Hierarchical Region-Aware Multi-Granularity Mamba for White Matter Lesion Segmentation

Dahye Lee[1] and Kwanseok Oh[1,2*]

[1] DEEPNOID Inc., Seoul, Korea
[2] Department of Artificial Intelligence, Korea University, Seoul, Korea
{dhlee,gsoh}@deepnoid.com, ksohh@korea.ac.kr

Abstract. Accurate white matter (WM) lesion segmentation is essential for diagnosing and monitoring neurological disorders such as white matter hyperintensities and multiple sclerosis. However, this task remains highly challenging owing to the small and spatially scattered nature of lesions and their low contrast with surrounding tissues. Recently, the Segment Anything Model (SAM) and its medical variants, often augmented with CNNs or Mamba, have shown promising potential for medical image segmentation, yet they struggle to reliably capture subtle lesion patterns and indistinct boundaries that demand fine-grained regional discrimination. In this work, we propose a novel Hierarchical Region-Aware Multi-Granularity Mamba (**HiRAM**) incorporated into the SAM decoder. HiRAM explicitly decomposes the features into region-specific representations (*i.e.*, interior, boundary, and background regions) and performs confidence-driven, non-causal state-space propagation. By incorporating region-aware modeling that prioritizes semantically reliable features and hierarchically integrates multi-granularity learning, our method enhances boundary sensitivity and fine-grained lesion delineation while preserving global contextual consistency. Comprehensive evaluations across two public challenge datasets reveal that HiRAM outperforms state-of-the-art methods, achieving unmatched segmentation accuracy and robustness. Code is available at `https://github.com/deepnoid-ai/HiRAM`.

Keywords: WM Lesion Segmentation · SAM · Mamba

1 Introduction

White matter (WM) lesion segmentation, including white matter hyperintensities (WMH) and multiple sclerosis (MS), plays a crucial role in the assessment of various neurological disorders [2, 4]. These lesions typically appear as small, spatially scattered hyperintense regions on FLAIR MRI, often exhibiting heterogeneous morphology and indistinct boundaries. Advances in deep learning, particularly convolutional neural networks (CNNs) and Vision Transformers (ViTs), have led to significant progress in WMH [8, 28, 31] and MS [24, 25, 34] segmentation. However, such data-driven approaches depend on large-scale, high-quality annotations, which are labor-intensive and time-consuming to acquire.

*Corresponding author.

Segment Anything Model (SAM) [15], pre-trained on extensive datasets, has recently demonstrated impressive generalization capability across diverse tasks and domains [1,21,37]. As a prominent foundation model for segmentation, SAM has attracted growing attention in the medical imaging community, with several studies exploring its applicability to various medical images [21, 23]. However, existing evaluations consistently report inferior performance of SAM in medical domains compared with natural scenes, particularly when segmenting targets characterized by low contrast, faint edges, and diffused lesions [9]. These limitations are largely attributable to the loss of fine-grained local information owing to SAM's large token size [11]. To this end, recent SAM-variants [18,19,27] integrate CNN-based subnetworks and employ self- or cross-attention for their feature fusion. Although these strategies enhance local awareness, attention-based fusion primarily aggregates global features rather than performing contextual propagation, limiting its ability to model subtle and spatially distributed targets. Besides, the quadratic computational complexity of attention, particularly when applied across multi-scale features, imposes substantial overhead on MRI.

Mamba [5], built upon state-space models (SSMs) [14], has emerged as an efficient alternative for modeling long-range dependencies. By leveraging structured state propagation, Mamba enables effective sequence modeling with linear complexity and reduced memory usage, often outperforming attention-based architectures and ViT [22,32,33]. Due to such favorable properties, SAM equipped with Mamba has been explored for medical image segmentation [3,17,26,30,36]. However, most existing SAM-Mamba methods adopt a causal state-space formulation, where 2D (or 3D) images are flattened into 1D token sequences according to predefined scanning rules. Under this paradigm, each token interacts only with previously scanned tokens via a compressed hidden state, creating an information flow that is inherently misaligned with the non-causal spatial structure of medical images. As a consequence, pixels cannot fully exploit spatially diffused contextual cues, resulting in underutilization of lesion-related information. Moreover, causal state propagation is prone to long-range interaction decay, whereby distant yet clinically relevant regions exert progressively weaker influence on the current state. This limits effective modeling of spatially separated but semantically correlated lesion patterns, commonly observed in WMH and MS.

In this work, we propose a novel ***Hi****erarchical* ***R****egion-****A****ware* ***M****ulti-Granularity Mamba* (**HiRAM**) and integrate it into a SAM-CNN hybrid framework to enhance the WM lesion segmentation (see Fig. 1). Compared to Mamba-equipped SAM variants, HiRAM decomposes SAM and CNN features into region-specific representations corresponding to interior, boundary, and background regions, and performs confidence-driven, non-causal modeling to enable hierarchical feature modulation. To be specific, it comprises two key components: (1) a region-tailored prior generator that constructs semantically coherent region-wise priors, and (2) a multi-granularity Mamba that performs region-specific SSM conditioned on these priors and then conducts holistic propagation to enforce global context. By modeling region-level semantics and prioritizing high-confident information, this design enhances fine-grained discrimination, particularly for low-

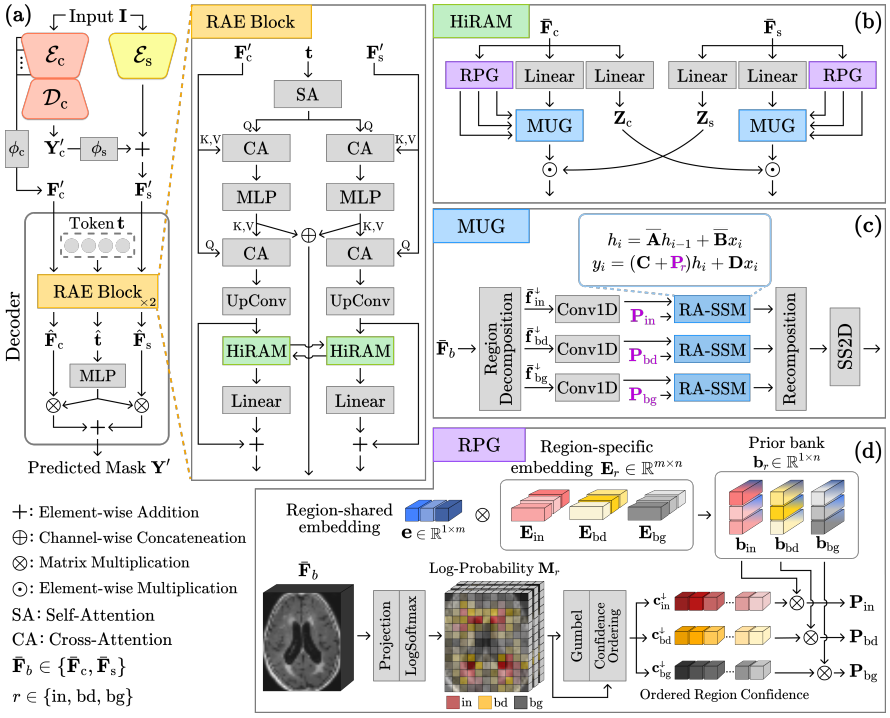

Fig. 1: Overall flowchart of the proposed framework with HiRAM.

contrast and diffuse lesions. Extensive experiments show that our method consistently outperforms ten competitive baselines on two public challenge datasets.

2 Methodology

As shown in Fig. 1(a), our proposed framework derives two complementary representations and refines them within a unified decoder using Region-Awareness Enhancement (RAE) blocks for WM lesion segmentation. Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, a CNN encoder–decoder $(\mathcal{E}_c, \mathcal{D}_c)$ extracts multi-scale features $\{\mathbf{F}_c^k\}_{k=1}^K = \mathcal{E}_c(\mathbf{I})$ from K encoder stages and yields a coarse segmentation mask $\mathbf{Y}'_c = \mathcal{D}_c(\mathbf{F}_c^1, ..., \mathbf{F}_c^K)$. In parallel, the SAM encoder $\mathcal{E}_s$ generates a global semantic features $\mathbf{F}_s = \mathcal{E}_s(\mathbf{I}) \in \mathbb{R}^{H_s\times W_s\times C}$. To align these two representations, all CNN features are resized into $H_s \times W_s$, concatenated $(\|)$, and fused via a learnable projection ϕ_c as $\mathbf{F}'_c := \phi_c(\|_{k=1}^K \text{Resize}(\mathbf{F}_c^k))$. Meanwhile, the mask $\mathbf{Y}'_c$ is downscaled via ϕ_s and merged into the SAM feature space through addition with $\mathbf{F}_s$ as $\mathbf{F}'_s := \phi_s(\mathbf{Y}'_c) + \mathbf{F}_s$. Both ϕ_c and ϕ_s are implemented as convolutional blocks with LayerNorm and GELU activation. Subsequently, the decoder takes $\mathbf{F}'_c$ and $\mathbf{F}'_s$ together with a learnable token $\mathbf{t}$ and processes them using two stacked RAE blocks. The final segmentation mask $\mathbf{Y}'$ is obtained via a lightweight integration

head as

$$\mathbf{Y}' = \hat{\mathbf{F}}_{\mathrm{c}} \otimes \mathrm{MLP}(\hat{\mathbf{t}})[:C'/2] + \hat{\mathbf{F}}_{\mathrm{s}} \otimes \mathrm{MLP}(\hat{\mathbf{t}})[C'/2:], \tag{1}$$

where $\hat{\mathbf{F}}_{\mathrm{c}}$, $\hat{\mathbf{F}}_{\mathrm{s}}$, and $\hat{\mathbf{t}}$ denote the outputs of two stacked RAE blocks, $\otimes$ represents matrix multiplication, and C' denotes the embedding dimension.

2.1 Mask Decoder with RAE Blocks

Our decoder follows the original SAM decoder paradigm [15], in which a learnable token iteratively interacts with image features via self-attention (SA) and cross-attention (CA). However, this mechanism lacks explicit region-aware modeling and treats spatial features uniformly. To alleviate this issue, we design the RAE block, which retains the SAM token–feature interaction while incorporating HiRAM for flexible and hierarchical region-tailored propagation.

Within the RAE block in Fig. 1(a), the token and branch-specific features are updated bidirectionally. The learnable token $\mathbf{t}$ is first refined via the SA layer, producing an updated token $\bar{\mathbf{t}}$. It is then split channel-wise into two branch tokens, $\bar{\mathbf{t}}_{\mathrm{c}}$ and $\bar{\mathbf{t}}_{\mathrm{s}}$, corresponding to the CNN and SAM branches, respectively. For each branch $b \in \{\mathrm{c}, \mathrm{s}\}$, the branch token $\bar{\mathbf{t}}_b$ interacts with its corresponding input branch features $\mathbf{F}'_b$ via CA layer, where $\bar{\mathbf{t}}_b$ serves as the query and $\mathbf{F}'_b$ offers keys and values. The resulting tokens are subsequently refined by an MLP: $\hat{\mathbf{t}}_b = \mathrm{MLP}\left(\mathrm{CA}(\bar{\mathbf{t}}_b, \mathbf{F}'_b)\right)$. The branch features $\mathbf{F}'_b$ are further attended by another CA layer, where $\mathbf{F}'_b$ acts as the query and the refined $\hat{\mathbf{t}}_b$ provides keys and values. The resulting features are then processed by a transpose convolution (UpConv): $\bar{\mathbf{F}}_b = \mathrm{UpConv}\big(\mathrm{CA}(\mathbf{F}'_b, \hat{\mathbf{t}}_b)\big)$. The refined branch features $\bar{\mathbf{F}}_b \in \{\bar{\mathbf{F}}_{\mathrm{c}}, \bar{\mathbf{F}}_{\mathrm{s}}\}$ are fed into HiRAM, and their outputs are finally projected through linear layers and integrated with residual additions to yield output features. In parallel, the RAE block's output token is derived by concatenating the refined tokens $\hat{\mathbf{t}}_{\mathrm{c}}$ and $\hat{\mathbf{t}}_{\mathrm{s}}$.

2.2 Hierarchical Region-Aware Multi-Granularity Mamba

Unlike conventional Mamba approaches that apply a predefined causal sequence over the entire set of features, the proposed HiRAM reorganizes features into region-specific sequences and performs confidence-driven, non-causal propagation. By prioritizing high-certainty features rather than relying solely on spatial adjacency, this confidence-aware sequencing establishes more semantically reliable states while preserving structured state evolution, thereby improving robustness. In Fig. 1(b), HiRAM symmetrically operates on $\bar{\mathbf{F}}_{\mathrm{c}}$, $\bar{\mathbf{F}}_{\mathrm{s}}$, and consists of two core components: a region-tailored prior generator (RPG) and multi-granularity Mamba (MUG). For brevity, we denote $\bar{\mathbf{F}}_{b\in\{\mathrm{c,s}\}}$ as $\bar{\mathbf{F}}$ hereafter.

RPG. In Fig. 1(d), the RPG aims to decompose the input features into interior (in), boundary (bd), and background (bg) regions and generate region-tailored priors that capture semantically coherent region groups to guide a subsequent MUG module. For each region $r \in \{\mathrm{in,bd,bg}\}$, we construct a prior bank

$\mathbf{b}_r \in \mathbb{R}^{1\times n}$. Instead of learning each prior independently, we adopt a semantic decoupling strategy to improve interpretability and parameter efficiency:

$$\mathbf{b}_r = \mathbf{e} \otimes \mathbf{E}_r, \quad \mathbf{e} \in \mathbb{R}^{1\times m}, \quad \mathbf{E}_r \in \mathbb{R}^{m\times n}. \tag{2}$$

Here, $\mathbf{e}$ is a region-shared embedding that establishes a common latent semantic basis, while $\mathbf{E}_r$ offers region-specific transformation coefficients. This ensures that all regions reside in a unified semantic space while allowing region-aware specialization. We next decompose the input $\bar{\mathbf{F}}$ through a linear projection followed by the LogSoftmax to obtain a log-probability $\mathbf{M}_r \in \mathbb{R}^{H'\times W'}$. To enable differentiable region assignment, we apply the Gumbel-Softmax [13] along the region dimension of $\mathbf{M}_r$, yielding a one-hot routing map $\bar{\mathbf{M}}_r$. For each region r, we flatten the $\mathbf{M}_r$ and identify the set of pixel indices assigned to that region:

$$\mathcal{S}_r = \{i \in \{1,\dots,N\} \mid \bar{\mathbf{m}}_{r,i} = 1\}, \quad \bar{\mathbf{m}}_{r,i} \in \{0,1\}, \tag{3}$$

where $N = H'W'$ and $\bar{\mathbf{m}}_r = \mathrm{vec}(\bar{\mathbf{M}}_r)$. Using this index set $\mathcal{S}_r$, we retrieve the corresponding log-probability values from the flattened map $\mathbf{m}_r = \mathrm{vec}(\mathbf{M}_r)$ to construct the region-specific confidence sequence as $\mathbf{c}_r = \{\mathbf{m}_{r,i} \mid i \in \mathcal{S}_r\}$. This $\mathbf{c}_r$ is then sorted in descending order. We define a permutation as

$$\pi_r : \{1,\dots,|\mathcal{S}_r|\} \to \mathcal{S}_r, \quad \text{s.t. } \mathbf{m}_{r,\pi_r(1)} \geq \mathbf{m}_{r,\pi_r(2)} \geq \cdots \geq \mathbf{m}_{r,\pi_r(|\mathcal{S}_r|)}. \tag{4}$$

Accordingly, the ordered $\mathbf{c}_r$ is represented as $\mathbf{c}_r^{\downarrow} = \left[\mathbf{m}_{r,\pi_r(1)},\dots,\mathbf{m}_{r,\pi_r(|\mathcal{S}_r|)}\right]$, and the $\mathbf{c}_r^{\downarrow}$ is combined with the prior bank $\mathbf{b}_r$ defined in Eq. (2) through matrix multiplication ($\otimes$) to produce the final region-tailored priors:

$$\mathbf{P}_r := \mathbf{c}_r^{\downarrow} \otimes \mathbf{b}_r. \tag{5}$$

MUG. In Fig. 1(c), our MUG first flattens the input features as $\bar{\mathbf{f}} = \mathrm{vec}(\bar{\mathbf{F}})$, and partitions it into region-specific subsets according to π_r from Eq. (4). This ordering ensures that high-certainty features are processed earlier in the state-space propagation. As SSMs accumulate information recurrently over time steps, prioritizing reliable patterns stabilizes early hidden states. For multi-granularity propagation, each ordered regional sequence $\bar{\mathbf{f}}_r^{\downarrow}$ is then fed into a region-adaptive SSM (RA-SSM) after 1D convolution. We briefly recall that a discretized SSM [5, 20] evolves as $h_t = \bar{\mathbf{A}}h_{t-1} + \bar{\mathbf{B}}x_t$, $y_t = \mathbf{C}h_t + \mathbf{D}x_t$, where x_t, h_t, and y_t are the input, hidden state, and output at time step t, respectively. Here, $\bar{\mathbf{A}} = \exp(\Delta\mathbf{A})$ and $\bar{\mathbf{B}} = \Delta\mathbf{B}$ arise from discretization, while $\mathbf{C}$ and $\mathbf{D}$ correspond to the output matrix and residual connection. Building on this formulation, the RA-SSM adequately injects the $\mathbf{P}_r$ from Eq. (5) into the output projection:

$$h_t = \bar{\mathbf{A}}h_{t-1} + \bar{\mathbf{B}}x_t, \quad y_t = (\mathbf{C} + \mathbf{P}_r)\, h_t + \mathbf{D}x_t. \tag{6}$$

Inspired by [7, 35], the $\mathbf{P}_r$ functions as a region-conditioned readout bias that adaptively modulates the contribution of latent states according to semantic region-aware priors without altering the underlying state transition dynamics.

Table 1: Quantitative results (%) on the MWSC and MSLesSeg. Best results are **boldfaced**, while the best results among comparative methods are underlined.

Methods	Venues	MWSC				MSLesSeg			
		DSC↑	AVD↓	TPR↑	PPV↑	DSC↑	AVD↓	TPR↑	PPV↑
nnUNet	Nat.Methods'21	74.9±6.5	23.3±15.2	70.0±8.6	80.3±5.2	68.7±5.7	30.1±16.7	69.3±6.2	70.7±8.2
SwinUNETR	CVPR'22	69.7±5.7	26.6±14.0	69.5±9.8	72.3±5.2	67.5±5.9	29.4±16.5	68.8±6.3	69.5±8.7
SAM	ICCV'23	35.8±9.2	68.5±16.8	35.5±6.3	53.8±5.3	27.6±5.7	69.8±16.1	49.3±6.2	49.1±8.7
MedSAM	Nat.Commun'24	56.9±7.1	57.1±15.6	39.7±6.7	65.7±4.6	45.1±9.6	42.4±11.4	50.7±7.2	54.6±8.9
SAMUS	MICCAI'24	74.0±5.9	18.9±12.4	75.5±7.0	74.6±6.5	63.8±7.1	33.2±17.9	63.4±7.8	68.7±9.0
SAMCT	IEEE TMI'25	75.8±5.7	21.5±14.3	79.6±6.7	74.4±6.5	66.5±7.2	25.3±13.2	66.0±7.8	70.3±8.3
TP-Mamba	MICCAI'24	74.6±5.9	20.6±13.9	74.8±7.8	77.4±6.0	66.9±5.3	26.0±9.7	67.3±6.6	69.6±8.4
MaGuP	MICCAI'25	72.0±6.8	21.8±12.9	71.9±8.7	75.6±5.6	63.8±7.3	29.9±11.0	63.1±7.8	67.6±9.6
WMH-seg	NeuroImage'25	77.0±5.6	18.1±14.2	81.8±6.9	76.1±5.8	65.4±6.9	34.4±19.2	65.4±7.1	69.5±9.3
Deep-MS	ICCVW'25	75.1±5.9	20.8±13.5	79.2±7.4	73.3±6.1	70.2±5.0	25.8±16.3	72.5±6.2	70.4±7.6
Ours$^{\text{SAM}}$	This Paper	82.1±4.5	**13.1±6.8**	82.5±6.3	**82.5±4.4**	**73.5±2.4**	**16.5±6.3**	**75.3±6.7**	**73.1±4.2**
Ours$^{\text{MedSAM}}$	This Paper	**82.7±2.1**	13.8±7.5	**87.0±4.3**	80.9±4.2	72.1±4.9	19.1±7.9	72.8±3.5	71.1±5.0

After RA-SSM propagation, the enhanced regional sequences are remapped to their original flattened positions using the inverse permutation π_r^{-1}, and recomposed into a unified feature map aligned with the dimension of $\bar{\mathbf{F}}_b$. To further enforce spatial coherence, the recomposed feature map is processed by SS2D [20], which performs holistic SSM over the full spatial domain. Following the MUG, the resulting features from each branch are finally modulated by the cross-branch embeddings $\mathbf{Z}_\text{c}$ and $\mathbf{Z}_\text{s}$ (see Fig. 1(b)), enabling adaptive feature interaction.

Loss Functions. For segmentation supervision, we adopt the combination of dice (Dice) and binary cross-entropy (BCE) losses for both the final prediction $\mathbf{Y}'$ and CNN backbone prediction $\mathbf{Y}'_\text{c}$. To further promote structured region-aware representation learning, we introduce an auxiliary region-decomposition loss. For the l-th RPG and each branch b, the log-probability $\mathbf{M}^l_{b,r}$ is optimized by a region-separated mask $\mathbf{Y}_r$, which is derived from ground-truth (GT) $\mathbf{Y}$ via region parcellation. As the spatial resolution of $\mathbf{Y}_r$ does not necessarily match that of $\mathbf{M}^l_{b,r}$, we resize $\mathbf{Y}_r$ to (H_l, W_l) using an interpolation function $g(\cdot, l)$. We then compute the KL divergence between $g(\mathbf{Y}_r, l)$ and $\mathbf{M}^l_{b,r}$. To emphasize foreground structural learning, the loss is evaluated only over interior and boundary regions:

$$\mathcal{L}_\text{dec} = \sum_l \sum_b \sum_{r\in\{\text{in,bd}\}} \frac{1}{|\Omega^l_r|} \sum_{i\in\Omega^l_r} \text{KL}\left(g(\mathbf{Y}_r(i), l)||\mathbf{M}^l_{b,r}(i)\right), \tag{7}$$

where Ω^l_r is the set of spatial positions belonging to region r at layer l. The overall training objective is defined as: $\mathcal{L} = \mathcal{L}_\text{DiceBCE}(\mathbf{Y}', \mathbf{Y}) + \mathcal{L}_\text{DiceBCE}(\mathbf{Y}'_\text{c}, \mathbf{Y}) + \mathcal{L}_\text{dec}$.

3 Experiments and Results

3.1 Experimental Settings

Datasets and Evaluation Metrics. We assess the performance of all methods on WMH and MS segmentation. For WMH segmentation, we use the 2017 MICCAI WMH Segmentation Challenge (MWSC) dataset [16]. The dataset consists

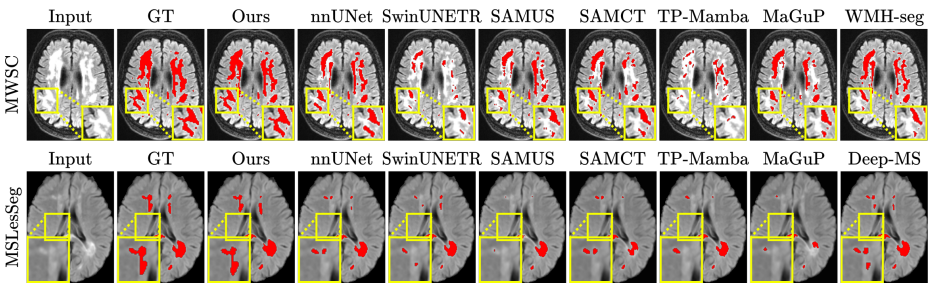

Fig. 2: Qualitative comparison on the MWSC and MSLesSeg datasets.

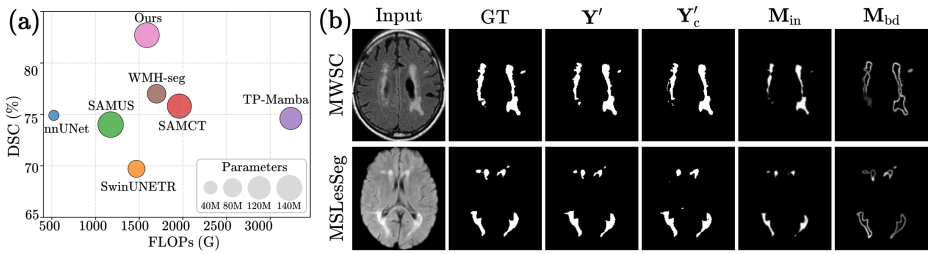

Fig. 3: Analysis of (a) computational cost and (b) module-wise output quality.

of 170 subjects, with 60 for training and 110 for testing. MS segmentation is evaluated on the publicly available MSLesSeg [6] dataset, which comprises 115 MRI scans acquired from 75 patients. The dataset is divided into 93 for training and 22 for testing. Among the multiple modalities in both datasets, we utilize only the FLAIR modality for both tasks. For quantitative evaluation, we employ the Dice Similarity Coefficient (DSC) and the Absolute Volume Difference (AVD), while voxel-wise True Positive Rate (TPR) and Positive Predictive Value (PPV) are additionally reported to measure voxel-level classification accuracy.

Baselines and Implementation Details. We compare our method against 10 state-of-the-art (SOTA) approaches: nnUNet [12], SwinUNETR [29], SAM [15], MedSAM [21], SAMUS [19], SAMCT [18], TP-Mamba [30], MaGuP [3], WMH-seg [8], and Deep-MS [24]. All experiments were implemented in PyTorch and executed on a single NVIDIA Quadro RTX 8000 GPU with 48GB of memory. Our model is trained for 100 epochs using the Adam optimizer with an initial learning rate of 1×10^{-4} and a batch size of 4. We adopt nnUNet as the CNN backbone, and data preprocessing follows its self-configuration pipeline [12]. The LoRA [10] is applied to efficiently fine-tune the SAM encoder $\mathcal{E}_s$ based on ViT-B, initialized with pretrained weights from SAM (using **default** settings unless otherwise specified) or MedSAM. The channel dimension C' in Eq. (1) is 64, and the embedding dimensions m and n in Eq. (2) are set to 64 and 16, respectively.

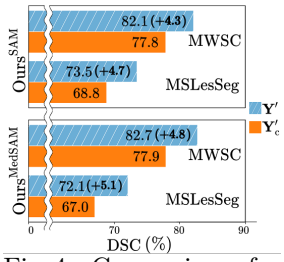

Fig. 4: Comparison for $\mathbf{Y}'$ and $\mathbf{Y}'_c$ DSC scores.

Table 2: Ablation results on key components.

Cases	HiRAM	RA-SSM	$\mathcal{L}_{dec}$	π_r	$\mathbf{P}_r$	MWSC		MSLesSeg	
						DSC↑	AVD↓	DSC↑	AVD↓
1	-	-	-	-	-	69.9	28.9	66.4	30.5
2	✓	-	-	-	-	70.2	26.4	66.3	31.4
3	✓	✓	-	-	-	74.6	20.1	68.5	27.9
4	✓	✓	✓	-	-	77.5	16.3	70.2	20.3
5	✓	✓	✓	✓	-	79.0	15.8	71.8	18.8
6	✓	✓	✓	-	✓	80.4	15.2	72.1	18.6
Ours	✓	✓	✓	✓	✓	**82.1**	**13.1**	**73.5**	**16.5**

3.2 Quantitative and Qualitative Evaluations

As reported in Table 1, our method (*i.e.*, **Ours**$^{\text{SAM/MedSAM}}$) outperforms all SOTA models on both datasets. In particular, we observe a clear improvement in DSC (approximately +4% over the best competing methods) and a notable reduction in AVD, suggesting more accurate lesion delineation and volume estimation. The qualitative results in Fig. 2 further support these findings. The SAM-CNN (SAMUS, SAMCT) and SAM-Mamba (TP-Mamba, MaGuP) methods tend to predict fragmented masks or mis-segmentation, where precise delineation is difficult. In contrast, our method produces more coherent, structurally sophisticated segmentation masks even with diluted or indistinct WM lesions.

Meanwhile, we investigate the computational complexity along with DSC scores to baselines on MWSC (see Fig. 3(a)). Notably, compared to models of similar size (*e.g.*, SAMCT and TP-Mamba), our method achieves the best performance while using FLOPs more efficiently. These indicate that HiRAM improves representation quality without introducing excessive computational overhead.

Impact of HiRAM. In Fig. 3(b), we analyze the effect of HiRAM and the quality for region-wise decomposition. We observe that $\mathbf{M}_{\text{in}}$ and $\mathbf{M}_{\text{bd}}$ are elaborately parcellated, with boundary responses concentrated around lesion contours. A comparison between $\mathbf{Y}'$ and $\mathbf{Y}'_c$ further verifies the benefit of incorporating HiRAM, as $\mathbf{Y}'$ exhibits improved lesion completeness and more accurate boundary delineation. This qualitative gain is also corroborated in Fig. 4, where HiRAM promotes the average DSC by +4.7% over $\mathbf{Y}'_c$, highlighting its effectiveness.

Ablation Study. Table 2 summarizes the contribution of each key component in our framework. The baseline without HiRAM (Case 1) exhibits relatively limited performance. Simply introducing SS2D (Case 2) yields only a marginal increase, suggesting that global SSM alone suffers from accurate lesion segmentation. Inclusion of RA-SSM with region decomposition (Case 3) improves substantially, with DSC increasing by ≈2–4% and AVD decreasing by ≈2–8% compared to Case 1. Further incorporating $\mathcal{L}_{\text{dec}}$ (Case 4) leads to additional gains thanks to enhanced region discrimination and boundary refinement. Employing

confidence-based permutation π_r (Case 5) or region-tailored priors $\mathbf{P}_r$ (Case 6) boosts the average DSC by +1.6% and +2.4%, respectively, compared to Case 4. Finally, the full model achieves the best performance, confirming that each design choice contributes complementary benefits to segmentation accuracy.

4 Conclusion

In summary, HiRAM introduces a hierarchical region-aware framework for WM lesion segmentation, combining region-tailored prior injection with multi-granularity propagation. Through confidence-driven region decomposition and adaptive SSM, our method effectively captures scattered and low-contrast lesions with indistinct boundaries. Extensive evaluations verify the effectiveness of our approach, coherently achieving SOTA performance, thereby establishing new benchmarks.

Acknowledgments. This work was supported by the Technology Innovation Program (or Industrial Strategic Technology Development Program) (RS-2025-02221011, Development of Medical-Specialized Multimodal Hyperscale Generative AI Technology for Global Integration) funded by the Ministry of Trade Industry & Energy (MOTIE, Korea).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chen, C., Miao, J., Wu, D., Zhong, A., Yan, Z., Kim, S., Hu, J., Liu, Z., Sun, L., Li, X., et al.: MA-SAM: Modality-agnostic SAM adaptation for 3D medical image segmentation. Med. Image Anal. **98**, 103310 (2024)
2. Debette, S., Markus, H.S.: The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. BMJ **341** (2010)
3. Dutta, T.K., Majhi, S., Nayak, D.R., Jha, D.: Mamba guided boundary prior matters: a new perspective for generalized polyp segmentation. In: MICCAI. pp. 380–391. Springer (2025)
4. García-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D.L., Collins, D.L.: Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. Med. Image Anal. **17**(1), 1–18 (2013)
5. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: COLM (2024)
6. Guarnera, F., Rondinella, A., Crispino, E., Russo, G., Di Lorenzo, C., Maimone, D., Pappalardo, F., Battiato, S.: MSLesSeg: baseline and benchmarking of a new multiple sclerosis lesion segmentation dataset. Sci. Data **12**(1), 920 (2025)
7. Han, D., Wang, Z., Xia, Z., Han, Y., Pu, Y., Ge, C., et al.: Demystify mamba in vision: A linear attention perspective. NeurIPS **37**, 127181–127203 (2024)
8. He, H., Jiang, J., Peng, S., He, C., Sun, T., Fan, F., Song, H., Sun, D., Xu, Z., Wu, S., et al.: A robust automated segmentation method for white matter hyperintensity of vascular-origin. NeuroImage **315**, 121279 (2025)

9. He, S., Bao, R., Li, J., Stout, J., Bjornerud, A., Grant, P.E., Ou, Y.: Computer-vision benchmark segment-anything model (sam) in medical images: Accuracy in 12 datasets. arXiv preprint arXiv:2304.09324 (2023)
10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022)
11. Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., et al.: Segment anything model for medical images? Med. Image Anal. **92**, 103061 (2024)
12. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
13. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. In: ICLR (2017)
14. Kalman, R.E.: A new approach to linear filtering and prediction problems. Journal of Basic Engineering **82**(1), 35–45 (1960)
15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., et al.: Segment anything. In: ICCV. pp. 4015–4026 (2023)
16. Kuijf, H.J., Biesbroek, J.M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M.J., et al.: Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge. IEEE Trans. Med. Imaging **38**(11), 2556–2568 (2019)
17. Liang, P., Shi, L., Pu, B., Wu, R., Chen, J., Zhou, L., Xu, L., Chen, Z., Chang, Q., Li, Y.: MambaSAM: A visual mamba-adapted SAM framework for medical image segmentation. IEEE J. Biomed. Health. Inf. **29**(8), 5824–5835 (2025)
18. Lin, X., Xiang, Y., Wang, Z., et al.: SAMCT: Segment any CT allowing labor-free task-indicator prompts. IEEE Trans. Med. Imaging **44**(3), 1386–1399 (2025)
19. Lin, X., Xiang, Y., Yu, L., Yan, Z.: Beyond adapting SAM: Towards end-to-end ultrasound image segmentation via auto prompting. In: MICCAI. pp. 24–34. Springer (2024)
20. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: Vmamba: Visual state space model. NeurIPS **37**, 103031–103063 (2024)
21. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nat. Commun. **15**(1), 654 (2024)
22. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
23. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment anything model for medical image analysis: an experimental study. Med. Image Anal. **89**, 102918 (2023)
24. Rondinella, A., Guarnera, F., Rundo, F., Battiato, S.: DeepSeg-MS: A 3D network based on hierarchical multi-scale learning for MRI multiple sclerosis lesion segmentation. In: ICCVW. pp. 7367–7376 (2025)
25. Salem, M., Hassan, S., Papineni, V.R.K., Elsayed, A., Yaqub, M.: DEFUSE-MS: Deformation field-guided spatiotemporal graph-based framework for multiple sclerosis new lesion detection. In: MICCAI. pp. 257–267. Springer (2025)
26. Shi, R., Pang, Q., Ma, L., Duan, L., Huang, T., Jiang, T.: Shapemamba-EM: Fine-tuning foundation model with local shape descriptors and mamba blocks for 3D EM image segmentation. In: MICCAI. pp. 731–741. Springer (2024)
27. Song, D., Huang, W., Liu, J., Islam, M.J., Yang, H., Wang, S., Zheng, H., Wang, S.: Optimized vessel segmentation: a structure-agnostic approach with small vessel enhancement and morphological correction. IEEE Trans. Image Process. (2025)

28. Sundaresan, V., Zamboni, G., Rothwell, P.M., Jenkinson, M., Griffanti, L.: Triplanar ensemble U-Net model for white matter hyperintensities segmentation on MR images. Med. Image Anal. **73**, 102184 (2021)
29. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3D medical image analysis. In: CVPR. pp. 20730–20740 (2022)
30. Wang, H., Lin, Y., Ding, X., Li, X.: Tri-plane mamba: Efficiently adapting segment anything model for 3D medical images. In: MICCAI. pp. 636–646. Springer (2024)
31. Wu, Y., Dong, Z., Li, H.B., Chong, Y.F., Ji, F., Chong, J.S.X., Tang, N.R.J., Hilal, S., Fu, H., Chen, C.L.H., et al.: WMH-dualtasker: A weakly supervised deep learning model for automated white matter hyperintensities segmentation and visual rating prediction. Hum. Brain Mapp. **46**(6), e70212 (2025)
32. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: SegMamba: Long-range sequential modeling mamba for 3D medical image segmentation. In: MICCAI. pp. 578–588. Springer (2024)
33. Yang, Z., Zhang, J., Wang, G., Kalra, M.K., Yan, P.: Cardiovascular disease detection from multi-view chest X-rays with bi-mamba. In: MICCAI. pp. 134–144. Springer (2024)
34. