International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(11s) 2026 pp. 1035-1070
https://doi.org/10.70917/ijcisim-2026-3831
Article

PVRSAP-Net: Probabilistic Voxel Refinement and Spatial Attention Pyramid Learning for Three-Dimensional Multiregion Brain Tumor Segmentation

P. Sree Lakshmi[1], Arpita Gupta[2]

[1,2]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad - 500075, Telangana, India.

Abstract: Accurate three-dimensional delineation of glioma subregions from multimodal magnetic resonance imaging remains difficult when tumor margins are weak, lesion volumes are small, acquisition intensity varies across subjects, or a coarse segmentation prior contains spatially isolated errors. The supplied research synopsis proposes a voxel-based framework that joins probabilistic threshold refinement, adaptive multiscale descriptors, spatial-attention pyramid learning, residual convolution, and multiregion classification. This paper converts that concept into a single testable method named PVRSAP-Net. Its first mechanism is bounded edge-aware probabilistic voxel refinement, which updates uncertain tumor probabilities with neighborhood evidence, image-derived conductance, projection to the valid probability interval and nested-region constraint set, and explicit relations among enhancing tumor, tumor core, and whole tumor. The refined probabilities join normalized multimodal intensity, local texture, and contextual descriptors. A windowed spatial-attention pyramid then models nonlocal dependence at several resolutions, and a residual three-dimensional decoder predicts nested tumor regions with confidence estimates. Training uses a region-weighted Dice, cross-entropy, boundary, and calibration objective. Three algorithms, three conditional mathematical propositions, computational analysis, safeguards, and a leakage-resistant BraTS evaluation protocol are specified. Since raw clinical volumes and executable baseline predictions were not supplied, no clinical superiority claim is made. A seeded controlled simulation of 20 synthetic four-channel multimodal volumes, each represented on a 48^3 voxel grid, tests only the internal refinement behavior. In that simulation, PVRSAP-Net refinement attained pooled Dice 0.9091, IoU 0.8382, sensitivity 0.9289, HD95 1.2316 voxels, Brier score 0.0040, and expected calibration error 0.0142. The corresponding raw probabilistic prior attained Dice 0.8551 and HD95 7.1811 voxels. Removal of edge conductance reduced Dice to 0.8844; removal of anatomical hierarchy projection reduced Dice to 0.8802. These results support numerical and structural behavior under declared synthetic assumptions, not patient-level efficacy.

Keywords: brain tumor segmentation, multimodal MRI, probabilistic voxel refinement, three-dimensional segmentation, spatial attention, residual learning, uncertainty calibration, BraTS.

1. INTRODUCTION

Gliomas are a heterogeneous group of central nervous system (CNS) tumours whose clinical description, at least in adults, now includes histologic and molecular features following the 2021 World Health Organization classification [1]. Magnetic resonance (MR) contrasts are analogous to those of computed tomography (CT) in that they can be used to visualize abnormal morphology without ionizing radiation. In the traditional BraTS scenario, T1-weighted, contrast-enhanced T1-weighted (T1ce), T2-weighted, and fluid attenuated inversion recovery (FLAIR) sequences reveal different parts of a tumor along with its surrounding tissue [2]–[4]. T1ce highlights active enhancement while FLAIR and T2 demonstrate edema and infiltrative modification, and T1 adds structural information. Commonly, segmentation focuses on enhancing tumor (ET), tumor core (TC), and whole tumor (WT), forming the anatomical relationship $ET \subseteq TC \subseteq WT$ for analysis. These areas facilitate volumetric measurement and investigation workflows, but an automatic mask is not a diagnosis and must be reviewed by a specialist before any clinical application.

Manual segmentation remains time-consuming and subject to interobserver variability, especially for diffuse boundaries, small enhancing foci, postoperative cavities, and heterogeneous necrotic areas. The BraTS baseline (now called standard) setting was established to provide a uniform multimodal evaluation framework and paved the way to expose both improvements and remaining challenges [2]–[4]. A high mean overlap can hide a miss with clinical meaning on a small component. A small quantity of false-positive voxels can lead to a severe score penalty if the ground-truth region is empty. Quality of embedded systems, scanner protocol, motion, bias field, resampling, annotation policy, and tumor prevalence affect performance. Therefore, segmentation research has to disentangle three questions: does a model learn a meaningful representation, do the training and testing protocols avoid leakage, and do the observed behaviors transfer to other institutions or tumor populations.

Early auto matic methods are mainly based on intensity thresholds, region growing, deformable contours, probabilistic graphical models, and hand-crafted texture. While such processes involvederasable operations that were easier to understand, they still relied heavily on initialization and intensity homogeneity. Learning hierarchical features directly from raw image patches using convolutional neural networks has transformed the predominant design in the area. U-Net [7], 3D U-Net [8] and V-Net [9] introduced encoder-decoder architectures with skip connections and volumetric output. Brain-tumor-specific networks were then introduced by combining local pathways, cascades, conditional random fields, multi-scale patches, and ensembles [10]-[12] . Residual connections [13] and attention gates [14] fostered better optimization and more selective feature delivery. The self-adapting nnU-Net framework demonstrated that preprocessing, patch size, resampling, loss, augmentation, inference, and post-processing can be as important as a named architectural block [15]. This observation is crucial to make the comparisons fair: one cannot proclaim a method to be better based on a plot in which the results from different splits, ensembles, metrics definitions, or post-processing rules are aggregated. Transformer-based volumetric segmentation was motivated by the representation of long-range spatial dependence. TransBTS introduces transformer-based processing within a convolutional encoder-decoder framework [17]. UNETR leverages a transformer encoder as a bridge to a convolutional decoder [18]. Swin UNETR takes advantage of hierarchical shifted-window attention and achieved competitive results on BraTS 2021 [19]. nnFormer [20], MedNeXt [21], UNETR++ [22], as well as future state-space architectures like SegMamba [23] further investigated the tradeoff between context, memory, boundary precision, and data efficiency. Recent BraTS submissions report strong source-reported overlap values [24]–[30], but vary in terms of cohorts, folds, internal splits, ensembles, lesion-level rules, and computational budgets. A convincing novel study requires either a common reimplementation or a clearly labeled source-reported comparison to a mixed ranking.

The provided synopsis of research for this paper defines a smaller technical gap. A preceding probabilistic detector can attract candidate tumor tissue, but the result may have jagged borders, ambiguous transition voxels, isolated false positives, and nonnested subregion probabilities. A deterministic encoder that ignores the confidence values of the prior, thus forfeiting the latter's confidence values, would be discarding information about ambiguity. Isolated errors can~be mitigated by uniform smoothing, but small objects are eroded, and boundaries are displaced across image edges. Global attention can capture distant relations but the computational cost of full attention for a three-dimensional volume grows quadratically as the number of tokens increases. A simple residual decoder can retain local detail, but it does not, on its own, enforce ET-TC-WT consistency or probability calibration.

PVRSAP-Net fills this gap with a staged design, based on the synopsis. The first step normalizes each MRI modality within a brain mask. The second stage does bounded edge-aware probabilistic voxel refinement (BE-PVR). Rather than performing a free Laplacian update, BE-PVR gates the update by voxel uncertainty, limits diffusion across high image-gradients, projects each update into 0,1, and projects the three region probabilities into their anatomical nesting set. In the third stage, adaptive voxel descriptors are constructed based on intensity, texture, neighbourhood context and refined probability. (xiv) the fourth stage applies a spatial-attention pyramid former (SAPFormer) via windowed attention at different scales with nonnegative fusion weights that add up to one; and (xv) the fifth stage employs residual 3D convolution and a multi-head output for ET, TC and WT. The fifth stage features a multi-head output for ET, TC and WT utilizing residual 3D convolution. The final objective combines overlap, voxel classification, boundary and calibration terms.

This paper tests three hypotheses. H1 says that bounded edge-aware refinement enhances the internal consistency and boundary behavior of a noisy probability prior as compared to no refinement, neighborhood averaging, isotropic diffusion in a controlled data-generation process. H2 says that edge conductance and anatomical hierarchy projection provide distinct, quantifiable contributions, measured by removal ablations. H3 says that the complete PVRSAP-Net can be tested relatively on recorded BraTS volumes via patient-separated folds, using training-only preprocessing fits, a frozen test protocol, confidence intervals, and publicly available baseline implementations.

The contributions are as follows.

1. A bounded probabilistic refinement rule extends the synopsis's neighborhood and Laplacian updates with uncertainty gating, multimodal edge conductance, interval projection, and nested-region projection. The mechanism is paired with a stability proposition, stopping rule, component ablation, and failure safeguards.
2. An adaptive multiscale voxel representation joins normalized MRI intensities, local texture moments, spatial context, positional encoding, and refined probabilities without treating hand-crafted descriptors as a substitute for learned features.
3. A windowed SAPFormer module joins fine and coarse context through simplex-constrained scale fusion. A bounded-output proposition states the conditions under which the fused representation remains controlled.
4. A residual volumetric decoder and composite objective predict ET, TC, and WT with overlap, boundary, classification, and calibration supervision. A conditional descent result is supplied for the clipped-gradient training update rather than an unsupported claim of global optimality.
5. A reproducible controlled simulation tests the refinement mechanism with a public script, fixed seed, machine-readable results, eight generated figures, ablations, parameter sensitivity, calibration diagnostics, and error analysis. Its evidence boundary is stated throughout.
6. A recorded-data protocol based on the source synopsis defines BraTS 2020/2021 development, recent common baselines, external-shift testing, subject-level uncertainty, resource reporting, and a protocol-aware literature comparison. Missing patient data, author metadata, ethics information, and clinical runs remain visible submission items.

The remainder of the paper proceeds as follows. Section II synthesizes classical, convolutional, attention-based, probabilistic, and recent volumetric segmentation methods. Section III defines the data model, assumptions, nesting constraints, research questions, and evidence levels. Section IV gives the PVRSAP-Net architecture, equations, algorithms, propositions, complexity, and safeguards. Section V defines recorded-data and controlled-simulation protocols. Section VI reports the simulation and protocol-aware literature results. Section VII discusses mechanisms, clinical interpretation limits, reproducibility, and failure cases. Section VIII states limitations and required validation. Section IX concludes the paper.

2. RELATED WORK

A. Multimodal MRI Data and Evaluation Protocols

The original BraTS challenge integrated multimodal glioma MRI, expert annotations, and standardized evaluation to assess segmentation techniques [2]. Subsequent compilations provided larger, multi-institutional cohorts with expert derived labels [3], [4]. The benchmark typically offers T1, T1ce, T2 and FLAIR images registered, skull stripped and resampled, but details differ in each release. ET, TC and WT are overlapping evaluation regions rather than separate classes. ET is the enhancing tumor, TC includes the enhancing and nonenhancing or necrotic core tissue, and WT is the core with edema or invaded tissue. A model that independently estimates these regions can break nesting. Region-based output heads, logical projections or label conversions can be used to repair consistency.

The ATLAS is also named as a lesion-segmentation transfer target in the synopsis. ATLAS Release 1 provided 304 T1-weighted MRIs with manually traced stroke lesions [5]. Release 2 expanded and cleaned the resource across cohorts [6]. The class labels and domain shifts dealt with in ATLAS are wholly different from BraTS: it is about stroke lesions, mostly T1-weighted imaging, different lesion biology, different class structure, and a different acquisition mix. Suitable for a stress or transfer experiment, but not a pooled extension of the BraTS test data. A glioma MRI-trained model cannot be claimed as stroke lesion segmentation model with no separate adaptation and evaluation. In this paper, we consider ATLAS as a future domain-shift benchmark to examine whether BE-PVR can generalize to uncertain lesion boundaries when the segmentation backbone is retrained.

Dice score, IoU, sensitivity, specificity and 95th-percentile Hausdorff distance (HD95) are widely used segmentation metrics. Dice can still be high for a large WT area due to a boundary displacement, and HD95 can be highly influenced by a far away false positive component. The implementation of segmentations metrics have to declare if: (a) empty masks are given a special score, (b) distances are expressed in voxels or millimetres, (c)

aggregation is made over subject, lesion, region, or over all of the voxels pooled. Review of the characteristics and the computational aspects of the evaluation metrics for medical segmentation are given in [40] by Taha and Hanbury. The present results first report the region-level values, and then the means are reported for the subject-level In the simulated controlled setting we report the HD95 in units of voxels, as no physical spacing is assigned. The recorded-data protocol will report in millimeters after image spacing is preserved.

B. Classical and Probabilistic Segmentation

Thresholding and region growing rely on intensity homogeneity, but suffer when normal tissue and tumor share similar intensities. Boundary or smoothness terms are added in active contours and graph cuts but they are still sensitive to the weights of energies and initialization. Fuzzy clustering embodies partial membership association and can describe intensity overlapping but it needs to be regularized with a spatial distance to control outliers. Markov and conditional random fields tie neighboring labels together at the expense of inference complexity. These methods are still useful as inductive biases and for post processing.

There is unsurprisingly a close connection between voxelwise probabilistic labeling and the current paper. Rather than hard thresholding a probability map and replacing it with a binary mask, a probabilistic approach is able to retain confidence and enable local refinement. The synopsis suggests a convex combination of a voxel's prior probability and the mean of its neighborhood, followed by a laplacian update. This may eliminate isolated noise, but an unrestrained diffusion step can go beyond actual boundaries. It may leave the valid probability interval if projection is not employed. Nested ET, TC and WT predictions — there is no guarantee of that with this approach. PVRSAP-Net preserves the probabilistic premise and incorporates edge conductance, uncertainty gating, interval projection and anatomical projection.

MRI preprocessing is no exception. Magnitude MRI is typically assumed to be corrupted by Rician or noncentral-χ noise rather than by plain additive Gaussian noise [32]. Bias fields induces slow intensity variation in images. The N4 bias field correction is a popular approach to estimate low-frequency inhomogeneity [31]. There is no universally harm-free preprocessing method. Denoising, bias correction, registration, and resampling may modify borders of tiny lesions. The result is a recorded-data protocol that thereby customizes or configures preprocessing on the training folds and stores its parameters. The controlled simulation adds smooth bias, correlated noise, modality-specific contrast, and false positive blobs just to challenge the refinement rule. It is not a full scanner physics simulation.

C. Convolutional Volumetric Networks

U-Net introduced a symmetric encoder-decoder architecture with skip connections to capture localization and semantic context [7]. The 3D U-Net extended this concept to the three dimensional domain of volumetric kernels [8]. V-Net combined the 3-D convolutional architecture with a Dice-based optimization function [9]. Brain tumor networks then integrated multiple pathways, patch scales, cascades, and conditional random fields. Havaei et al. applied local and global pathways for brain tumor segmentation [10]. Pereira et al. employed small convolutional kernels and dealt with label imbalance [11]. Kamnitsas et al. presented a multi-scale 3D network integrated with a fully connected conditional random field (FC-CRF) [12]. These results demonstrated the power of volumetric context, but revealed memory bottlenecks and patch-boundary effects.

Deep mappings become easier to optimize with residual connections since they provide identity paths to propagate information [13]. In a volumetric decoder, these identity paths are instrumental in preserving the low level structural information and helping the flow of gradients, but such paths can transmit background noise. Attention U-Net included gates to prevent irrelevant features from being combined [14]. Worldwide attention gates are local selection units; they do not necessarily discover long-distance dependencies over an entire tumor. PVRSAP-Net comprises residual blocks following multiscale attention and utilizes a learned gate in the stage of feature fusion. The probabilistic prior still is a separate input, it is just not manifested as an explicit attention map. nnU-Net represented a different type of innovation: the full automatic configuration of the pipeline based on the properties of the dataset [15]. Its rules prescribe target spacing, patch size, batch size, depth of the architecture, normalization, augmentation, loss, inference, and post-processing. This recipe often beats off-the-shelf specialized models when they have shallower training recipes. A PVRSAP-Net evaluation must start with nnU-Net as a strong baseline and disclose whether the two models are provided with the same folds, input channels, augmentation, inference overlap, and component filtering. The original synopsis cited a modified V-Net, 128 cubed input, batch size two, Adam, 100 epochs, and a environment based on NVIDIA A100 with 40 GB. These values represent a rudimentary protocol, not evidence of an executed experiment.

In a top-ranked BraTS solution [16], Myronenko employed an asymmetric encoder-decoder with variational autoencoder regularization (although not termed as such). This work highlights how auxiliary representation constraints can be advantageous. Yet others braTS systems leverage model ensembles to reduce variance. Deep ensembles offer opportunities for improving predictive uncertainty but also increase training and inference cost [39]. STAPLE may combine segmentations across models [41]. PVRSAP-Net is characterized as a single model for the primary analysis. An optional ensemble can be reported separately, not merged with the single-model row.

