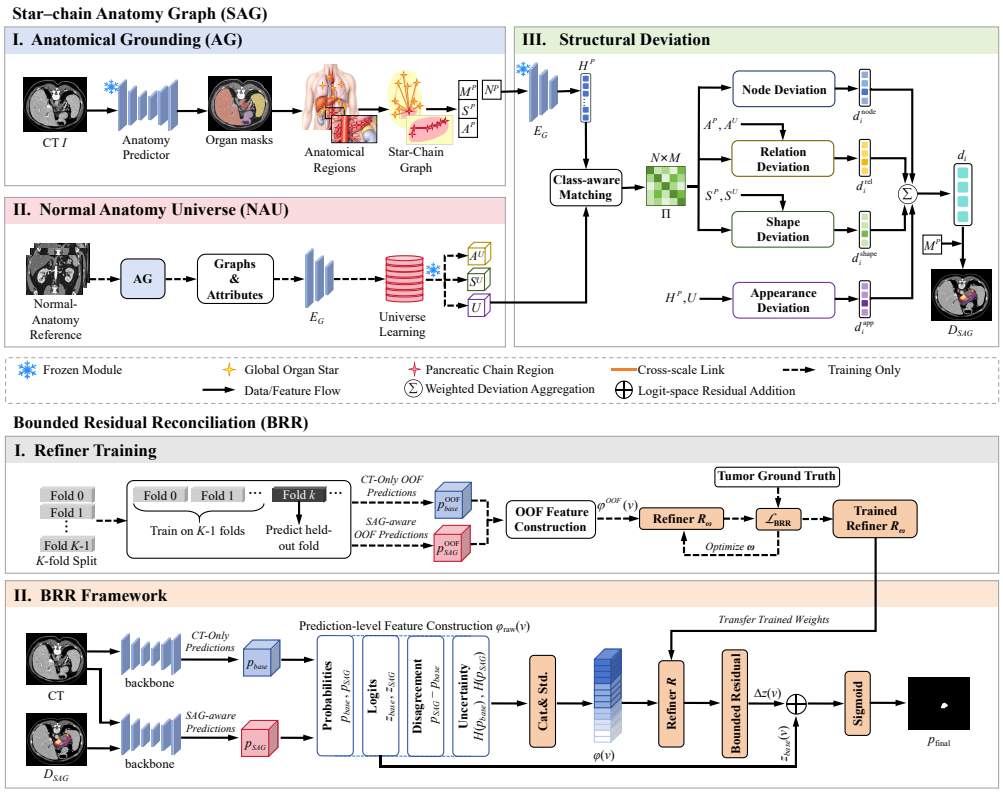

Figure 1: Overview of SAG and BRR.

prototypes are defined as:

$$\gamma_{ik} = \frac{\cos(h_i^P, u_k)}{\tau}, \tag{3}$$

where k satisfies $\{k \mid \ell_k = c_i\}$; $H^P = E_G(N^P) = [h_1^P, \ldots, h_N^P]^\top$ and τ is the matching temperature.

We normalize the matching scores over the candidate prototypes to obtain the assignment matrix Π, where Π_{ik} is the matching weight between patient node i and prototype u_k:

$$\Pi_{ik} = \begin{cases} \dfrac{\exp(\gamma_{ik})}{\sum_{k' \in \mathcal{K}_i} \exp(\gamma_{ik'})}, & k \in \mathcal{K}_i, \\ 0, & k \notin \mathcal{K}_i. \end{cases} \tag{4}$$

Using Π, we obtain the patient-specific normal relations $\hat{A}^P$ and shapes $\hat{S}^P$:

$$\hat{A}^P = \Pi A^U \Pi^\top, \tag{5}$$

$$\hat{S}^P = \Pi S^U = [\hat{s}_1^P, \ldots, \hat{s}_N^P]^\top. \tag{6}$$

Structural deviation. As shown in Fig. 2, we compute four deviations for each patient node. **Node deviation** measures how node i matches one of its candidate normal prototypes using the largest assignment weight and normalized assignment entropy:

$$\bar{H}(\Pi_i) = \begin{cases} H(\Pi_i)/\log K_i, & K_i > 1, \\ 0, & K_i = 1, \end{cases} \tag{7}$$

$$d_i^{\text{node}} = \frac{1}{2}(1 - \Pi_i^{\max}) + \frac{1}{2}\bar{H}(\Pi_i). \tag{8}$$

Here, Π_i denotes the i-th row of Π, and $H(\Pi_i)$ is the entropy of its assignment weights. $K_i = |\mathcal{K}_i|$ and $\Pi_i^{\max} = \max_{k \in \mathcal{K}_i} \Pi_{ik}$.

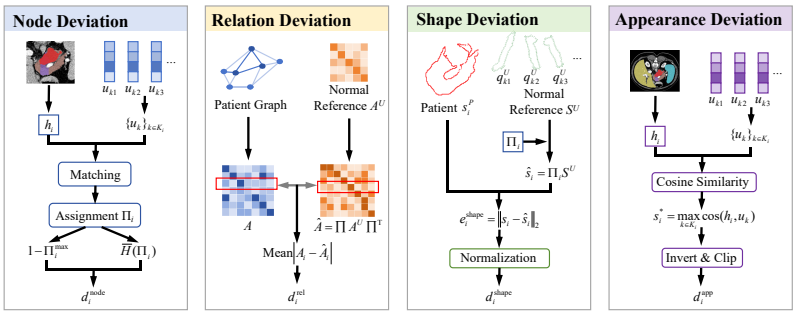

Figure 2: Structural deviation components. Node, relation, shape, and appearance deviations characterize complementary differences between the patient-specific anatomy and its corresponding normal anatomical reference.

Relation deviation measures the difference between the observed relations A^P and the patient-specific normal relations $\hat{A}^P$:

$$d_i^{\text{rel}} = \frac{1}{N-1} \sum_{j \neq i} \left| A_{ij}^P - \hat{A}_{ij}^P \right|. \tag{9}$$

Shape deviation measures the difference between the observed shape s_i^P and its normal reference $\hat{s}_i^P$:

$$e_i^{\text{shape}} = \|s_i^P - \hat{s}_i^P\|_2, \tag{10}$$

$$d_i^{\text{shape}} = \frac{e_i^{\text{shape}}}{\max_{j \in \mathcal{V}_{\text{valid}}} e_j^{\text{shape}} + \epsilon}. \tag{11}$$

Here, $\mathcal{V}_{\text{valid}}$ is the set of valid nodes and $\epsilon > 0$ is used for numerical stability.

Appearance deviation measures the difference between node i and its closest normal prototype in the embedding space:

$$d_i^{\text{app}} = \text{clip}\left(\frac{1 - \max_{k \in \mathcal{K}_i} \cos(h_i^P, u_k)}{2}, 0, 1 \right). \tag{12}$$

We combine the four deviations using weights w_n, w_r, w_s, and w_a into a single score d_i for each node:

$$d_i = \frac{w_n d_i^{\text{node}} + w_r d_i^{\text{rel}} + w_s d_i^{\text{shape}} + w_a d_i^{\text{app}}}{w_n + w_r + w_s + w_a}. \tag{13}$$

Finally, we project the node scores to the CT volume using the voxel masks M_i^P:

$$D_{\text{SAG}}(v) = \frac{\sum_i m_i^{\text{proj}} M_i^P(v) d_i}{\sum_i m_i^{\text{proj}} M_i^P(v) + \epsilon}. \tag{14}$$

Here, m_i^{proj} is set to 1 for valid pancreatic chain nodes and 0 otherwise. Only pancreatic chain nodes are projected to voxel space, while their deviation scores are computed from the complete star–chain anatomical structure. Therefore, D_{SAG} incorporates pancreatic anatomical information together with global anatomical context. Dataset-specific deviation-field export details are provided in Appendix A.4.

3.4 Bounded Residual Reconciliation (BRR)

As illustrated in Fig. 1, BRR is explained in two parts: Refiner Training and the BRR Framework.

Refiner Training. We generate K-fold OOF predictions on the training set. In each round, one fold is held out, while the CT-only and SAG-aware branches are trained on the remaining $K - 1$

folds. The trained branches then predict the held-out fold, producing CT-only and SAG-aware OOF predictions for each case. We use these predictions to construct OOF features $\varphi^{\mathrm{OOF}}(v)$ and train the refiner module R_ω with the corresponding tumor ground truth:

$$\mathcal{L}_{\mathrm{BRR}} = \mathcal{L}_{\mathrm{BCE}} + \mathcal{L}_{\mathrm{Dice}} + \lambda_\Delta \mathcal{L}_\Delta + \lambda_{\mathrm{anchor}} \mathcal{L}_{\mathrm{anchor}}. \tag{15}$$

Here, $\mathcal{L}_{\mathrm{BCE}}$ and $\mathcal{L}_{\mathrm{Dice}}$ are the weighted binary cross-entropy and soft Dice losses for tumor segmentation, while $\mathcal{L}_\Delta$ and $\mathcal{L}_{\mathrm{anchor}}$ penalize large logit corrections and excessive deviations from the CT-only prediction, respectively. Detailed loss definitions and training settings are provided in Appendix C.2. After training, R_ω is used in the BRR Framework, while the OOF branch models are not used at test time. The complete OOF training protocol is provided in Appendix C.3.

BRR Framework. Given the input CT volume I and its corresponding deviation field D_{SAG}, the CT-only branch uses only I to produce p_{base}. The SAG-aware branch uses both I and D_{SAG} to produce p_{SAG}. The two branches have the same backbone architecture.

These two predictions are then used to construct voxel-level features for BRR. For each voxel v, let p^t_{base} and p^t_{SAG} denote the tumor probabilities, and p^p_{base} and p^p_{SAG} the pancreas probabilities. We define the tumor logits as $z^t = \mathrm{logit}(p^t)$, and the uncertainty as $H_{\mathrm{bin}}(p) = -p\log p - (1-p)\log(1-p)$. We construct a 10-D raw feature vector $\varphi_{\mathrm{raw}}(v)$ containing probabilities, logits, disagreement, maximum response, and uncertainty. The features are standardized using statistics from the OOF training set to obtain $\varphi(v)$. The detailed feature definition and standardization are provided in Appendix C.1.

