STRUCTURED ROUTING OVER PROPAGATION AXES AND RECEPTIVE SUPPORTS FOR VOLUMETRIC MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Volumetric medical image segmentation requires context that adapts to anatomical direction, spatial location, and semantic class. We introduce VERSA, a structured adaptive routing framework that learns how contextual evidence propagates and which spatial supports inform each prediction. In latent space, V-BAR RWKV learns spatially varying mixtures of bidirectional recurrent responses along three orthogonal axes and integrates them with anisotropic multi-support context. In prediction space, DiVERS constructs nested point, axial, planar-union, and volumetric support experts over decoder features and combines them with boundary-enhanced and pooled-context experts through class- and location-conditioned routing. Normalized expert disagreement regulates the bounded gain of contextual corrections to a point-supported anchor. With 3.67M parameters, VERSA outperforms 19 existing state-of-the-art methods on all 12 public benchmarks spanning CT, MRI, and OCT. Controlled ablations show that learned routing outperforms uniform aggregation with identical candidate axes and support experts, demonstrating the value of adaptive context selection beyond receptive-field expansion.

1 INTRODUCTION

Volumetric medical image segmentation requires contextual reasoning without sacrificing fine anatomical detail. U-shaped convolutional networks capture local and multiscale structure (Çiçek et al., 2016; Isensee et al., 2021; Roy et al., 2023), while Transformers and state-space or recurrent models extend contextual interactions (Hatamizadeh et al., 2022b; Xing et al., 2024; Peng et al., 2023; Duan et al., 2025). Yet access to broader context does not determine how it should be used. Elongated structures benefit from directional continuity, whereas thin boundaries demand selective aggregation. The central question is therefore not only how much context to provide, but along which axes to propagate it and over which supports to form predictions (Figure 1).

These choices concern complementary stages of volumetric computation. In latent space, fixed scan orders and spatially invariant axis fusion do not explicitly represent location-dependent preferences among propagation directions. At the prediction stage, fixed-support classifiers leave the choice among differently shaped neighborhoods implicit. Directional anatomy and anisotropic sampling make these choices consequential: overly restricted context can miss structural continuity, while broad aggregation can mix evidence across boundaries. We therefore investigate input-conditioned routing over predefined propagation axes and receptive supports.

We introduce VERSA, a Volumetric Explicit Routing framework over Supports and Axes that unifies these routing decisions within a U-shaped architecture. Its bottleneck module, Volumetric Bidirectional Axis-Routed RWKV (V-BAR RWKV), forms bidirectional recurrent responses along depth, height, and width with axis-specific token mixing and recurrent dynamics. Spatially varying routing combines these responses, while update-conditioned fusion integrates recurrent and anisotropic multi-support context. At the prediction stage, Disagreement-Regulated Volumetric Expert Routing over Supports (DiVERS) constructs nested point, axial, planar-union, and volumetric experts over decoder features, supplemented by boundary-enhanced and pooled-context experts. Class-specific global evidence and class-shared spatial scores produce class- and location-conditioned expert mixtures. DiVERS refines a point-supported anchor with a routed residual whose gain is bounded

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.

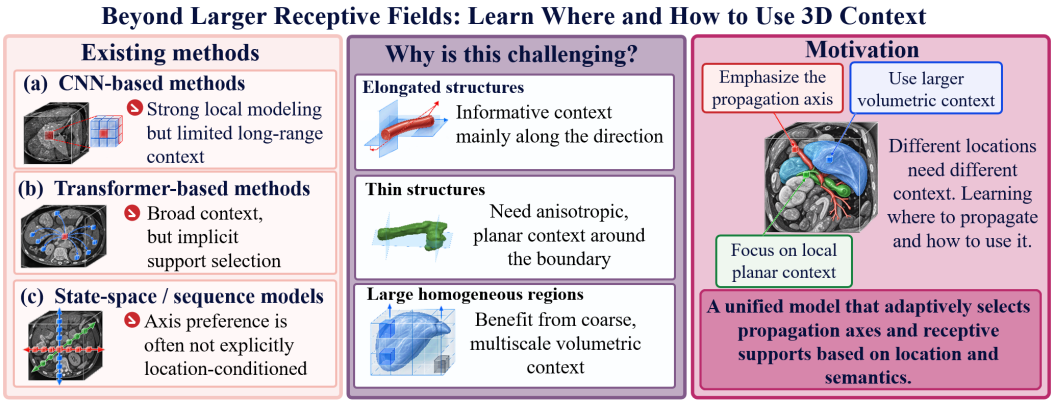

Figure 1: Context selection beyond receptive-field expansion.

and modulated by normalized expert disagreement. A learned class-specific coefficient determines whether disagreement amplifies or attenuates this refinement. Our contributions are threefold:

- We formulate volumetric context selection as structured adaptive routing over propagation axes in latent space and receptive supports in prediction space, unified within VERSA.
- We develop V-BAR RWKV for spatial axis routing and DiVERS for class-conditioned support routing with disagreement-regulated, point-anchored refinement.
- We evaluate VERSA against 19 methods on 12 public benchmarks. Controlled ablations retaining the same candidate axes or support experts favor learned routing over uniform aggregation, supporting the value of context selection beyond receptive-field expansion.

2 Related Work

3D Medical Image Segmentation. Volumetric medical image segmentation requires voxel-wise delineation while preserving fine structures and anatomical coherence in anisotropically sampled volumes. 3D U-Net (Çiçek et al., 2016) and V-Net (Milletari et al., 2016) established skip-connected encoder–decoder architectures that capture local and multiscale structure for volumetric segmentation. UNETR (Hatamizadeh et al., 2022b), SwinUNETR (Hatamizadeh et al., 2022a), and nnFormer (Zhou et al., 2023) extend contextual interactions through global or hierarchical self-attention, while SegMamba (Xing et al., 2024) provides efficient long-range modeling via multidirectional state-space scans. Nevertheless, high-resolution volumes, anisotropic voxel spacing, and substantial variations in target morphology still pose a central representation-learning challenge: preserving fine local geometry while propagating information over anatomically relevant spatial extents.

RWKV for Medical Image Segmentation. RWKV (Peng et al., 2023) combines recurrent weighted key–value aggregation with parallelizable training and linear scaling in sequence length, while Vision-RWKV (Duan et al., 2025) adapts it to dense visual representations. U-RWKV (Cai et al., 2026) introduces spatial shifts and tri-axis WKV scans with learned feature fusion. MMSegR-WKV (Cao et al., 2026) factorizes spatial and through-slice dependencies using dual-view WKV and a dynamic short-time window. This leaves underexplored the spatially conditioned coordination of anatomical-axis preference, forward–reverse propagation, and local–global context, V-BAR RWKV treats axis preference as a spatial routing decision and combines it with channel-wise bidirectional calibration and update-conditioned fusion of recurrent and multi-support context.

Mixture-of-Experts for Volumetric Segmentation. Mixture-of-experts (MoE) models combine specialized computations through learned routing (Shazeer et al., 2017). SwinSMT (Płotka et al., 2024) replaces deep feed-forward layers with Soft MoE. SAM-Med3D-MoE (Wang et al., 2024) combines general and task-specific decoders for downstream adaptation. Mamba-HoME (Plotka et al., 2025) organizes Soft-MoE routing into local and global stages. These methods primarily specialize feature transformations. DiVERS defines prediction experts over nested point, axial, planar-union, and volumetric supports on decoder features, complemented by boundary-enhanced and pooled-context experts. Dense class- and location-conditioned routing combines logits, with expert disagreement modulating the bounded gain of residual corrections to a point-supported anchor.

3 METHOD

3.1 THE STRUCTURE OF VERSA

Preserving fine anatomical details while aggregating relevant context is challenging in volumetric medical image segmentation, as contextual requirements vary in direction and spatial support across locations and classes. To address this challenge, we propose VERSA, a volumetric segmentation framework that unifies spatially adaptive propagation-axis routing in latent space with class- and location-conditioned receptive-support routing in prediction space (Figure 2). At the bottleneck, V-BAR RWKV integrates bidirectional propagation along the depth, height, and width axes with anisotropic multi-support context through spatially adaptive routing. DiVERS produces voxel-wise logits through class- and location-conditioned routing over six support experts, with normalized expert disagreement regulating the bounded gain of residual corrections to a point-supported anchor.

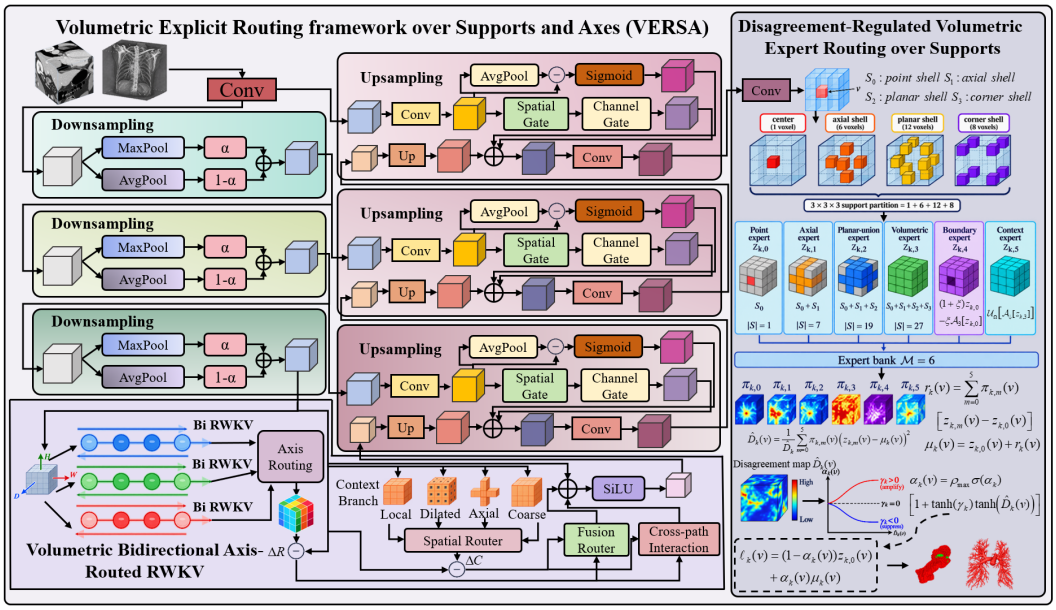

Figure 2: Overview of VERSA for volumetric medical image segmentation, integrating spatially adaptive bidirectional axis routing via V-BAR RWKV with class- and location-conditioned receptive-support routing and disagreement-regulated, point-anchored refinement via DiVERS.

3.2 VOLUMETRIC BIDIRECTIONAL AXIS-ROUTED RWKV

Volumetric anatomy is both directional and scale dependent. Structures may remain continuous along one coordinate axis but change abruptly along another, while the evidence required at a voxel may range from a compact boundary neighborhood to coarse organ-level context. A single rasterized sequence obscures this geometry by imposing an arbitrary traversal order, whereas a fixed convolution applies the same receptive support everywhere. We therefore formulate volumetric context modeling as structured selection over two finite geometric bases: axis-aligned propagators that transport information and receptive-field operators that aggregate it. V-BAR RWKV is placed at the encoder bottleneck and contains two parallel paths. A bidirectional tri-axis RWKV path models long-range dependencies along depth, height, and width lines; an anisotropic pyramid path constructs local, dilated, axis-elongated, and coarse-context responses. Spatial routers combine the candidate axes and supports, after which the two paths are reconciled in residual-update space.

Let $\mathbf{X} \in \mathbb{R}^{B\times C\times D\times H\times W}$ denote the input. We first form a compact recurrent representation and an unpooled local response,

$$\begin{aligned} \mathbf{Z}_0 &= \mathcal{P}_{\text{in}}(\mathrm{GN}(\mathbf{X})), \quad \mathbf{Z} = \mathbf{Z}_0 + \mathcal{K}_{\text{pos}}^{3\times3\times3,\text{dw}}(\mathbf{Z}_0), \quad \mathbf{L} = \mathcal{P}_{\text{loc}}\Big[\phi\Big(\mathrm{GN}\Big(\mathcal{K}_{\text{loc}}^{3\times3\times3,g}(\mathbf{Z})\Big)\Big)\Big], \\ \widetilde{\mathbf{Z}} &= \mathrm{AAP}_{\widetilde{D}\times\widetilde{H}\times\widetilde{W}}(\mathbf{Z}), \quad (\widetilde{D},\widetilde{H},\widetilde{W}) = \big(\min(D,M),\min(H,M),\min(W,M)\big), \end{aligned} \tag{1}$$

where ϕ denotes SiLU, dw and g denote depthwise and grouped convolutions, and $M = 12$ by default. Only the recurrent path is pooled; $\mathbf{L}$ remains on the original bottleneck grid.

For $a \in \mathcal{A} = \{d, h, w\}$, the operator $\mathcal{S}_a$ rearranges $\widetilde{\mathbf{Z}}$ into independent coordinate lines of length $L_a \in \{\widetilde{D}, \widetilde{H}, \widetilde{W}\}$. Their batch dimensions are $B\widetilde{H}\widetilde{W}$, $B\widetilde{D}\widetilde{W}$, and $B\widetilde{D}\widetilde{H}$, respectively. Let $\mathcal{J}_{\rightarrow} = \mathcal{I}$ and $\mathcal{J}_{\leftarrow} = \mathcal{J}$ denote identity and sequence reversal, and let $\mathcal{T}^{-}$ replicate the first token and shift the remaining sequence by one position. For direction $\delta \in \{\rightarrow, \leftarrow\}$, the axis-specific token mixing and shared projections are

$$\begin{aligned} &\mathbf{Q}_a^{\delta} = \mathcal{J}_{\delta}\,\mathrm{LN}\Big(\mathcal{S}_a(\widetilde{\mathbf{Z}})\Big), \quad \mathbf{M}_{a,q}^{\delta} = \boldsymbol{\mu}_{a,q} \odot \mathbf{Q}_a^{\delta} + (\mathbf{1} - \boldsymbol{\mu}_{a,q}) \odot \mathcal{T}^{-}(\mathbf{Q}_a^{\delta}), \quad q \in \{k, v, r\}, \\ &\mathbf{K}_a^{\delta} = \mathbf{M}_{a,k}^{\delta}\mathbf{W}_K, \qquad \mathbf{V}_a^{\delta} = \mathbf{M}_{a,v}^{\delta}\mathbf{W}_V, \qquad \mathbf{R}_a^{\delta} = \sigma\big(\mathbf{M}_{a,r}^{\delta}\mathbf{W}_R\big). \end{aligned} \tag{2}$$

The vectors $\boldsymbol{\mu}_{a,q}$ are axis specific and unconstrained, whereas $\mathbf{W}_K$, $\mathbf{W}_V$, and $\mathbf{W}_R$ are shared across axes and directions. Hence, axis specialization is introduced through traversal geometry, token mixing, and recurrent dynamics rather than duplicated projections.

V-BAR RWKV evaluates two causal recurrences on each coordinate line. For scan position i, recurrent channel c, and $w_{a,\delta,c} = -\exp(\theta_{a,\delta,c}) < 0$, the kernel and bidirectional axis response are

$$\mathrm{WKV}_{a,\delta}(\mathbf{K}, \mathbf{V})_{n,i,c} = \frac{e^{u_{a,\delta,c} + K_{n,i,c}} V_{n,i,c} + \sum_{j<i} e^{K_{n,j,c} + (i-j-1)w_{a,\delta,c}} V_{n,j,c}}{e^{u_{a,\delta,c} + K_{n,i,c}} + \sum_{j<i} e^{K_{n,j,c} + (i-j-1)w_{a,\delta,c}} + \varepsilon}, \tag{3}$$

$$\mathbf{O}_a^{\delta} = \mathcal{J}_{\delta}^{-1}\big\{\big[\mathbf{R}_a^{\delta} \odot \mathrm{WKV}_{a,\delta}(\mathbf{K}_a^{\delta}, \mathbf{V}_a^{\delta})\big]\mathbf{W}_O\big\}, \quad \mathbf{Y}_a = \mathcal{S}_a^{-1}[\boldsymbol{\beta} \odot \mathbf{O}_a^{\rightarrow} + (\mathbf{1} - \boldsymbol{\beta}) \odot \mathbf{O}_a^{\leftarrow}],$$

Where $\varepsilon = 10^{-9}$, $\boldsymbol{\beta} = \sigma(\mathbf{b}_{\mathrm{bi}})$. The forward and backward recurrences have independent decay and current-token parameters $(\theta_{a,\delta}, u_{a,\delta})$ but share token-mixing vectors and linear projections. The fusion vector $\boldsymbol{\beta} \in (0, 1)^{C_r}$ is an input-independent channel calibration shared across axes.

The restored axis responses are combined by an input-conditioned simplex field predicted on a grid of at most 6^3 cells and interpolated to the recurrent grid. A static channel gate then balances the routed coordinate-line response against the local branch:

$$\begin{aligned} &\boldsymbol{\alpha} = \mathrm{Tri}_{\widetilde{D},\widetilde{H},\widetilde{W}}\Big[\mathrm{Softmax}_a\Big(\mathcal{G}_{\mathrm{axis}}\Big(\mathrm{AAP}_{\le 6^3}(\widetilde{\mathbf{Z}})\Big)\Big)\Big], \qquad \mathbf{Y}_{\mathrm{axis}} = \sum_{a \in \mathcal{A}} \alpha_a \odot \mathbf{Y}_a, \\ &\mathbf{X}_1 = \mathbf{X} + \mathrm{DropPath}[\boldsymbol{\gamma}_{\mathrm{r}} \odot \mathcal{P}_{\mathrm{out}}(\mathbf{g}_{\mathrm{lr}} \odot \mathbf{L} + (\mathbf{1} - \mathbf{g}_{\mathrm{lr}}) \odot \mathrm{Tri}_{D,H,W}(\mathbf{Y}_{\mathrm{axis}}))], \quad \mathbf{g}_{\mathrm{lr}} = \sigma(\mathbf{b}_{\mathrm{lr}}), \\ &\mathbf{X}_{\mathrm{R}} = \mathbf{X}_1 + \mathrm{DropPath}[\boldsymbol{\gamma}_{\mathrm{f}} \odot \mathrm{SE}(\mathrm{FFN}(\mathrm{GN}(\mathbf{X}_1)))]. \end{aligned} \tag{4}$$

Here α_a varies spatially. Consequently, V-BAR RWKV learns which axis-aligned propagator to emphasize at each bottleneck region without learning an unconstrained traversal path.

In parallel, the anisotropic pyramid bridge projects $\mathbf{X}$ to $\mathbf{Z}_c = \phi(\mathrm{GN}(\mathcal{P}_c(\mathbf{X})))$ and constructs four fixed receptive-field bases: a local grouped 3^3 response $\mathbf{B}_0$, a dilation-two grouped 3^3 response $\mathbf{B}_1$, the normalized mean of $5 \times 1 \times 1$, $1 \times 5 \times 1$, and $1 \times 1 \times 5$ responses $\mathbf{B}_2$, and a coarse response $\mathbf{B}_3$ computed on at most 4^3 cells and interpolated back. Their spatial selection is

$$\begin{aligned} \eta_s(\mathbf{p}) &= \frac{\exp([\mathcal{P}_{\eta}(\mathbf{Z}_c)]_s(\mathbf{p}))}{\sum_{t=0}^{3} \exp([\mathcal{P}_{\eta}(\mathbf{Z}_c)]_t(\mathbf{p}))}, \quad \mathbf{B}_{\mathrm{route}} = \sum_{s=0}^{3} \eta_s \odot \mathbf{B}_s, \\ \mathbf{X}_{\mathrm{C}} &= \phi\{\mathbf{X} + \boldsymbol{\lambda}_{\mathrm{C}} \odot \mathrm{GN}(\mathcal{P}_{\mathrm{C}}(\mathbf{B}_{\mathrm{route}}))\}. \end{aligned} \tag{5}$$

Thus, the bridge learns a spatial combination of predefined support bases rather than continuous offsets or deformable support geometry.

Finally, V-BAR RWKV routes the two paths by their updates relative to the common input. Defining $\boldsymbol{\Delta}_{\mathrm{C}} = \mathbf{X}_{\mathrm{C}} - \mathbf{X}$ and $\boldsymbol{\Delta}_{\mathrm{R}} = \mathbf{X}_{\mathrm{R}} - \mathbf{X}$, we obtain

$$\begin{aligned} [\omega_{\mathrm{C}}, \omega_{\mathrm{R}}] &= \mathrm{Softmax}_2\{\mathcal{G}_f[\boldsymbol{\Delta}_{\mathrm{C}} \| \boldsymbol{\Delta}_{\mathrm{R}} \| (\boldsymbol{\Delta}_{\mathrm{C}} \odot \boldsymbol{\Delta}_{\mathrm{R}})]\}, \\ \mathrm{Y}(\mathbf{X}) &= \phi\{\mathrm{GN}[\mathbf{X} + \omega_{\mathrm{C}} \odot \boldsymbol{\Delta}_{\mathrm{C}} + \omega_{\mathrm{R}} \odot \boldsymbol{\Delta}_{\mathrm{R}} + \sigma(s_{\times})\,\mathcal{P}_{\times}(\boldsymbol{\Delta}_{\mathrm{C}} \odot \sigma(\boldsymbol{\Delta}_{\mathrm{R}}))]\}. \end{aligned} \tag{6}$$

The router uses the two updates and their Hadamard-product agreement, while the last term introduces an asymmetric cross interaction with $s_{\times} = -3$ at initialization. The WKV operator admits

linear recurrent evaluation along each coordinate line; the short-sequence fallback evaluates the same kernel in parallel with an $L_a \times L_a$ intermediate, whose cost is bounded by M.

Consequently, V-BAR RWKV realizes a spatially non-stationary volumetric propagation operator in which recurrent memory, propagation orientation, anatomical-axis preference, and local–context balance are learned at distinct structural levels without constructing dense voxel-to-voxel attention.

3.3 Disagreement-Regulated Volumetric Expert Routing over Supports

The terminal predictor of a volumetric segmentation network faces a support-selection problem. Evidence concentrated at the target voxel is necessary for retaining weak responses and thin boundaries, whereas evidence distributed along one, two, or all three anatomical axes is required to resolve elongated structures, sheet-like interfaces, and region-level semantics. The reliable support nevertheless varies jointly with semantic class and spatial location; a fixed classifier cannot determine how much neighboring evidence should enter each logit, while a generic mixture of transformations leaves the geometry of its experts implicit. We introduce DiVERS, which parameterizes receptive support by the number of jointly displaced spatial axes, routes class logits over the resulting ordered support basis, and expresses contextual predictions as regulated corrections to a point-supported anchor.

Let $F \in \mathbb{R}^{C\times D\times H\times W}$ denote the final decoder features after an optional pointwise projection, with spatial domain Ω and batch indexing omitted. We partition the 3^3 neighborhood into shells $\mathcal{S}_j = \{q \in \{-1,0,1\}^3 : \|q\|_0 = j\}$ and construct class-k logits through normalized shell accumulation:

$$z_{k,m}(v) = \frac{1}{\sqrt{m+1}} \sum_{j=0}^{m} \left[b_{k,j} + \sum_{c=1}^{C} \sum_{q\in\mathcal{S}_j} W_{k,j,c}(q) F_c(v+q) \right], \qquad m \in \{0,1,2,3\}. \tag{7}$$

The nested point, axial, planar-union, and volumetric supports contain $1, 7, 19, 27$ sites. Their trainable shell responses are computed jointly by one zero-padded masked convolution and reused across experts. Boundary-enhanced and pooled-context transformations complete the $M = 6$ experts:

$$\big(z_{k,4}, z_{k,5}\big) = \big((1+\xi) z_{k,0} - \xi \mathcal{A}_3[z_{k,0}],\ \mathcal{U}_\Omega[\mathcal{A}_\downarrow[z_{k,3}]]\big)\,. \tag{8}$$

Here $\xi \geq 0$ is the boundary strength, $\mathcal{A}_3$ denotes zero-padded 3^3 averaging with unit stride, $\mathcal{A}_\downarrow$ denotes stride-two average pooling with ceiling-mode size rounding and singleton axes unchanged, and $\mathcal{U}_\Omega$ is trilinear interpolation to Ω.

A class-shared spatial field, $s = \mathcal{U}_\Omega[\mathcal{R}_\theta(\mathcal{P}_g F)]$, complements class-specific volumetric evidence. $\mathcal{P}_g$ adaptively average-pools each axis to at most $g = 8$ sites, and $\mathcal{R}_\theta$ comprises two pointwise convolutions separated by group normalization and SiLU. Writing $\langle\cdot\rangle_\Omega$ for spatial averaging, define

$$\begin{aligned} G_{k,m}(v) = \log\max\Big\{\epsilon,\ \langle |z_{k,m}|\rangle_\Omega + \eta \max_{u\in\Omega} |z_{k,m}(u)|\Big\} \\ + \beta \tanh\left(B_{k,m} - \frac{1}{M}\sum_{j=0}^{M-1} B_{k,j} \right) + \delta \tanh\left(s_m(v) - \frac{1}{M}\sum_{j=0}^{M-1} s_j(v) \right), \end{aligned} \tag{9}$$

$$\pi_{k,m}(v) = \frac{\lambda}{M} + (1-\lambda)\frac{\exp(G_{k,m}(v)/\tau)}{\sum_{j=0}^{M-1}\exp(G_{k,j}(v)/\tau)}.$$

The learned $B_{k,m}$ encodes class–expert preferences; β and δ bound the centered score corrections, η weights peak evidence, $\epsilon > 0$ is a numerical stabilizer, $\tau > 0$ is the temperature, and $\lambda \in [0,1)$ sets a uniform routing floor.

Taking $z_{k,0}$ as the anchor, we compute the routed correction and normalize the weighted expert variance by its class-specific spatial mean:

$$\begin{aligned} r_k(v) &= \sum_{m=0}^{M-1} \pi_{k,m}(v)\big[z_{k,m}(v) - z_{k,0}(v)\big], \qquad \mu_k(v) = z_{k,0}(v) + r_k(v), \\ \widehat{D}_k(v) &= \frac{\sum_{m=0}^{M-1} \pi_{k,m}(v)\big[z_{k,m}(v) - \mu_k(v)\big]^2}{\epsilon + \dfrac{1}{|\Omega|}\displaystyle\sum_{u\in\Omega}\sum_{m=0}^{M-1} \pi_{k,m}(u)\big[z_{k,m}(u) - \mu_k(u)\big]^2}. \end{aligned} \tag{10}$$

The output class logit is

$$\begin{aligned}\ell_k(v) &= z_{k,0}(v) + \underbrace{\rho_{\max}\sigma(a_k)\left[1+\tanh(\gamma_k)\tanh\Big(\widehat{D}_k(v)\Big)\right]}_{\alpha_k(v)\in(0,2\rho_{\max})} r_k(v) \\ &= (1-\alpha_k(v))z_{k,0}(v) + \alpha_k(v)\mu_k(v),\end{aligned} \tag{11}$$

where σ is the sigmoid, and the learned scalars a_k and γ_k control residual strength and the sign of disagreement modulation, respectively; γ_k is initialized to zero.