D. Transformers, Large-Kernel Networks, and State-Space Models

Self-attention calculates the relationships between pairs of tokens [42]. The Vision Transformer demonstrated that images can be divided into patch tokens for processing by transformers [43]. Swin Transformer achieves linear computational complexity by applying multi-head self-attention within shifted windowed partitions of the input feature with a hierarchical algorithm [44]. These concepts were applied to 3D medical segmentation. TransBTS leverages convolutional encoding, a transformer bottleneck and convolutional decoding [17]. UNETR bridges transformer representations at multiple depths to a U- shaped decoder [18]. It integrates shifted-window attention and hierarchical features [19]. For the five-fold experiments of the Swin UNETR article, the average Dice values for ET, WT, and TC were 0.891, 0.933, and 0.917, respectively; for the official BraTS 2021 test values were different, which demonstrates the difference between cross-validation and challenge testing [19].

For volumetric segmentation, nnFormer melds local and global attention [20]. MedNeXt investigates large-kernel ConvNeXt-style blocks and compound scaling without the use of global quadratic attention [21]. UNETR++ utilizes efficient paired attention to reduce complexity [22]. SegMamba brings in state-space modeling for long-range sequential dependence in 3D medical images [23]. No single context operator seems to dominate in these model classes. Full attention can model wide dependence, but is expensive. Windows can be controlled in cost, but will have window seams. Large kernels retain convolutional bias, but require appropriate scaling. State-space models provide near linear sequence cost, but are scan order and fusion design dependent.

The PVRSAP-Net SAPFormer sub-module employs three design constraints in response. First, attention works within three-dimensional windows rather than over every voxel. Second, the pyramid achieves multiple receptive fields at different pooled scales. Thirdly, scale weights live in a simplex preventing unbounded additive growth. Residual convolution is performed after attention to recover local inductive bias. This amalgam is close in spirit to hybrid networks, but its chief hypothesis is on the interplay of a sharpened probability field and multiscale context, rather than the claim that attention per se is novel.

E. Recent BraTS Studies and Comparison Risks

Recent works still are very diverse in the demonstration mode. 2023 challenge-winning pipeline was based on synthetic augmentation and ensembles and published lesion-aware validation results with the new BraTS metric [24]. Dice 89.77% was reported by an effective Swin-based method but the data and clustering in the publication need to be read before comparison [25]. LIU-Net incorporates inception-style multiscale convolution and a Dice–focal objective, it achieved BraTS 2021 test Dice of 0.8121,0.8856, and 0.8444 for ET, WT, and TC [26]. multiPI-TransBTS incorporates modality-specific encoding, adaptive feature fusion, and task-specific decoding [27]. HybridMamba achieved BraTS 2023 internal-split Dice scores of 94.10%, 92.84%, and 88.83% for WT, TC, and ET in [28]. The 2026 SegGuidedNet preprint achieved mean Dice 0.905 on a held-out BraTS 2021 subset and revealed a 70/10/20 split [29]. The 2026 MHMamba preprint achieved mean Dice 0.9102 on its internal BraTS 2021 split [30].

Those figures cannot simply be put into a single ordered contest table. A few are test-server results, a few are cross-validation means, a few internal splits, and some are preprints. Some rely on ensembles, some on a single network, and some on different BraTS versions. Lesion-aware BraTS 2023 scores are not meaningful in terms of legacy region overlap. In this paper, we review them at a protocol-aware table level with dataset, partition, model multiplicity, and evidence mode. There is no recorded-data row for PVRSAP-Net yet. This processing prevents literature from turning into an artificial leaderboard.

F. Losses, Calibration, and Reliability

Severe class imbalances exist in volumetric tumor segmentation. Generalized Dice accounts for region contribution per class volume [35]. Focal loss downweights the relative importance of easy examples [36]. Boundary loss exploits distance information to penalize contour errors [37]. Cross-entropy provides classification gradients at the voxel level. There is no one loss term to fix all failures. Boundary weight too large can destroy smooth overlap.

High Dice weight can cause unstable gradients for empty classes. PVRSAP-Net adopts a scheduled composite objective and models empty-class behavior.

Probability calibration is different from overlap. Guo et al. demonstrated that state-of-the-art neural networks can be poorly calibrated even when highly accurate [38]. A segmentation confidence of 0.9 for a sampling unit on the basis of which an empirical frequency is approxi- mately 0.9 without any sampling unit information should be considered as 0.9 under a defined sampling unit. Background may also dominate voxel-level calibration. Region-balanced ECE, Brier score, reliability diagrams and risk-coverage curves deliver complementary insights. In the present controlled simulation report Brier score and traditional ECE, then mention class-imbalance limitation. The recorded-data design entails a region-balanced calibration and a subject-level selective risk.

G. Research Gap

The literature favours simple strong volumetric encoders, long-range context modules, and flexible training pipelines. Four vacancies are lined up along with the summary. First, a coarse probabilistic tumor map is thresholded or concatenated with a controlled spatial update. Second, constant smoothing fails differentiate uncertain tissue from high-confidence interior tissue and is not aware of image edges. Thirdly, the separate algorithmically predicted probabilities for ET, TC, and WT can break nesting. Fourth, a host of papers compare figures from protocols that are not compatible and lack calibration, confidence intervals, cost of resources, or analysis of failure.

PVRSAP-Net addresses these gaps with an audit-able probability refinement operator, nested projection, multiscale window attention, a residual decoder, calibrated loss, and evidence-separated evaluation. The novelty lies in the joint mathematical constraints and their tests. There are no claims that standard normalization, Adam optimization [33], batch normalization [34], or the U-shaped decoder are novel.

3. PROBLEM FORMULATION AND DESIGN REQUIREMENTS

A. Data Model

Let the recorded multimodal MRI dataset be

$$\mathcal{D} = \{(\mathbf{I}_s, \mathbf{Y}_s, g_s, d_s)\}_{s=1}^{S},$$

where s indexes a subject, $\mathbf{I}_s \in \mathbb{R}^{H_s \times W_s \times D_s \times M}$ is a registered MRI volume with $M = 4$ modalities, $\mathbf{Y}_s$ is the reference segmentation, g_s is a subject identifier, and d_s identifies a site, scanner, or acquisition domain when available. The four-channel voxel vector at position $v_i = (x_i, y_i, z_i)$ is

$$\mathbf{x}_{s,i} = \left[I_{s,i}^{\mathrm{T1}}, I_{s,i}^{\mathrm{T1ce}}, I_{s,i}^{\mathrm{T2}}, I_{s,i}^{\mathrm{FLAIR}}\right]^{\top}.$$

The target is represented by nested binary regions

$$\mathbf{Y}_s = \{Y_s^{\mathrm{ET}}, Y_s^{\mathrm{TC}}, Y_s^{\mathrm{WT}}\}, \qquad Y_s^{\mathrm{ET}} \subseteq Y_s^{\mathrm{TC}} \subseteq Y_s^{\mathrm{WT}}.$$

This region form avoids ambiguity when a release uses different integer labels. A deterministic conversion function maps the release-specific labels to the three regions before training. The conversion is versioned and tested on known label combinations.

The unit of inference is the subject, not a slice, patch, voxel, or lesion. Every patch from one subject stays in a single partition. The split is

$$\mathcal{D} = \mathcal{D}_{\mathrm{tr}} \ \dot{\cup} \ \mathcal{D}_{\mathrm{va}} \ \dot{\cup} \ \mathcal{D}_{\mathrm{cal}} \ \dot{\cup} \ \mathcal{D}_{\mathrm{te}},$$

with disjoint subject identifiers. $\mathcal{D}_{\mathrm{va}}$ supports optimization and early stopping. $\mathcal{D}_{\mathrm{cal}}$ supports temperature or threshold calibration. $\mathcal{D}_{\mathrm{te}}$ remains frozen until every architecture, loss weight, component threshold, and post-processing rule is fixed. A five-fold subject-level analysis can replace a single development split, yet calibration remains separated inside each training fold.

B. Normalization and Initial Probability Prior

Each modality is standardized inside a training-defined brain mask:

$$\hat{x}_{s,i,m} = \frac{x_{s,i,m} - \mu_{s,m}^{(B)}}{\sigma_{s,m}^{(B)} + \epsilon}, \qquad m \in \{1, \ldots, M\},$$

where $\mu_{s,m}^{(B)}$ and $\sigma_{s,m}^{(B)}$ are the subject-level brain-mask mean and standard deviation. When a global clipping quantile is used, it is estimated only from $\mathcal{D}_{\mathrm{tr}}$. The constant $\epsilon > 0$ prevents division by zero. Bias correction, if selected, is configured before the test lock and applied identically to all models.

A coarse localizer f_{θ_0} produces region probabilities

$$\mathbf{P}_s^{(0)} = f_{\theta_0}(\hat{\mathbf{I}}_s) = \{P_{s,r}^{(0)}(v_i)\}_{r\in\mathcal{R}}, \qquad \mathcal{R} = \{\mathrm{ET}, \mathrm{TC}, \mathrm{WT}\},$$

where every value lies in $[0,1]$. The coarse localizer may be the K-maximum-likelihood probabilistic segmenter described in the broader thesis, a compact 3D U-Net, or an exported probability map. Its identity must remain constant across the refinement ablation. If ground-truth information is used to create the prior, the experiment becomes invalid.

C. Feasible Probability Set

PVRSAP-Net refines probabilities within the feasible set

$$\mathcal{C} = \{\mathbf{p} \in [0,1]^{3N} : p_i^{\mathrm{ET}} \le p_i^{\mathrm{TC}} \le p_i^{\mathrm{WT}} \ \forall i\}.$$

The Euclidean projection $\Pi_{\mathcal{C}}$ enforces interval and hierarchy constraints. A simple closed-form isotonic projection can be applied voxelwise to the three values. In implementation, the ordered values are obtained with a three-element pool-adjacent-violators procedure. The shorter approximation

$$\tilde{p}_i^{\mathrm{WT}} = p_i^{\mathrm{WT}}, \quad \tilde{p}_i^{\mathrm{TC}} = \min(p_i^{\mathrm{TC}}, \tilde{p}_i^{\mathrm{WT}}), \quad \tilde{p}_i^{\mathrm{ET}} = \min(p_i^{\mathrm{ET}}, \tilde{p}_i^{\mathrm{TC}})$$

is fast and used in the controlled simulation. The exact isotonic projection is preferred in recorded-data training since it minimizes the projection distance.

D. Optimization Target

The network F_Θ maps the standardized images and refined prior to region probabilities:

$$\hat{\mathbf{P}}_s = F_\Theta(\hat{\mathbf{I}}_s, \mathbf{P}_s^{(R)}).$$

The primary recorded-data objective is mean subject-level Dice across ET, TC, and WT. Secondary endpoints are IoU, sensitivity, specificity, HD95 in millimeters, average symmetric surface distance, Brier score, region-balanced ECE, negative log likelihood, parameter count, peak memory, and sliding-window inference time. No single metric defines success.

The training problem is

$$\Theta^\star = \underset{\Theta\in\mathcal{T}}{\operatorname{argmin}} \frac{1}{|\mathcal{D}_{\mathrm{tr}}|} \sum_{s\in\mathcal{D}_{\mathrm{tr}}} \mathcal{L}\left(\mathbf{Y}_s, F_\Theta(\hat{\mathbf{I}}_s, \mathbf{P}_s^{(R)})\right) \lambda_w \parallel \Theta \parallel_2^2,$$

subject to finite memory, valid probabilities, nested outputs, and a fixed inference budget. The feasible parameter set $\mathcal{T}$ is implicit in the architecture, gradient clipping, and finite-precision safeguards.

E. Assumptions

The mathematical results in Section IV use narrower assumptions than the empirical model:

1. A refinement neighborhood graph is undirected with nonnegative symmetric weights.
2. The normalized graph Laplacian is positive semidefinite and has largest eigenvalue $\lambda_{\max} \le 2$.
3. The uncertainty gate and edge conductance lie in $[0,1]$.
4. The projection onto $\mathcal{C}$ is Euclidean and nonexpansive.
5. The feature tensors entering a proposition have finite norm.
6. A local training-descent result assumes an L-smooth objective over the examined update interval and bounded clipped gradients.

These assumptions do not make the full deep objective convex. They support bounded updates, controlled fusion, and conditional local descent only.

F. Research Questions and Evidence Levels

The study asks:

RQ1: Does BE-PVR improve a corrupted probability field in the controlled simulation?

RQ2: Which refinement components explain any measured change?

RQ3: Is the result stable across diffusion coefficient, tumor size, and noise severity?

RQ4: Under a common recorded-data protocol, does PVRSAP-Net improve ET/TC/WT performance over nnU-Net, TransBTS, Swin UNETR, and SegMamba?

RQ5: Does any overlap gain retain calibration and acceptable compute cost?

RQ6: Does a BraTS-trained mechanism transfer after retraining to ATLAS lesion data, or does domain shift invalidate the learned prior?

RQ1–RQ3 receive controlled-simulation evidence. RQ4–RQ6 remain planned recorded-data questions. Literature values are source-reported evidence and do not answer RQ4.

G. Nomenclature

Symbol	Meaning
I_s	Four-modality MRI volume for subject s
v_i	Voxel coordinate
R	Region set {ET, TC, WT}
P^(0)	Initial coarse probability prior
P^(R)	Refined probability field
u_i	Uncertainty gate
c_ij	Edge conductance between neighboring voxels
L_c	Conductance-weighted graph Laplacian
C	Valid nested probability set
F^(l)	Feature tensor at pyramid level l
w_l	Nonnegative scale-fusion weight
Theta	Trainable parameters
L	Composite training loss
N	Number of voxels or tokens, as stated
K_l	Tokens in one attention window at level l

4. PROPOSED PVRSAP-NET

A. Architecture and Stage Alignment

PVRSAP-Net adheres to the fixed order shown in Fig. 1. This order is repeated throughout the procedures described in algorithms, equations, ablation studies and the interpreting the results. The model takes a four standardized MRI channels, plus three coarse region-probability channels. BE-PVR updates the probabilities. Adaptive voxel descriptors combine image and probability data. SAPFormer models spatial context at multiple scales. Residual decoding recovers detail. Three sigmoid heads are used to produce the ET, TC, and WT probabilities respectively, which are then followed by nested projection.

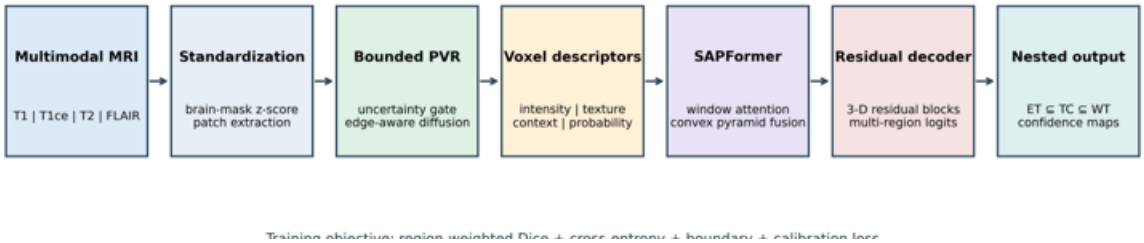

Fig. 1. PVRSAP-Net architecture. The diagram is an original schematic of the proposed stage order.

The model has seven functional stages:

1. multimodal standardization and patch sampling;
2. bounded edge-aware probabilistic voxel refinement;
3. adaptive intensity–texture–context–probability descriptors;
4. multiresolution window tokenization;
5. spatial-attention pyramid fusion;
6. residual three-dimensional decoding;
7. nested prediction, calibration, and composite optimization.

The synopsis supplies the core concepts: neighborhood probability mixing, Laplacian refinement, intensity/texture/context/probability descriptors, query–key–value attention, convex multiscale fusion, residual CNN learning, and Dice–cross-entropy optimization. PVRSAP-Net retains those concepts and changes the unconstrained portions that could produce invalid probabilities, boundary leakage, or inconsistent regions.

B. Bounded Edge-Aware Probabilistic Voxel Refinement

For each region r, construct a six- or 26-neighbor graph $G = (V, E)$. The fused structural intensity is

$$J_i = \sum_{m=1}^{M} a_m \hat{x}_{i,m}, \qquad a_m \geq 0, \quad \sum_{m=1}^{M} a_m = 1.$$

The conductance between adjacent voxels is

$$c_{ij} = \exp\left(-\frac{\| \hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j \|_2^2}{\kappa_x^2} - \frac{|J_i - J_j|^2}{\kappa_j^2} \right), \qquad (i,j) \in E.$$

Large multimodal differences reduce c_{ij}, limiting probability flow across a likely boundary. A minimum conductance $c_{\min}$ may be imposed to prevent disconnected numerical islands.

The uncertainty gate is

$$u_i^{(t)} = 4 p_i^{(t)} \left(1 - p_i^{(t)}\right),$$

which is zero at probabilities zero and one and maximal at 0.5. The gate focuses neighborhood correction on uncertain voxels and leaves high-confidence interiors closer to their prior value. The conductance-weighted neighborhood mean is

$$\bar{p}_i^{(t)} = \frac{\sum_{j\in\mathcal{N}_i} c_{ij}\, p_j^{(t)}}{\sum_{j\in\mathcal{N}_i} c_{ij} + \epsilon}.$$

The update before projection is

$$q_i^{(t+1)} = p_i^{(t)} + \beta\big[L_c \mathbf{p}^{(t)}\big]_i + \gamma u_i^{(t)}\big(\bar{p}_i^{(t)} - p_i^{(t)}\big),$$

where the sign convention defines $L_c\mathbf{p}$ as neighbor-minus-center diffusion. The implementation uses a normalized discrete operator so that a declared step-size bound applies. The interval update is

$$\tilde{\mathbf{p}}^{(t+1)} = \Pi_{[0,1]^N}\big(\mathbf{q}^{(t+1)}\big).$$

After independently forming the three region fields, a joint projection gives

$$\mathbf{P}^{(t+1)} = \Pi_{\mathcal{C}}\big(\tilde{\mathbf{p}}_{\mathrm{ET}}^{(t+1)}, \tilde{\mathbf{p}}_{\mathrm{TC}}^{(t+1)}, \tilde{\mathbf{p}}_{\mathrm{WT}}^{(t+1)}\big).$$

The iteration stops when

$$\frac{\| \mathbf{P}^{(t+1)} - \mathbf{P}^{(t)} \|_2}{\sqrt{3N}} < \delta \quad \text{or} \quad t = R_{\max}.$$

The ledger records the update norm, clipped voxel count, hierarchy violations before projection, numerical exceptions, and connected-component counts. If the update norm rises for three consecutive iterations, the algorithm returns the best finite iterate and flags the case.

Algorithm 1: Bounded Edge-Aware Probabilistic Voxel Refinement (BE-PVR)

Input: standardized volume $\hat{\mathbf{I}}$; initial region probabilities $\mathbf{P}^{(0)}$; neighborhood graph G; $\beta, \gamma, \kappa_x, \kappa_J, \delta, R_{\max}, \epsilon$.
Output: refined nested probabilities $\mathbf{P}^{(R)}$; audit ledger $\mathcal{A}$.

1. Validate finite inputs, modality order, probability range, and spatial shape.
2. Compute fused intensity J and conductances c_{ij} using (11)–(12).
3. Initialize $t \leftarrow 0$, $\mathbf{P}^{(t)} \leftarrow \Pi_{\mathcal{C}}\big(\mathbf{P}^{(0)}\big)$, and an empty ledger.
4. For each region $r \in \{\mathrm{ET}, \mathrm{TC}, \mathrm{WT}\}$, compute uncertainty $u_{i,r}^{(t)}$ from (13).
5. Compute conductance-normalized neighborhood means $\bar{p}_{i,r}^{(t)}$ from (14).
6. Update $q_{i,r}^{(t+1)}$ with diffusion and gated correction using (15).
7. Project every region to $[0,1]^N$ using (16); replace a nonfinite value by its previous finite value and record the event.
8. Jointly project ET, TC, and WT into $\mathcal{C}$ using (17).
9. Record RMS change, clipped count, hierarchy correction, and component count.
10. If RMS change is below δ, return the current probabilities and ledger.
11. If the update grows for three iterations, halve β and γ; restore the best finite iterate.
12. Set $t \leftarrow t + 1$ and repeat until $R_{\max}$.
13. Return $\mathbf{P}^{(R)}$ and $\mathcal{A}$.

Lines 1-3 guard the stage against malformed priors and introduce an nested initialization. Lines 4-6 run the synopsis's probabilistic neighborhood refinement with two modifications: uncertain voxels undergo larger correction, and cross-edge pairs are given lesser coupling. Lines 7-8 make probability and anatomy constraints explicit. 9–12 add a stopping diagnostic and a conservative recovery rule. The following ablation are on the removal side: "no conductance","no uncertainty gate" and "no hierarchy projection." The sensitivity to the parameters is changed by β, γ, κx and the number of iterations. A failure mode is a true lesion for which no intensity boundary exists; conductance cannot maintain an edge that the input fails to express.

Proposition 1 (Nonexpansive Bounded Refinement Under a Frozen Gate)

Let A be the linear operator obtained from a fixed conductance matrix and a fixed diagonal gate U, and let $T(\mathbf{p}) = \Pi_{\mathcal{C}}(A\mathbf{p} + \mathbf{b})$. If $\| A \|_2 \leq 1$, then T maps $\mathcal{C}$ to itself and is nonexpansive. If $\| A \|_2 = \rho < 1$, T is a contraction with a unique fixed point.

Derivation. For $\mathbf{p}, \mathbf{q} \in \mathcal{C}$,

$$\| T(\mathbf{p}) - T(\mathbf{q}) \|_2 = \| \Pi_{\mathcal{C}}(A\mathbf{p} + \mathbf{b}) - \Pi_{\mathcal{C}}(A\mathbf{q} + \mathbf{b}) \|_2,$$

$$\| T(\mathbf{p}) - T(\mathbf{q}) \|_2 \leq \| A(\mathbf{p} - \mathbf{q}) \|_2,$$

$$\| A(\mathbf{p} - \mathbf{q}) \|_2 \leq \| A \|_2 \| \mathbf{p} - \mathbf{q} \|_2,$$

$$\| T(\mathbf{p}) - T(\mathbf{q}) \|_2 \leq \rho \| \mathbf{p} - \mathbf{q} \|_2.$$

For the normalized Laplacian convention, write

$$A = I - \beta L_c - \gamma U(I - S_c),$$

where S_c is the conductance-normalized averaging operator. Since $\lambda(L_c) \subseteq [0,2]$ and $0 \preccurlyeq U \preccurlyeq I$,

$$\| I - \beta L_c \|_2 = \max_{\lambda \in [0,2]} |1 - \beta\lambda|,$$

and for $0 \leq \beta \leq 1$,

$$\| I - \beta L_c \|_2 \leq 1.$$

If the gated correction is averaged with a step satisfying $\gamma \| U(I - S_c) \|_2 \leq 1 - \| I - \beta L_c \|_2$, then

$$\| A \|_2 \leq \| I - \beta L_c \|_2 + \gamma \| U(I - S_c) \|_2 \leq 1.$$

Projection gives $T(\mathbf{p}) \in \mathcal{C}$. When the final inequality is strict, Banach's fixed-point theorem yields a unique frozen-gate fixed point and geometric convergence:

$$\| \mathbf{p}^{(t)} - \mathbf{p}^{\star} \|_2 \leq \rho^t \| \mathbf{p}^{(0)} - \mathbf{p}^{\star} \|_2.$$

The actual gate changes with $\mathbf{p}$, so the proposition does not prove global convergence of the nonlinear iteration. The runtime safeguard monitors the update norm and reduces the step if the frozen-operator condition is numerically violated.

The time cost of BE-PVR is $O(R|E||\mathcal{R}|)$, which becomes $O(RN|\mathcal{R}|)$ for a fixed-size 3D neighborhood. Conductance storage costs $O(|E|)$, but weights may be computed on the fly to reduce memory to $O(N|\mathcal{R}|)$.

C. Adaptive Multiscale Voxel Descriptors

The synopsis represents a voxel with local intensity, texture, context, and probability. PVRSAP-Net makes each term explicit. The local vector is

$$\mathbf{f}_i^{\text{int}} = \hat{\mathbf{x}}_i.$$

Texture begins with directional local differences and moments rather than a single undefined GLCM scalar:

$$\mathbf{f}_i^{\text{tex}} = \left[\mu_{\mathcal{N}_i}, \sigma_{\mathcal{N}_i}, \text{MAD}_{\mathcal{N}_i}, \| \nabla J_i \|_2, \text{LoG}(J_i), h_i^{(1)}, \ldots, h_i^{(K)}\right],$$

where $h_i^{(k)}$ are local co-occurrence statistics derived from quantized intensities. Haralick et al. established the use of co-occurrence features for texture [48]. Quantization bins, offsets, and physical distances are fixed from the training protocol. To avoid nondifferentiable overhead in the main network, the default implementation replaces explicit per-voxel GLCM with learned depthwise 3^3 and 5^3 convolutions initialized to gradient and Laplacian filters. A separate "explicit texture" ablation retains the synopsis's descriptor.

The contextual vector uses pooled neighborhoods:

$$\mathbf{f}_i^{\text{ctx}} = \oplus_{\rho \in \{1,2,4\}} \left[\text{AvgPool}_{\rho}(\hat{\mathbf{I}})_i, \text{MaxPool}_{\rho}(\hat{\mathbf{I}})_i\right],$$

and the probability descriptor is

$$\mathbf{f}_i^{\mathrm{prob}} = \left[\mathbf{P}_i^{(R)},\ 4\mathbf{P}_i^{(R)} \odot \left(1 - \mathbf{P}_i^{(R)}\right),\ \nabla \mathbf{P}_i^{(R)}\right].$$

The concatenated descriptor is

$$\mathbf{f}_i = \mathrm{LN}\left(W_f\left[\mathbf{f}_i^{\mathrm{int}} \,\|\, \mathbf{f}_i^{\mathrm{tex}} \,\|\, \mathbf{f}_i^{\mathrm{ctx}} \,\|\, \mathbf{f}_i^{\mathrm{prob}}\right] + \mathbf{b}_f\right),$$

where LN is layer normalization and W_f maps the variable descriptor to d_0 channels. A learned positional encoding is added after patch embedding. This stage does not label each voxel independently; it prepares local evidence for multiscale attention.

D. Spatial-Attention Pyramid Former

At pyramid level ℓ, a strided convolution or average pooling maps the descriptor tensor to $\mathbf{F}^{(\ell)} \in \mathbb{R}^{N_\ell \times d_\ell}$. The volume is partitioned into windows of K_ℓ tokens. Within a window,

$$\mathbf{Q}_h^{(\ell)} = \mathbf{F}^{(\ell)} W_{Q,h}^{(\ell)}, \quad \mathbf{K}_h^{(\ell)} = \mathbf{F}^{(\ell)} W_{K,h}^{(\ell)}, \quad \mathbf{V}_h^{(\ell)} = \mathbf{F}^{(\ell)} W_{V,h}^{(\ell)}.$$

The attention head is

$$\mathbf{A}_h^{(\ell)} = \mathrm{softmax}\left(\frac{\mathbf{Q}_h^{(\ell)}\left(\mathbf{K}_h^{(\ell)}\right)^{\top}}{\sqrt{d_h}} + \mathbf{B}_h^{(\ell)}\right),$$

$$\mathbf{Z}_h^{(\ell)} = \mathbf{A}_h^{(\ell)} \mathbf{V}_h^{(\ell)}.$$

Shifted windows in alternating blocks permit cross-window exchange. The head outputs are concatenated, projected, and added to a residual path:

$$\mathbf{Z}^{(\ell)} = \mathbf{F}^{(\ell)} + W_O^{(\ell)} \mathrm{Concat}_{h=1}^{H_a} \mathbf{Z}_h^{(\ell)}.$$

Each scale is restored to the finest decoder grid by a learned upsampler U_ℓ. The fusion logits a_ℓ give simplex weights

$$w_\ell = \frac{\exp(a_\ell/\tau)}{\sum_{k=1}^{L} \exp\left(a_k/\tau\right)}, \qquad \sum_{\ell=1}^{L} w_\ell = 1, \quad w_\ell \geq 0,$$

and

$$\mathbf{F}_{\mathrm{SAP}} = \sum_{\ell=1}^{L} w_\ell \, U_\ell\left(\mathbf{Z}^{(\ell)}\right).$$

The scale temperature τ is bounded below by 0.1 to avoid near-discrete collapse during early training. Entropy regularization may be applied to w for the first training phase, then annealed.

Algorithm 2: Spatial-Attention Pyramid Fusion (SAPFormer)

Input: descriptor tensor **F**; pyramid scales $\{s_\ell\}_{\ell=1}^{L}$; window sizes $\{K_\ell\}$; attention parameters; scale logits **a**; temperature τ.
Output: fused representation $\mathbf{F}_{\mathrm{SAP}}$; scale ledger $\mathcal{S}$.

1. Validate that each spatial dimension supports the chosen scale and window; pad reflectively when required.
2. For level $\ell = 1, \ldots, L$, construct $\mathbf{F}^{(\ell)}$ by strided convolution and normalization.
3. Partition $\mathbf{F}^{(\ell)}$ into 3D windows and add positional bias.
4. Compute query, key, and value projections for every head using (33).
5. Compute scaled window attention and head outputs using (34)–(35).
6. Merge heads, apply output projection, residual addition, and feed-forward refinement using (36).

7. Shift the window partition in the next block; reverse padding after attention.
8. Upsample each scale output $U_\ell(\mathbf{Z}^{(\ell)})$ to the decoder grid.
9. Compute simplex weights w_ℓ with temperature-bounded softmax using (37).
10. Fuse levels through the convex sum in (38).
11. Record scale weights, attention entropy, activation norms, and padding.
12. If a nonfinite activation occurs, skip the affected residual branch and return the finite identity path with an error flag.
13. Return $\mathbf{F}_{\mathrm{SAP}}$ and $\mathcal{S}$.

Line 1–3 are a valid multiresolution window. Lines 4–7 describe local attention and cross-window interaction. Line 8–10 re-grid and fuse scales under a convex constraint . Line 11 provides diagnostics for scale collapse or attention saturation. Line 12 stops a single nonfinite branch from invalidating the entire volume. The main ablation substitutes SAPFormer by equal-weight convolutional pyramid pooling. Sensitivity on parameters by changing window size, number of levels, attention heads and fusion temperature.

Proposition 2 (Bounded Convex Pyramid Output)

Suppose every restored scale feature satisfies $\| U_\ell(\mathbf{Z}^{(\ell)}) \|_F \le B_\ell$, and the weights in (37) lie on the simplex. Then the fused representation satisfies $\| \mathbf{F}_{\mathrm{SAP}} \|_F \le \sum_\ell w_\ell B_\ell \le \max_\ell B_\ell$.

Derivation.

$$\mathbf{F}_{\mathrm{SAP}} = \sum_{\ell=1}^{L} w_\ell \, U_\ell(\mathbf{Z}^{(\ell)}),$$

$$\| \mathbf{F}_{\mathrm{SAP}} \|_F = \left\| \sum_{\ell=1}^{L} w_\ell \, U_\ell(\mathbf{Z}^{(\ell)}) \right\|_F,$$

$$\| \mathbf{F}_{\mathrm{SAP}} \|_F \le \sum_{\ell=1}^{L} \| w_\ell U_\ell(\mathbf{Z}^{(\ell)}) \|_F,$$

$$\sum_{\ell=1}^{L} \| w_\ell U_\ell(\mathbf{Z}^{(\ell)}) \|_F = \sum_{\ell=1}^{L} w_\ell \, \| U_\ell(\mathbf{Z}^{(\ell)}) \|_F,$$

$$\sum_{\ell=1}^{L} w_\ell \, \| U_\ell(\mathbf{Z}^{(\ell)}) \|_F \le \sum_{\ell=1}^{L} w_\ell \, B_\ell,$$

$$\sum_{\ell=1}^{L} w_\ell \, B_\ell \le \left(\max_k B_k \right) \sum_{\ell=1}^{L} w_\ell,$$

$$\left(\max_k B_k \right) \sum_{\ell=1}^{L} w_\ell = \max_k B_k,$$

$$\therefore \quad \| \mathbf{F}_{\mathrm{SAP}} \|_F \le \max_\ell B_\ell.$$

If each restored branch is K_ℓ-Lipschitz with respect to the input and the fusion weights are treated as fixed for the bound, then

$$\| \mathbf{F}_{\mathrm{SAP}}(\mathbf{X}) - \mathbf{F}_{\mathrm{SAP}}(\mathbf{Y}) \|_F \le \sum_{\ell=1}^{L} w_\ell \, K_\ell \, \| \mathbf{X} - \mathbf{Y} \|_F \le K_{\max} \, \| \mathbf{X} - \mathbf{Y} \|_F.$$

The proposition supports controlled fusion; it does not bound the entire trained network unless every preceding and following operator has a known finite Lipschitz constant.

Window attention at level ℓ costs $O(N_\ell K_\ell d_\ell + N_\ell d_\ell^2)$, compared with $O(N_\ell^2 d_\ell)$ for global attention. Peak attention memory is $O(N_\ell K_\ell)$. The pyramid sum adds the cost across levels. This cost remains material for 128^3 patches and motivates mixed precision, checkpointing, and bounded window sizes.

E. Residual Volumetric Decoder and Nested Prediction

The fused feature enters a residual decoder. At decoder level k, the feature is

$$\mathbf{R}_k = \phi\big(\mathrm{Norm}\big[H_k(\mathbf{R}_{k+1}) + G_k(\mathbf{E}_k) + U_k\big(\mathbf{P}^{(R)}\big)\big]\big),$$

where H_k upsamples the deeper decoder feature, G_k gates the matching encoder feature, U_k embeds the refined prior at that resolution, Norm is group or instance normalization for small batches, and ϕ is GELU or leaky ReLU. A residual block has

$$\mathrm{ResBlock}(\mathbf{X}) = \mathbf{X} + W_2\phi\big(\mathrm{Norm}(W_1\mathbf{X})\big).$$

Group normalization is preferred when batch size is one or two; batch normalization [34] is retained only when the effective batch supports stable statistics. This choice is fixed during development.

Three output heads produce logits $\mathbf{z}_r$:

$$\mathbf{z}_r = W_r\mathbf{R}_0 + b_r, \qquad \hat{\mathbf{p}}_r = \sigma(\mathbf{z}_r/T_r),$$

where $T_r > 0$ is a region-specific temperature estimated on $\mathcal{D}_{\mathrm{cal}}$. During training, $T_r = 1$. At inference, temperatures adjust confidence but do not change ranking. Nested projection gives the released probability field

$$\widehat{\mathbf{P}} = \Pi_{\mathcal{C}}(\hat{\mathbf{p}}_{\mathrm{ET}}, \hat{\mathbf{p}}_{\mathrm{TC}}, \hat{\mathbf{p}}_{\mathrm{WT}}).$$

The binary mask at threshold τ_r is

$$\hat{Y}_{i,r} = \mathbb{1}\big[\hat{p}_{i,r} \geq \tau_r\big],$$

where τ_r is chosen on $\mathcal{D}_{\mathrm{cal}}$. The primary analysis uses $\tau_r = 0.5$ unless calibration shows a prespecified benefit. A component-size filter is permitted only when its threshold is fixed on training or calibration data and is applied to every baseline under the same rule.

F. Composite Objective

For region r, the smoothed soft Dice loss is

$$\mathcal{L}_{\mathrm{Dice},r} = 1 - \frac{2\sum_i y_{i,r}\,\hat{p}_{i,r} + \epsilon}{\sum_i y_{i,r} + \sum_i \hat{p}_{i,r} + \epsilon}.$$

The weighted binary cross-entropy is

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N|\mathcal{R}|}\sum_{i,r}\big[\omega_r^+ y_{i,r}\log\big(\hat{p}_{i,r} + \epsilon\big) + \omega_r^-\big(1 - y_{i,r}\big)\log\big(1 - \hat{p}_{i,r} + \epsilon\big)\big].$$

The generalized Dice formulation [35] may replace the mean in (53) when an agreed class-weighting rule is used. Focal loss [36] is a planned sensitivity analysis rather than part of the primary objective.

Let $\phi_{i,r}$ be a signed distance transform of the reference boundary, clipped to $[-d_{\max}, d_{\max}]$. A boundary term is

$$\mathcal{L}_{\mathrm{bd}} = \frac{1}{N|\mathcal{R}|}\sum_{i,r}\hat{p}_{i,r}\,\phi_{i,r},$$

following the principle of distance-aware boundary supervision [37]. For an empty reference region, the term uses a defined zero map and the CE term handles false positives.

The Brier calibration term is

$$\mathcal{L}_{\mathrm{Br}} = \frac{1}{N|\mathcal{R}|}\sum_{i,r}\left(\hat{p}_{i,r} - y_{i,r}\right)^2.$$

An explicit hierarchy penalty before projection is

$$\mathcal{L}_{\mathrm{hier}} = \frac{1}{N}\sum_{i}\left[\mathrm{ReLU}\left(\hat{p}_{i,\mathrm{ET}} - \hat{p}_{i,\mathrm{TC}}\right) + \mathrm{ReLU}\left(\hat{p}_{i,\mathrm{TC}} - \hat{p}_{i,\mathrm{WT}}\right)\right].$$

This term teaches the raw heads to respect nesting instead of relying entirely on inference projection. The total loss is

$$\mathcal{L}_{\mathrm{total}} = \lambda_D \frac{1}{|\mathcal{R}|}\sum_{r}\mathcal{L}_{\mathrm{Dice},r} + \lambda_C\mathcal{L}_{\mathrm{CE}} + \lambda_B\mathcal{L}_{\mathrm{bd}} + \lambda_R\mathcal{L}_{\mathrm{Br}} + \lambda_H\mathcal{L}_{\mathrm{hier}} + \lambda_W \parallel \Theta \parallel_2^2.$$

The source summary was based on Dice plus cross-entropy and weight regularization. (58) preserves these terms and adds boundary, calibration and hierarchy terms associated with declared end-points. Loss weights are chosen within nested validation. They are not adjusted based on the test set.

G. End-to-End Training

Training alternates between stages no separate optimization stages are performed after freezing the coarse localizer. BE-PVR can be realized as a differentiable unroll layer or as precomputation outside the net- work. The main protocol precomputes the refined prior from a frozen coarse localizer, in order to isolate its effect and to save memory. A secondary end-to-end scheme unrolls a handful of refinement iterations.

Algorithm 3: Leakage-Resistant PVRSAP-Net Training and Calibrated Inference

Input: subject-partitioned dataset; frozen coarse localizer f_{θ_0}; model F_Θ; folds; optimizer settings; loss weights; random seed.
Output: trained parameters $\Theta^\star$; temperatures $\mathbf{T}$; thresholds $\boldsymbol{\tau}$; test predictions; audit record.

1. Fix seeds; validate subject-level partition disjointness and input hashes.
2. Fit clipping, normalization policy, and optional bias-correction configuration using training data only.
3. For each training subject, obtain $\mathbf{P}^{(0)} = f_{\theta_0}(\hat{\mathbf{I}})$; run Algorithm 1 and store the ledger.
4. Initialize PVRSAP-Net parameters with a recorded seed; set optimizer, scheduler, mixed-precision policy, and gradient limit G.
5. For every epoch, sample tumor-aware and background patches from training subjects only.
6. Construct descriptors by (28)–(32); run Algorithm 2; decode probabilities by (48)–(51).
7. Compute the composite loss in (58), including defined empty-region handling.
1. Backpropagate; replace a nonfinite loss by a skipped update; clip the gradient to norm G.
2. Update Θ with Adam or AdamW; record loss terms, learning rate, gradient norm, memory, and time.
3. Evaluate subject-level validation metrics; save the checkpoint selected by a prespecified rank rule.
4. Stop after the epoch budget or patience limit; never inspect test labels.
5. On the calibration partition, estimate region temperatures T_r and any nondefault thresholds τ_r.
6. Lock code, configuration, checkpoint hash, and post-processing.
7. Run test inference once with fixed sliding-window overlap and nested projection; save probabilities before masks.
8. Compute subject-level metrics, bootstrap confidence intervals, calibration, resource use, and failure summaries.
9. Return the model, calibration values, predictions, and audit record.

Lines 1–3 define the data boundary and render the prior in the probability reproducible. Lines 4–10 establish the criteria for optimization and checkpoint selection. Lines 11–13 separate calibration from testing and introduce an analysis lock. Lines 14–16 free probability maps, masks, metrics, uncertainty and a traceable record. The ablation removal freezes all parameters and substitutes BE-PVR with the raw prior, SAPFormer with convolutional pyramid pooling, or the composite loss with Dice–CE.A failure case is an out-of-distribution acquisition, where all four modalities are outside the development range; the degradation is reported, not hidden, by protocol.

Proposition 3 (Conditional Descent for a Clipped Gradient Step)

Let $\mathcal{L}$ be differentiable and L-smooth in a neighborhood of Θ_t. Let $\mathbf{g}_t = \nabla\mathcal{L}(\Theta_t)$, and let the clipped direction be $\tilde{\mathbf{g}}_t = \alpha_t \mathbf{g}_t$, where $\alpha_t = \min(1, G/\| \mathbf{g}_t \|_2)$. For $0 < \eta_t < 2/(L\alpha_t)$, the update $\Theta_{t+1} = \Theta_t - \eta_t \tilde{\mathbf{g}}_t$ does not increase $\mathcal{L}$.

Derivation. The descent lemma gives

$$\mathcal{L}(\Theta_{t+1}) \le \mathcal{L}(\Theta_t) + \langle \nabla\mathcal{L}(\Theta_t), \Theta_{t+1} - \Theta_t \rangle + \frac{L}{2} \| \Theta_{t+1} - \Theta_t \|_2^2.$$

Substitute the update:

$$\mathcal{L}(\Theta_{t+1}) \le \mathcal{L}(\Theta_t) - \eta_t \langle \mathbf{g}_t, \tilde{\mathbf{g}}_t \rangle + \frac{L\eta_t^2}{2} \| \tilde{\mathbf{g}}_t \|_2^2.$$

Since $\tilde{\mathbf{g}}_t = \alpha_t \mathbf{g}_t$,

$$\langle \mathbf{g}_t, \tilde{\mathbf{g}}_t \rangle = \alpha_t \| \mathbf{g}_t \|_2^2,$$
$$\| \tilde{\mathbf{g}}_t \|_2^2 = \alpha_t^2 \| \mathbf{g}_t \|_2^2.$$

Hence

$$\mathcal{L}(\Theta_{t+1}) - \mathcal{L}(\Theta_t) \le -\eta_t \alpha_t \| \mathbf{g}_t \|_2^2 + \frac{L\eta_t^2 \alpha_t^2}{2} \| \mathbf{g}_t \|_2^2,$$

$$\mathcal{L}(\Theta_{t+1}) - \mathcal{L}(\Theta_t) \le -\eta_t \alpha_t \left(1 - \frac{L\eta_t \alpha_t}{2}\right) \| \mathbf{g}_t \|_2^2.$$

If $0 < \eta_t < 2/(L\alpha_t)$, then

$$1 - \frac{L\eta_t \alpha_t}{2} > 0,$$

so

$$\mathcal{L}(\Theta_{t+1}) \le \mathcal{L}(\Theta_t).$$

For $\eta_t \le 1/(L\alpha_t)$, the decrease is at least

$$\mathcal{L}(\Theta_t) - \mathcal{L}(\Theta_{t+1}) \ge \frac{\eta_t \alpha_t}{2} \| \mathbf{g}_t \|_2^2.$$

Conditional result applies to exact clipped gradient and local smoothness interval. Adam [33], stochastic batches, mixed precision, and nonconvexity are at odds with elements of the deterministic setting. The training ledger verifies finite losses and gradient norms; validation selects the chosen checkpoint.Neither global convergence nor global optimality is claimed.

H. Computational Complexity

Let N be the voxels in a patch, C_k the channels at decoder level k, q_k the kernel width, R the BE-PVR iterations, L the attention scales, K_ℓ the tokens per window, and d_ℓ the attention width. The dominant costs are summarized below.

Stage	Time complexity	Peak activation or state memory	Main control

Standardization	O(MN)	O(MN)	brain mask, clipping
BE-PVR	O(RN|R|)	O(N|R|)	iterations, neighborhood
Descriptor convolutions	O(N d0 q^3)	O(N d0)	channels, kernel
SAPFormer	sum_l O(N_l K_l d_l + N_l d_l^2)	sum_l O(N_l K_l)	windows, levels, heads
Residual decoder	sum_k O(N_k C_k^2 q_k^3)	sum_k O(N_k C_k)	channels, patch
Three output heads	O(N C0 |R|)	O(N |R|)	regions

Inference applies sliding windows if a full volume is larger than memory. Overlap weighting softens seams at the expense of extra cost. The resulting paper should provide the number of parameters, multiply–accumulate operations, peak allocated GPU memory, per-volume latency, and throughput on the specified hardware. Those measures are not in the provided materials, and are still in the works.

I. Numerical Safeguards and Failure Conditions

Each division is calculated using ϵ. Each probability is clipped prior to taking logarithms. Empty reference regions are assigned a fixed Dice and boundary treatment. Image and label orientation is verified following resampling. Nonfinite activations cause a skipped residual branch at runtime and a hard crash at locked testing. A modality-order hash disallows silent T1/T1ce/T2/FLAIR permutation. The test stage saves probabilities prior to thresholding so that the filtering of components can be audited.

BE-PVR is prone to breakdown when the initial prior misses an entire lesion, as local refinement is unable to generate evidence away from nonzero support. It can retain a spurious edge caused by an artifact. Hierarchy projection may also strengthen consistency but spread a WT error to the admissible set. SAPFormer may fail to detect a structure smaller than its patch embedding or overfit a site-specific global pattern. Residual skip paths can reintroduce background. Calibration drifts under domain shift. Each condition corresponds to a planned diagnostic: a prior-recall analysis, an artifact stress testing, a projection correction volume, a small-lesion stratification, an attention entropy, and a site-wise reliability.

5. EXPERIMENTAL SETUP AND EVALUATION PROTOCOL

A. Evidence Separation

The evaluation uses three evidence categories. First, the uploaded synopsis is a source document for the research objective, architecture concepts, proposed BraTS 2020 protocol, modified V-Net dimensions, optimizer settings, and preliminary charts. The charts lack raw predictions, subject counts, folds, uncertainty, and baseline implementation details; they are not used as validated head-to-head measurements. Second, a controlled simulation executes the BE-PVR portion on generated data. It supports a statement about internal algorithm behavior under declared assumptions. Third, recent-paper values are transcribed from primary publications or preprints and labeled source-reported. They describe prior studies under their own protocols.

No recorded BraTS or ATLAS volume was attached. No trained checkpoint, code repository, patient-level prediction, or experiment log was supplied. The recorded-data portion of this section is a complete protocol, not a completed experiment. The abstract and conclusion retain that boundary. Table II summarizes the evidence.

Evidence item	Data availability	Permitted conclusion	Prohibited conclusion
Uploaded synopsis	Text, equations, images, charts	source proposes the listed stages and protocol	charts prove clinical or benchmark superiority
Controlled simulation	generated volumes, seed, script, CSV outputs	refinement behavior under the synthetic process	patient efficacy, BraTS rank, clinical validity

Recent papers	publication-level reported values	literature reports a value under its protocol	a common leaderboard across protocols
Recorded BraTS study	planned only	protocol can test H3	PVRSAP-Net outperforms a recorded-data baseline
ATLAS transfer	planned only	protocol can test domain transfer after retraining	glioma training works directly for stroke

B. Planned Recorded-Data Cohorts

The main development cohort is BraTS 2021 when available under the data-use terms [4]. BraTS 2020 can be employed for reproduction of synopsis configuration and for investigation of temporal benchmark drift. There is a clear manifest mapping the four MRI modalities. The conversion of the label yields ET, TC, and WT. Participants with unreadable files, contradictory affine metadata, missing necessary modalities, or invalid labels are not quietly excluded but instead are reported in a PRISMA diagram.

A locked subject-level five-fold split for model selection is suggested. After stratification by tumor-volume quartile and site (subjects may have more than one), ten percent of subjects from each outer training fold act as calibration subset. The official validation or test server, if available, continues to be an external final evaluation. When only the official training labels are available, a hidden 20% test split is generated once prior to development. Splits are stored as lists of subject IDs and hashed.

The additional external-shift study is conducted using either BraTS 2020, BraTS 2023 GLI, or an alternate institutional dataset, but not pooled with development data. The ATLAS study [5], [6] constitutes a separate experiment with a one-channel T1 input adapter, a stroke-lesion output, and a newly trained network. What is common to ask is whether BE-PVR can refine an uncertain lesion probability map, and not whether glioma is semantics transmutable.

C. Preprocessing and Patch Construction

The benchmark volumes are verified for orientation, affine consistency, voxel spacing and modality alignment. If a BraTS version is already registered and skull-stripped, the pipeline will not apply a second rigid registration. Optional N4 correction [31] is compared with a preprocessing ablation. Within-brain z-score normalization follows (5). Intensities are clipped to training derived percentiles, each modality at a time.

The synopsis is to propose to crop non informative edges slices and downscale from 240×240×155 to 1283. Direct anisotropic resizing, however, could distort anatomy and physical distance. The main protocol, in contrast, resamples to the prescribed spacing, crops around the nonzero brain bounding box, and samples 1283 training patches. A source-faithful 1283 resize is preserved as a sensitivity condition. Third-order spline interpolation is used for images and nearest neighbor for labels. Each transformation is sufficiently invertible to map predictions back to the reference grid.

Patch sampling is implemented with three queues: tumor-centered, boundary-centered, and random brain patches. A background prior of 0:4=0:4=0:2 mixture is default to keep background representation without overpowering. The boundary queue is generated from training labels only. Augmentation is composed of random flips, small rotations, scaling, gamma change, Gaussian or Rician-like noise, smoothed bias field, modality dropout. The same random parameter distribution is used for all baselines unless an official method specifies a documented difference.

D. Coarse Prior Construction

The preferred coarse prior is a frozen compact 3D U-Net trained on the same training subjects. It produces ET, TC, and WT probabilities before thresholding. A second condition uses the K-maximum-likelihood probabilistic segmenter from the broader thesis once its code is available. The prior model is trained inside each outer fold; out-of-fold probabilities are required for training PVRSAP-Net. Generating a prior with a model trained on the same target subject's label would leak information.

Four prior conditions isolate dependence:

1. oracle-free recorded prior from the frozen coarse model;
2. degraded prior with controlled noise and false-positive components;
3. raw prior without BE-PVR;
4. zero or constant prior to test whether PVRSAP-Net merely becomes a standard image model.

Prior recall is measured at permissive thresholds. If the coarse prior has zero support around a missed lesion, BE-PVR is not expected to recover it. That case is reported separately.

The synopsis proposes a modified V-Net with 128^3 input, filters $\{8,16,32,64,128\}$, 5^3 kernels, batch size two, initial learning rate 5×10^{-4}, 100 epochs, Adam, learning-rate decay, and dropout between 0.1 and 0.3. PVRSAP-Net adopts the patch size and initial learning rate as starting values but compares 3^3 and 5^3 kernels. Large 5^3 kernels have 4.63 times as many kernel elements as 3^3 kernels, so they need a measured accuracy–memory benefit.

The initial proposed configuration uses encoder widths $\{24,48,96,192\}$, group normalization, GELU, two residual blocks per resolution, three SAPFormer levels, four attention heads, windows of 4^3 or 6^3, and six BE-PVR iterations. The decoder mirrors the encoder. Deep supervision is applied at two intermediate scales with weights 0.25 and 0.125. The main output weight is one. Dropout is 0.1 in the decoder and attention-dropout is 0.0 unless validation supports a change.

Loss weights start at $\lambda_D = 1$, $\lambda_C = 1$, $\lambda_B = 0.1$, $\lambda_R = 0.05$, $\lambda_H = 0.1$, and $\lambda_W = 10^{-5}$. The boundary weight rises linearly from zero during the first ten epochs to avoid unstable early contours. Hyperparameters are selected by nested validation rank over mean Dice, HD95, and ECE. A single metric cannot dictate every choice.

F. Training Environment and Reproducibility

The source synopsis names an NVIDIA A100 40 GB GPU through Google Colab Pro Plus. That statement is treated as the proposed environment. The final run must record the actual GPU model, count, driver, CUDA, cuDNN, operating system, Python, framework, MONAI version, mixed-precision mode, CPU, RAM, and storage. MONAI supplies tested medical-imaging transforms and sliding-window inference when used [45]. Adam [33] or AdamW is run with gradient clipping. The primary epoch budget is 300 with patience 50; a 100-epoch source-faithful condition tests whether the longer budget changes conclusions.

G. Baselines

The planned reproduced baselines are:

3D U-Net [8], representing a direct volumetric encoder-decoder;

V-Net [9], matching the source synopsis's comparison;

nnU-Net [15], representing a strong self-configured convolutional pipeline;

TransBTS [17], representing transformer bottleneck fusion;

Swin UNETR [19], representing hierarchical window attention;

MedNeXt [21], representing large-kernel convolution;

SegMamba [23], representing state-space long-range modeling;

PVRSAP-Net without BE-PVR;

PVRSAP-Net without SAPFormer;

full PVRSAP-Net.

All reproduced methods use the same subject folds, label conversion, modality manifest, evaluation code, and output grid. Official code and documented configurations receive priority. A method whose code cannot be reproduced is moved to the source-reported table. Ensembles and single models occupy separate rows.

H. Metrics and Statistical Plan

For binary reference Y_r and prediction $\hat{Y}_r$,

$$\mathrm{Dice}_r = \frac{2|Y_r \cap \hat{Y}_r|}{|Y_r| + |\hat{Y}_r|}, \qquad \mathrm{IoU}_r = \frac{|Y_r \cap \hat{Y}_r|}{|Y_r \cup \hat{Y}_r|}.$$

Sensitivity and specificity are

$$\mathrm{Sens}_r = \frac{\mathrm{TP}_r}{\mathrm{TP}_r + \mathrm{FN}_r}, \qquad \mathrm{Spec}_r = \frac{\mathrm{TN}_r}{\mathrm{TN}_r + \mathrm{FP}_r}.$$

HD95 is the 95th percentile of the bidirectional surface distances computed on the original physical grid. The empty-label rules match the benchmark and are reported below. Brier score, negative log likelihood, adaptive region-balanced ECE, reliability curves and risk–coverage are considered for calibration. Efficiency is reported in terms of the parameter count and the number of FLOPs or MACs under a specified counter, peak memory, latency, and throughput.

The unit of independent sampling is the subject. Means have 95% subject-bootstrap confidence intervals with a minimum of 2000 resamples. Model comparisons are pairwise and use subject-paired Dice and HD95 differences. When differences are nonnormal, a Wilcoxon signed-rank test is prespecified; Holm correction is applied to control for multiple regional comparisons. The effect size is reported as the median paired difference with bootstrap interval. Magnitude of effects should not be substituted for p-values. Unless otherwise stated, all site-wise and tumor-volume-quartile analyses are descriptive, and are conducted only if sample sizes allow prespecified statistical tests.

I. Controlled Simulation

2 Algorithm 1 is tested independently in a controlled simulation. It generates 20 synthetic 483 volumes with seed 20260727. Each volume is nested irregular combination of ellipsoids for WT, TC and ET. Four image channels with modality-specific contrasts, smooth multiplicative-style bias fields, locally correlated textures, and bounded noises are produced. This setting generates MRI-like contrast patterns but no realistic anatomy. Fig. 2 shows a generated case.

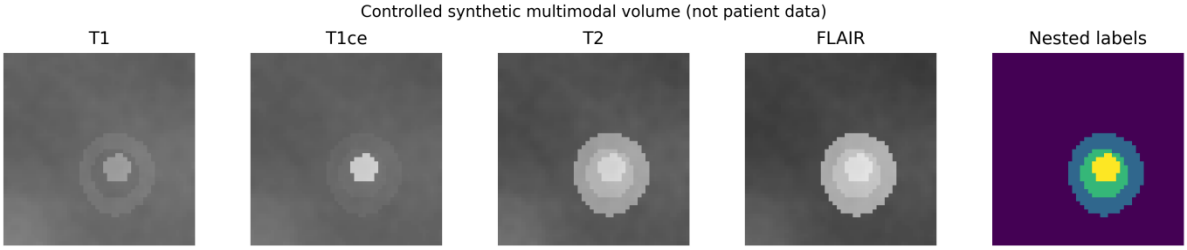

Fig. 2. Controlled synthetic multimodal volume and nested labels. These are generated arrays, not patient images.

For region r, let $d_r(v)$ be positive inside the region and negative outside. The uncorrupted probability is

$$p_r^{\star}(v) = \sigma(d_r(v)/t_r).$$

The corrupted logit is

$$\ell_r(v) = d_r(v)/t_r + \xi_r(v) + 0.22\, b(v),$$

where ξ_r is a standardized correlated random field and b is a smooth bias field. Region-specific noise amplitudes are 0.45, 0.62, and 0.78 for WT, TC, and ET. Between two and four false-positive Gaussian blobs are added. The initial prior is $\sigma(\ell_r)$ after blob insertion.

Four methods receive identical priors:

1. raw prior;
2. one-step neighborhood averaging with prior weight 0.55;
3. ten-step isotropic diffusion;

4. six-step PVRSAP-Net BE-PVR with $\beta = 0.04$, $\gamma = 0.18$, $\kappa = 0.55$, and nested projection.

A connected-component threshold of 32 voxels is applied equally to binary masks. It is a declared simulation rule. Metrics are computed for every case and region. The ablation removes conductance, uncertainty gating, or hierarchy projection. Sensitivity varies β from 0 to 0.08. The script saves case metrics, summaries, ablations, sensitivity values, calibration bins, environment metadata, and figures.

Simulation parameter	**Value**
Seed	20260727
Number of volumes	20
Spatial grid	48 × 48 × 48
Image channels	4
Regions	WT, TC, ET
BE-PVR iterations	6
	0.04
	0.18
Edge scale	0.55
Mask threshold	0.5
Minimum component	32 voxels
Evidence level	controlled simulation

The simulation performs a parameter-recovery sanity check: the generated labels are nested, the prior is bounded, projection removes hierarchy violations, and repeated runs with the same environment and seed reproduce the CSV values. It does not contain MRI acquisition physics, patient anatomy, rare multifocal lesions, surgical cavities, scanner shift, or expert annotation variation.

6. RESULTS AND INTERPRETATION

A. Qualitative Refinement Behavior

Figure 3 shows a single WT slice. The raw probability is along the desired region but has a jagged and uncertain border. Local variation is smoothed by mean filtering in the Neighb and isotropic diffusion. The PVRSAP update preserves a sharper edge-homogeneity transition in regions where the image exhibits a contrast edge. The error map focuses residual error at the outer boundary. This example demonstrates a technique; it is not proof that the same visual pattern appears on a patient scan.

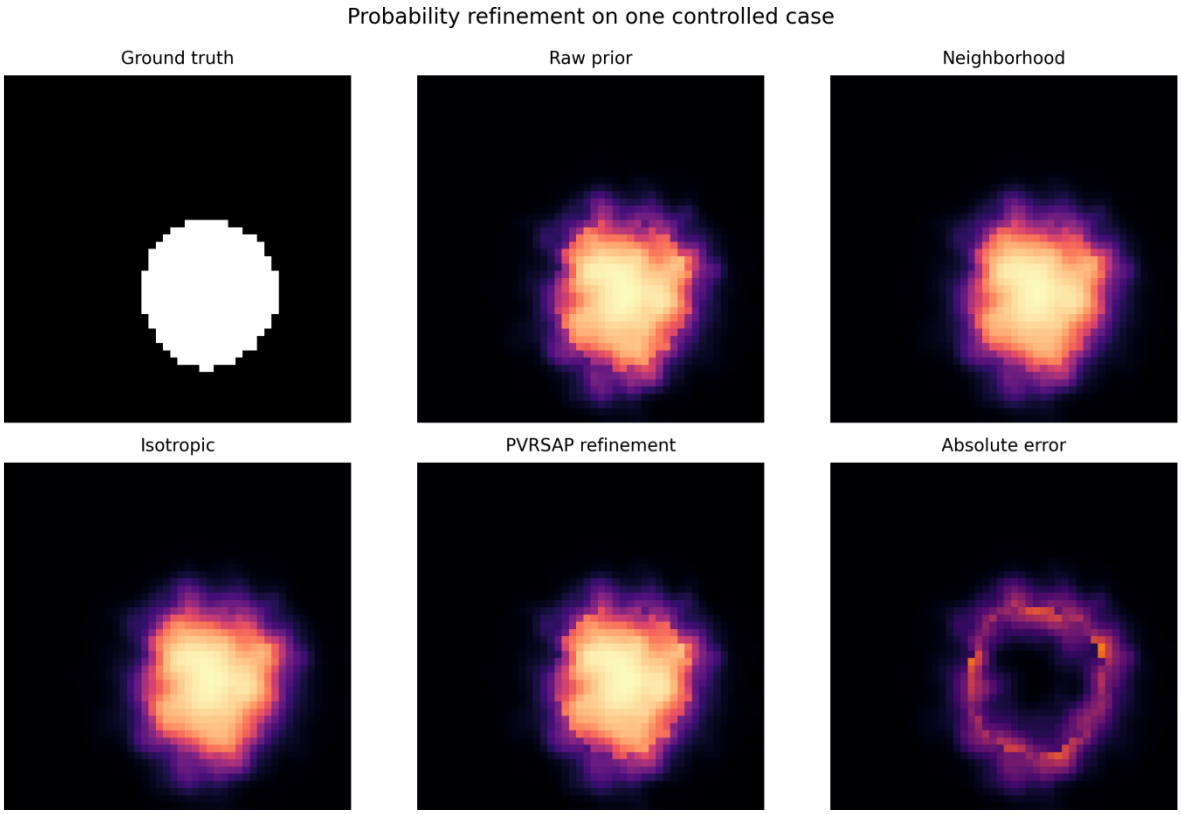

Fig. 3. Probability refinement for one controlled case. All panels arise from the seeded synthetic process.

The example reveals an anticipated limitation. BE-PVR treats the image gradient as evidence. When a synthetic boundary aligns with an image edge, conductance approaches unity along the contour. In case of disagreement, the update may retain the wrong feature. The analysis of recorded data must stratify the cases according to local contrast, and then compare predicted errors with modality gradients.

B. Primary Controlled-Simulation Results

Table IV gives case–region means pooled. Here pooling means each of 20 cases provides three region observations; no pooling of voxels occur. PVRSAP-Net obtained Dice 0.9091, compared with 0.8551 for the raw prior, 0.8722 for neighborhood averaging, and 0.8634 for isotropic diffusion. The difference of Dice from the raw prior was 0.0540. The IoU improved from 0.7625 to 0.8382. Sensitivity was almost identical to the but prior one, 0.9289 versus 0.9290, suggesting that the change of overlap was mainly driven by fewer false-positive or boundary errors, rather than higher recall.

Method	Dice	IoU	Sensitivity	Specificity	HD95 (voxels)	Brier	ECE
Raw prior	0.8551	0.7625	0.9290	0.9983	7.1811	0.0046	0.0158
Neighborhood averaging	0.8722	0.7837	0.9098	0.9986	5.0885	0.0047	0.0163
Isotropic diffusion	0.8634	0.7699	0.8960	0.9985	4.9216	0.0048	0.0161
PVRSAP-Net refinement	0.9091	0.8382	0.9289	0.9989	1.2316	0.0040	0.0142

The HD95 decreased from 7.1811 at voxels for the raw prior to 1.2316 voxels for PVRSAP-Net. The dramatic reduction is due to the elimination of distant components and the nested correction in the generated environment. Neighborhood averaging and isotropic diffusion reduced HD95 at the cost ofinsensitivity (loss of sensitivity). Specificity was above 0.998 for all methods, which is partly explained by the large synthetic background. This ceiling causes specificity to be a poor discriminator in this experiment.

The Brier score decreased from 0.0046 to 0.0040. Standard ECE decreased from 0.0158 to 0.0142. Both values are influenced by background prevalence. They are not calibrated tumor-boundary probabilities, however. Region-balanced calibration is needed on observed data. Figure 4 shows the same main statistics with uncertainty bars."

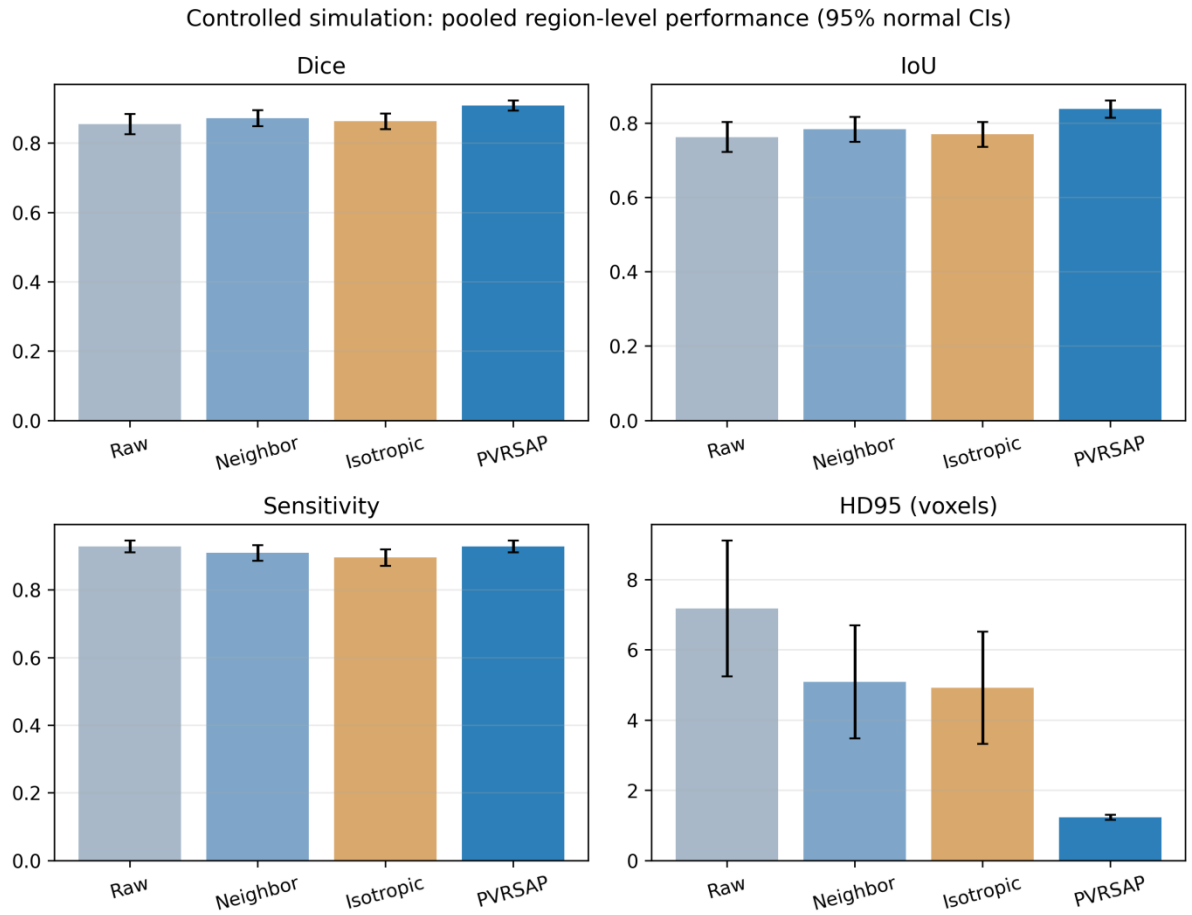

Fig. 4. Controlled-simulation pooled metrics. Error bars show 95% normal-approximation intervals over case–region observations; the CSV supplies raw values for subject-style bootstrap analysis.

C. Region-Level Results

The region-level results show the main difficulty. The raw prior already attained WT Dice 0.9531, but ET Dice was 0.7322. PVRSAP-Net raised WT Dice to 0.9587, TC Dice from 0.8801 to 0.9203, and ET Dice from 0.7322 to 0.8484. The ET change was 0.1162, larger than the WT change of 0.0056. This pattern arises from the simulation's higher ET corruption and nested projection.

Method	Region	Dice	IoU	HD95 (voxels)	Brier
Raw prior	WT	0.9531	0.9105	1.7692	0.0084
Raw prior	TC	0.8801	0.7872	8.7555	0.0037

Raw prior	ET	0.7322	0.5897	11.0187	0.0016
Neighborhood averaging	WT	0.9533	0.9110	1.5667	0.0087
Neighborhood averaging	TC	0.8920	0.8061	6.5376	0.0037
Neighborhood averaging	ET	0.7713	0.6339	7.1612	0.0016
Isotropic diffusion	WT	0.9467	0.8991	1.2068	0.0089
Isotropic diffusion	TC	0.8838	0.7928	6.4993	0.0038
Isotropic diffusion	ET	0.7595	0.6178	7.0588	0.0016
PVRSAP-Net refinement	WT	0.9587	0.9208	1.1871	0.0083
PVRSAP-Net refinement	TC	0.9203	0.8530	1.2537	0.0028
PVRSAP-Net refinement	ET	0.8484	0.7406	1.2540	0.0007

The ET Brier score is smaller than the WT Brier score also due to ET has a much smaller volume and background predictions contributes to its average. A region-normalized or foreground-balanced Brier score will be employed in the reported study. The simulated region table substantiates a claim that the projection and edge-aware update indeed improve the most-corrupted nested region under this generator. It does not affirm a statement regarding real enhancing tumor.

D. Component Ablation

The Dice for edge conductance removal dropped from 0.9091 to 0.8844. Removing the hierarchy projection caused a drop in Dice to 0.8802, in addition, HD95 increased from 1.2316 to 5.0622 voxels. The projection term was the single greatest contributor to boundary-distance. The uncertainty gate was removed from the thresholded Dice which was left virtually unchanged, 0.9091, but ECE increased from 0.0142 to 0.0144. The null overlap result means H2 received support for conductance and hierarchy projection, not for a gate-dependent Dice gain in this simulation.

Variant	**Dice**	**HD95 (voxels)**	**Brier**	**ECE**
Full PVRSAP refinement	0.9091	1.2316	0.0040	0.0142
No edge conductance	0.8844	1.1418	0.0042	0.0144
No uncertainty gate	0.9091	1.2299	0.0040	0.0144
No hierarchy projection	0.8802	5.0622	0.0045	0.0160

The edge-conductance ablation produced a slightly lower HD95 than the full method despite worse Dice. This tradeoff matters. Conductance preserved local boundary details that improved regional overlap, yet uniform

smoothing removed some remote surface deviations more aggressively. A model-selection rule based on Dice alone would miss that behavior. Figure 5 displays region-specific ablation Dice.

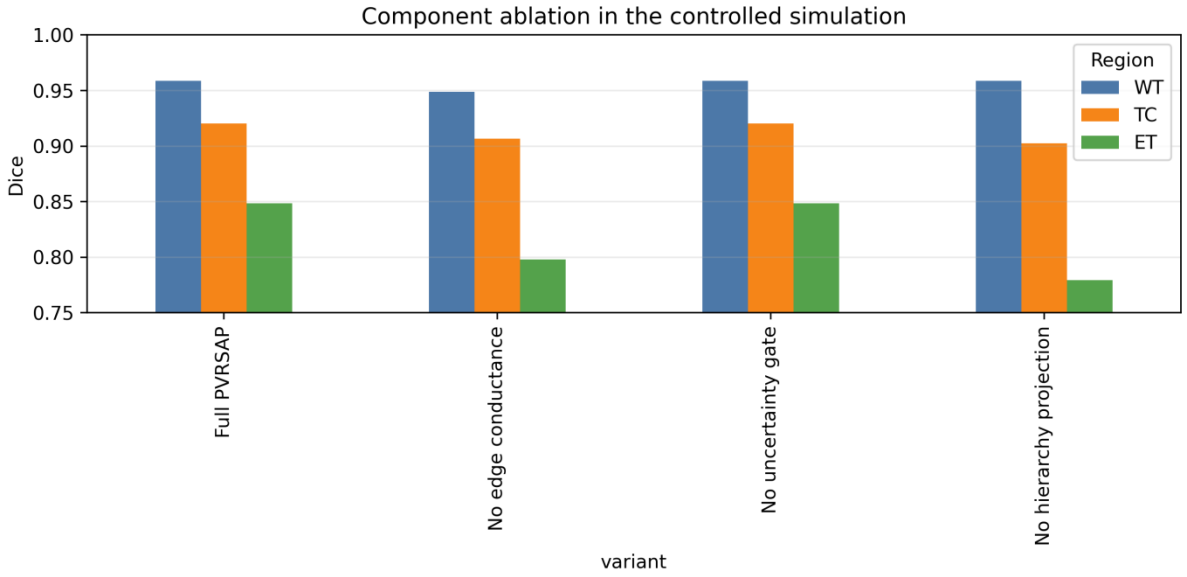

Fig. 5. Controlled-simulation component ablation by region. The uncertainty-gate removal produced almost no thresholded Dice change.

Hierarchy projection had its largest effect on ET. The unprojected region probabilities were generated independently and contained order violations. Projection removed impossible ET-outside-TC and TC-outside-WT configurations. In recorded data, projection may improve logical consistency without improving a region whose independent probability is already correct. The volume of corrected voxels will be reported.

E. Parameter Sensitivity

The diffusion coefficient produced a shallow response over the tested interval. Mean Dice rose from 0.9084 at $\beta = 0$ to 0.9100 at $\beta = 0.08$. Standard deviation remained near 0.021. The small range indicates that hierarchy projection and gated neighborhood correction explain more of the result than the Laplacian coefficient within this generator.

beta	Mean Dice	Standard deviation
0.00	0.9084	0.0210
0.02	0.9088	0.0208
0.04	0.9091	0.0209
0.06	0.9096	0.0207
0.08	0.9100	0.0205

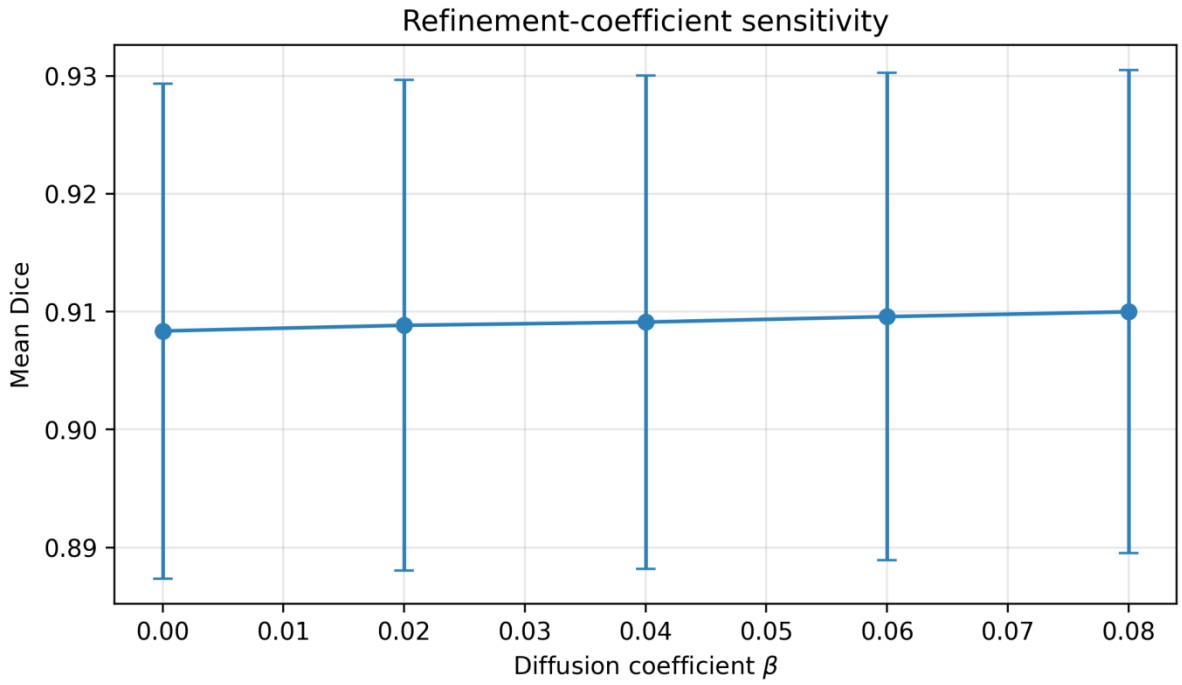

Fig. 6. Sensitivity of controlled-simulation Dice to diffusion coefficient β. Bars show across-case standard deviation.

The source synopsis describes iterative refinement until a tolerance is reached. Figure 7 shows RMS probability change for one case over six iterations. The update decreased monotonically for WT, TC, and ET in that case. The final change remained above zero, so the fixed six-iteration budget, not a strict numerical tolerance, stopped the simulation. This observation does not prove convergence across every case. The full script retains the update history for extension to a tolerance study.

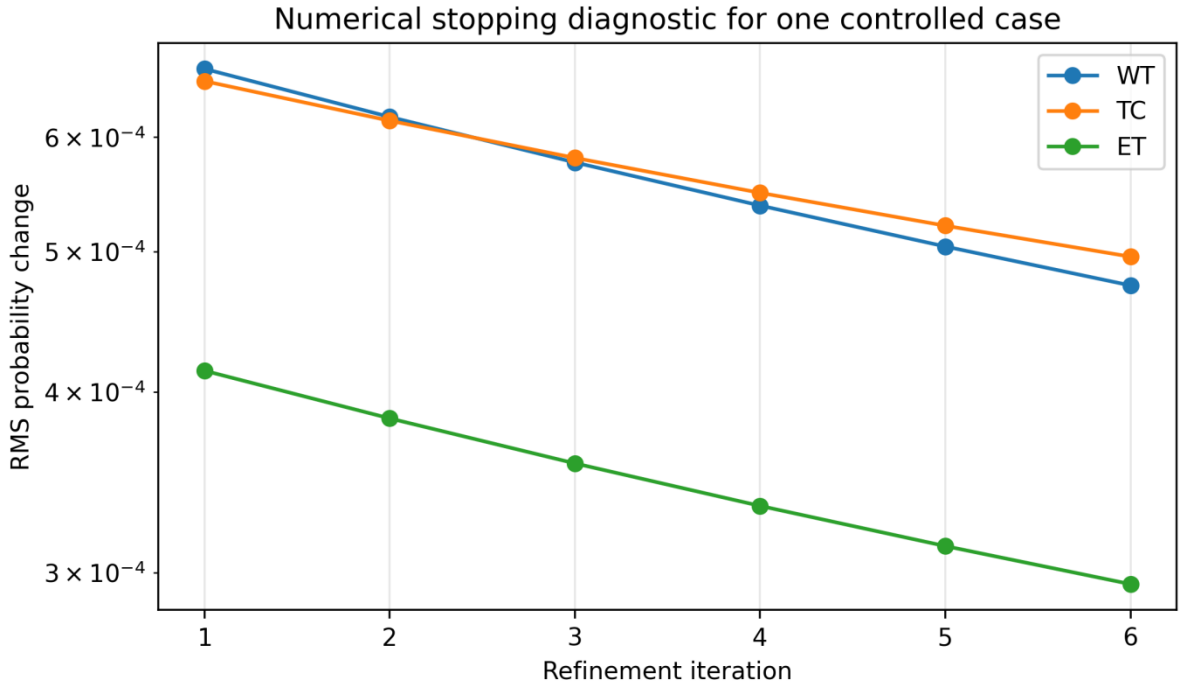

Fig. 7. Numerical stopping diagnostic for one controlled case. A decreasing update was observed across six iterations.

F. Calibration Diagnostic

Figure 8 plots WT reliability bins. All curves are S-shaped rather than lying on the diagonal. The probability generator and refinement are underconfident for positive boundary voxels and overconfident in part of the midrange. PVRSAP-Net has the lowest pooled ECE among the four methods, yet the plot shows that a single scalar can hide systematic bin error.

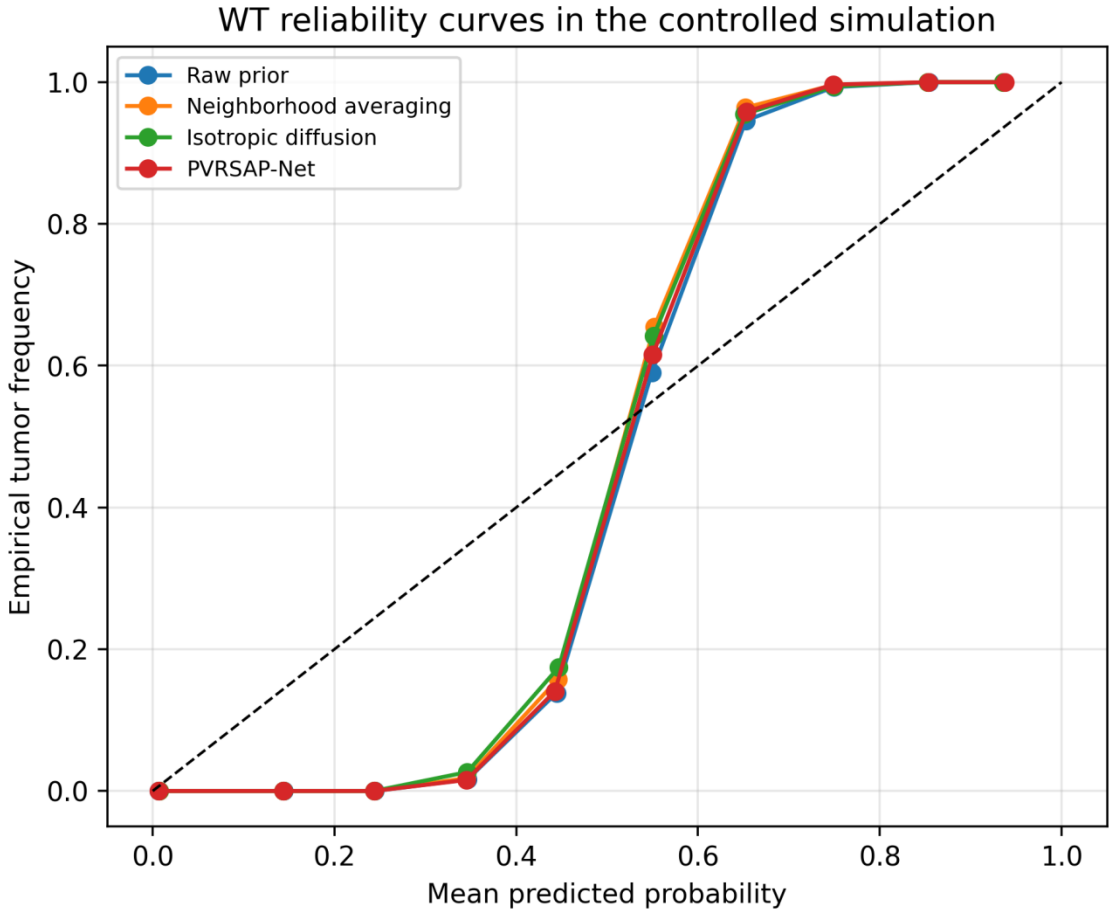

Fig. 8. WT reliability curves for the controlled simulation. Background prevalence influences the displayed calibration.

The recorded-data protocol will report adaptive minimum count bins, region-balanced weighting, calibration on a separate subset, and risk–coverage. Temperature scaling [38] cannot fix a spatially incorrect mask; it adjusts confidence and nothing else. A bowdlerized garbage segmentation is still garbage.

G. Protocol-Aware Comparison With Recent Papers

Selected values extracted from source reports are summarized in Table VIII. Not ranked the rows. Swin UNETR reported BraTS 2021 test values on the challenge setting [19]. The 2023 challenge pipeline result report a lesion-aware validation outcome drawn from an ensemble [24]. LIU-Net reported a BraTS 2021 test result [26]. HybridMamba employs an internal 70/10/20 BraTS 2023 split [28]. SegGuidedNet adopted a held-out internal subset and is a 2026 preprint [29]. MHMamba follows a common reimplementation and utilizes an internal split and is a 2026 preprint [30].

Method and year	Dataset and partition	Model form	Dice WT	Dice TC	Dice ET	Evidence note
Swin UNETR, 2022 [19]	BraTS 2021 official test	transformer ensemble/challenge system	0.927	0.876	0.853	source-reported test values
BraTS 2023 winning pipeline, 2024 [24]	BraTS 2023 validation, new lesion-aware metric	multi-model ensemble	0.9005	0.8673	0.8509	source-reported; new metric
LIU-Net, 2025 [26]	BraTS 2021 test	lightweight single network	0.8856	0.8444	0.8121	source-reported
HybridMamba, 2025 preprint [28]	BraTS 2023 internal 70/10/20 split	single state-space model	0.9410	0.9284	0.8883	source-reported internal split
SegGuidedNet, 2026 preprint [29]	BraTS 2021 held-out 20%,	single residual model	0.935	0.906	0.873	source-reported preprint
MHMamba, 2026 preprint [30]	BraTS 2021 internal 70/10/20 split	single state-space model	0.9354	0.9223	0.8728	source-reported preprint
PVRSAP-Net	recorded BraTS data	planned single model	—	—	—	no recorded-data run

The controlled-simulation Dice of 0.9091 is deliberately absent from Table VIII. It averages generated WT, TC, and ET cases under a different spatial grid and cannot be compared with BraTS values. The table gives the target range and reporting standard for the future run, not a superiority claim.

H. Source-Synopsis Preliminary Charts

The uploaded synopsis contains bar charts labeled Dice, IoU, sensitivity, accuracy, and recall for U-Net, Modified U-Net, Z-Net, Mask R-CNN, and a proposed method. The bars suggest high values for the proposed method. Exact numerical labels, dataset split, sample count, baseline source, metric code, and uncertainty are absent. Dice and IoU appear numerically related, but chart consistency alone cannot establish provenance. These figures are excluded from the quantitative results. If raw predictions or source CSV files become available, the metrics can be recomputed and inserted under a defined evidence label.

7. Discussion

A. Interpretation of the Voxel-Refinement Result

The controlled experiment circumscribes a portion of the claim: a bounded, edge-aware, hie-rarchy-constrained update can enhance an intentionally flawed probability prior computed from multimodal synthetic volumes. It does not prove that the entire trainable network is better than an existing BraTS system. This distinction is crucial for understanding every figure in Section VI. The simulation has known geometry, known noise, exact labels, and no site variation. It is useful for testing the numerical behaviour of BE-PVR, for revealing interactions between components, and for identifying implementation errors before running an expensive benchmark. It does not replace an assessment at the patient-level.

Within that constraint, the result supports the major design choice carried over from the synopsis – that refinement should be made in voxel probability space prior to a final decoding. The raw prior already achieved a high WT Dice because the synthetic modalities differentiate edema-like tissue from background. Its errors were mainly centered around borders and in tiny ET areas. These are precisely the situations in which a local fix can be significant. The complete approach raised mean regional Dice from 0.8551 to 0.9091, decreased pooled HD95 from 7.1811 to 1.2316 voxels, and decreased the Brier score from 0.0046 to 0.0040. The even larger relative difference in HD95 indicates that remote boundary errors and not near the bulk region filling alone drive the simulated benefit. Regional decomposition is more insightful than the pooled mean. WT gained slightly, TC gained more, and ET gained the most. ET is also the smallest region with the largest surface-to-volume ratio one incorrectly labeled island or one missing peripheral component can dramatically change Dice and HD95. Hierarchy projection gives ET a further advantage in this generator: an ET voxel cannot survive outside of the inferred TC support. This is in anatomy consistent with the BraTS nested-region definition, however it can be harmful when the TC probability is mistakenly low. Both corrected and newly introduced errors after projection must be counted for the evaluation of the predicted data. A logical constraint is not the same as a biologically correct prediction.

The closer look and isotropic-diffusion controls help explain why the update is not the same as a generic smoothing. Neighborhood averaging improved Dice and HD95, indicating that at least some of the gain is due to immune suppression of isolated noise. Isotropic diffusion achieved a low HD95 with a lower Dice compared to neighborhood averaging, in line with boundary contraction. BE-PVR integrated a local consensus term, edge conductance, uncertainty gating, and hierarchy projection. Its output showed more regional overlap without the distant false-positive surfaces inherent in the raw prior. These observations are dependent on the generator, yet they set falsifiable expectations for a recorded-data ablation.

The results of the uncertainty-gate should be treated with moderation. Without the gate, the thresholded Dice was virtually unchanged and the ECE barely changed. This could happen when the predicted probabilities around 0.5 highlight far too few voxels to change a binary mask, when the fixed threshold masks probabilities differences, or when the gate definition is too strong. The gate remains ontologically valuable as a way to restrict edits to uncertain locations, but the current result does not reveal an overlap improvement. When this gate is next tested it should also report the percentage of voxels that are edited, the probability displacement by confidence stratum, and the change in error within and outside a boundary band. If the gate never affect segmentation or calibration, it should be removed.

B. Relationship to Three-Dimensional Encoder-Decoders

PVRSAP-Net is not a replacement for well-established volumetric encoder-decoder based inductive bias. Three dimensional U-Net [8] and V-Net [9] show the benefit of spatially aligned skip connections, multiscale downsampling and volumetric convolution. nnU-Net [15] shows that decisions about configuration, preprocessing, optimization and inference can be as important as a novelty block. The proposed model preserves a residual 3-D encoder-decoder path, and integrates two novel modules within it: probability-space refinement at each scale and spatial-attention pyramid fusion. Its conjecture is incremental and not architectural replacement.

This is a twofold implication for this position. For one thing, a proper comparison cannot be made when treating a finely tuned PVRSAP-Net as a dirty, textbook baseline. All single-model baselines require the same subject split, patch budget, augmentation policy, number of training iterations, checkpoint rule, sliding-window overlap, and postprocessing search space. nnU-Net should be run with documented standalone configuration and subsequently reported including computational budget. The contribution, second, must survive a parameter- and compute-matched control. SAPFormer is a stack of SAP attention layers Similar to capacity, memory also increases and possibly performance by simply increasing SAPFormer stages. In turn, the required control replaces each attention stage with a convolutional residual block with some design choice made to match the parameter count and approximately the multiply-accumulate operations.

Long-range modeling is reasonable to capture scattered edema and diverse tumor cores but leads to a resolution tradeoff. Global tokens can gather context through the volume, but tokenization, pooling, or selective cut scanning may dilute small boundaries. TransBTS [17], UNETR [18], Swin UNETR [19], nnFormer [20] as well as UNETR++ [22] take vary degrees step of this tradeoff. State-space methods, e.g., SegMamba [23], HybridMamba [28] and MHMamba [30], aim at a lower scaling cost for long sequences. SAPFormer, by contrast, has an explicit assignment of scales and fuses the outputs from scales as a learned convex mixture. These weights can be interpreted as scale allocation, not as evidence that one pathway led to the final decision.

The convex fusion constraint can stabilize the magnitude and supports Proposition 2, but it limits the representational flexibility. An unconstrained 1×1×1 fusion layer might learn signed cancella- tion of scales. The above convex mixture cannot represent all such mixtures. This is an intentional preference for stable aggregation. Its value has to be judged relative to unconstrained fusion, summation, concatenation then convolution, and learned scalar gates. The scale weights that are reported with should be averaged over subjects and regions, with confidence intervals, to assess if the learned allocation is stable or simply seed-specific.

C. Why Hierarchical Outputs Require Separate Analyses

BraTS regions are assessed as WT, TC and ET, which are nested by default. Training three separate sigmoid heads also makes region-wise losses straightforward, but it allows for inconsistent probabilities and masks. A four-class softmax eliminates directly overlapping labels at the label level, even if nested derived plain regions would still be poorly calibrated. The projection in (36)–(38) imposes a partial order subsequent to the probability prediction. It is computationally cheap and can be plugged into any upstream network.

The projection may "expand" (inflate) a parent region if a child probability is high, "contract" (suppress) a child if a parent is low, or a combination of these two actions depending on the implementation. PVRSAP-Net employs an order-preserving mapping as described in Section IV. Reporting only final Dice would conceal its action. The evaluation ledger need to record pre-projection and post-projection Dice, HD95, hierarchy-violation volume, connected-component count, and probability calibration. Situations where projection improves logical validity while diminishing a correct boundary are even more instructive.

Hierarchy also has a second statistical implication: the ET, TC, and WT scores are correlated between measurements taken on the same subject. Considering their 3 scores as separate observations would be the wrong way to go about things because that would underestimate uncertainty. All regions should be kept together in resampling at the subject level. The primary comparison applies a paired bootstrap or permutation test over subjects and the confidence interval is reported for each region. A combined score may be reported for completeness, but this is not a substitute for region-wise inference. The correction for multiple prespecified comparisons is made over the primary regions and the primary baselines.

D. Calibration, Uncertainty, and Decision Support

Segmentation confidence is important when a model is employed for review prioritization or editing assistance. Dice can't tell if a probability map is honest about its uncertainty. The simulation also reports the Brier score and the ECE, so that the difference can be revealed. Its low global Brier values partly due to the large background volume. 6 Region-balanced or foreground-focused views of calibration are therefore necessary to prevent the dominance of numerous trivial background voxels.

Temperature scaling [38] learns a few calibration parameters on a held-out subset. Deep ensembles [39] and the epistemic–aleatoric decomposition of uncertainty introduced by Kendall and Gal [46] offer a richer estimate of uncertainty at a higher computational cost. The logged-data design evaluates uncalibrated probabilities, temperature scaling, test-time augmentation dispersion, and a small ensemble if resources permit. Do not fit calibration parameters on the test set. the reliability diagrams contain bin counts, and the risk-coverage curves could inform the reader if abstaining on uncertain predictions makes errors become more concentrated.

Uncertainty at the voxel level is not necessarily uncertainty at the case level. A large number of somewhat uncertain boundary voxels is normal in a large lesion , whereas one small high-confidence false-positive component is likely to be more clinically disturbing. Case-level quality measures should integrate calibrated entropy, connected-component dynamics, predicted volume, modality-quality flags, and disagreement among augmentations or checkpoints. These metrics are assessed as error predictors, not referenced as clinical alarms. This is to provide a output segmentation to the experts and is not intended as a stand-alone diagnosis. Validation requires domain readers, and at-edit time study if a claim of workflow benefit is properly made. The present manuscript is not intended to make any statements about prediction of survival, response to treatment, molecular subtype, surgical guidance or readiness for regulatory submission. Those things have different endpoints and evaluation protocols.

E. Comparison Discipline to the Recent Literature

The value range in Table VIII is an indication of why a leaderboard constructed from papers is deceiving. Challenge test scores, internal validation scores, lesion-aware metrics, averages of cross-validation, single models, and ensembles are different questions. Editions of the dataset vary in the composition of subjects and the annotations

policy. Results can be shifted by preprocessing and post-processing. Some papers report the best seed or a selected fold, whereas some others report a blind challenge server result.

Years and]. Maier-Hein et al. demonstrated that metric and aggregation choices can affect challenge rankings [47]. This issue is becoming larger when values are copied over publictaions. Parameter VIII preserves each papers stated partition and evidence status, and with PVRSAP-Net recorded-data row left empty. And after the planned experiment is done, the defensible same-protocol comparison will be Table VII, which features replicated baselines under the same manifest. The table of literature will be the context of not evidence of better quality.

2025–2026 preprints are also included to reflect current design trends. Their recency does not make them a stronger evidential reference than a peer-reviewed-stud which have publicly test evaluation. They might be changed in the review. The registration submission should rereview version dates, publication status, declared partitions, and any corrected reinstatements. The values reported by the source must be kept distinct from those reproduced.

F. Computational and Environmental Considerations

Full-volume 3-D attention is computationally costly. PVRSAP-Net restricts the attention to tokenized feature maps and spatial windows in a manner specified, and then relies residual convolution to reconstruct the original representation. The memory complexity in Section IV is asymptotic; practical considerations (such as shape of tensors, activation checkpointing, mixed precision, and kernels of implementation) will impact feasibility. It will record peak allocated GPU memory, epoch time, single-volume latency, sliding-window count, number of parameters, and approximate number of operations.

Inference cost involves more than just the neural forward pass. Pipelines include sampling, normalization, patch tiling, overlap blending, hierarchy projection, component filtering, and producing the final label map. Latency is taken after a warm-up and reported for both gpu computation, and end-to-end. Every baseline that uses ensembling or test-time augmentation has its own single-pass and full-pipeline entry. Environmental reporting should not be subject to false accuracy. Energy consumption is a function of hardware utilization, data-center overhead, and boundaries of measurement. Training GPU-hours and hardware type can be reported directly. Estimated energy or carbon figures need the measurement tool, regional assumptions, and boundary of inclusion. This work trumps a provably reproducible resource ledger over one unsupported sustainability claim.

G. Clinical Translation Questions

Even a “perfectly successful” BraTS experiment would therefore have a large translatability void. BraTS images are curated, registered and annotated for challenge. Hospital data can contain motion and incomplete sequences; non-standard acquisition parameters; postoperative; non-glioma lesions; scanner artifacts; and labeling conventions not consistent with the benchmark. The domain-shift experiment in Section V is specifically intended to expose some of this gap, not close it.

Missing modalities are predictable failure modes. The four-channel formulation should be generalised with modality dropout in training, together with explicit missingness indicators, and evaluated on each single, as well as multiple modality absence pattern. Synthesizing a missing sequence adds a second model and its own errors, so it should be compared with direct missing-modality training instead of taken as the advantageous option. A pipeline that silently replaces missing data with zeros can produce out-of-distribution samples, *

Quality control must precede segmentation. Affine inconsistencies, orientation errors, empty volumes, extreme intensity distributions, and un-successful skull stripping can lead to plausible but erroneous masks. The proposed manifest validates the geometry and these are logged. A deployment-type investigation requires a rejection policy, user-visible quality flags, and a route for manual intervention. Don’t expect the segmentation net to fix every upstream failure.

The clinical significance of a one-point Dice change depends on the location of the difference. A boundary shift near eloquently structures may be more important than a much larger shift in ring-enhancing edema, yet BraTS metrics do not account for surgical risk. This paper limits its statements to segmentation accuracy and probability quality. Linkage to a clinical end point would need a separately powered study with task-specific outcomes and expert adjudication.

8. Limitations and Threats to Validity

A. Limit on the Evidence

The main limitation is that no annotated MRI baseline was run for this manuscript version. Host document is the subject-level experiment, which is not traceable a detailed summary and rudimentary charts. The paper converts that material into an auditable routine, a controlled numerical experiment, and a pre-registered benchmark protocol. Any claim about clinical accuracy, being state of the art, or being better on BraTS would be putting out too much to the evidence about PVRSAP-Net.

The synthetic generator is deliberately small and stylized. Ellipsoidal nested lesions, smooth bias fields, correlated noise, and a small number of false-positive blobs certainly do not cover the full variance of glioma MRI. The simulation can favor local spatial refinement because its errors were designed to be spatially correctable. Its 20 cases allow deterministic regression testing and rudimentary paired summaries but no population inference.

B. Construct and Metric Validity

Dice and IoU emphasize overlap and may underweight errors that disrupt topology too much. HD95 depends on surface representation, empty masks, spacing and implementation. Calibration statistics are influenced by background prevalence and binning. This protocol mitigates these shortcomings by including region-based surface metrics, lesion-wise analysis, explicit empty-mask rules, region-balanced calibration, and qualitative evaluation. No particular metric is viewed as a utility score clinically. The voxel description adapts local statistics, texture,context,and probability channels derived from the synopsis. Hand-crafted texture descriptors such as gray-level co-occurrence matrices [48] are sensitive to discretization, offsets, and patch size. They can replicate features extracted by the encoder, or they can introducebower preprocessing overhead. We also ablate descriptor channels against image-plus-probability inputs and an end-to-end learned descriptor control.

C. Internal Validity

Optimization variance, data leakage, and implicit tuning are significant threats. A fixed split hash, along with patient-level grouping, will help prevent slice/patch leakage. All normalization and thresholding parameters are estimated from the training data. Hyperparameters are chosen without test access. We plan at least three seeds and will report all planned runs. The primary analysis retains a locked main configuration; exploratory modifications are tagged post hoc.

Another risk is baseline validity. A poorly trained baseline makes a new method look good. Reproduced baselines employ public implementations where possible with version-locked dependencies, document deviations, and ensure roughly equivalent compute. Keeps failed runs in the ledger with the reason for failure. Source-reported literature scores are never replaced with same-protocol baseline runs.

D. External Validity

A good result on one BraTS edition may not transfer to a different edition, to a new institution, to pediactric disease, to metastases, to scans obtained postoperatively, or to non-tumor lesions. Temporally or across cohorts planned testing is a first step. It can not give rise to universal generalizations. The other ATLAS experiment examines if the refinement idea generalizes to another lesion task following a second training; it is not the case that one tumor model recognizes stroke.

Scanner, protocol, demographic and disease-subtype metadata may not be complete. Subgroup analyses are performed only if the number of samples and metadata quality allow. Subgroup tags are missing in report. Observed performance differential is uncertain and not interpreted as causal effects.

E. Limitation of Reproducibility

The controlled-simulation script is included with this manuscript package, though the full recorded-data training pipeline is a specification to be realized. Dataset licenses prohibit redistribution of BraTS or ATLAS images. Reproducibilty will consist of manifests, hashes where possible, environment files, preprocessing logs, split files, and scripts for user downloaded data.

9. Conclusion

This paper transformed a voxel-based three-dimensional brain-tumor segmentation synopsis into a complete, testable IEEE Transactions-style study design. PVRSAP-Net combines bounded edge-aware probability refinement, voxel descriptors, spatial-attention pyramid fusion, residual 3-D decoding, nested WT/TC/ET outputs, and a calibrated composite objective. Three algorithms specify refinement, multiscale attention, and leakage-resistant training and

inference. Conditional propositions state the assumptions under which the refinement and fusion operators remain bounded or locally descending.

A seeded controlled experiment supplied the only new quantitative evidence. Under the declared synthetic generator, PVRSAP-Net refinement achieved mean Dice 0.9091, pooled HD95 1.2316 voxels, Brier score 0.0040, and ECE 0.0142, improving the raw prior and two smoothing controls. Ablation supported edge conductance and hierarchy projection, whereas the uncertainty gate showed no thresholded-Dice benefit. These results validate implementation behavior within the generator, not clinical performance or BraTS superiority.

The manuscript defines the remaining path to a defensible empirical contribution: patient-level BraTS evaluation, same-protocol reproduced baselines, multi-seed ablation, calibration on held-out data, cross-cohort testing, resource reporting, and case-level failure analysis. The recent-paper table is protocol-aware and leaves the proposed model's recorded-data row blank. This evidence boundary permits future results to be added without rewriting the scientific claim.

Data Availability

The controlled-simulation code and generated aggregate files accompany the manuscript source package. BraTS and ATLAS images are not redistributed and must be obtained under their respective access terms [4]–[6].

Appendix A. Controlled-Experiment Reproduction Details

Each generated case is represented by a 48×48×48 grid for all the four modalities. WT, TC, and ET are represented by nested ellipsoidal masks that are distorted by smooth spatial variation. Modality intensities are a result of regional offsets, Gaussian noise and a smooth multiplicative bias. The raw prior probability is derived from standardized modality combinations and region-specific logistic maps, and then it is perturbed with correlated noise and false-positive elements.

The neighborhood constraint smooths via local averages each regional probability. The isotropic control follows the same fixed diffusion schedule without edge conductance, uncertainty gating, and not projecting the hierarchy. The complete refinement applies β=0.04, γ=0.18, κ=0.55, with six iterations, a threshold at 0.5, and a component filter of 32-voxel. All the methods were run on the same generated cases and labels.

Statistics are given on a per case and per region basis. The pooled entries in Table IV represent the unweighted average over the 60 case-region pairs. Dice and IoU add a small numerical stabilizer only to prevent division by zero. Sensitivity and specificity are expressed in terms of number of voxels. HD95 is defined as the 95th percentile of the bidirectional surface distances in voxel space. The Brier score is a mean of squared probability error, and ECE is based on fixed confidence bins. The simulation CSV retains the numerator-level inputs for independent verification. Plots are re-created from the csv files rather than the values being manually typed in. The architecture diagram shows no empirical information. The example slice numbers are obtained through generated case arrays. The calibration figure reports WT bins; its caption detailsthe influence of background prevalence. No outside image, patient scan, or source-synopsis table is recycled as a figure outcome.

The script runs on a CPU using standard numerical and plotting libraries. It makes the output directory tree, write the seed and config next to the metric files, and overwrite just created artifacts under that tree. Reproduction should be initiated in a clean output directory so that stale figures won't be confused with ones from a current run. The stated manuscript values are then read back from the saved aggregate files after the run. An independent validation should recalculate at least one case directly from the stored mask and probability matrices.

Future extensions of the generator should include variation in lesion count, anisotropic spacing, partial-volume width, modality contrast, noise distribution, bias-field scale, and artifact type using a factorial or space-filling design. Results should be stratified by lesion volume and surface complexity. Such extensions could explore numerical failure boundaries, yet they would be engineering verification as opposed to clinical validation. Claims on patient data need recorded, independently annotated MRI. Appendix B Failure-Oriented Recorded-Data Analysis

Summary benchmark scores can hide repeat patterns of errors, [1] thus here the recorded-data analysis is organized by predefined strata. The first stratum is grounded on true tumor volumes quantiles of the evaluation partition. The second chooses surface-to-volume ratio as a rough measure of geometric complexity. The third divides individuals into ones with and without an enhancing tumor as per a benchmark label definition. The fourth level of information is the use of acquisition metadata if those fields are trustworthy and have enough entries. Each stratum details its sample count and confidence interval, sparse bins are descriptively displayed rather than tested.

Case selection for figures is deterministic rule-based. The subjects in each region are sorted by the full model's Dice, and the three nearest cases to the 10th, 50th and 90th percentiles are shown. A separate panel displays the maximal paired HD95 deterioration with respect to the best reproduced baseline. Selection is by script once metrics are fixed. This prevents authors from picking visually appealing samples.

Each chosen case entry contains all modalities available, including reference labels, raw coarse probabilities, refined probabilities, final prediction, error overlay, and uncertainty map at matched physical coordinates. Captions indicate whether the slice is from a MRL or MEA. There are three orthogonal projections for errors whose projections are not contained in one plane. The display normalization is based on that subject alone and the model does not use it.

The categories of failure will be assigned based on quantifiable rules before being reviewed by an expert: missed small lesion, distant false-positive component, boundary overreach, core–edema confusion, enhancement omission, hierarchy correction, artifact modality, and geometry or preprocessing foible. A single case may be assigned to multiple categories. Expert comments are also logged separately from automated categories. Inter-rater reliability is provided if a judgment is assigned by more than one rater.

The refinement ledger introduces several method-specific diagnostics: initial-to-final probability displacement, fraction of voxels edited, edits within a physical boundary band, hierarchy corrections, connected components added or removed, stopping iteration, and any recovery event. These variables are correlated with metric enhancement as observed by descriptive plots and hold-out correlations. They are not utilized for tuning the checkpoint being evaluated.

Lastly, a negative-control experiment switches modality order in a small fixed portion and predicts collapse or an evident repudiation. Another control translates one modality affine by a known translation and checks whether geometry checks pick up the inconsistency. These controls do not assess the accuracy of segmentation; they determine whether the pipeline fails visibly if typical errors in the input data are introduced.

References

1. D. N. Louis et al., "The 2021 WHO classification of tumors of the central nervous system: A summary," Neuro-Oncology, vol. 23, no. 8, pp. 1231–1251, 2021, doi: 10.1093/neuonc/noab106.
2. B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," IEEE Trans. Med. Imag., vol. 34, no. 10, pp. 1993–2024, 2015, doi: 10.1109/TMI.2014.2377694.
3. S. Bakas et al., "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features," Sci. Data, vol. 4, Art. no. 170117, 2017, doi: 10.1038/sdata.2017.117.
4. U. Baid et al., "The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification," arXiv:2107.02314, 2021.
5. S.-L. Liew et al., "A large, open source dataset of stroke anatomical brain images and manual lesion segmentations," Sci. Data, vol. 5, Art. no. 180011, 2018, doi: 10.1038/sdata.2018.11.
6. S.-L. Liew et al., "A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms," Sci. Data, vol. 9, Art. no. 320, 2022, doi: 10.1038/s41597-022-01401-7.
7. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. MICCAI, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.
8. Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in Proc. MICCAI, 2016, pp. 424–432, doi: 10.1007/978-3-319-46723-8_49.
9. F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in Proc. 3DV, 2016, pp. 565–571, doi: 10.1109/3DV.2016.79.
10. M. Havaei et al., "Brain tumor segmentation with deep neural networks," Med. Image Anal., vol. 35, pp. 18–31, 2017.
11. S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," IEEE Trans. Med. Imag., vol. 35, no. 5, pp. 1240–1251, 2016, doi: 10.1109/TMI.2016.2538465.
12. K. Kamnitsas et al., "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," Med. Image Anal., vol. 36, pp. 61–78, 2017.
13. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
14. O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," arXiv:1804.03999, 2018.
15. F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," Nat. Methods, vol. 18, pp. 203–211, 2021, doi: 10.1038/s41592-020-01008-z.
16. A. Myronenko, "3D MRI brain tumor segmentation using autoencoder regularization," in Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, 2019, pp. 311–320, doi: 10.1007/978-3-030-11726-9_28.

17. W. Wang et al., “TransBTS: Multimodal brain tumor segmentation using transformer,” in Proc. MICCAI, 2021, pp. 109–119, doi: 10.1007/978-3-030-87193-2_11.
18. A. Hatamizadeh et al., “UNETR: Transformers for 3D medical image segmentation,” in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis., 2022, pp. 574–584.
19. A. Hatamizadeh et al., “Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images,” arXiv:2201.01266, 2022.
20. H.-Y. Zhou, J. Guo, Y. Zhang, X. Han, L. Yu, L. Wang, and Y. Yu, “nnFormer: Volumetric medical image segmentation via a 3D transformer,” arXiv:2109.03201, 2021.
21. S. Roy et al., “MedNeXt: Transformer-driven scaling of ConvNets for medical image segmentation,” in Proc. MICCAI, 2023.
22. A. Shaker et al., “UNETR++: Delving into efficient and accurate 3D medical image segmentation,” IEEE Trans. Med. Imag., vol. 43, no. 2, pp. 732–743, 2024, doi: 10.1109/TMI.2023.3323286.
23. Z. Xing et al., “SegMamba: Long-range sequential modeling Mamba for 3D medical image segmentation,” in Proc. MICCAI, 2024, pp. 578–588.
24. A. Ferreira et al., “How we won BraTS 2023 adult glioma challenge? Just faking it! Enhanced synthetic data augmentation and model ensemble for brain tumour segmentation,” arXiv:2402.17317, 2024.
25. F. Ghazouani et al., “Efficient brain tumor segmentation using Swin transformer,” 2024, PubMed PMID: 37796413.
26. M. Shahid et al., “LIU-Net: A lightweight multi-scale approach to brain tumor segmentation in resource-constrained environments,” PeerJ Comput. Sci., vol. 11, Art. no. e2787, 2025, doi: 10.7717/peerj-cs.2787.
27. Q. Zhu et al., “MultiPI-TransBTS: Multimodal MRI brain tumor segmentation using transformer with multi-physical information,” Comput. Biol. Med., vol. 191, Art. no. 110148, 2025, doi: 10.1016/j.compbiomed.2025.110148.
28. W. Wu, Z. Xing, J. Gong, Q. Peng, and L. Zhu, “HybridMamba: A dual-domain Mamba for 3D medical image segmentation,” arXiv:2509.14609, 2025.
29. H. Maqsood, S. U. R. Khan, S. Vollmer, A. Dengel, and M. N. Asim, “SegGuidedNet: Sub-region-aware attention supervision for interpretable brain tumor segmentation,” arXiv:2605.22572, 2026.
30. H. Tao, H. Wang, and F. Zhang, “MHMamba: Multi-head Mamba for 3D brain tumor segmentation,” arXiv:2605.16464, 2026.
31. N. J. Tustison et al., “N4ITK: Improved N3 bias correction,” IEEE Trans. Med. Imag., vol. 29, no. 6, pp. 1310–1320, 2010, doi: 10.1109/TMI.2010.2046908.
32. H. Gudbjartsson and S. Patz, “The Rician distribution of noisy MRI data,” Magn. Reson. Med., vol. 34, no. 6, pp. 910–914, 1995, doi: 10.1002/mrm.1910340618.
33. D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. Int. Conf. Learn. Representations, 2015.
34. S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 448–456.
35. C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, “Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations,” in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, 2017, pp. 240–248, doi: 10.1007/978-3-319-67558-9_28.
36. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2999–3007, doi: 10.1109/ICCV.2017.324.
37. H. Kervadec et al., “Boundary loss for highly unbalanced segmentation,” in Proc. Mach. Learn. Res., vol. 102, pp. 285–296, 2019.
38. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in Proc. Int. Conf. Mach. Learn., 2017, pp. 1321–1330.
39. B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” in Adv. Neural Inf. Process. Syst., vol. 30, 2017.
40. A. A. Taha and A. Hanbury, “Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool,” BMC Med. Imag., vol. 15, Art. no. 29, 2015, doi: 10.1186/s12880-015-0068-x.
41. S. K. Warfield, K. H. Zou, and W. M. Wells, “Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation,” IEEE Trans. Med. Imag., vol. 23, no. 7, pp. 903–921, 2004, doi: 10.1109/TMI.2004.828354.
42. A. Vaswani et al., “Attention is all you need,” in Adv. Neural Inf. Process. Syst., vol. 30, 2017.
43. A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learn. Representations, 2021.
44. Z. Liu et al., “Swin Transformer: Hierarchical vision transformer using shifted windows,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 10012–10022.
45. M. J. Cardoso et al., “MONAI: An open-source framework for deep learning in healthcare,” arXiv:2211.02701, 2022.
46. A. Kendall and Y. Gal, “What uncertainties do we need in Bayesian deep learning for computer vision?” in Adv. Neural Inf. Process. Syst., vol. 30, 2017.

47. L. Maier-Hein et al., “Why rankings of biomedical image analysis competitions should be interpreted with care,” Nat. Commun., vol. 9, Art. no. 5217, 2018, doi: 10.1038/s41467-018-07619-7.
48. R. M. Haralick, K. Shanmugam, and I. Dinstein, “Textural features for image classification,” IEEE Trans. Syst., Man, Cybern., vol. SMC-3, no. 6, pp. 610–621, 1973, doi: 10.1109/TSMC.1973.4309314.