The trained refiner R_ω maps $\varphi(v)$ to a tumor-logit correction, which is bounded by $\Delta_{\max}$ through the $\tanh$ function. The bounded correction $\Delta z(v)$ is added to the CT-only tumor logit to obtain the final prediction:

$$\Delta z(v) = \Delta_{\max} \tanh(R_\omega(\varphi(v))), \tag{16}$$

$$p_{\mathrm{final}}(v) = \mathrm{sigmoid}\big(z^t_{\mathrm{base}}(v) + \Delta z(v)\big). \tag{17}$$

Here, $\Delta_{\max} = 3.0$ limits the correction magnitude. Thus, BRR limits the change from the CT-only tumor logit while allowing SAG-aware evidence to influence the correction when needed.

4 Experiments

4.1 Datasets

We evaluate the proposed DARA model on PanTSMini, MSD07, and PANORAMA by using fixed subject-level train/validation/test splits, with all test subjects held out until final evaluation. PanTSMini is constructed from PanTS (Li et al., 2025) with 7,901/500/500 train/validation/test cases and 1,000 reference cases; MSD07 (Antonelli et al., 2022) uses 161/40/40 cases and 40 reference cases; PANORAMA (Alves et al., 2026) uses 1,164/200/200 cases and 400 reference cases. PanTSMini and PANORAMA use separate tumor-negative reference cohorts. Since MSD07 has no independent healthy cohort, its reference is built from normal-appearing anatomy after excluding the annotated tumor and a 5-mm margin around it. Reference construction and leakage safeguards are described in Appendix B.1, and complete split statistics are provided in Appendix B.2.

4.2 Evaluation Protocol

The main comparison focuses on tumor-level performance. We report Dice similarity coefficient (DSC), intersection over union (IoU), normalized surface Dice (NSD), 95th-percentile Hausdorff distance (HD95), and Recall in the main tables. DSC, IoU, NSD, and Recall are reported in percentage (%), while HD95 is measured in millimeters (mm). Higher values are better for DSC, IoU, NSD, and Recall, whereas lower values are preferred for HD95. Tumor metrics are computed for each tumor-positive test subject. For each random seed, subject-level results are averaged over the test subjects, and we report the mean $\pm$ standard deviation across three seeds. All methods use the same predefined test split. Detailed implementation and training settings are provided in Appendix A, metric definitions in Appendix A.1, and pancreas segmentation results in Appendix D.1.

Table 1: Quantitative comparison of tumor segmentation. Results are reported as mean ± standard deviation across three random seeds, with each seed evaluated over tumor-positive test subjects. The **best** and second-best performance are shown in bold and underlined.

Method	Source	DSC(%)↑	IoU(%)↑	NSD(%)↑	HD95(mm)↓	Recall(%)↑
		PanTSMini				
nnU-Net (Isensee et al., 2021)	Nature'21	30.60± 2.84	22.09± 2.15	27.08± 3.02	33.97± 4.51	32.30± 3.12
SwinUNETR (Hatamizadeh et al., 2022)	MICCAI'22	20.59± 1.53	13.50± 1.24	17.49± 1.68	50.87± 3.87	21.99± 1.95
MedNeXt (Roy et al., 2023)	MICCAI'23	25.24± 3.47	17.64± 2.98	20.41± 3.62	143.91± 12.45	46.08± 4.12
SegMamba (Xing et al., 2024)	MICCAI'24	14.32± 1.89	9.07± 1.42	10.59± 1.95	152.14± 9.36	22.75± 2.14
UNETR++ (Shaker et al., 2024)	TMI'24	17.62± 2.95	12.10± 2.31	13.02± 3.14	72.65± 7.24	25.26± 3.57
DiffUNet (Xing et al., 2025)	MedIA'25	21.04± 1.68	13.81± 1.35	18.34± 1.82	84.16± 5.68	26.58± 2.05
MASS (Gao et al., 2026)	CVPR'26	31.18± 3.12	22.30± 2.74	28.52± 3.45	29.03± 3.87	33.01± 3.64
DINO-Med3D (Hu et al., 2026)	MICCAI'26	21.57± 1.45	15.48± 1.12	18.02± 1.58	30.08± 2.14	21.87± 1.76
SegMamba-V2 (Xing et al., 2026)	TMI'26	19.63± 2.76	13.09± 2.18	15.41± 2.95	122.00± 8.45	27.63± 3.21
VoCo (Wu et al., 2026)	TPAMI'26	34.25± 2.31	25.16± 1.98	30.80± 2.54	**25.05± 2.87**	35.12± 2.65
DARA	Ours	**38.21± 3.05**	**28.11± 2.76**	**34.02± 3.24**	31.70± 3.98	**48.01± 3.42**
		MSD07				
nnU-Net (Isensee et al., 2021)	Nature'21	51.77± 2.95	40.89± 2.54	45.64± 3.12	**9.15± 2.34**	49.44± 2.87
SwinUNETR (Hatamizadeh et al., 2022)	MICCAI'22	40.20± 1.42	29.11± 1.18	33.47± 1.65	100.83± 6.54	43.72± 1.98
MedNeXt (Roy et al., 2023)	MICCAI'23	55.79± 2.18	44.29± 1.95	50.89± 2.45	11.43± 2.15	54.38± 2.36
SegMamba (Xing et al., 2024)	MICCAI'24	37.15± 3.24	26.91± 2.87	30.06± 3.54	51.96± 4.87	43.79± 3.45
UNETR++ (Shaker et al., 2024)	TMI'24	35.13± 1.65	25.76± 1.34	27.29± 1.89	26.99± 2.54	39.96± 1.82
DiffUNet (Xing et al., 2025)	MedIA'25	43.01± 2.76	31.73± 2.42	34.96± 2.98	33.43± 3.65	46.53± 2.95
MASS (Gao et al., 2026)	CVPR'26	48.99± 1.85	35.65± 1.52	40.60± 2.14	43.28± 3.12	56.74± 2.18
DINO-Med3D (Hu et al., 2026)	MICCAI'26	39.96± 3.12	28.72± 2.85	30.98± 3.45	32.00± 3.54	44.30± 3.24
SegMamba-V2 (Xing et al., 2026)	TMI'26	40.25± 1.54	29.41± 1.28	32.74± 1.75	58.97± 4.12	46.84± 1.95
VoCo (Wu et al., 2026)	TPAMI'26	55.15± 2.42	42.40± 2.15	49.14± 2.68	14.99± 2.45	53.93± 2.54
DARA	Ours	**59.65± 2.87**	**47.47± 2.54**	**51.80± 3.12**	9.34± 2.18	**65.44± 2.76**
		PANORAMA				
nnU-Net (Isensee et al., 2021)	Nature'21	62.87± 2.45	52.22± 2.18	61.67± 2.87	**10.79± 2.34**	63.68± 2.56
SwinUNETR (Hatamizadeh et al., 2022)	MICCAI'22	33.98± 1.24	25.03± 1.08	25.61± 1.45	23.41± 1.87	27.81± 1.35
MedNeXt (Roy et al., 2023)	MICCAI'23	63.05± 1.85	52.72± 1.62	60.52± 2.14	11.81± 1.95	66.16± 1.78
SegMamba (Xing et al., 2024)	MICCAI'24	50.55± 3.12	40.64± 2.84	47.06± 3.45	14.78± 2.87	48.82± 3.24
UNETR++ (Shaker et al., 2024)	TMI'24	57.34± 1.54	46.08± 1.32	48.47± 1.78	12.39± 1.65	**70.42± 1.62**
DiffUNet (Xing et al., 2025)	MedIA'25	46.59± 2.87	37.79± 2.54	42.79± 3.12	13.11± 2.45	44.38± 2.95
MASS (Gao et al., 2026)	CVPR'26	49.58± 1.42	38.27± 1.25	44.51± 1.68	25.89± 2.14	52.52± 1.56
DINO-Med3D (Hu et al., 2026)	MICCAI'26	53.32± 2.65	42.09± 2.31	48.20± 2.87	39.13± 3.45	56.23± 2.74
SegMamba-V2 (Xing et al., 2026)	TMI'26	43.68± 1.35	35.80± 1.18	41.46± 1.54	13.31± 1.42	41.79± 1.28
VoCo (Wu et al., 2026)	TPAMI'26	59.58± 2.95	48.69± 2.68	58.48± 3.14	10.89± 2.56	58.34± 2.87
DARA	Ours	**63.71± 1.42**	**52.94± 1.25**	**62.54± 1.68**	12.20± 1.54	67.92± 1.35

4.3 Comparison with Other Popular Models

As shown in Table 1, DARA achieves the highest DSC, IoU, and NSD on all three datasets. Compared with the second-best model, DARA improves DSC by 3.96%, 3.86%, and 0.66% on PanTSMini, MSD07, and PANORAMA, respectively. DARA also achieves the highest Recall on PanTSMini and MSD07 and the second-highest Recall on PANORAMA. Additional pancreas segmentation, patient-level tumor recovery, tumor-size and case-level Dice are provided in Appendix D.

4.4 Ablation Studies

All ablation studies are conducted on PanTSMini using the 36-node SAG setting and the same nnU-Net backbone and data split. In the main-component ablation, settings without BRR report the corresponding segmentation prediction before reconciliation.

Main components. Table 2 evaluates the contributions of AG, structural deviation, and BRR. Using AG alone gives a DSC/NSD/Recall of 30.07/25.66/34.58, while AG with structural deviation obtains gains of 2.17/2.5/1.28 DSC/NSD/Recall. When BRR further includes, 5.97/5.86/12.15

Table 2: Ablation of the main components.

AG	Dev.	BRR	DSC(%) ↑	NSD(%) ↑	Recall(%) ↑
✓			30.07	25.66	34.58
✓	✓		32.24	28.16	35.86
✓		✓	33.60	30.16	39.75
✓	✓	✓	**38.21**	**34.02**	**48.01**

Table 3: Ablation of structural deviation.

N	R	S	A	DSC(%) ↑	NSD(%) ↑	Recall(%) ↑
✓				37.40	32.90	47.81
	✓			36.73	32.11	46.16
		✓		36.79	32.60	47.20
			✓	36.31	31.91	46.30
✓	✓	✓		37.64	33.11	47.52
✓	✓	✓	✓	**38.21**	**34.02**	**48.01**

Table 4: Ablation of star-chain anatomical structure.

Configuration	DSC(%) ↑	NSD(%) ↑	Recall(%) ↑
Pancreas anatomical prior	36.04	31.52	46.73
Global organ nodes	36.37	31.88	45.87
Pancreatic chain nodes	37.71	33.19	**50.41**
Star-chain	**38.21**	**34.02**	48.01

Table 5: Layer-wise contribution of SAG deviation.

Layer	Real-D R^2	Zero-D R^2	ΔR^2
Early encoder	0.871	0.335	+0.536
Deep encoder	0.565	0.528	+0.037
Bottleneck	0.143	0.130	+0.013
Final decoder	0.741	0.603	+0.139

Table 6: Comparison of prediction reconciliation strategies.

Reconciliation strategy	DSC(%) ↑	NSD(%) ↑	Recall(%) ↑
SAG-aware only	32.24	28.16	35.86
Simple average	32.95	28.46	36.34
Weighted average	30.81	26.88	33.40
Maximum	36.54	32.11	43.44
Learned fusion w/o OOF	33.83	28.77	46.27
OOF MLP	34.89	30.80	47.63
BRR w/o structural deviation	33.60	30.16	39.75
BRR (Ours)	**38.21**	**34.02**	**48.01**

DSC/NSD/Recall gains can be achieved. The above results show that each component contributes to the model improvement.

Deviation components. Table 3 evaluates the node, relation, shape, and appearance deviations in SAG. Combining node, relation, and shape deviations improves DSC/NSD/Recall to 37.64/33.11/47.52. Further adding appearance deviation increases DSC/NSD/Recall to 38.21/34.02/48.01, which means 0.57/0.91/0.49 gains are obtained. These results show that jointly using the four deviation components is more effective than using any individual deviation alone.

Star-chain anatomical structure. Table 4 evaluates different anatomical structures in SAG. The global organ nodes and pancreatic chain nodes individually improve the results over the pancreas anatomical prior. Combining them achieves the best DSC and NSD of 38.21 and 34.02, respectively. Notably, when using pancreatic chain nodes alone, DSC and NSD is not the optimal although Recall is the best. Thus, the complete star-chain anatomical structure is more effective than using either the global organ nodes or pancreatic chain nodes alone.

Layer-wise contribution of SAG deviation. Table 5 compares network representations obtained with and without D_{SAG}, denoted as Real-D and Zero-D, respectively. R^2 denotes the coefficient of determination used in the layer-wise analysis, where a higher value indicates that more information can be captured from the corresponding network, and ΔR^2 measures the difference between them. With D_{SAG}, R^2 reaches 0.871 at the early encoder and 0.741 at the final decoder, compared with 0.335 and 0.603 without D_{SAG}. The largest difference appears at the early encoder ($\Delta R^2 = 0.536$), while a clear difference remains at the final decoder ($\Delta R^2 = 0.139$). These results show that the SAG-derived deviation field provides substantial information to the segmentation network.

Prediction reconciliation. Table 6 compares different strategies for reconciling the CT-only and SAG-aware prediction. The compared strategies include fixed fusion rules, learned fusion without OOF, an OOF MLP, and BRR without structural deviation. BRR achieves the best result, which are 38.21 DSC, 34.02 NSD and 48.01 Recall. When structural deviation is removed, DSC/NSD/Recall decrease to 33.60/30.16/39.75, further showing the importance of structural deviation for BRR.

Visualizations of the ablation studies in Tables 3–6 are provided in Appendix E.

Regional structural deviation analysis. Figure 3 examines whether D_{SAG} is larger in tumor regions than in non-tumor pancreatic regions. In Fig. 3(a), we compare the mean D_{SAG} in three

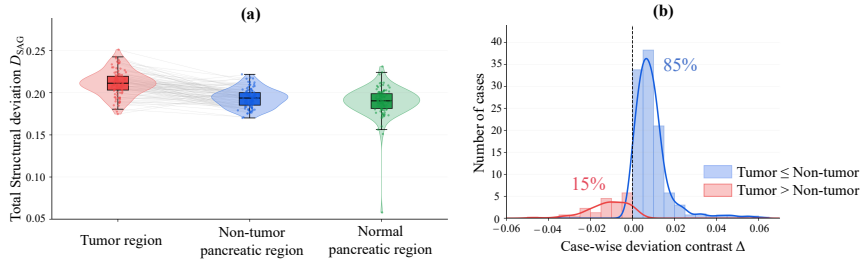

Figure 3: Regional behavior of structural deviation.

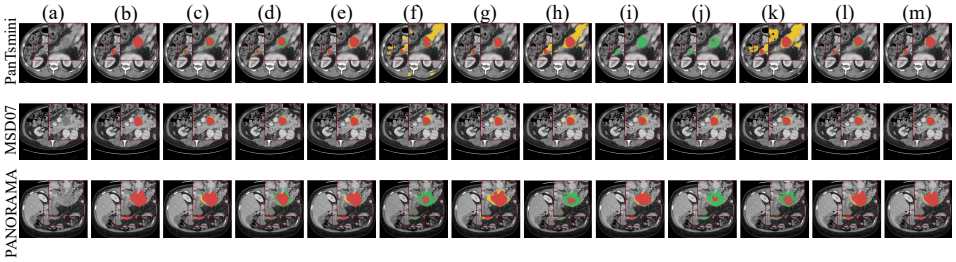

Figure 4: Visualization of results. (a) CT image, (b) ground truth, (c) nnU-Net, (d) SwinUNETR, (e) MedNeXt, (f) SegMamba, (g) UNETR++, (h) DiffUNet, (i) MASS, (j) DINO-Med3D, (k) SegMamba-V2, (l) VoCo, and (m) DARA.

regions: the tumor region of tumor-positive patients, the non-tumor pancreatic region of the same patients, and the pancreatic region of normal subjects. Tumor regions show higher deviation values than the two non-tumor groups. In Fig. 3(b), we further compare the tumor region with the non-tumor pancreatic region within each tumor-positive case. A positive difference means that the tumor region has a higher D_{SAG}, which is observed in 85% of the cases. These results show that the SAG-derived deviation field tends to assign larger deviation values to tumor regions.

4.5 Visualization of Results

Fig. 4 shows representative cases from all three datasets. Compared with other methods, DARA produces predictions that better match the tumor regions in the displayed cases. In the segmentation overlays, red denotes prediction–ground-truth overlap, green denotes ground-truth tumor missed by the prediction, and yellow denotes predicted regions outside the ground truth. Additional qualitative analysis of structural deviation and BRR is provided in Appendix E.1.

5 Conclusion

We introduce DARA for CT pancreatic tumor segmentation by modeling differences between patient anatomy and normal anatomy. DARA consists of two components: SAG and BRR. SAG first constructs a patient-specific anatomical representation from the CT volume. It then compares the patient anatomy with the normal anatomy universe, which represents plausible normal anatomical patterns. SAG projects the resulting patient-to-normal anatomical differences into image space as a deviation field. The CT-only prediction is produced from the CT volume alone, and the SAG-aware prediction is produced from the CT volume together with the deviation field. BRR uses these two predictions to estimate a bounded correction, applies it to the CT-only prediction, and obtains the final tumor prediction. Effectiveness of the proposed model has been validated in three datasets: PanTSMini, MSD07, and PANORAMA. Experimental results show that DARA achieves the highest mean DSC, IoU, and NSD on all three datasets. It also achieves the highest Recall on PanTSMini and MSD07 and the second-highest Recall on PANORAMA. Future work will further explore extensions to cross-disease, cross-organ, and cross-modality settings.

AI USE STATEMENT

Large Language Models (LLMs) were employed exclusively for language editing during manuscript preparation. Their use was strictly limited to improving the clarity and readability of the manuscript, including rephrasing sentences, correcting grammar, and refining textual flow. **All research-related work was performed independently and exclusively by the authors.** In particular, the formulation of research questions, development of scientific ideas, design of the methodology, implementation of the proposed methods, design and execution of experiments, data processing and analysis, interpretation of results, and formulation of scientific conclusions were all carried out solely by the authors, without involvement of LLMs. LLMs did not contribute to any scientific, methodological, experimental, computational, or analytical aspect of the study. All LLM-assisted language edits were subsequently reviewed and verified by the authors, who take full responsibility for the accuracy, originality, and integrity of the manuscript.

REFERENCES

S. Mazdak Abulnaga, Andrew Hoopes, Neel Dey, Malte Hoffmann, Bruce Fischl, John Guttag, and Adrian V. Dalca. Multimorph: On-demand atlas construction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30906–30917, 2025.

Natalia Alves, Megan Schuurmans, Dawid Rutkowski, Anindo Saha, Pierpaolo Vendittelli, Nancy Obuchowski, Marjolein H. Liedenbaum, Ingfrid S. Haldorsen, Anders Molven, Derya Yakar, Jeroen Geerdink, Sebastiaan van Koeverden, Deniece M. Riviere, Wulphert Venderink, Robbert de Haas, Namkug Kim, J.-Matthias Löhr, Garima Suman, Klaus H. Maier-Hein, Horst K. Hahn, Weichung Wang, Alan L. Yuille, Avinash Kambadakone, Elliot K. Fishman, Caroline Verbeke, Geert Litjens, John J. Hermans, Henkjan Huisman, and PANORAMA consortium. Artificial intelligence and radiologists in pancreatic cancer detection using standard of care ct scans (panorama): an international, paired, non-inferiority, confirmatory, observational study. *The Lancet Oncology*, 27(1):116–124, 2026. doi: 10.1016/S1470-2045(25)00567-4.

Michela Antonelli, Annika Reinke, Spyridon Bakas, et al. The medical segmentation decathlon. *Nature Communications*, 13:4128, 2022. doi: 10.1038/s41467-022-30695-9.

Weiwei Cao, Jianpeng Zhang, Zhongyi Shui, Sinuo Wang, Zeli Chen, Xi Li, Le Lu, Xianghua Ye, Qi Zhang, Tingbo Liang, and Ling Zhang. Boosting vision semantic density with anatomy normality modeling for medical vision-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23041–23050, 2025.

Qi Chen, Xiaoxi Chen, Haorui Song, Zhiwei Xiong, Alan Yuille, Chen Wei, and Zongwei Zhou. Towards generalizable tumor synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11147–11158, 2024.

Kaiqi Dong, Peijun Hu, Yu Tian, Yan Zhu, Xiang Li, Tianshu Zhou, Xueli Bai, Tingbo Liang, and Jingsong Li. Position-aware representation learning with anatomical priors for enhanced pancreas tumor segmentation. *Neurocomputing*, 616:128881, 2025. doi: 10.1016/j.neucom.2024.128881.

Alessandro Fontanella, Grant Mair, Joanna Wardlaw, Emanuele Trucco, and Amos Storkey. Diffusion models for counterfactual generation and anomaly detection in brain images. *IEEE Transactions on Medical Imaging*, 44(9):3574–3585, 2025. doi: 10.1109/TMI.2024.3460391.

Yunhe Gao, Yabin Zhang, Chong Wang, Jiaming Liu, Maya Varma, Jean-Benoit Delbrouck, Akshay Chaudhari, and Curtis Langlotz. Learning generalizable 3d medical image representations from mask-guided self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13744–13754, 2026.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, volume 12962 of *Lecture Notes in Computer Science*, pp. 272–284. Springer, 2022. doi: 10.1007/978-3-031-08999-2_22.

Christian Hinge, Anders Bertil Rodell, Sven Zuehlsdorff, Bruce Spottiswoode, Kirsten Korsholm, Barbara Malene Fischer, Claes Nøhr Ladefoged, and Flemming Littrup Andersen. Normal twin pet: personalized generative modeling for confounder correction and anomaly detection in whole-body pet/ct. *Scientific Reports*, 15:42662, 2025. doi: 10.1038/s41598-025-26827-y.

Haoyu Hu, Xiyao Ma, Shiqi Liu, Linsen Zhang, Xiaoliang Xie, Xiaohu Zhou, and Zeng-Guang Hou. Dino-med3d: Bridging dimension and domain gaps in volumetric segmentation via progressive adaptation. *arXiv preprint arXiv:2606.18886*, 2026. Accepted at MICCAI 2026.

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

Young Seok Jeon, Hongfei Yang, Huazhu Fu, and Mengling Feng. Teaching ai the anatomy behind the scan: Addressing anatomical flaws in medical image segmentation with learnable prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24024–24033, 2025. doi: 10.1109/ICCV51701.2025.02227.

Chengkang Li, Yishen Mao, Shuyu Liang, Ji Li, Yuanyuan Wang, and Yi Guo. Deep causal learning for pancreatic cancer segmentation in ct sequences. *Neural Networks*, 175:106294, 2024. doi: 10.1016/j.neunet.2024.106294.

Wenxuan Li, Xinze Zhou, Qi Chen, Tianyu Lin, Pedro R. A. S. Bassi, Xiaoxi Chen, Chen Ye, Zheren Zhu, Kai Ding, Heng Li, Kang Wang, Yang Yang, Yucheng Tang, Daguang Xu, Alan L. Yuille, and Zongwei Zhou. Pants: The pancreatic tumor segmentation dataset. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1065.

Yuzhen Li, Dong Liang, Xingyu Qiu, Xiangyu Li, Kuanquan Wang, and Gongning Luo. Masked graph convolutional neural network for medical image segmentation with anatomical priors. *Neurocomputing*, 669:132464, 2026. doi: 10.1016/j.neucom.2025.132464.

Peirong Liu, Ana Lawry Aguila, and Juan E. Iglesias. Unraveling normal anatomy via fluid-driven anomaly randomization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10455–10465, 2025. doi: 10.1109/CVPR52734.2025.00978.

Xinyu Ma, Jinhui Ma, Shiqi He, Xin Che, Hon Yiu So, and Lingyang Chu. Matchgen: Detecting medical abnormal region by generating matched normal regions. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15975 of *Lecture Notes in Computer Science*, pp. 321–331. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-05325-1_31.

Andrea Moglia, Matteo Cavicchioli, Luca Mainardi, and Pietro Cerveri. Deep learning for pancreas segmentation on computed tomography: A systematic review. *Artificial Intelligence Review*, 58: 220, 2025. doi: 10.1007/s10462-024-11050-4.

Marianne Rakic, Andrew Hoopes, Mazdak S. Abulnaga, Mert R. Sabuncu, John V. Guttag, Adrian V. Dalca, and Alzheimer's Disease Neuroimaging Initiative. Atlasmorph: Learning conditional deformable templates for brain mri. *Medical Image Analysis*, 110:103893, 2026. doi: 10.1016/j.media.2025.103893.

Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F. Jaeger, and Klaus H. Maier-Hein. Mednext: Transformer-driven scaling of convnets for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14223 of *Lecture Notes in Computer Science*, pp. 405–415. Springer Nature Switzerland, 2023. doi: 10.1007/978-3-031-43901-8_39.

Nico Schulthess and Ender Konukoglu. Anomaly detection by clustering dino embeddings using a dirichlet process mixture. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15962 of *Lecture Notes in Computer Science*, pp. 46–56. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-04947-6_5.

Abdelrahman M. Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Unetr++: Delving into efficient and accurate 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 43(9):3377–3390, 2024. doi: 10.1109/TMI.2024.3398728.

Linshan Wu, Jiaxin Zhuang, and Hao Chen. Large-scale 3d medical image pre-training with geometric context priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3): 3801–3818, 2026. doi: 10.1109/TPAMI.2025.3639593.

Madeleine K. Wyburd, Nicola K. Dinsdale, Mark Jenkinson, and Ana I. L. Namburete. Anatomically plausible segmentations: Explicitly preserving topology through prior deformations. *Medical Image Analysis*, 97:103222, 2024. doi: 10.1016/j.media.2024.103222.

Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15008 of *Lecture Notes in Computer Science*, pp. 578–588. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72111-3_54.

Zhaohu Xing, Liang Wan, Huazhu Fu, Guang Yang, Yijun Yang, Lequan Yu, Baiying Lei, and Lei Zhu. Diff-unet: A diffusion embedded network for robust 3d medical image segmentation. *Medical Image Analysis*, 105:103654, 2025. doi: 10.1016/j.media.2025.103654.

Zhaohu Xing, Tian Ye, Yijun Yang, Du Cai, Baowen Gai, Xiao-Jian Wu, Feng Gao, and Lei Zhu. Segmamba-v2: Long-range sequential modeling mamba for general 3-d medical image segmentation. *IEEE Transactions on Medical Imaging*, 45(1):4–15, 2026. doi: 10.1109/TMI.2025.3589797.

Xin You, Junjun He, Jie Yang, and Yun Gu. Learning with explicit shape priors for medical image segmentation. *IEEE Transactions on Medical Imaging*, 44(2):927–940, 2025. doi: 10.1109/TMI.2024.3469214.

Hantao Zhang, Yuhe Liu, Jiancheng Yang, Shouhong Wan, Xinyuan Wang, Wei Peng, and Pascal Fua. Lefusion: Controllable pathology synthesis via lesion-focused diffusion models. In *International Conference on Learning Representations*, 2025.

A Evaluation Metrics and Additional Implementation Details

A.1 Evaluation Metrics and Reporting Protocol

We distinguish *voxel-level segmentation metrics* from *patient-level tumor recovery measures*. For each subject, let P and T denote the predicted and ground-truth foreground voxel sets, respectively. We define

$$\mathrm{TP} = |P \cap T|, \qquad \mathrm{FP} = |P \setminus T|, \qquad \mathrm{FN} = |T \setminus P|, \tag{18}$$

where TP, FP, and FN denote the numbers of true-positive, false-positive, and false-negative voxels.

Overlap metrics. The Dice similarity coefficient (DSC), intersection over union (IoU), Precision, and Recall are computed as

$$\mathrm{DSC} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \qquad \mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{19}$$

and

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \qquad \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{20}$$

These quantities are defined at the voxel level. Tumor Recall in the segmentation tables denotes the fraction of ground-truth tumor voxels recovered by the prediction. It is different from the patient-level sensitivity defined below. Percentage values are obtained by multiplying the corresponding fractions by 100.

Surface metrics. Let ∂P and ∂T denote the predicted and ground-truth surfaces. For a surface point x and a surface S, the minimum physical distance is

$$d(x, S) = \min_{y \in S} \|x - y\|_2, \tag{21}$$

where y denotes a point on S. The distance is measured in millimeters using the image spacing. We collect the bidirectional surface distances as

$$\mathcal{D}(P, T) = \{d(x, \partial T) : x \in \partial P\} \cup \{d(y, \partial P) : y \in \partial T\}. \tag{22}$$

The 95th-percentile Hausdorff distance is

$$\mathrm{HD95}(P, T) = Q_{0.95}(\mathcal{D}(P, T)), \tag{23}$$

where $Q_{0.95}$ denotes the 95th percentile. Normalized surface Dice (NSD) with tolerance τ_{NSD} is

$$\mathrm{NSD}_{\tau_{\mathrm{NSD}}}(P, T) = \frac{\sum_{x \in \partial P} \mathbf{1}[d(x, \partial T) \le \tau_{\mathrm{NSD}}] + \sum_{y \in \partial T} \mathbf{1}[d(y, \partial P) \le \tau_{\mathrm{NSD}}]}{|\partial P| + |\partial T|}. \tag{24}$$

Here, $\mathbf{1}[\cdot]$ denotes the indicator function. We use $\tau_{\mathrm{NSD}} = 2$ mm in all experiments.

For a tumor-positive subject with an empty tumor prediction, NSD is set to zero, while standard HD95 is non-finite because the predicted tumor surface is empty. We do not assign an artificial finite penalty to such cases. For each random seed, HD95 is averaged only over tumor-positive subjects with finite HD95 values. The main table then reports the mean $\pm$ standard deviation across the three seed-level averages. Because the number of empty tumor predictions can differ across methods, the number of subjects contributing to HD95 can also differ. We therefore report the corresponding finite-HD95 count, $N_{\mathrm{HD95}}^{\mathrm{finite}}$, for each method. HD95 should be interpreted together with this count rather than as a standalone basis for strict method ranking.

Patient-level tumor recovery. Let $\mathcal{S}_+$ denote the set of tumor-positive test subjects and let $N_+ = |\mathcal{S}_+|$. For subject $n \in \mathcal{S}_+$, let P_n and T_n denote its predicted and ground-truth tumor voxel sets. We define the recovery indicator r_n as

$$r_n = \mathbf{1}[|P_n \cap T_n| > 0]. \tag{25}$$

Table 7: Anatomical node configurations across datasets. All three datasets use 12 ordered pancreatic chain nodes, while the available global organ nodes differ across datasets.

Dataset	Global organ nodes	Pancreatic chain nodes	Global anatomical structures
PanTSMini	24	12	Pancreas, left/right adrenal glands, aorta, bladder, celiac artery, colon, common bile duct, duodenum, gall bladder, IVC, left/right kidneys, liver, left/right lungs, pancreatic duct, portal/splenic vein, prostate, SMA, SMV, spleen, stomach, and veins.
MSD07	0	12	None.
PANORAMA	5	12	Pancreas, veins, arteries, pancreatic duct, and common bile duct.

A tumor-positive subject is therefore counted as recovered when the prediction overlaps the ground-truth tumor by at least one voxel. This measure is used only as a supplementary patient-level recovery indicator. Patient-level sensitivity is

$$\text{Patient Sensitivity} = \frac{1}{N_+} \sum_{n \in \mathcal{S}_+} r_n, \tag{26}$$

and the number of completely missed tumor-positive subjects is

$$N_{\text{miss}} = \sum_{n \in \mathcal{S}_+} (1 - r_n). \tag{27}$$

Voxel-level tumor Recall measures how much of the ground-truth tumor volume is recovered. Patient-level sensitivity measures whether a tumor-positive subject is recovered at all. A subject can therefore have low tumor Recall while still being counted as recovered at the patient level.

The finite-HD95 count is

$$N_{\text{HD95}}^{\text{finite}} = \sum_{n \in \mathcal{S}_+} \mathbf{1}[\text{HD95}(P_n, T_n) < \infty]\,. \tag{28}$$

Here, $N_{\text{HD95}}^{\text{finite}}$ denotes the number of tumor-positive subjects with finite HD95. It is different from patient-level sensitivity because HD95 can remain finite whenever the tumor prediction is non-empty, even if the predicted tumor does not overlap the ground-truth tumor.

False-positive lesions. Predicted tumor lesions are defined as 3D connected components using 6-connectivity. A predicted component is counted as a false-positive lesion when it has no voxel overlap with the ground-truth tumor. FP lesions/case is the mean number of such components over all evaluated test subjects, including tumor-negative subjects.

Metric calculation and reporting. For each random seed, DSC, IoU, NSD, and Recall are averaged over all tumor-positive test subjects, whereas HD95 is averaged only over tumor-positive subjects with finite values. The main tables report the mean $\pm$ standard deviation across the three seed-level averages. Pancreas segmentation metrics are computed over all usable test subjects. Patient-level sensitivity is reported as a cohort-level proportion.

A.2 Anatomical Grounding and Node Construction

The anatomical representation depends on the anatomical information available in each dataset. PanTSMini uses 24 global organ nodes and 12 pancreatic chain nodes. PANORAMA uses five global organ nodes and 12 pancreatic chain nodes. MSD07 uses only the 12 pancreatic chain nodes, as summarized in Table 7. Tumor labels are never included as normal-reference nodes.

Anatomical Grounding uses a dataset-specific 3D full-resolution nnU-Net anatomy predictor to generate organ masks. Each predictor is trained using the anatomical annotations available in the corresponding training data. No ground-truth organ masks or tumor annotations are used to construct SAG inputs at inference.

The 12 pancreatic chain nodes are constructed from the predicted pancreas mask. Pancreatic voxels are ordered along the longest physical axis and divided into 12 consecutive regions. Each node has an 11-D node state containing spatial, volumetric, bounding-box, and CT-intensity statistics. Each node also has a 9-D shape descriptor containing volume, extent, surface-to-volume ratio, aspect ratios, and occupancy. Missing anatomical structures are marked as invalid and are excluded from patient-to-normal comparison.

A.3 Normal Anatomy Universe Training

The Normal Anatomy Universe is learned only from the reference data described in Appendix B.1. For each valid anatomical node, the 11-D node state is encoded by a two-layer MLP ($11 \to 64 \to 64$) with LayerNorm, GELU, dropout, and final ℓ_2 normalization. The output is a normalized node embedding. We use four prototypes for each anatomical class and a matching temperature of $\tau = 0.16$.

NAU parameterization. The learnable NAU contains node prototypes $U = \{u_k\}_{k=1}^{M}$, prototype-level relations A^U, and 9-D shape prototypes S^U. Here, u_k is the k-th node prototype and M is the total number of node prototypes. The relation matrix A^U is obtained from learnable relation logits by row-wise softmax, removal of self-relations, and row normalization. Node prototypes are initialized from $\mathcal{N}(0, 0.02^2)$. The relation logits and shape prototypes are initialized to zero. Class-compatible matching follows the assignment matrix Π defined in the main text, so each patient node is compared only with prototypes from the same anatomical region.

Relational consistency and reconstruction. For a reference subject b, let $A^{(b)}$ denote its anatomical relation matrix and let $\Pi^{(b)}$ denote its node-to-prototype assignment matrix. We denote the relation matrix projected to the NAU space by $C^{(b)}$ and its reconstruction in the patient node space by $\widehat{A}^{(b)}$:

$$C^{(b)} = {\Pi^{(b)}}^{\top} A^{(b)} \Pi^{(b)}. \tag{29}$$

$$\widehat{A}^{(b)} = \Pi^{(b)} A^U {\Pi^{(b)}}^{\top}. \tag{30}$$

Let $\mathcal{B}$ denote a mini-batch of reference subjects. Let $\mathcal{P} = \{(b, b') \mid b < b',\, b, b' \in \mathcal{B}\}$ denote all subject pairs in the mini-batch. The relational consistency loss is

$$\mathcal{L}_{\text{cons}} = \frac{1}{|\mathcal{P}|} \sum_{(b,b') \in \mathcal{P}} \text{MSE}\Big(C^{(b)}, C^{(b')}\Big). \tag{31}$$

The relation reconstruction loss is

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \text{MSE}\Big(\widehat{A}^{(b)}, A^{(b)}\Big). \tag{32}$$

Here, $\text{MSE}(\cdot, \cdot)$ denotes mean squared error.

Prototype regularization. For subject b, let $\mathcal{V}_b$ denote its valid nodes and let $\mathcal{K}_i$ denote the class-compatible prototypes for node i. Let $N_v = \sum_{b \in \mathcal{B}} |\mathcal{V}_b|$ denote the total number of valid nodes in the mini-batch, and let $\epsilon = 10^{-8}$. The assignment confidence loss is

$$\mathcal{L}_{\text{conf}} = -\frac{1}{N_v} \sum_{b \in \mathcal{B}} \sum_{i \in \mathcal{V}_b} \sum_{k \in \mathcal{K}_i} \Pi_{ik}^{(b)} \log\Big(\Pi_{ik}^{(b)} + \epsilon\Big). \tag{33}$$

For node i in subject b, let $h_i^{(b)}$ denote its normalized node embedding and let $s_i^{(b)}$ denote its 9-D shape descriptor. Let s_k^U denote the shape prototype paired with node prototype u_k. The assignment-weighted shape and node references are

$$\widehat{s}_i^{(b)} = \sum_{k \in \mathcal{K}_i} \Pi_{ik}^{(b)} s_k^U, \tag{34}$$

$$\widehat{h}_i^{(b)} = \text{norm}\left(\sum_{k \in \mathcal{K}_i} \Pi_{ik}^{(b)} \,\text{norm}(u_k)\right). \tag{35}$$

Here, $\mathrm{norm}(\cdot)$ denotes ℓ_2 normalization. The shape reconstruction loss is

$$\mathcal{L}_{\mathrm{shape}} = \frac{1}{N_v} \sum_{b \in \mathcal{B}} \sum_{i \in \mathcal{V}_b} \mathrm{SmoothL1}\Big(\widehat{s}_i^{(b)}, s_i^{(b)}\Big). \tag{36}$$

The prototype compactness loss is

$$\mathcal{L}_{\mathrm{comp}} = \frac{1}{N_v} \sum_{b \in \mathcal{B}} \sum_{i \in \mathcal{V}_b} \mathrm{SmoothL1}\Big(h_i^{(b)}, \widehat{h}_i^{(b)}\Big). \tag{37}$$

Here, $\mathrm{SmoothL1}(\cdot,\cdot)$ denotes the smooth ℓ_1 loss.

Let $\mathcal{C}$ denote the set of anatomical classes. For class $c \in \mathcal{C}$, let $\mathcal{K}_c$ denote its prototype set. The indices k and l denote two prototypes in the same class. We use the cosine-margin diversity loss

$$\mathcal{L}_{\mathrm{div}} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{1}{|\mathcal{K}_c|(|\mathcal{K}_c|-1)} \sum_{\substack{k,l \in \mathcal{K}_c \\ k \neq l}} \left[\max(0, \cos(u_k, u_l) - m)\right]^2, \tag{38}$$

where $m = 0.3$ is the cosine margin.

We also regularize prototype usage in each mini-batch. Let a_k denote the total assignment weight received by prototype k:

$$a_k = \sum_{b \in \mathcal{B}} \sum_{i \in \mathcal{V}_b} \Pi_{ik}^{(b)}. \tag{39}$$

For prototype k in class c, let $q_{c,k}$ denote its normalized usage within that class:

$$q_{c,k} = \frac{a_k}{\sum_{l \in \mathcal{K}_c} a_l}, \qquad k \in \mathcal{K}_c. \tag{40}$$

Let q_c denote the vector $\{q_{c,k}\}_{k \in \mathcal{K}_c}$, and let $\mathbf{1}$ denote an all-one vector of length $|\mathcal{K}_c|$. The usage loss is

$$\mathcal{L}_{\mathrm{use}} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathrm{MSE}\Big(q_c, \frac{1}{|\mathcal{K}_c|}\mathbf{1}\Big). \tag{41}$$

Training objective. The full NAU objective is

$$\begin{aligned} \mathcal{L}_U = \mathcal{L}_{\mathrm{cons}} + \mathcal{L}_{\mathrm{rec}} + 0.01\,\mathcal{L}_{\mathrm{conf}} + 0.10\,\mathcal{L}_{\mathrm{shape}} + 0.10\,\mathcal{L}_{\mathrm{comp}} \\ + 0.05\,\mathcal{L}_{\mathrm{div}} + 0.10\,\mathcal{L}_{\mathrm{use}}. \end{aligned} \tag{42}$$

Class compatibility is enforced directly by the prototype-matching mask, so no additional class-matching loss is used. NAU is trained for 80 epochs with a mini-batch size of 8 using AdamW. The learning rate is 2×10^{-4}, the weight decay is 1×10^{-5}, the encoder dropout is 0.1, and the random seed is 2026. After training, the node encoder, node prototypes, normal relation matrix, and shape prototypes are frozen. They are then used to construct patient-specific normal references and structural deviation. Tumor segmentation labels are not used to optimize NAU.

A.4 Deviation Field

Structural deviation scores are projected to voxel space only through valid pancreatic chain nodes. Global organ nodes are not projected directly. However, they still affect the pancreatic chain node deviations because each deviation score is computed from the complete patient-specific star–chain anatomical structure. We use deviation weights $(w_n, w_r, w_s, w_a) = (0.25, 0.75, 0.25, 1.0)$ in all experiments, with normalization by their sum as defined in the main text.

The anatomy predictor provides a voxel-wise confidence map. When softmax probabilities are available, the confidence at each voxel is the probability assigned to its predicted anatomical label. The confidence of a node is the mean voxel confidence inside its mask. For PanTSMini and PANORAMA, node confidence is applied before projection. We then apply normalized masked Gaussian smoothing within the valid projection region with a standard deviation of 2.5 mm. For MSD07, confidence weighting is applied at both the node and voxel levels without Gaussian smoothing. The resulting deviation field is used as the additional input to the SAG-aware segmentation model. Unless otherwise specified, the four deviation components are combined into one deviation field rather than used as separate input channels.

Table 8: Normal Reference construction across datasets.

Dataset	Reference cohort	Reference construction
PanTSMini	1,000 cases	Separate tumor-negative cohort.
MSD07	40 cases	Normal-appearing reference with the tumor and a 5-mm margin around it excluded.
PANORAMA	400 cases	Separate tumor-negative cohort.

Table 9: Dataset split statistics. Numbers of total, usable, empty, and tumor-positive cases in each split.

Dataset	Split	Total	Usable	Empty	Tumor-positive
PanTSMini	Train	7,901	7,474	427	–
	Val.	500	488	12	250
	Test	500	487	13	250
	Reference	1,000	1,000	0	0
MSD07	Train	161	161	0	161
	Val.	40	40	0	40
	Test	40	40	0	40
	Reference	40	40	0	40
PANORAMA	Train	1,164	1,164	0	–
	Val.	200	200	0	100
	Test	200	200	0	100
	Reference	400	400	0	0

B Normal Reference Construction and Dataset Splits

B.1 Reference Construction and Leakage Safeguards

The Normal Anatomy Universe is learned only from the designated reference split and is frozen before validation and test evaluation. The reference subjects are separate from the segmentation training, validation, and test splits. No validation or test subject is used to learn or update the Normal Anatomy Universe.

PanTSMini and PANORAMA use separate tumor-negative reference cohorts. For MSD07, an independent healthy cohort is unavailable. We therefore construct a normal-appearing reference by excluding the annotated tumor and a 5-mm margin around the tumor from the designated reference cases. This margin is used only for reference construction. The MSD07 reference is therefore referred to as a normal-appearing reference rather than an independent healthy cohort. Table 8 summarizes the reference construction.

The reference cases are used only to learn the Normal Anatomy Universe. They are not used as additional tumor-segmentation training cases. For PanTSMini and PANORAMA, the reference cases are tumor-negative and are used to model normal anatomical variation. For MSD07, only normal-appearing regions outside the tumor and its 5-mm margin are used. Thus, the reference data provide normal anatomical information for patient-to-normal comparison rather than additional tumor supervision.

At inference, Anatomical Grounding uses the anatomy predictor to produce organ masks from the input CT volume. SAG then constructs the patient-specific anatomy from these predicted masks. No ground-truth organ masks or tumor annotations are used to construct SAG inputs at inference.

B.2 Split Composition

Table 9 summarizes the dataset split. **Usable** denotes cases retained after data validity checks, and **Empty** denotes cases marked as empty in the split records. The reference split is not included in validation or test evaluation. For PANORAMA, cases overlapping with PanTSMini or MSD07 are removed before split construction and evaluation.

C BRR Feature Construction and Out-of-Fold Training

C.1 BRR Feature Construction

BRR uses information from the CT-only and SAG-aware predictions. For each voxel v, let p^t_{base} and p^t_{SAG} denote the tumor probabilities. Let p^p_{base} and p^p_{SAG} denote the pancreas probabilities. The subscripts base and SAG refer to the CT-only and SAG-aware predictions, respectively. The superscripts t and p denote tumor and pancreas.

For a tumor probability p^t, its logit is $z^t = \mathrm{logit}(p^t) = \log(p^t/(1-p^t))$. For a probability p, the binary entropy is

$$H_{\mathrm{bin}}(p) = -p\log p - (1-p)\log(1-p). \tag{43}$$

For numerical stability, probabilities are clipped to $[10^{-5}, 1-10^{-5}]$ before logits and entropy are computed.

The 10-D raw feature vector is

$$\begin{aligned}\varphi_{\mathrm{raw}}(v) = \Big[&p^t_{\mathrm{base}}(v),\, p^t_{\mathrm{SAG}}(v),\, p^p_{\mathrm{base}}(v),\, p^p_{\mathrm{SAG}}(v),\, z^t_{\mathrm{base}}(v),\, z^t_{\mathrm{SAG}}(v),\, p^t_{\mathrm{SAG}}(v) - p^t_{\mathrm{base}}(v),\\ &\max\big(p^t_{\mathrm{base}}(v), p^t_{\mathrm{SAG}}(v)\big)\,,\, H_{\mathrm{bin}}\big(p^t_{\mathrm{base}}(v)\big)\,,\, H_{\mathrm{bin}}\big(p^t_{\mathrm{SAG}}(v)\big)\Big]^\top.\end{aligned} \tag{44}$$

Here, $\varphi_{\mathrm{raw}}(v)$ denotes the raw BRR feature vector and $^\top$ denotes transpose.

The features are standardized using statistics from the OOF training samples:

$$\varphi(v) = \frac{\varphi_{\mathrm{raw}}(v) - \mu_{\mathrm{OOF}}}{\sigma_{\mathrm{OOF}}}. \tag{45}$$

Here, μ_{OOF} and σ_{OOF} are the feature-wise mean and standard deviation computed only from the OOF training samples. The standardized feature vector is denoted by $\varphi(v)$.

C.2 BRR Training Details

The refiner R_ω is a $10 \to 32 \to 32 \to 1$ MLP, where ω denotes its trainable parameters. LayerNorm is applied after the first hidden layer, and GELU is used after both hidden layers. The residual bound is $\Delta_{\max} = 3.0$.

Let $\mathcal{Q}$ denote a sampled BRR training mini-batch and let $N = |\mathcal{Q}|$ denote its number of voxels. For voxel $v \in \mathcal{Q}$, let $y_v \in \{0,1\}$ denote the ground-truth tumor label, let $\hat{p}_v$ denote the BRR tumor probability, and let $\Delta z(v)$ denote the bounded logit correction defined in the main text.

Let ρ denote the fraction of tumor-positive voxels in the BRR training samples. The positive class weight α is

$$\alpha = \min\left(\frac{1-\rho}{\rho}, 20\right). \tag{46}$$

The weighted binary cross-entropy loss is

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{N}\sum_{v\in\mathcal{Q}} \left[\alpha y_v \log \hat{p}_v + (1-y_v)\log(1-\hat{p}_v)\right]. \tag{47}$$

The soft Dice loss is

$$\mathcal{L}_{\mathrm{Dice}} = 1 - \frac{2\sum_{v\in\mathcal{Q}} \hat{p}_v y_v + 10^{-6}}{\sum_{v\in\mathcal{Q}} \hat{p}_v + \sum_{v\in\mathcal{Q}} y_v + 10^{-6}}. \tag{48}$$

The residual loss is

$$\mathcal{L}_{\Delta} = \frac{1}{N}\sum_{v\in\mathcal{Q}} |\Delta z(v)|. \tag{49}$$

The anchor loss is

$$\mathcal{L}_{\mathrm{anchor}} = \frac{1}{N}\sum_{v\in\mathcal{Q}} \big(\hat{p}_v - p^t_{\mathrm{base}}(v)\big)^2. \tag{50}$$

The four losses are combined using Eq. 15 in the main text. We set $\lambda_{\Delta} = 0.01$ and $\lambda_{\text{anchor}} = 0.05$. Thus, $\mathcal{L}_{\Delta}$ limits the size of the logit correction, while $\mathcal{L}_{\text{anchor}}$ keeps the final tumor probability close to the CT-only prediction.

For BRR training, we sample 20,000 voxels from each case. Among these samples, 35% are tumor voxels and 10% are background voxels outside the candidate region. The remaining 55% are non-tumor voxels inside the candidate region. The candidate region contains voxels with a pancreas probability of at least 0.5 or a tumor probability of at least 0.05 in either prediction. Ground-truth tumor voxels are also included in the candidate region. Sampling uses replacement when a group contains fewer voxels than requested.

BRR is trained for 500 epochs with four steps per epoch and a batch size of 65,536 sampled voxels. We use AdamW with a learning rate of 1×10^{-3} and a weight decay of 1×10^{-4}.

C.3 Out-of-Fold Training Protocol

Five-fold OOF inference is used only to generate training data for the Refiner. The five folds are constructed entirely within the training split. For each fold, the CT-only and SAG-aware segmentation models are trained on the other four folds and generate predictions only for the held-out fold. The held-out predictions from all five folds are then pooled to train the Refiner.

The anatomy predictor is trained once using the available organ annotations and is then frozen. The same anatomy predictor is used across all OOF folds and is not cross-fitted. The Normal Anatomy Universe is learned separately from the reference split and is also frozen before OOF prediction generation.

The validation and test splits are not involved in the OOF procedure. The OOF models are used only to construct Refiner training data and are not ensembled at test time.

D Additional Quantitative Results

This section provides additional quantitative results for DARA. We report pancreas segmentation and patient-level tumor recovery results. We also analyze tumor segmentation across different tumor sizes and case-level Dice thresholds. All PANORAMA results use the split defined in Appendix B.2, after removing cases that overlap with PanTSMini or MSD07.

D.1 Pancreas Segmentation

Table 10 reports pancreas segmentation across the same methods and datasets used in the main tumor comparison. Results are averaged over three random seeds.

DARA shows consistent pancreas segmentation performance across the three datasets. On PanTSMini, it achieves the highest DSC, IoU, Precision, and Recall. On MSD07, the proposed model achieves the best DSC, IoU, Precision, and HD95. On MSD07, it achieves the highest DSC, IoU, and Precision and the lowest HD95. On PANORAMA, it achieves the highest DSC, IoU, Precision, and Recall and the second-lowest HD95.

D.2 Patient-Level Tumor Recovery

Table 11 reports patient-level tumor recovery and false-positive results. Patient-level sensitivity indicates whether a tumor-positive subject is recovered. The missed-positive count reports the number of completely missed tumor-positive subjects. FP lesions/case reports the average number of false-positive tumor lesions per test case. The finite-HD95 count reports the number of tumor-positive subjects with finite HD95. The metric definitions are given in Appendix A.1.

On PanTSMini, DARA achieves the highest patient-level sensitivity of 82.00% with 45 missed tumor-positive subjects. On PANORAMA, DARA also achieves the highest patient-level sensitivity of 92.00% with 8 missed tumor-positive subjects. For PanTSMini and PANORAMA, FP lesions/case is computed over the complete evaluated test cohort and therefore also includes tumor-negative subjects. All 40 MSD07 test subjects are tumor-positive, so its missed-positive count is directly determined from patient-level sensitivity.

Table 10: Pancreas segmentation comparison across datasets. Results are reported as mean ± standard deviation across three random seeds. Best and second-best mean values within each dataset and metric are shown in bold and underlined, respectively.

Method	**Source**	DSC(%)↑	IoU(%)↑	Precision(%)↑	HD95(mm)↓	Recall(%)↑
		PanTSMini				
nnU-Net (Isensee et al., 2021)	Nature'21	84.32± 0.18	74.99± 0.26	84.41± 0.21	9.36± 0.32	87.34± 0.19
SwinUNETR (Hatamizadeh et al., 2022)	MICCAI'22	80.74± 0.08	69.93± 0.11	79.56± 0.09	20.45± 0.24	85.45± 0.10
MedNeXt (Roy et al., 2023)	MICCAI'23	72.93± 0.31	60.30± 0.42	66.60± 0.36	97.05± 0.52	86.90± 0.34
SegMamba (Xing et al., 2024)	MICCAI'24	69.42± 0.12	55.97± 0.16	62.96± 0.14	97.09± 0.21	83.62± 0.13
UNETR++ (Shaker et al., 2024)	TMI'24	73.01± 0.27	60.67± 0.37	70.70± 0.32	40.05± 0.45	80.21± 0.29
DiffUNet (Xing et al., 2025)	MedIA'25	76.60± 0.15	64.57± 0.21	73.56± 0.18	56.09± 0.34	84.34± 0.16
MASS (Gao et al., 2026)	CVPR'26	83.42± 0.22	73.30± 0.31	84.23± 0.25	8.37± 0.18	84.90± 0.23
DINO-Med3D (Hu et al., 2026)	MICCAI'26	79.17± 0.09	67.38± 0.13	78.13± 0.11	10.77± 0.15	83.15± 0.10
SegMamba-V2 (Xing et al., 2026)	TMI'26	72.07± 0.35	59.15± 0.47	66.44± 0.41	81.59± 0.56	84.32± 0.38
VoCo (Wu et al., 2026)	TPAMI'26	84.39± 0.14	75.02± 0.20	84.13± 0.16	**7.38± 0.12**	87.15± 0.15
DARA	Ours	**84.58± 0.19**	**75.32± 0.28**	**84.45± 0.22**	10.37± 0.29	**87.78± 0.20**
		MSD07				
nnU-Net (Isensee et al., 2021)	Nature'21	85.78± 0.13	75.35± 0.19	86.02± 0.15	5.21± 0.14	**87.38± 0.12**
SwinUNETR (Hatamizadeh et al., 2022)	MICCAI'22	75.48± 0.34	62.66± 0.47	71.03± 0.42	98.12± 0.58	84.86± 0.36
MedNeXt (Roy et al., 2023)	MICCAI'23	85.27± 0.07	74.78± 0.10	85.43± 0.08	5.49± 0.09	87.30± 0.08
SegMamba (Xing et al., 2024)	MICCAI'24	76.88± 0.28	63.14± 0.39	76.30± 0.35	22.28± 0.44	80.87± 0.31
UNETR++ (Shaker et al., 2024)	TMI'24	74.70± 0.41	60.57± 0.55	69.67± 0.49	17.17± 0.48	85.30± 0.43
DiffUNet (Xing et al., 2025)	MedIA'25	80.34± 0.16	67.73± 0.23	81.23± 0.19	7.80± 0.21	82.43± 0.18
MASS (Gao et al., 2026)	CVPR'26	82.50± 0.24	70.89± 0.34	82.22± 0.29	10.12± 0.33	85.08± 0.26
DINO-Med3D (Hu et al., 2026)	MICCAI'26	82.35± 0.11	70.57± 0.16	82.40± 0.13	7.55± 0.17	84.74± 0.12
SegMamba-V2 (Xing et al., 2026)	TMI'26	75.22± 0.37	61.35± 0.51	72.62± 0.45	29.73± 0.54	82.71± 0.40
VoCo (Wu et al., 2026)	TPAMI'26	83.06± 0.10	71.55± 0.15	81.44± 0.12	5.98± 0.13	87.33± 0.11
DARA	Ours	**85.79± 0.08**	**75.38± 0.12**	**86.15± 0.10**	**5.14± 0.11**	87.27± 0.09
		PANORAMA				
nnU-Net (Isensee et al., 2021)	Nature'21	93.22± 0.06	86.15± 0.09	92.67± 0.07	4.05± 0.10	94.02± 0.06
SwinUNETR (Hatamizadeh et al., 2022)	MICCAI'22	85.77± 0.21	76.13± 0.30	82.44± 0.26	68.11± 0.38	90.58± 0.23
MedNeXt (Roy et al., 2023)	MICCAI'23	93.78± 0.04	88.67± 0.06	93.63± 0.05	5.23± 0.07	94.38± 0.04
SegMamba (Xing et al., 2024)	MICCAI'24	91.65± 0.14	84.97± 0.21	93.28± 0.18	7.49± 0.19	90.50± 0.16
UNETR++ (Shaker et al., 2024)	TMI'24	92.33± 0.09	86.15± 0.14	91.15± 0.12	5.40± 0.13	94.01± 0.10
DiffUNet (Xing et al., 2025)	MedIA'25	92.54± 0.07	86.58± 0.11	93.25± 0.09	**3.04± 0.08**	92.39± 0.08
MASS (Gao et al., 2026)	CVPR'26	88.90± 0.19	80.56± 0.28	89.52± 0.24	5.81± 0.25	89.07± 0.20
DINO-Med3D (Hu et al., 2026)	MICCAI'26	89.28± 0.12	81.82± 0.18	87.40± 0.15	9.33± 0.22	91.90± 0.13
SegMamba-V2 (Xing et al., 2026)	TMI'26	92.18± 0.10	86.03± 0.15	93.20± 0.13	3.98± 0.12	91.82± 0.11
VoCo (Wu et al., 2026)	TPAMI'26	92.09± 0.08	85.85± 0.12	92.11± 0.10	4.35± 0.11	92.70± 0.09
DARA	Ours	**93.97± 0.05**	**89.15± 0.08**	**93.88± 0.06**	3.71± 0.08	**94.66± 0.05**

Tumor-size analysis. Figure 5 compares tumor segmentation performance across different tumor-size groups. DARA shows improvements across multiple groups on PanTSMini and MSD07, with larger differences in the intermediate-size groups. The differences are smaller at the extreme groups, and very small tumors remain challenging. These results show that the benefit of structural deviation is observed across different tumor sizes rather than being limited to large tumors.

Case-level DSC distribution. Figure 6 shows the percentage of tumor-positive cases whose DSC is higher than different thresholds. On PanTSMini and MSD07, DARA has a higher percentage of cases over a wide range of thresholds. The difference is smaller on PANORAMA, which is consistent with the smaller difference in mean DSC on this dataset. This suggests that the improvement of DARA is distributed across many cases rather than being caused by only a few high-DSC cases.

E Additional Qualitative Ablation Analysis

This section provides qualitative results for the ablation studies in Sec. 4.4. All examples are from PanTSMini and use the same setting as the quantitative ablations. The same representative cases are used across all figures to facilitate comparison across variants. The boxes show enlarged tumor regions.

Table 11: Patient-level tumor recovery and false-positive behavior across methods. Best and second-best results for patient sensitivity and missed-positive count are shown in bold and underlined, respectively. FP lesions/case is computed over all evaluated test subjects. HD95 finite n reports the number of tumor-positive cases with finite HD95.

Method	Source	Patient sens.(%)↑	Missed positive↓	FP lesions/case↓	HD95 finite n
PanTSMini					
nnU-Net (Isensee et al., 2021)	Nature'21	72.40	69	0.48	223
SwinUNETR (Hatamizadeh et al., 2022)	MICCAI'22	74.00	65	3.64	242
MedNeXt (Roy et al., 2023)	MICCAI'23	<u>78.00</u>	<u>55</u>	27.66	250
SegMamba (Xing et al., 2024)	MICCAI'24	67.60	81	25.93	248
UNETR++ (Shaker et al., 2024)	TMI'24	55.60	111	4.41	213
DiffUNet (Xing et al., 2025)	MedIA'25	77.60	56	9.20	248
MASS (Gao et al., 2026)	CVPR'26	75.20	62	0.70	229
DINO-Med3D (Hu et al., 2026)	MICCAI'26	54.00	115	0.57	175
SegMamba-V2 (Xing et al., 2026)	TMI'26	71.60	71	29.78	248
VoCo (Wu et al., 2026)	TPAMI'26	71.60	71	**0.32**	215
DARA	Ours	**82.00**	**45**	<u>0.42</u>	242
MSD07					
nnU-Net (Isensee et al., 2021)	Nature'21	90.00	4	<u>0.13</u>	36
SwinUNETR (Hatamizadeh et al., 2022)	MICCAI'22	87.50	5	5.85	39
MedNeXt (Roy et al., 2023)	MICCAI'23	87.50	5	**0.10**	38
SegMamba (Xing et al., 2024)	MICCAI'24	82.50	7	2.83	40
UNETR++ (Shaker et al., 2024)	TMI'24	82.50	7	1.60	36
DiffUNet (Xing et al., 2025)	MedIA'25	85.00	6	0.93	39
MASS (Gao et al., 2026)	CVPR'26	**97.50**	**1**	1.60	40
DINO-Med3D (Hu et al., 2026)	MICCAI'26	90.00	4	1.48	40
SegMamba-V2 (Xing et al., 2026)	TMI'26	87.50	5	2.38	40
VoCo (Wu et al., 2026)	TPAMI'26	<u>95.00</u>	<u>2</u>	0.38	38
DARA	Ours	92.50	3	0.18	39
PANORAMA					
nnU-Net (Isensee et al., 2021)	Nature'21	<u>91.00</u>	<u>9</u>	0.21	94
SwinUNETR (Hatamizadeh et al., 2022)	MICCAI'22	70.00	30	0.35	81
MedNeXt (Roy et al., 2023)	MICCAI'23	85.00	15	0.31	92
SegMamba (Xing et al., 2024)	MICCAI'24	82.00	18	0.86	86
UNETR++ (Shaker et al., 2024)	TMI'24	84.00	16	0.28	84
DiffUNet (Xing et al., 2025)	MedIA'25	73.00	27	<u>0.13</u>	78
MASS (Gao et al., 2026)	CVPR'26	90.00	10	1.73	97
DINO-Med3D (Hu et al., 2026)	MICCAI'26	87.00	13	1.77	97
SegMamba-V2 (Xing et al., 2026)	TMI'26	70.00	30	0.67	75
VoCo (Wu et al., 2026)	TPAMI'26	88.00	12	0.36	92
DARA	Ours	**92.00**	**8**	**0.11**	95

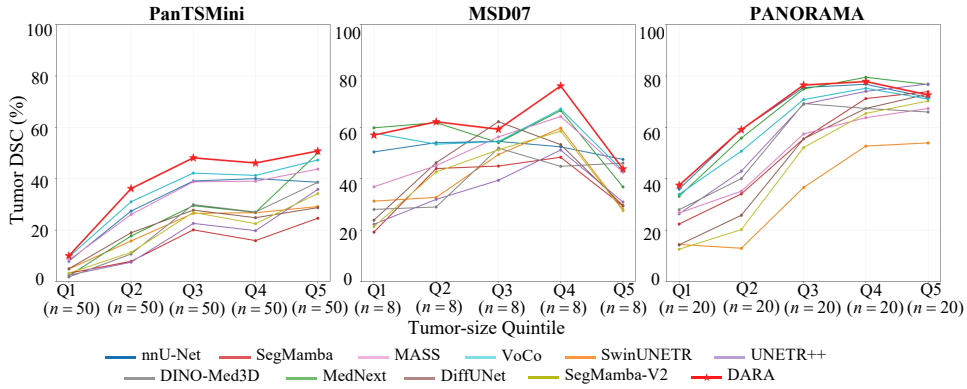

Figure 5: Tumor segmentation across tumor-size groups. Cases are grouped from Q1 (smallest) to Q5 (largest).

E.1 Effect of Structural Deviation

Figure 7 shows the effect of structural deviation in DARA. Compared with the CT-only prediction, the SAG-aware prediction contains additional tumor regions in the displayed cases. BRR then reconciles the CT-only and SAG-aware predictions to produce the final prediction. The segmentation

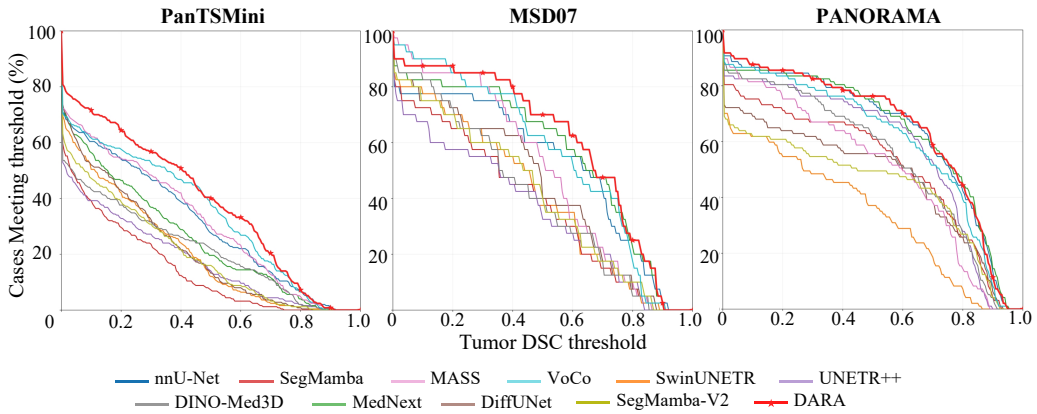

Figure 6: Case-level tumor segmentation across Dice thresholds. The x-axis denotes the DSC threshold, and the y-axis denotes the percentage of tumor-positive cases exceeding that threshold. Each curve shows the percentage of tumor-positive cases with a Dice score above the corresponding threshold.

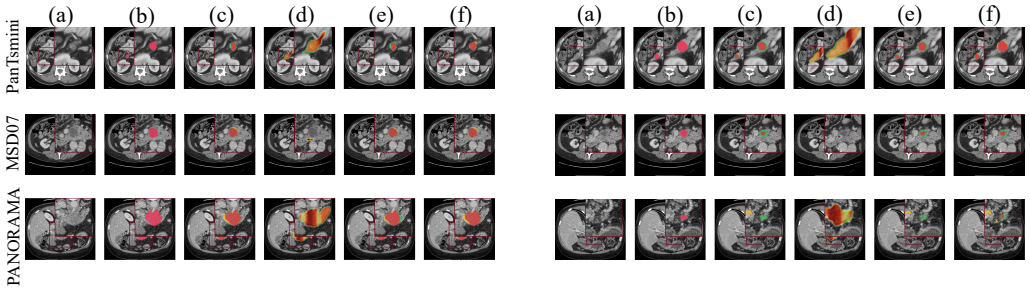

Figure 7: Qualitative effect of structural deviation and BRR. (a) CT image, (b) ground truth, (c) CT-only prediction, (d) deviation field D_{SAG}, (e) SAG-aware prediction, and (f) final prediction. Red denotes prediction–ground-truth overlap, green denotes missed ground-truth tumor, and yellow denotes prediction outside the ground truth.

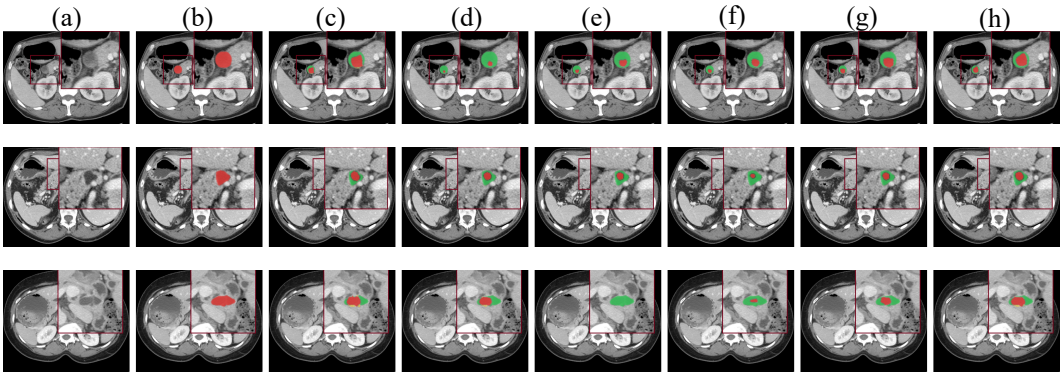

Figure 8: Qualitative comparison of structural deviation components. (a) CT image, (b) ground truth, (c) node deviation only, (d) relation deviation only, (e) shape deviation only, (f) appearance deviation only, (g) N+R+S, and (h) N+R+S+A. Here, N, R, S, and A denote node, relation, shape, and appearance deviations, respectively.

panels use the same red, green, and yellow convention as Fig. 4. The deviation field D_{SAG} is shown as a heatmap.

E.2 Deviation Components

Figure 8 compares the four deviation components using the same representative cases. Node, relation, shape, and appearance deviations produce different segmentation results. Their combination gives the result shown by N+R+S+A. The corresponding quantitative results are reported in Table 3.

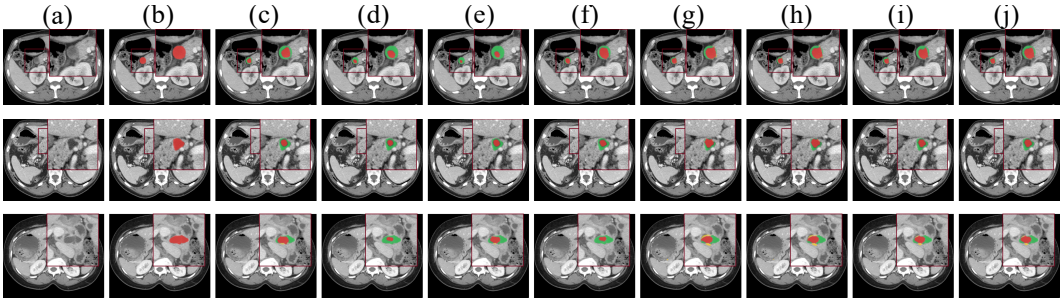

Figure 9: Qualitative comparison of prediction reconciliation strategies. (a) CT image, (b) ground truth, (c) SAG-aware prediction, (d) simple average, (e) weighted average, (f) maximum, (g) learned fusion without OOF, (h) OOF MLP, (i) BRR without structural deviation, and (j) BRR (ours).

E.3 Prediction Reconciliation

Figure 9 shows the prediction results of the reconciliation strategies compared in Table 6. The figure includes simple average, weighted average, maximum, learned fusion without OOF, OOF MLP, and BRR. These examples show how different reconciliation strategies change the tumor prediction. The quantitative comparison is reported in Table 6.