DiVERS converts receptive-field choice from an implicit architectural constant into an ordered, class–space-factorized, and disagreement-regulated prediction geometry in which every broader support enters as a bounded residual around fine-grained point evidence.

4 Experiment

4.1 Datasets and Implementation details

We evaluate VERSA on 12 public volumetric segmentation benchmarks: MSD Lung Tumour (Antonelli et al., 2022), MSD Pancreas Tumour (Antonelli et al., 2022), MSD Hepatic Vessel (Antonelli et al., 2022), MSD Colon Cancer (Antonelli et al., 2022), BreastDCEDL ISPY1 (Fridman et al., 2026), BreastDCEDL ISPY2 (Fridman et al., 2026), OIMHS (Ye et al., 2023), COVID-19 CT Seg (Ma et al., 2020), CrossMoDA2021 (Dorent et al., 2023), Parse2022 (Luo et al., 2023), ISLES2022 (Hernandez Petzsche et al., 2022), and Colorectal-Liver Metastases (Simpson et al., 2024). Additional details are provided in the appendix. All experiments use PyTorch 2.5.1 with Python 3.10 and CUDA 12.1 on a server equipped with eight NVIDIA RTX 4090 GPUs, with each model trained and tested on a single GPU. We use the AdamW optimizer with initial learning rate of 0.0001, a batch size of 1, and DiceCE loss. Each model is trained for 320,000 steps, and the random seed is fixed to 42. During training, volumes are randomly cropped into patches of 96^3 voxels. During inference, sliding-window prediction is performed with the same patch size and an overlap ratio of 0.7.

4.2 Segmentation Performance

We compare VERSA with 19 representative methods spanning convolutional networks, Transformers, state-space/recurrent models, and mixture-of-experts architectures on 12 public volumetric segmentation benchmarks. VERSA achieves state-of-the-art segmentation performance across all 12 datasets (Table 1). Notably, the strongest competing method varies across datasets, whereas VERSA maintains its overlap advantage across heterogeneous target geometries. This consistency demonstrates that its benefits are not confined to a particular anatomy. VERSA also achieves the lowest HD95 on most benchmarks, extending its advantages from foreground overlap to boundary localization. Together, these results establish VERSA as a broadly effective framework for volumetric segmentation, with complementary improvements in regional agreement and boundary accuracy.

Qualitative comparisons further highlight VERSA's ability to suppress spurious predictions while preserving small foreground structures (Figure 3). In the middle example, VERSA avoids prominent extraneous regions produced by several competing methods; in the bottom example, it retains small, spatially separated components alongside the dominant target. These examples illustrate two complementary strengths: excluding irrelevant foreground responses without sacrificing disconnected targets. This balance is consistent with VERSA's design, which combines adaptive contextual propagation with support-selective, point-anchored prediction refinement.

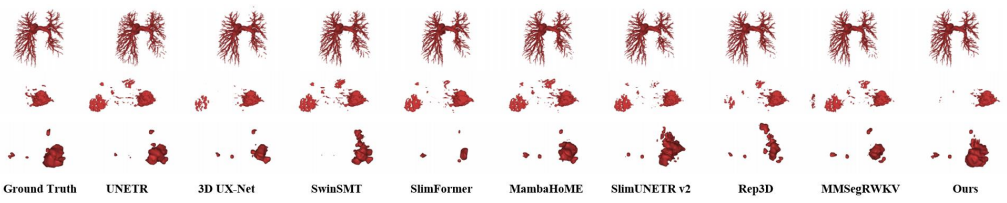

Figure 3: Qualitative comparison of 3D segmentation results.

Table 1: Quantitative comparison with SoTA methods on 12 public 3D segmentation benchmarks.

Method	Venue	MSD Lung Tumour			MSD Pancreas Tumour			MSD Hepatic Vessel			MSD Colon Cancer		
		IoU	Dice	HD95	IoU	Dice	HD95	IoU	Dice	HD95	IoU	Dice	HD95
3D U-Net Çiçek et al. (2016)	MICCAI 2016	54.31	68.82	80.35	54.14	67.99	26.36	49.21	63.65	39.54	42.90	55.69	96.14
UNETR Hatamizadeh et al. (2022b)	WACV 2022	42.34	54.12	161.68	38.80	51.69	25.63	39.88	53.77	54.52	18.91	28.35	127.05
SwinUNETR Hatamizadeh et al. (2022a)	MICCAI 2022	52.13	65.37	80.13	50.39	63.43	44.79	49.15	63.79	47.79	33.84	46.41	130.48
3D UX-Net Lee et al. (2023)	ICLR 2023	51.78	65.35	81.77	51.53	64.21	40.67	49.42	64.13	49.47	30.23	41.40	88.66
nnFormer Zhou et al. (2023)	TIP 2023	32.65	43.90	112.67	47.83	60.99	24.57	42.92	57.71	45.98	19.31	29.95	214.26
MedNeXt Roy et al. (2023)	MICCAI 2023	51.26	65.43	111.59	53.49	67.25	11.02	49.43	63.77	30.81	33.75	46.35	63.54
UNesT Yu et al. (2023)	MedIA 2023	55.65	68.64	58.94	51.40	65.04	14.17	48.65	62.54	**27.45**	39.03	50.76	63.78
SegMamba Xing et al. (2024)	MICCAI 2024	53.25	66.10	80.56	54.85	67.69	17.67	50.89	65.57	32.03	38.99	52.61	95.38
SwinSMT Płotka et al. (2024)	MICCAI 2024	54.29	67.76	78.40	49.29	62.67	31.93	47.14	61.53	46.29	35.03	46.94	97.71
VSmTrans Liu et al. (2024)	MedIA 2024	54.18	67.17	43.34	55.09	68.53	42.56	48.12	62.43	86.58	41.26	54.56	85.15
SuperLightNet Yu et al. (2025)	CVPR 2025	54.01	66.89	48.95	53.26	67.09	15.20	46.52	60.75	42.37	35.64	48.77	84.79
WaveFormer Al Hasan et al. (2025)	MICCAI 2025	47.13	59.51	114.72	50.20	63.05	25.99	48.60	63.26	31.08	33.24	45.95	68.85
SlimFormer3D Hong et al. (2025)	MICCAI 2025	50.49	63.33	105.55	51.60	64.67	12.32	47.67	62.36	31.09	34.07	46.72	96.65
MambaHoME Płotka et al. (2025)	NeurIPS 2025	48.56	62.67	105.95	48.30	61.85	23.55	45.21	59.52	53.35	25.81	37.47	99.45
SlimUNETR V2 Pang et al. (2026)	TMI 2026	56.28	70.35	45.36	52.00	65.86	-	49.14	63.73	36.34	34.97	48.91	151.55
LightUNETR-L Liu et al. (2026)	TPAMI 2026	55.86	68.67	35.72	55.21	68.57	12.91	50.53	65.12	39.91	37.29	49.74	91.58
Rep3D Lee et al. (2026)	ICML 2026	47.42	62.29	95.10	52.96	66.31	16.44	46.19	60.73	103.89	34.23	47.45	131.39
MMSegRWKV Cao et al. (2026)	IoTJ 2026	49.97	63.25	59.52	52.13	65.01	18.82	49.34	63.71	33.03	32.16	42.98	127.28
HCRT Zheng et al. (2026)	PR 2026	54.60	68.67	62.35	51.53	63.94	-	47.71	62.37	30.67	33.95	48.46	131.21
VERSA (Ours)	-	**59.63**	**73.04**	**35.23**	**56.86**	**69.95**	**9.62**	**51.24**	**65.81**	36.59	**43.96**	**57.28**	**53.97**

Method	Venue	BreastDCEDL ISPY1			BreastDCEDL ISPY2			OIMHS			COVID-19 CT Seg		
		IoU	Dice	HD95	IoU	Dice	HD95	IoU	Dice	HD95	IoU	Dice	HD95
3D U-Net Çiçek et al. (2016)	MICCAI 2016	73.99	83.46	9.92	73.28	83.61	10.27	87.69	93.18	2.91	71.50	81.92	51.96
UNETR Hatamizadeh et al. (2022b)	WACV 2022	70.57	80.95	12.43	69.80	81.06	21.23	85.65	91.85	4.36	62.38	74.20	61.26
SwinUNETR Hatamizadeh et al. (2022a)	MICCAI 2022	71.67	81.61	13.49	74.13	84.19	14.10	88.15	93.44	3.34	70.29	80.97	81.29
3D UX-Net Lee et al. (2023)	ICLR 2023	74.36	83.79	10.68	74.26	84.20	11.60	88.37	93.55	2.87	71.44	81.75	57.31
nnFormer Zhou et al. (2023)	TIP 2023	61.52	74.57	21.19	65.81	78.36	15.55	73.03	82.27	19.25	67.52	79.07	50.66
MedNeXt Roy et al. (2023)	MICCAI 2023	69.64	80.26	9.72	73.38	83.66	10.24	87.19	92.78	**2.20**	70.06	80.55	34.54
UNesT Yu et al. (2023)	MedIA 2023	73.59	83.35	8.75	74.29	84.48	10.89	88.69	93.76	2.72	69.81	80.39	45.62
SegMamba Xing et al. (2024)	MICCAI 2024	74.45	83.59	10.54	73.52	83.79	12.80	88.83	93.83	2.66	71.73	81.84	49.95
SwinSMT Płotka et al. (2024)	MICCAI 2024	72.28	82.30	10.48	73.40	83.55	13.34	88.13	93.40	2.92	68.81	79.55	46.39
VSmTrans Liu et al. (2024)	MedIA 2024	70.48	81.03	19.22	74.38	84.40	11.20	88.56	93.70	2.62	64.31	75.53	81.06
SuperLightNet Yu et al. (2025)	CVPR 2025	72.83	82.79	8.66	74.07	84.14	13.41	83.85	90.62	3.50	72.62	82.87	36.32
WaveFormer Al Hasan et al. (2025)	MICCAI 2025	71.68	81.90	14.33	74.82	84.74	12.53	88.55	93.67	2.68	66.63	77.94	62.30
SlimFormer3D Hong et al. (2025)	MICCAI 2025	67.27	78.69	11.57	69.94	81.41	13.77	87.51	93.03	2.92	68.46	78.68	45.94
MambaHoME Płotka et al. (2025)	NeurIPS 2025	73.33	83.09	12.64	74.59	84.51	14.08	88.28	93.51	2.71	71.29	82.26	59.26
SlimUNETR V2 Pang et al. (2026)	TMI 2026	68.27	79.61	8.96	71.73	82.63	11.29	87.32	92.89	2.74	70.15	80.73	38.48
LightUNETR-L Liu et al. (2026)	TPAMI 2026	74.69	83.86	9.51	73.38	83.61	11.80	86.06	92.09	3.36	70.76	81.07	53.85
Rep3D Lee et al. (2026)	ICML 2026	72.56	82.36	15.80	70.39	81.25	23.44	87.90	93.28	3.19	68.73	78.90	40.63
MMSegRWKV Cao et al. (2026)	IoTJ 2026	72.72	82.57	13.20	74.59	84.56	12.56	88.55	93.69	2.76	69.14	79.54	39.42
HCRT Zheng et al. (2026)	PR 2026	70.06	80.98	10.22	70.84	81.94	10.42	81.92	89.77	50.73	51.57	66.54	123.45
VERSA (Ours)	-	**75.81**	**84.56**	**8.61**	**76.60**	**85.73**	**9.64**	**89.79**	**94.40**	2.58	**73.61**	**83.23**	**15.22**

Method	Venue	CrossMoDA2021			Parse2022			ISLES2022			Colorectal-Liver Metastases		
		IoU	Dice	HD95	IoU	Dice	HD95	IoU	Dice	HD95	IoU	Dice	HD95
3D U-Net Çiçek et al. (2016)	MICCAI 2016	67.95	79.76	1.97	72.36	83.77	5.68	61.24	73.92	5.26	75.98	84.44	10.00
UNETR Hatamizadeh et al. (2022b)	WACV 2022	61.29	74.39	10.95	65.21	78.77	11.76	59.75	72.64	9.33	71.25	79.94	19.89
SwinUNETR Hatamizadeh et al. (2022a)	MICCAI 2022	69.18	80.49	6.36	72.48	83.88	5.54	61.99	74.62	4.77	77.14	85.19	10.18
3D UX-Net Lee et al. (2023)	ICLR 2023	66.79	78.76	2.60	72.34	83.78	15.36	62.10	74.72	6.51	77.75	85.91	12.74
nnFormer Zhou et al. (2023)	TIP 2023	20.00	26.04	-	64.10	77.92	10.93	55.68	68.33	-	69.37	78.70	24.41
MedNeXt Roy et al. (2023)	MICCAI 2023	67.25	78.93	1.84	73.20	84.36	5.00	60.77	73.11	4.35	76.07	84.54	15.79
UNesT Yu et al. (2023)	MedIA 2023	68.61	80.15	1.92	74.03	84.91	4.19	59.90	72.99	9.79	77.62	85.58	9.43
SegMamba Xing et al. (2024)	MICCAI 2024	67.65	79.29	10.46	74.12	84.93	4.29	63.28	75.64	4.14	77.36	85.24	10.05
SwinSMT Płotka et al. (2024)	MICCAI 2024	69.28	80.80	15.35	71.59	83.28	6.82	60.20	72.84	6.56	76.72	84.87	9.02
VSmTrans Liu et al. (2024)	MedIA 2024	67.32	79.20	14.83	73.54	84.60	4.26	56.89	69.45	6.90	77.80	85.79	13.73
SuperLightNet Yu et al. (2025)	CVPR 2025	64.09	76.38	23.96	70.34	82.38	7.63	64.44	76.43	**3.59**	76.17	84.47	9.80
WaveFormer Al Hasan et al. (2025)	MICCAI 2025	69.02	80.54	8.98	73.15	84.33	4.70	60.73	73.73	5.12	78.01	86.23	8.45
SlimFormer3D Hong et al. (2025)	MICCAI 2025	68.62	80.25	4.29	71.89	83.50	6.45	55.97	69.04	6.95	76.63	84.79	13.85
MambaHoME Płotka et al. (2025)	NeurIPS 2025	68.30	79.76	3.89	71.85	83.45	7.83	63.27	75.61	5.21	74.55	83.20	15.22
SlimUNETR V2 Pang et al. (2026)	TMI 2026	66.35	78.48	14.31	71.10	82.92	5.39	62.39	74.52	7.12	77.24	85.55	11.33
LightUNETR-L Liu et al. (2026)	TPAMI 2026	69.90	81.00	1.94	71.80	83.38	5.05	62.34	74.71	4.22	77.27	85.61	12.21
Rep3D Lee et al. (2026)	ICML 2026	62.79	75.63	2.68	73.82	84.73	4.29	57.47	70.57	7.43	76.96	85.35	17.37
MMSegRWKV Cao et al. (2026)	IoTJ 2026	67.18	79.21	2.63	72.32	83.74	6.40	60.67	73.37	5.50	77.40	85.65	9.76
HCRT Zheng et al. (2026)	PR 2026	68.18	79.65	1.88	72.43	83.83	11.34	61.94	73.91	3.80	75.31	83.81	22.30
VERSA (Ours)	-	**70.50**	**81.63**	**1.57**	**75.22**	**85.64**	**3.85**	**64.51**	**76.80**	4.17	**78.81**	**86.68**	**6.65**

4.3 Model Complexity and Inference Efficiency

Table 2 shows that VERSA uses 3.67M parameters and 107.67 GFLOPs, achieving 37.39 patches/s on a single RTX 4090 with 96^3-voxel inputs. Together with strong segmentation performance, these results highlight a favorable balance between accuracy, compactness, and inference efficiency.

Table 2: Comparison of model parameters, FLOPs, and inference throughput.

Method	Venue	Params↓	FLOPs↓	Throughput↑	Method	Venue	Params↓	FLOPs↓	Throughput↑
3D U-Net	MICCAI 2016	5.75 M	135.88 G	75.25 patches/s	UNETR	WACV 2022	92.62 M	82.58 G	58.98 patches/s
SwinUNETR	MICCAI 2022	61.99 M	329.46 G	20.37 patches/s	3D UX-Net	ICLR 2023	53.00 M	631.76 G	18.83 patches/s
nnFormer	TIP 2023	149.10 M	224.36 G	52.60 patches/s	MedNeXt	MICCAI 2023	11.64 M	174.10 G	9.78 patches/s
UNesT	MedIA 2023	84.88 M	258.11 G	39.86 patches/s	SegMamba	MICCAI 2024	64.24 M	655.62 G	19.03 patches/s
SwinSMT	MICCAI 2024	15.51 M	83.71 G	28.95 patches/s	VSmTrans	MedIA 2024	49.86 M	358.89 G	15.93 patches/s
SuperLightNet	CVPR 2025	2.75 M	19.52 G	38.36 patches/s	WaveFormer	MICCAI 2025	16.86 M	371.38 G	18.64 patches/s
SlimFormer3D	MICCAI 2025	17.02 M	78.35 G	55.71 patches/s	MambaHoME	NeurIPS 2025	170.1 M	654.10 G	12.29 patches/s
SlimUNETR V2	TMI 2026	22.03 M	54.75 G	116.92 patches/s	LightUNETR-L	TPAMI 2026	3.93 M	54.33 G	29.79 patches/s
Rep3D	ICML 2026	65.83 M	766.60 G	2.37 patches/s	MMSegRWKV	IoTJ 2026	37.28 M	282.88 G	24.27 patches/s
HCRT	PR 2026	6.26 M	74.64 G	57.91 patches/s	VERSA (Ours)	–	3.67 M	107.67 G	37.39 patches/s

4.4 Ablation Study

Cross-Architecture Effectiveness of V-BAR RWKV. To assess its utility beyond VERSA, we augment eight segmentation architectures with V-BAR RWKV on BreastDCEDL ISPY2 (Table 3). All eight achieve higher IoU and Dice and lower HD95. The accompanying HD95 reductions indi-

cate that improved overlap does not come at the expense of boundary localization. Together, these results support V-BAR RWKV as a reusable volumetric context module.

Table 3: Cross-architecture evaluation of V-BAR RWKV.

Variant	IoU	Dice	HD95	Variant	IoU	Dice	HD95
UNETR Hatamizadeh et al. (2022b)	69.80	81.06	21.23	SwinUNETR Hatamizadeh et al. (2022a)	74.13	84.19	14.10
+ V-BAR RWKV	**72.36**	**82.93**	**14.99**	+ V-BAR RWKV	**75.09**	**84.84**	**12.54**
SuperLightNet Yu et al. (2025)	74.07	84.14	13.41	SlimFormer3D Hong et al. (2025)	69.94	81.41	13.77
+ V-BAR RWKV	**75.28**	**84.93**	**10.35**	+ V-BAR RWKV	**70.17**	**81.62**	**12.87**
MambaHoME Plotka et al. (2025)	74.59	84.51	14.08	SlimUNETR V2 Pang et al. (2026)	71.73	82.63	11.29
+ V-BAR RWKV	**74.91**	**84.72**	**12.61**	+ V-BAR RWKV	**72.36**	**83.09**	**11.18**
MMSegRWKV Cao et al. (2026)	74.59	84.56	12.56	HCRT Zheng et al. (2026)	70.84	81.94	10.42
+ V-BAR RWKV	**75.16**	**84.99**	**10.43**	+ V-BAR RWKV	**71.45**	**82.44**	**10.19**

Propagation Geometry and Axis Routing. We ablate V-BAR RWKV on MSD Pancreas Tumour (Table 4). Single-axis bidirectional variants underperform the full model, highlighting complementary context across axes. Learned axis weighting outperforms uniform aggregation over identical candidate axes. Equal-weight bidirectional fusion underperforms forward-only propagation, whereas learned channel-wise fusion outperforms both, highlighting the value of directional calibration. Sharing recurrent time parameters across axes yields the largest HD95 increase with a modest Dice decrease, underscoring the role of axis-specific dynamics in boundary localization.

Table 4: Ablation of propagation geometry and axis routing in V-BAR RWKV.

Variant	IoU	Dice	HD95	Variant	IoU	Dice	HD95
Depth-only bidirectional propagation	54.58	67.70	10.31	Height-only bidirectional propagation	55.69	68.64	10.25
Width-only bidirectional propagation	55.11	68.10	10.09	Forward-only tri-axis propagation	56.24	69.74	10.36
Equal-weight bidirectional fusion	55.01	68.01	13.27	Uniform axis aggregation	54.52	68.13	11.18
Axis-shared recurrent time parameters	55.76	69.25	15.35	V-BAR RWKV (Full)	**56.86**	**69.95**	**9.62**

Receptive-Support Routing and Cross-Path Interaction. We ablate support branches, learned support weighting, and cross-path interaction in V-BAR RWKV on Colorectal-Liver Metastases (Table 5). All tested support removals degrade performance. Removing the axis-elongated branch causes the largest overlap loss with tri-axis propagation retained, highlighting the complementary role of axial aggregation. Coarse-context removal more than doubles HD95 with a smaller Dice decrease. Learned weighting outperforms uniform averaging over identical support responses, demonstrating the value of adaptive support selection. The cross-path interaction projection improves both overlap and boundary accuracy beyond update-conditioned fusion alone.

Table 5: Ablation of receptive-support routing and cross-path interaction in V-BAR RWKV.

Variant	IoU	Dice	HD95	Variant	IoU	Dice	HD95
w/o dilated context branch	77.71	85.82	7.61	w/o axis-elongated context branch	77.25	85.25	8.74
w/o coarse-context branch	77.62	85.77	13.66	Equal-weight support averaging	77.54	85.68	10.24
w/o cross-path interaction projection	77.35	85.29	9.10	V-BAR RWKV (Full)	**78.81**	**86.68**	**6.65**

Comparison of V-BAR-RWKV with Alternative Modules We compare V-BAR RWKV with alternatives spanning feature recalibration, convolution, attention, and state-space/recurrent modeling on ISLES2022 (Table 6). V-BAR RWKV achieves the best results on all three metrics. The improvements over both RWKV and Vision-RWKV further favor the complete volumetric adaptation over an RWKV-family replacement alone. These results support the combined design of spatially routed bidirectional axis propagation and multi-support context.

Table 6: Comparison of V-BAR RWKV with alternative modules in VERSA.

Replacement module	IoU	Dice	HD95	Replacement module	IoU	Dice	HD95
SE Hu et al. (2018)	64.09	76.29	4.28	CBAM Woo et al. (2018)	63.21	75.16	6.00
ResConv3D He et al. (2016)	62.46	75.00	6.61	ConvNeXt3D Lee et al. (2023)	64.22	76.11	4.70
Transformer Vaswani et al. (2017)	63.87	75.82	4.98	Swin Transformer Liu et al. (2021)	64.34	76.33	4.27
Efficient Paired Attention Shaker et al. (2024)	64.33	76.31	4.44	Mamba layer Ma et al. (2024)	64.17	76.17	4.19
Vision Mamba Zhu et al. (2024)	63.86	75.87	4.47	RWKV Peng et al. (2023)	63.12	74.96	5.66
Vision-RWKV Duan et al. (2025)	63.81	76.00	4.73	V-BAR RWKV (Ours)	**64.51**	**76.80**	**4.17**

Cross-Architecture Generality of DiVERS. We integrate DiVERS into eight segmentation architectures to evaluate the generality of adaptive receptive-support selection (Table 7). DiVERS improves both IoU and Dice for every model. Gains on MambaHoME and MMSegRWKV further suggest that prediction-stage support selection complements existing expert-based and recurrent context modeling. Together, these results support DiVERS as a reusable prediction head whose explicit, adaptive selection of receptive supports.

Table 7: Cross-architecture evaluation of DiVERS on MSD Lung Tumour.

Variant	IoU	Dice	HD95	Variant	IoU	Dice	HD95
UNETR Hatamizadeh et al. (2022b)	42.34	54.12	161.68	SwinUNETR Hatamizadeh et al. (2022a)	52.13	65.37	80.13
+ DiVERS	**44.88**	**55.68**	**101.43**	+ DiVERS	**53.94**	**67.75**	**20.63**
SuperLightNet Yu et al. (2025)	54.01	66.89	48.95	SlimFormer3D Hong et al. (2025)	50.49	63.33	105.55
+ DiVERS	**55.31**	**68.34**	**46.33**	+ DiVERS	**53.68**	**65.56**	**47.90**
MambaHoME Plotka et al. (2025)	48.56	62.67	105.95	SlimUNETR V2 Pang et al. (2026)	56.28	70.35	**45.36**
+ DiVERS	**51.85**	**63.78**	**45.81**	+ DiVERS	**56.93**	**71.14**	60.10
MMSegRWKV Cao et al. (2026)	49.97	63.25	59.52	HCRT Zheng et al. (2026)	54.60	68.67	62.35
+ DiVERS	**51.33**	**63.74**	**51.84**	+ DiVERS	**55.84**	**69.24**	**44.20**

Receptive-Support Design and Auxiliary Experts. We ablate receptive supports and auxiliary experts in DiVERS on OIMHS (Table 8). The volumetric-only head lowers HD95 but slightly reduces overlap relative to the point-only head. Combining point and volumetric experts improves all metrics over either head; adding axial and planar-union experts yields further gains without enlarging the maximum support. Random-partition experts achieve higher overlap but worse HD95 than geometric experts with independent kernels; both geometry and parameter sharing differ. Removing pooled context reduces overlap more than removing boundary enhancement. Full DiVERS achieves the highest results, supporting diverse receptive supports and both auxiliary experts.

Table 8: Ablation of receptive-support construction and auxiliary experts in DiVERS

Variant	IoU	Dice	HD95	Variant	IoU	Dice	HD95
Point-only head	88.77	93.82	2.82	Volumetric-only head	88.63	93.73	2.76
Point-volume experts	88.88	93.87	2.60	Geometric experts	89.00	93.95	2.58
Random-partition experts	89.10	94.01	2.71	w/o context expert	88.77	93.81	2.73
w/o boundary expert	89.42	94.22	2.70	DiVERS (Full)	**89.79**	**94.40**	**2.58**

Factorized Routing and Disagreement-Regulated Refinement. We ablate DiVERS routing and refinement on CrossMoDA2021 with identical support experts (Table 9). Combining class-conditioned global and class-shared spatial scores outperforms uniform and single-factor routing. Point-anchored scaling alone improves upon direct expert mixing; normalized variance provides further gains over unmodulated, entropy-based, and unnormalized-variance refinement. Spatial shuffling preserves the variance distribution but lowers overlap below the unmodulated baseline while keeping HD95 close to full DiVERS, highlighting the importance of spatial alignment for overlap accuracy. Spatial-only routing and a class-shared disagreement gain worsen HD95 with modest Dice changes, underscoring the role of class specificity in boundary localization at both stages.

Table 9: Ablation of factorized routing and disagreement-regulated refinement in DiVERS.

Variant	IoU	Dice	HD95	Variant	IoU	Dice	HD95
Uniform expert routing	69.84	81.04	1.68	Class-conditioned global routing only	69.85	81.17	1.75
Class-shared spatial routing only	69.82	81.19	4.90	Anchored scaling, no disagreement	70.19	81.37	1.62
Direct expert mixture	69.62	80.96	2.08	Mean-normalized routing entropy	69.95	81.07	1.60
Unnormalized weighted expert variance	70.27	81.34	1.71	Spatially shuffled normalized variance	69.80	80.96	1.59
Class-shared disagreement gain	70.16	81.23	4.84	DiVERS (Full)	**70.50**	**81.63**	**1.57**

Comparison with Alternative Segmentation Heads. We compare DiVERS with alternative segmentation heads on BreastDCEDL ISPY2, keeping the rest of VERSA and the training protocol unchanged (Table 10). DiVERS achieves the best results across all metrics. These results demonstrate joint improvements in regional overlap and boundary localization with support-adaptive prediction.

Table 10: Comparison of DiVERS with alternative segmentation heads in VERSA.

Replacement module	IoU	Dice	HD95	Replacement module	IoU	Dice	HD95
ASEM Huang et al. (2024)	75.34	84.88	9.87	DLK Yang et al. (2026)	75.70	85.28	11.54
DMRFB Rahman et al. (2025)	75.67	85.32	10.95	EMCAD-MSCAM Rahman et al. (2024)	73.51	83.75	13.48
PaR Huang et al. (2025)	75.04	84.65	10.93	DiVERS (Ours)	**76.60**	**85.73**	**9.64**

5 Conclusion

We introduced VERSA, a volumetric medical image segmentation framework unifying structured adaptive routing over propagation axes in latent space and receptive supports in prediction space. V-BAR RWKV learns spatially varying mixtures of bidirectional axis-wise responses, while DiVERS combines class- and location-conditioned support routing with disagreement-regulated, point-anchored residual refinement. With 3.67M parameters, VERSA achieves the highest result across 12 public benchmarks. Controlled ablations show that learned routing outperforms uniform aggregation over identical candidate axes and supports, demonstrating the value of adaptive context selection beyond receptive-field expansion.

AI USE STATEMENT

AI tools were used only for language polishing and grammar checking to improve the clarity and readability of the manuscript.

REPRODUCIBILITY STATEMENT

Upon acceptance, we will publicly release the complete source code on GitHub to facilitate reproducibility and further research. The released repository will include the full training and inference pipelines, model implementations, data loading and preprocessing modules, configuration files, evaluation scripts, and all other components required to reproduce the experiments presented in this paper.

REFERENCES

Md Mahfuz Al Hasan, Mahdi Zaman, Abdul Jawad, Alberto Santamaria-Pang, Ho Hin Lee, Ivan Tarapov, Kyle B See, Md Shah Imran, Antika Roy, Yaser Pourmohammadi Fallah, et al. Waveformer: A 3d transformer with wavelet-driven feature representation for efficient medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 684–694. Springer, 2025.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature Communications*, 13(1):4128, 2022.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical image analysis*, 84:102680, 2023.

Hongyu Cai, Yifan Wang, Liu Wang, Jian Zhao, and Zhejun Kuang. U-rwkv: Accurate and efficient volumetric medical image segmentation via rwkv. *IEEE Transactions on Image Processing*, 35: 1025–1038, 2026.

Yitong Cao, Yuanqing Xia, Ke Tian, Di-Hua Zhai, and Danyang Liu. Mmsegrwkv: Enhancing multimodal mri segmentation for internet of medical things-enabled healthcare with rwkv-inspired architectures. *IEEE Internet of Things Journal*, 13(9):19191–19204, 2026.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 424–432. Springer, 2016.

Reuben Dorent, Aaron Kujawa, Marina Ivory, Spyridon Bakas, Nicola Rieke, Samuel Joutard, Ben Glocker, Jorge Cardoso, Marc Modat, Kayhan Batmanghelich, et al. Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation. *Medical Image Analysis*, 83:102628, 2023.

Yuchen Duan, Weiyun Wang, Zhe Chen, Xizhou Zhu, Lewei Lu, Tong Lu, Yu Qiao, Hongsheng Li, Jifeng Dai, and Wenhai Wang. Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures. In *International Conference on Learning Representations*, volume 2025, pp. 83166–83182, 2025.

Naomi Fridman, Bubby Solway, Tomer Fridman, Itamar Barnea, and Anat Goldstein. Breastdcedl: A standardized deep learning-ready breast dce-mri dataset of 2,070 patients. *Scientific Data*, 13 (1):264, 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 *International MICCAI Brainlesion Workshop*, pp. 272–284. Springer, 2022a.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1748–1758. IEEE, 2022b.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Moritz R Hernandez Petzsche, Ezequiel De La Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes, Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, et al. Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific Data*, 9(1):762, 2022.

Yang Hong, Lei Zhang, Xujiong Ye, and Jianqing Mo. Slimformer-3d: A layer-adaptive lightweight transformer for efficient 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 537–546. Springer, 2025.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

Xingru Huang, Jian Huang, Kai Zhao, Tianyun Zhang, Zhi Li, Changpeng Yue, Wenhao Chen, Ruihao Wang, Xuanbin Chen, Qianni Zhang, et al. Sasan: Spectrum-axial spatial approach networks for medical image segmentation. *IEEE Transactions on Medical Imaging*, 43(8):3044–3056, 2024.

Xingru Huang, Jian Huang, Yihao Guo, Tianyun Zhang, Zhao Huang, Yaqi Wang, Ruipu Tang, Guangliang Cheng, Shaowei Jiang, Zhiwen Zheng, et al. Volumetric axial disentanglement enabling advancing in medical image segmentation. In *IJCAI*, pp. 1197–1205, 2025.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

Ho Hin Lee, Shunxing Bao, Yuankai Huo, and Bennett A Landman. 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. In *The Eleventh International Conference on Learning Representations*, 2023.

Ho Hin Lee, Quan Liu, Shunxing Bao, Yuankai Huo, and Bennett A Landman. Spatially-adaptive gradient re-parameterization for 3d large kernel optimization. *arXiv preprint arXiv:2505.19603*, 2026.

Tiange Liu, Qingze Bai, Drew A Torigian, Yubing Tong, and Jayaram K Udupa. Vsmtrans: A hybrid paradigm integrating self-attention and convolution for 3d medical image segmentation. *Medical Image Analysis*, 98:103295, 2024.

Xinyu Liu, Zhen Chen, Wuyang Li, Chenxin Li, and Yixuan Yuan. Harnessing lightweight transformer with contextual synergic enhancement for efficient 3d medical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3852–3867, 2026.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002. IEEE, 2021.

Gongning Luo, Kuanquan Wang, Jun Liu, Shuo Li, Xinjie Liang, Xiangyu Li, Shaowei Gan, Wei Wang, Suyu Dong, Wenyi Wang, et al. Efficient automatic segmentation for multi-level pulmonary arteries: The parse challenge. *arXiv preprint arXiv:2304.03708*, 2023.

Jun Ma, Cheng Ge, Yixin Wang, Xingle An, Jiantao Gao, Ziqi Yu, Minqing Zhang, Xin Liu, Xueyuan Deng, Shucheng Cao, et al. Covid-19 ct lung and infection segmentation dataset. Zenodo, 2020. Dataset, version 1.0.

Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571. IEEE, 2016.

Yan Pang, Jiaming Liang, Junming Yan, Ying Hu, Hao Chen, and Qiong Wang. Slim unetrv2: 3d image segmentation for resource-limited medical portable devices. *IEEE Transactions on Medical Imaging*, 45(2):542–553, 2026.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, et al. Rwkv: Reinventing rnns for the transformer era. In *Findings of the association for computational linguistics: EMNLP 2023*, pp. 14048–14077, 2023.

Szymon Płotka, Maciej Chrabaszcz, and Przemyslaw Biecek. Swin smt: Global sequential modeling for enhancing 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 689–698. Springer, 2024.

Szymon Plotka, Gizem Mert, Maciej Chrabaszcz, Ewa Szczurek, and Arkadiusz Sitek. Mamba goes home: Hierarchical soft mixture-of-experts for 3d medical image segmentation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, volume 38, pp. 97871–97909, 2025.

Md Mostafijur Rahman, Mustafa Munir, and Radu Marculescu. Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11769–11779. IEEE, 2024.

Md Mostafijur Rahman, Mustafa Munir, and Radu Marculescu. Efficientmednext: Multi-receptive dilated convolutions for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 196–206. Springer, 2025.

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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 405–415. Springer, 2023.

Abdelrahman 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.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Amber L Simpson, Jacob Peoples, John M Creasy, Gabor Fichtinger, Natalie Gangai, Krishna N Keshavamurthy, Andras Lasso, Jinru Shia, Michael I D'Angelica, and Richard KG Do. Preoperative ct and survival data for patients undergoing resection of colorectal liver metastases. *Scientific Data*, 11(1):172, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Guoan Wang, Jin Ye, Junlong Cheng, Tianbin Li, Zhaolin Chen, Jianfei Cai, Junjun He, and Bohan Zhuang. Sam-med3d-moe: Towards a non-forgetting segment anything model via mixture of experts for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 552–561. Springer, 2024.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *European Conference on Computer Vision*, pp. 3–19. Springer, 2018.

Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 578–588. Springer, 2024.

Jin Yang, Peijie Qiu, Yichi Zhang, Daniel S Marcus, and Aristeidis Sotiras. D-net: Dynamic large kernel with dynamic feature fusion for volumetric medical image segmentation. *Biomedical Signal Processing and Control*, 113:108837, 2026.

Xin Ye, Shucheng He, Xiaxing Zhong, Jiafeng Yu, Shangchao Yang, Yingjiao Shen, Yiqi Chen, Yaqi Wang, Xingru Huang, and Lijun Shen. Oimhs: An optical coherence tomography image dataset based on macular hole manual segmentation. *Scientific Data*, 10(1):769, 2023.

Feng Yu, Jiacheng Cao, Li Liu, and Minghua Jiang. Superlightnet: Lightweight parameter aggregation network for multimodal brain tumor segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5197–5206, 2025.

Xin Yu, Qi Yang, Yinchi Zhou, Leon Y Cai, Riqiang Gao, Ho Hin Lee, Thomas Li, Shunxing Bao, Zhoubing Xu, Thomas A Lasko, et al. Unest: local spatial representation learning with hierarchical transformer for efficient medical segmentation. *Medical Image Analysis*, 90:102939, 2023.

Lei Zheng, Yuzhong Zhang, Jiadong Zhang, Tao Zhou, Kun Sun, Lei Zhou, and Dinggang Shen. Hcrt: Hybrid network with correlation-aware region transformer for breast tumor segmentation in dce-mri. *Pattern Recognition*, 174:112934, 2026.

Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE Transactions on Image Processing*, 32:4036–4045, 2023.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

A Appendix

A.1 Additional Dataset and Implementation Details

We evaluate on 12 public datasets spanning computed tomography (CT), magnetic resonance imaging (MRI), and optical coherence tomography (OCT). Table A.1 summarizes their released segmentation annotations, together with the training/test partitions and intensity clipping intervals used in this study.

Table A.1: Overview of the datasets and segmentation targets used in our experiments.

Dataset	Modality	Anatomical target	Split		Intensity clipping
			Train	Test	
MSD Lung Tumour (Antonelli et al., 2022)	CT	Lung tumour	50	13	$[-1000, 1000]$
MSD Pancreas Tumour (Antonelli et al., 2022)	CT	Pancreas; Pancreatic tumour	243	38	$[-180, 500]$
MSD Hepatic Vessel (Antonelli et al., 2022)	CT	Hepatic vessels; Liver tumour	265	38	$[-50, 260]$
MSD Colon Cancer (Antonelli et al., 2022)	CT	Colon tumour	106	20	$[-200, 230]$
BreastDCEDL ISPY1 (Fridman et al., 2026)	DCE-MRI	Breast tumour	129	32	$[0, 800]$
BreastDCEDL ISPY2 (Fridman et al., 2026)	DCE-MRI	Breast tumour	786	196	$[0, 2000]$
OIMHS (Ye et al., 2023)	OCT	Retina; Macular hole; Macular edema; Choroid	100	25	$[0, 300]$
COVID-19 CT Seg (Ma et al., 2020)	CT	Left lung; Right lung; Infection	16	4	$[-1250, 250]$
CrossMoDA2021 (Dorent et al., 2023)	MRI	Vestibular schwannoma; Cochlea	84	21	$[0, 1800]$
Parse2022 (Luo et al., 2023)	CTPA	Pulmonary arteries	80	20	$[-800, 500]$
ISLES2022 (Hernandez Petzsche et al., 2022)	MRI	Ischemic stroke lesion	200	50	$[0, 2200]$
Colorectal–Liver Metastases (Simpson et al., 2024)	CT	Liver; Liver metastases	157	40	$[-100, 300]$

All datasets use the same online pipeline except for intensity clipping. Volumes are arranged in channel-first format and reoriented to RAS. For dataset d, intensities are clipped to $[a_d, b_d]$ and linearly mapped to $[0, 1]$:

$$\widetilde{x} = \frac{\min(\max(x, a_d), b_d) - a_d}{b_d - a_d}. \tag{A.1}$$

An image-derived foreground crop is then applied using the bounding box of positive normalized intensities. No voxel-spacing resampling is applied at this online preprocessing stage.

Training patches of $96 \times 96 \times 96$ voxels are sampled with equal foreground/background center-selection probabilities. Foreground comprises all nonzero labels, while background candidates are restricted to positive-intensity image regions. Each patch undergoes 3D elastic deformation ($p = 0.5$, Gaussian smoothing parameter $\sigma \in [5, 10]$, magnitude parameter $m \in [50, 150]$, zero padding), flipping along the third spatial axis ($p = 0.5$), and random affine transformation ($p = 1$). The affine transform uses rotation only about the third spatial axis, with $\theta \sim \mathcal{U}(-\pi/20, \pi/20)$, and independent per-axis scale factors $s_i \sim \mathcal{U}(0.85, 1.15)$. Spatial augmentations are applied jointly to images and labels, using trilinear and nearest-neighbour interpolation, respectively.

A.2 Multi-Seed Stability and Paired Comparisons

We evaluate sensitivity to training randomness on OIMHS and MSD Pancreas Tumour by training VERSA and eight baselines with seeds 41–45, using the same data partitions and training protocol. Tables A.2 and A.3 report individual runs and mean scores across seeds.

For each baseline and metric, let d_s denote the VERSA-minus-baseline score difference for seed s, measured in percentage points. With $S = 5$, we report the mean gain $\bar{d}$ and its pointwise 95% confidence interval,

$$\bar{d} = \frac{1}{S} \sum_{s=41}^{45} d_s, \qquad \mathrm{CI}_{95\%} = \bar{d} \pm t_{0.975, S-1} \frac{s_d}{\sqrt{S}}, \tag{A.2}$$

where s_d is the sample standard deviation of the seed-wise differences. Two-sided paired t-tests ($df = 4$) assess zero mean gain, with Holm correction applied jointly to all $2 \times 8 \times 2 = 32$ comparisons across datasets, baselines, and metrics. This analysis quantifies variation across training runs on the fixed data partitions.

VERSA achieves the highest IoU and Dice in every seed, with the smallest observed across-seed standard deviations among the evaluated methods. Its mean IoU/Dice is $89.75/94.38$ on OIMHS and $56.90/69.88$ on MSD Pancreas Tumour. Relative to the strongest baseline by five-seed mean

on each dataset, VERSA improves IoU/Dice by 1.36/0.82 percentage points over MMSegRWKV on OIMHS and by 3.70/3.06 percentage points over MedNeXt on MSD Pancreas Tumour. All 32 comparisons remain significant after Holm correction ($\max p_{\mathrm{H}} = 5.29 \times 10^{-3}$), supporting consistent overlap gains across the evaluated seeds.

Table A.2: Multi-seed performance and paired comparisons on OIMHS.

Method	Seed=41		Seed=42		Seed=43		Seed=44		Seed=45		Average	
	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
UNETR Hatamizadeh et al. (2022b)	85.18	91.54	85.65	91.85	85.23	91.57	85.32	91.61	85.53	91.76	85.38	91.67
SwinUNETR Hatamizadeh et al. (2022a)	87.91	93.27	88.15	93.44	87.80	93.21	88.13	93.43	88.41	93.59	88.08	93.39
MedNeXt Roy et al. (2023)	87.08	92.75	87.19	92.78	87.53	93.04	87.10	92.76	87.42	92.97	87.26	92.86
SuperLightNet Yu et al. (2025)	83.48	90.41	83.85	90.62	85.01	91.44	85.16	91.54	84.91	91.36	84.48	91.07
SlimFormer3D Hong et al. (2025)	87.14	92.80	87.51	93.03	87.38	92.95	86.94	92.66	86.63	92.49	87.12	92.79
MambaHoME Plotka et al. (2025)	87.36	92.93	88.28	93.51	87.13	92.79	88.49	93.63	86.81	92.59	87.61	93.09
SlimUNETR V2 Pang et al. (2026)	86.87	92.64	87.32	92.89	86.83	92.60	87.15	92.82	86.98	92.70	87.03	92.73
MMSegRWKV Cao et al. (2026)	88.30	93.51	88.55	93.69	87.97	93.24	88.45	93.63	88.68	93.74	88.39	93.56
VERSA (Ours)	**89.70**	**94.35**	**89.79**	**94.40**	**89.80**	**94.41**	**89.75**	**94.38**	**89.70**	**94.37**	**89.75**	**94.38**

Baseline	Mean gain Δ (pp)		95% CI (pp)		p_{raw}		p_{H}	
	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
UNETR	+4.37	+2.72	[4.12, 4.61]	[2.56, 2.88]	1.04×10^{-6}	1.19×10^{-6}	3.22×10^{-5}	3.58×10^{-5}
SwinUNETR	+1.67	+0.99	[1.34, 1.99]	[0.80, 1.19]	1.38×10^{-4}	1.45×10^{-4}	1.59×10^{-3}	1.59×10^{-3}
MedNeXt	+2.48	+1.52	[2.25, 2.72]	[1.37, 1.68]	8.44×10^{-6}	1.11×10^{-5}	2.02×10^{-4}	2.37×10^{-4}
SuperLightNet	+5.27	+3.31	[4.33, 6.20]	[2.67, 3.94]	9.82×10^{-5}	1.32×10^{-4}	1.28×10^{-3}	1.59×10^{-3}
SlimFormer3D	+2.63	+1.60	[2.24, 3.02]	[1.34, 1.85]	4.87×10^{-5}	6.36×10^{-5}	8.29×10^{-4}	9.54×10^{-4}
MambaHoME	+2.13	+1.29	[1.24, 3.02]	[0.73, 1.85]	2.64×10^{-3}	3.08×10^{-3}	5.29×10^{-3}	5.29×10^{-3}
SlimUNETR V2	+2.72	+1.65	[2.48, 2.96]	[1.50, 1.80]	6.26×10^{-6}	6.56×10^{-6}	1.63×10^{-4}	1.64×10^{-4}
MMSegRWKV	+1.36	+0.82	[0.99, 1.73]	[0.56, 1.08]	5.25×10^{-4}	9.44×10^{-4}	3.15×10^{-3}	3.25×10^{-3}

Table A.3: Multi-seed performance and paired comparisons on MSD Pancreas Tumour.

Method	Seed=41		Seed=42		Seed=43		Seed=44		Seed=45		Average	
	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
UNETR Hatamizadeh et al. (2022b)	40.08	53.65	38.80	51.69	38.33	51.17	39.24	52.39	39.48	52.75	39.19	52.33
SwinUNETR Hatamizadeh et al. (2022a)	51.39	64.35	50.39	63.43	51.20	64.27	51.15	64.08	51.31	64.30	51.09	64.09
MedNeXt Roy et al. (2023)	54.03	67.44	53.49	67.25	53.47	67.22	52.09	65.39	52.93	66.84	53.20	66.83
SuperLightNet Yu et al. (2025)	53.24	67.08	53.26	67.09	53.13	66.80	51.80	65.31	51.83	65.34	52.65	66.32
SlimFormer3D Hong et al. (2025)	51.59	64.58	51.60	64.67	51.91	64.71	50.09	63.18	51.61	64.59	51.36	64.35
MambaHoME Plotka et al. (2025)	47.54	60.37	48.30	61.85	48.61	61.87	47.80	61.14	48.36	61.96	48.12	61.44
SlimUNETR V2 Pang et al. (2026)	52.34	65.96	52.00	65.86	51.51	65.04	51.67	64.78	50.95	64.21	51.69	65.17
MMSegRWKV Cao et al. (2026)	51.81	65.74	52.13	65.01	51.93	64.72	51.23	63.98	51.03	64.42	51.63	64.77
VERSA (Ours)	**56.89**	**69.69**	**56.86**	**69.95**	**57.31**	**70.25**	**56.49**	**69.51**	**56.97**	**70.02**	**56.90**	**69.88**

Baseline	Mean gain Δ (pp)		95% CI (pp)		p_{raw}		p_{H}	
	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
UNETR	+17.72	+17.55	[16.68, 18.76]	[16.11, 19.00]	1.19×10^{-6}	4.56×10^{-6}	3.58×10^{-5}	1.23×10^{-4}
SwinUNETR	+5.82	+5.80	[5.24, 6.39]	[5.21, 6.39]	9.72×10^{-6}	1.08×10^{-5}	2.24×10^{-4}	2.37×10^{-4}
MedNeXt	+3.70	+3.06	[2.96, 4.45]	[2.19, 3.92]	1.60×10^{-4}	5.95×10^{-4}	1.59×10^{-3}	3.15×10^{-3}
SuperLightNet	+4.25	+3.56	[3.43, 5.08]	[2.47, 4.65]	1.40×10^{-4}	8.13×10^{-4}	1.59×10^{-3}	3.25×10^{-3}
SlimFormer3D	+5.54	+5.54	[4.95, 6.14]	[4.95, 6.12]	1.35×10^{-5}	1.25×10^{-5}	2.43×10^{-4}	2.38×10^{-4}
MambaHoME	+8.78	+8.45	[8.38, 9.18]	[7.81, 9.08]	4.37×10^{-7}	3.19×10^{-6}	1.40×10^{-5}	8.93×10^{-5}
SlimUNETR V2	+5.21	+4.71	[4.40, 6.02]	[3.67, 5.75]	5.86×10^{-5}	2.29×10^{-4}	9.38×10^{-4}	1.61×10^{-3}
MMSegRWKV	+5.28	+5.11	[4.73, 5.83]	[4.24, 5.98]	1.19×10^{-5}	8.31×10^{-5}	2.38×10^{-4}	1.16×10^{-3}

A.3 Consistency across Cross-Validation Folds

We evaluate VERSA and eight baselines using five-fold cross-validation on COVID-19 CT Seg (Table A.4). Each fold uses 16 volumes for training and four for evaluation, with identical fold assignments across methods. The five disjoint test folds cover the entire dataset. VERSA achieves the highest IoU and Dice in every fold. Figure A.1 illustrates its Dice advantage over the strongest baseline in each fold.

Across folds, VERSA achieves the highest mean Dice (84.99%) and the smallest observed sample standard deviation (3.21 percentage points) among the evaluated methods (Figure A.2). Its mean IoU is also the highest at 76.96%. Relative to SlimUNETR V2, the strongest baseline by both mean IoU and mean Dice, VERSA improves IoU and Dice by 1.88 and 1.42 percentage points, respectively, and reduces mean HD95 from 49.05 to 41.74, a relative reduction of 14.9%. VERSA also achieves the lowest HD95 in four of the five folds. Together, these results show consistent

overlap gains across the evaluated partitions, accompanied by lower observed fold-to-fold variation in Dice and improved average boundary accuracy.

Table A.4: Five-fold cross-validation on COVID-19 CT Seg. IoU and Dice are reported as percentages (higher is better); lower HD95 is better. Average denotes the arithmetic mean across the five folds. Bold indicates the best result in each column.

Method	Venue	Fold1			Fold2			Fold3		
		IoU	Dice	HD95	IoU	Dice	HD95	IoU	Dice	HD95
UNETR (Hatamizadeh et al., 2022b)	WACV 2022	62.38	74.20	61.26	64.21	74.37	51.38	73.23	82.01	79.96
SwinUNETR (Hatamizadeh et al., 2022a)	MICCAI 2022	70.29	80.97	81.29	68.48	78.76	49.78	76.99	84.79	54.46
nnFormer (Zhou et al., 2023)	TIP 2023	67.52	79.07	50.66	55.98	68.64	85.67	71.31	80.63	71.97
WaveFormer (Al Hasan et al., 2025)	MICCAI 2025	66.63	77.94	62.30	67.63	77.56	48.99	76.06	83.62	62.03
MambaHoME (Plotka et al., 2025)	NeurIPS 2025	71.29	82.26	59.26	68.09	78.17	56.33	78.35	85.52	63.53
SlimUNETR V2 (Pang et al., 2026)	TMI 2026	70.15	80.73	38.48	70.20	80.64	**47.79**	78.89	85.92	65.26
Rep3D Lee et al. (2026)	ICML 2026	68.73	78.90	40.63	62.84	73.08	50.33	70.45	78.56	63.47
MMSegRWKV Cao et al. (2026)	IoTJ 2026	69.14	79.54	39.42	66.96	77.13	55.33	77.63	85.11	58.95
VERSA (Ours)	–	**73.61**	**83.23**	**15.22**	**72.46**	**82.12**	59.72	**80.13**	**87.44**	**54.20**

Method	Venue	Fold4			Fold5			Average		
		IoU	Dice	HD95	IoU	Dice	HD95	IoU	Dice	HD95
UNETR (Hatamizadeh et al., 2022b)	WACV 2022	72.94	80.48	76.93	76.05	84.57	68.12	69.76	79.13	67.53
SwinUNETR (Hatamizadeh et al., 2022a)	MICCAI 2022	74.40	81.53	60.26	80.51	87.99	35.70	74.13	82.81	56.30
nnFormer (Zhou et al., 2023)	TIP 2023	68.86	77.87	96.58	70.69	81.08	77.14	66.87	77.46	76.40
WaveFormer (Al Hasan et al., 2025)	MICCAI 2025	73.24	80.70	77.69	81.09	88.44	30.47	72.93	81.65	56.30
MambaHoME (Plotka et al., 2025)	NeurIPS 2025	75.27	82.10	71.15	79.98	87.81	48.78	74.60	83.17	59.81
SlimUNETR V2 (Pang et al., 2026)	TMI 2026	75.58	82.32	53.43	80.61	88.25	40.30	75.09	83.57	49.05
Rep3D Lee et al. (2026)	ICML 2026	74.18	81.50	67.61	79.06	86.81	49.35	71.05	79.77	54.28
MMSegRWKV Cao et al. (2026)	IoTJ 2026	73.21	80.73	77.22	81.27	88.64	40.55	73.64	82.23	54.29
VERSA (Ours)	–	**76.43**	**82.80**	**53.16**	**82.19**	**89.36**	**26.42**	**76.96**	**84.99**	**41.74**

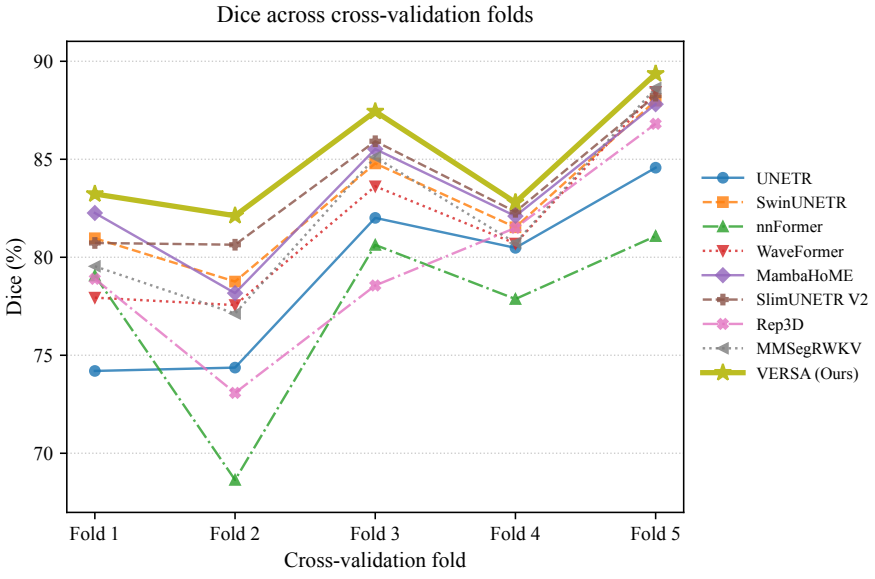

Figure A.1: Fold-wise Dice on COVID-19 CT Seg. All methods use identical five-fold partitions. VERSA achieves the highest Dice in every fold.

A.4 Cross-Dataset Generalization to LiTS2017

We assess cross-dataset generalization by directly evaluating models trained on Colorectal–Liver Metastases (Simpson et al., 2024) on LiTS2017 (Bilic et al., 2023). All 131 publicly annotated CT volumes from the official LiTS2017 training split constitute the external test cohort. This setting evaluates transfer from a metastasis-focused source dataset to a benchmark containing both primary and secondary liver tumours.

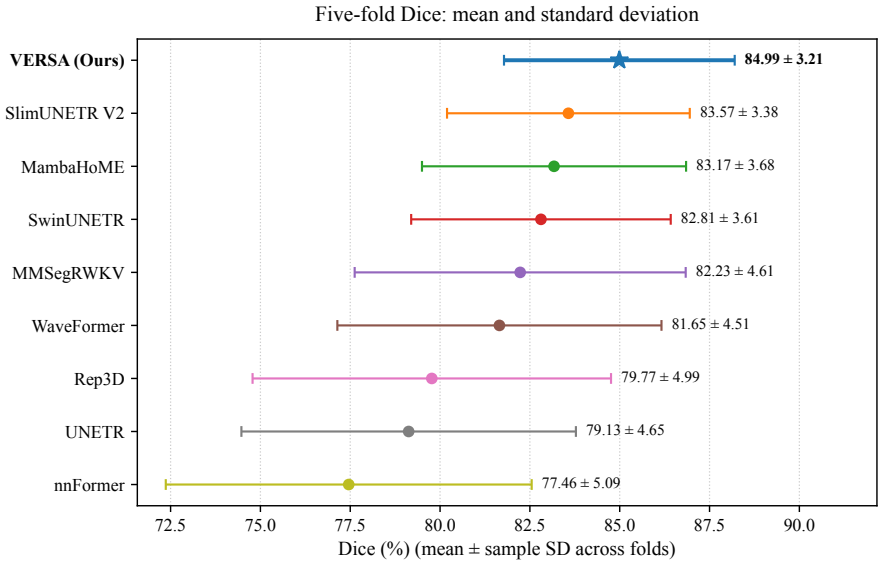

Figure A.2: Mean Dice and across-fold variation on COVID-19 CT Seg. Markers indicate the five-fold mean, and horizontal bars denote $\pm$ one sample standard deviation across folds. Methods are ordered by mean Dice. VERSA achieves the highest mean and smallest standard deviation.

As shown in Table A.5, VERSA achieves the highest IoU (59.34%) and Dice (68.10%) among all compared methods, exceeding SegMamba, the strongest baseline, by 0.27 and 0.52 percentage points, respectively. Figure A.3 shows close agreement with the reference liver contours and tumour extent. In the top row, VERSA recovers the annotated lesion while avoiding the additional tumour regions predicted by several baselines; in the bottom row, it preserves the peripheral lesion missed by nnFormer. These results support VERSA's cross-dataset generalization.

Table A.5: Cross-dataset generalization from Colorectal–Liver Metastases to LiTS2017. Models trained on the Colorectal–Liver Metastases dataset are evaluated on all 131 publicly annotated CT volumes from the LiTS2017 training split. IoU and Dice are reported as percentages (higher is better). Bold indicate the best.

Method	Venue	LiTS2017	
		IoU	Dice
3D U-Net Çiçek et al. (2016)	MICCAI 2016	57.02	65.80
UNETR Hatamizadeh et al. (2022b)	WACV 2022	52.79	61.47
SwinUNETR Hatamizadeh et al. (2022a)	MICCAI 2022	53.70	62.04
3D UX-Net Lee et al. (2023)	ICLR 2023	56.81	65.40
nnFormer Zhou et al. (2023)	TIP 2023	48.51	57.17
MedNeXt Roy et al. (2023)	MICCAI 2023	55.90	64.89
UNesT Yu et al. (2023)	MedIA 2023	56.03	64.50
SegMamba Xing et al. (2024)	MICCAI 2024	59.07	67.58
SwinSMT Płotka et al. (2024)	MICCAI 2024	55.86	64.57
VSmTrans Liu et al. (2024)	MedIA 2024	55.47	63.59
SuperLightNet Yu et al. (2025)	CVPR 2025	58.25	66.83
WaveFormer Al Hasan et al. (2025)	MICCAI 2025	58.92	67.21
SlimFormer3D Hong et al. (2025)	MICCAI 2025	51.85	60.39
MambaHoME Plotka et al. (2025)	NeurIPS 2025	55.99	65.21
SlimUNETR V2 Pang et al. (2026)	TMI 2026	56.90	65.98
LightUNETR-L Liu et al. (2026)	TPAMI 2026	57.00	65.79
Rep3D Lee et al. (2026)	ICML 2026	53.53	62.95
MMSegRWKV Cao et al. (2026)	IoTJ 2026	55.45	64.20
HCRT Zheng et al. (2026)	PR 2026	56.13	65.42
VERSA (Ours)	-	**59.34**	**68.10**

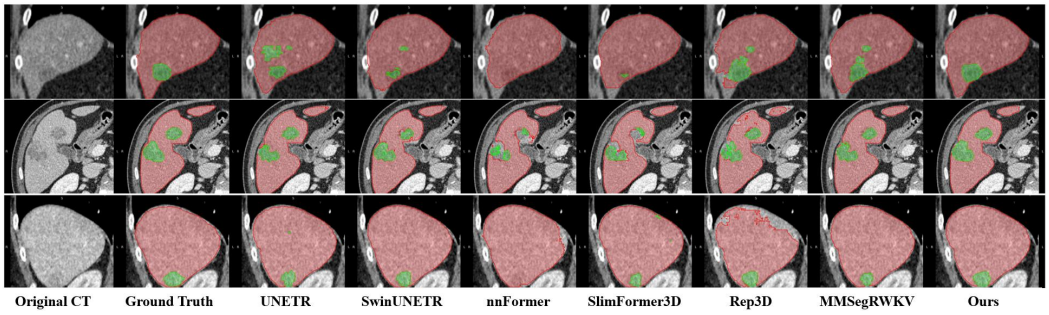

Figure A.3: Qualitative cross-dataset comparison on LiTS2017 using models trained on Colorectal–Liver Metastases. Each row shows a CT slice, its ground-truth annotation, and predictions from selected methods. Red and green denote liver and tumour regions, respectively. The examples highlight differences in lesion delineation, missed lesions, and spurious tumour predictions.

A.5 Paired Statistical Evidence and Effect Sizes

Test-set evaluation and case matching. We compare VERSA with 19 baselines using case-level predictions on the BreastDCEDL ISPY2 test set. Each test case receives equal weight, and the statistical analysis treats volumetric test cases as independent sampling units.

Paired significance testing. For baseline b, metric $m \in \{\mathrm{IoU}, \mathrm{Dice}\}$, and test case i, we define the paired difference as

$$\delta_{i,b,m} = s_{i,m}^{\mathrm{VERSA}} - s_{i,m}^{b}, \qquad i = 1, \ldots, n, \tag{A.3}$$

where n denotes the number of matched test-set cases and scores are expressed on the $[0, 1]$ scale. Positive values favor VERSA. We apply a two-sided exact Wilcoxon signed-rank test to each paired-difference vector. The null hypothesis specifies a paired-difference distribution symmetric about zero. Let r_i denote the average rank of $|\delta_{i,b,m}|$ among the nonzero absolute differences. The test statistic is

$$W_{b,m}^{+} = \sum_{\delta_{i,b,m}>0} r_i, \qquad W_{b,m}^{-} = \sum_{\delta_{i,b,m}<0} r_i, \qquad T_{b,m} = \min\{W_{b,m}^{+}, W_{b,m}^{-}\}. \tag{A.4}$$

Exact conditional p-values are computed using integer dynamic programming over the observed ranks. All observed paired differences are nonzero.

Joint multiple-comparison control. The $19 \times 2 = 38$ baseline–metric comparisons form a single testing family. We apply Holm's step-down correction at a family-wise significance level of $\alpha = 0.05$. For the ordered raw p-values $p_{(1)} \le \cdots \le p_{(38)}$, the adjusted values are

$$p_{(k)}^{\mathrm{H}} = \min\left\{1, \max_{1 \le j \le k} (39 - j) p_{(j)}\right\}, \qquad k = 1, \ldots, 38. \tag{A.5}$$

This joint correction covers all baseline comparisons for both IoU and Dice. The tables report the original p-values and their corresponding Holm-adjusted values.

Effect sizes and confidence intervals. We quantify improvement through the mean paired gain on the original metric scale and Cohen's standardized paired effect d_z. Specifically,

$$\bar{\delta}_{b,m} = \frac{1}{n} \sum_{i=1}^{n} \delta_{i,b,m}, \qquad s_{\delta,b,m} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(\delta_{i,b,m} - \bar{\delta}_{b,m}\right)^2}, \tag{A.6}$$

$$\Delta_{b,m} = 100 \bar{\delta}_{b,m}, \qquad d_{z,b,m} = \frac{\bar{\delta}_{b,m}}{s_{\delta,b,m}}. \tag{A.7}$$

Here, Δ is expressed in percentage points (pp), and d_z is scaled by the sample standard deviation of the paired differences. For both quantities, we construct pointwise 95% bias-corrected and accelerated (BCa) bootstrap confidence intervals. Each resample preserves the pairing between VERSA and the baseline and recomputes both the mean and the sample standard deviation. Within each comparison, the same bootstrap draws are used for Δ and d_z. These intervals characterize uncertainty in the effect estimates, while the Holm-adjusted tests provide joint multiple-comparison control.

Overall comparison. Tables A.6 and A.7 report all 38 comparisons on the BreastDCEDL ISPY2 test set. VERSA achieves 76.60% IoU and 85.73% Dice, reported as mean $\pm$ sample standard deviation across test-set cases. VERSA attains higher mean scores than all 19 baselines on both metrics, with paired improvements ranging from 1.78 to 10.79 pp in IoU and from 0.99 to 7.37 pp in Dice. All 38 comparisons remain statistically significant after joint Holm correction, with a maximum adjusted p-value of 2.160×10^{-6}. These results provide strong statistical evidence for VERSA's overlap advantage across the complete baseline set on the BreastDCEDL ISPY2 test set.

Comparison with the strongest baseline. WaveFormer achieves the highest baseline mean for both IoU and Dice on this test set. Relative to WaveFormer, VERSA improves IoU by 1.78 pp (95% BCa CI $[0.96, 2.59]$), with Cohen's $d_z = 0.306$ (95% BCa CI $[0.143, 0.445]$) and a Holm-adjusted p-value of 1.209×10^{-6}. The corresponding Dice improvement is 0.99 pp (95% BCa CI $[0.25, 1.61]$), with $d_z = 0.204$ (95% BCa CI $[0.018, 0.363]$) and an adjusted p-value of 2.160×10^{-6}. The confidence intervals for both the absolute gains and standardized paired effects lie above zero, supporting VERSA's improvement over the strongest competing method on both overlap measures.

Complementary mean-based analysis. Two-sided paired t-tests of zero mean paired difference further support the mean overlap improvements. After Holm correction within a separate family containing the same 38 comparisons, every comparison remains significant, with a maximum adjusted p-value of 0.01143. Together, the positive paired gains, quantified effect sizes, and agreement between rank-based and mean-based analyses provide convergent statistical evidence of improved segmentation overlap with VERSA on the BreastDCEDL ISPY2 test set, reinforcing the performance gains reported in the main comparison.

Table A.6: **Paired IoU comparisons on BreastDCEDL ISPY2 test set.** VERSA achieves $76.60 \pm 14.89\%$ (mean $\pm$ sample SD across cases). Δ is the mean paired gain in percentage points (pp); Cohen's d_z standardizes it by the sample SD of paired differences. Brackets show pointwise 95% BCa confidence intervals from 100,000 paired-case resamples. p_{raw} is the two-sided exact Wilcoxon signed-rank probability; p_{H} is Holm-adjusted jointly over all 38 IoU/Dice comparisons. All $p_{\text{H}} < 0.05$. Positive effects favor VERSA.

	Paired gain		Effect size		Significance	
Baseline	Δ (pp)	95% CI	Cohen's d_z	95% CI	p_{raw}	p_{H}
3D U-Net	+3.32	[2.67, 4.05]	0.673	[0.506, 0.817]	1.914×10^{-24}	4.594×10^{-23}
UNETR	+6.80	[5.74, 8.09]	0.810	[0.689, 0.923]	7.392×10^{-34}	2.587×10^{-32}
SwinUNETR	+2.47	[1.71, 3.37]	0.419	[0.295, 0.528]	1.340×10^{-9}	1.608×10^{-8}
3D UX-Net	+2.34	[1.37, 3.63]	0.293	[0.160, 0.402]	6.086×10^{-8}	4.260×10^{-7}
nnFormer	+10.79	[9.61, 11.88]	1.336	[0.934, 1.627]	4.158×10^{-40}	1.580×10^{-38}
MedNeXt	+3.22	[2.55, 3.91]	0.663	[0.386, 0.828]	4.769×10^{-26}	1.288×10^{-24}
UNesT	+2.31	[1.38, 3.18]	0.357	[0.178, 0.518]	1.925×10^{-11}	2.700×10^{-10}
SegMamba	+3.08	[2.36, 3.86]	0.571	[0.435, 0.697]	6.727×10^{-16}	1.480×10^{-14}
SwinSMT	+3.20	[2.35, 4.25]	0.474	[0.355, 0.582]	7.434×10^{-13}	1.338×10^{-11}
VSmTrans	+2.22	[1.20, 3.35]	0.288	[0.125, 0.411]	1.368×10^{-8}	1.231×10^{-7}
SuperLightNet	+2.53	[1.83, 3.40]	0.452	[0.324, 0.575]	2.220×10^{-12}	3.774×10^{-11}
WaveFormer	+1.78	[0.96, 2.59]	0.306	[0.143, 0.445]	4.422×10^{-7}	1.209×10^{-6}
SlimFormer3D	+6.66	[5.79, 7.53]	1.073	[0.863, 1.265]	1.031×10^{-33}	3.504×10^{-32}
MambaHoME	+2.01	[1.19, 2.93]	0.321	[0.172, 0.441]	1.725×10^{-7}	8.626×10^{-7}
SlimUNETR V2	+4.87	[3.94, 5.61]	0.821	[0.402, 1.104]	1.478×10^{-32}	4.760×10^{-31}
LightUNETR-L	+3.22	[2.38, 4.15]	0.508	[0.367, 0.628]	1.480×10^{-15}	3.108×10^{-14}
Rep3D	+6.21	[5.06, 7.66]	0.677	[0.574, 0.775]	6.673×10^{-27}	1.869×10^{-25}
MMSegRWKV	+2.01	[1.23, 2.81]	0.354	[0.171, 0.475]	7.683×10^{-9}	8.452×10^{-8}
HCRT	+5.76	[5.05, 6.48]	1.124	[0.914, 1.310]	2.184×10^{-35}	7.861×10^{-34}

Table A.7: **Paired Dice comparisons on BreastDCEDL ISPY2 test set.** VERSA achieves $85.73 \pm 12.45\%$ (mean $\pm$ sample SD across cases). Δ is the mean paired gain in percentage points (pp); Cohen's d_z standardizes it by the sample SD of paired differences. Brackets show pointwise 95% BCa confidence intervals from 100,000 paired-case resamples. p_{raw} is the two-sided exact Wilcoxon signed-rank probability; p_{H} is Holm-adjusted jointly over all 38 IoU/Dice comparisons. All $p_{\text{H}} < 0.05$. Positive effects favor VERSA.

	Paired gain		Effect size		Significance	
Baseline	Δ (pp)	95% CI	Cohen's d_z	95% CI	p_{raw}	p_{H}
3D U-Net	+2.12	[1.55, 2.69]	0.519	[0.282, 0.688]	8.574×10^{-23}	1.972×10^{-21}
UNETR	+4.67	[3.78, 5.73]	0.667	[0.504, 0.785]	2.724×10^{-31}	8.443×10^{-30}
SwinUNETR	+1.54	[0.92, 2.20]	0.335	[0.173, 0.460]	8.813×10^{-9}	8.813×10^{-8}
3D UX-Net	+1.53	[0.66, 2.77]	0.209	[0.056, 0.310]	1.725×10^{-7}	8.626×10^{-7}
nnFormer	+7.37	[6.30, 8.26]	1.051	[0.688, 1.375]	3.328×10^{-37}	1.231×10^{-35}
MedNeXt	+2.07	[1.35, 2.58]	0.484	[0.126, 0.752]	6.239×10^{-25}	1.560×10^{-23}
UNesT	+1.25	[0.41, 1.93]	0.231	[0.044, 0.407]	1.613×10^{-10}	2.097×10^{-9}
SegMamba	+1.94	[1.37, 2.52]	0.469	[0.274, 0.600]	1.898×10^{-14}	3.607×10^{-13}
SwinSMT	+2.18	[1.52, 2.99]	0.417	[0.288, 0.524]	3.166×10^{-12}	5.066×10^{-11}
VSmTrans	+1.33	[0.38, 2.29]	0.194	[0.004, 0.324]	6.367×10^{-8}	4.260×10^{-7}
SuperLightNet	+1.59	[1.02, 2.30]	0.351	[0.193, 0.481]	1.800×10^{-11}	2.700×10^{-10}
WaveFormer	+0.99	[0.25, 1.61]	0.204	[0.018, 0.363]	2.160×10^{-6}	2.160×10^{-6}
SlimFormer3D	+4.32	[3.57, 5.01]	0.839	[0.556, 1.054]	1.372×10^{-30}	4.116×10^{-29}
MambaHoME	+1.22	[0.44, 1.94]	0.225	[0.035, 0.364]	4.030×10^{-7}	1.209×10^{-6}
SlimUNETR V2	+3.10	[2.28, 3.67]	0.640	[0.255, 0.955]	2.579×10^{-29}	7.478×10^{-28}
LightUNETR-L	+2.12	[1.40, 2.88]	0.402	[0.226, 0.531]	6.063×10^{-15}	1.213×10^{-13}
Rep3D	+4.48	[3.56, 5.71]	0.592	[0.495, 0.682]	3.697×10^{-25}	9.611×10^{-24}
MMSegRWKV	+1.17	[0.38, 1.79]	0.237	[0.007, 0.400]	2.728×10^{-8}	2.182×10^{-7}
HCRT	+3.79	[3.26, 4.31]	1.005	[0.774, 1.204]	1.443×10^{-32}	4.760×10^{-31}

A.6 Accuracy–Efficiency Trade-offs

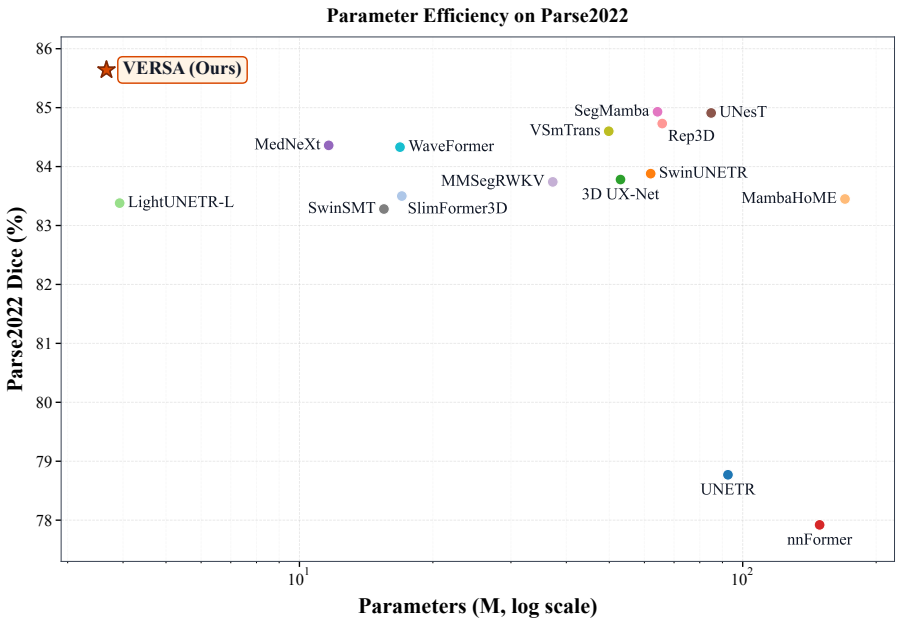

Figure A.4: Accuracy–parameter trade-off on Parse2022.

We complement the model-level statistics in Table 2 by examining Parse2022 dataset Dice against computational cost, parameter count, and measured inference throughput (Figures A.4, A.5, and A.6). The efficiency benchmark uses $96 \times 96 \times 96$-voxel input patches on a single NVIDIA

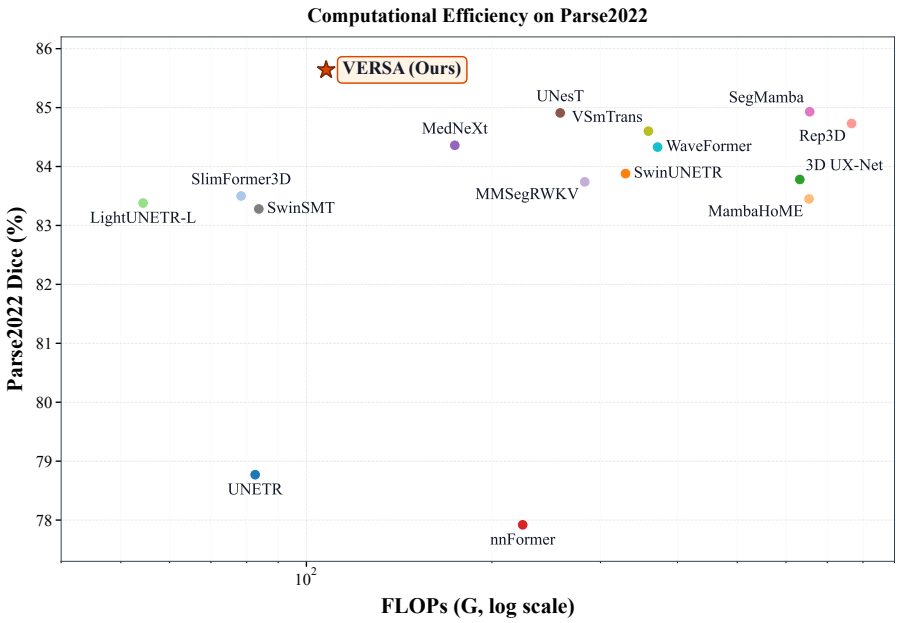

Figure A.5: Accuracy–computation trade-off on Parse2022.

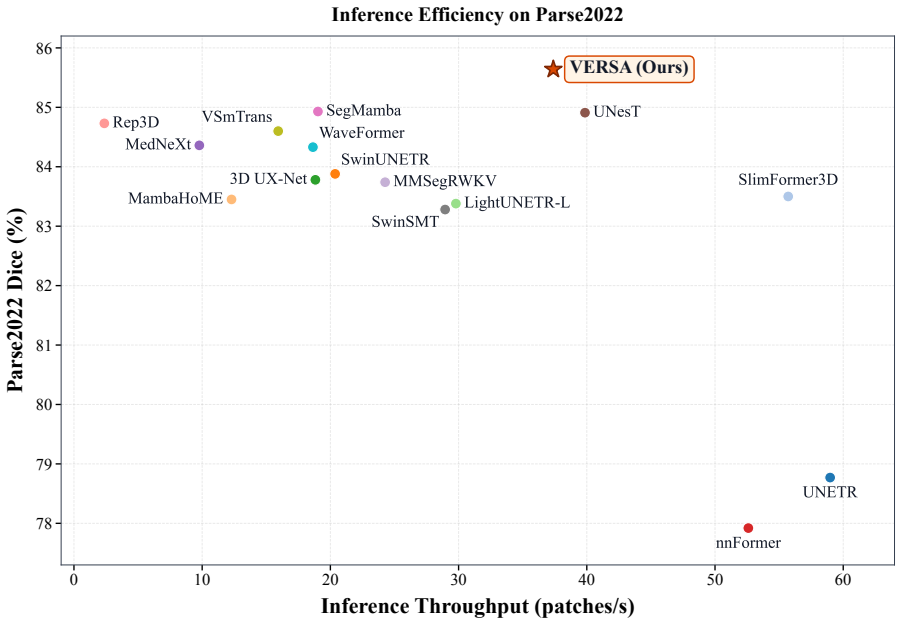

Figure A.6: Accuracy–throughput trade-off on Parse2022.

RTX 4090 GPU. The complete performance and efficiency comparisons are provided in Tables 1 and 2, respectively.

VERSA achieves 85.64% Dice with 3.67M parameters, combining the highest Dice with the comparatively small parameter count among the displayed models (Figure A.4). Relative to LightUNETR-L, the next-smallest displayed model, VERSA improves Dice by 2.26 percentage points while using 6.6% fewer parameters (3.67M versus 3.93M). Compared with SegMamba, the strongest baseline by Parse2022 Dice, VERSA improves Dice by 0.71 percentage points with 94.3% fewer parameters (3.67M versus 64.24M). These comparisons show that the strongest observed overlap does not require a large parameter budget. VERSA requires 107.67 GFLOPs per input patch

(Figure A.5). This is 83.6% lower than SegMamba (655.62 GFLOPs) and 38.2% lower than MedNeXt (174.10 GFLOPs), while Dice improves by 0.71 and 1.28 percentage points, respectively. The observed advantage is therefore a favorable high-accuracy operating point. VERSA reaches 37.39 patches/s (Figure A.6), corresponding to $1.96\times$, $3.82\times$, and $1.54\times$ the throughput of SegMamba, MedNeXt, and MMSegRWKV, respectively.

Under the reported measurements, VERSA is non-dominated in all three accuracy–efficiency views and dominates every displayed baseline in the Dice–parameter plane. This operating point is consistent with the compact realization of structured context selection: V-BAR RWKV operates at the bottleneck and shares projection matrices across axes and directions, while DiVERS reuses shell responses across nested support experts. Importantly, routing adapts the aggregation of computed candidates.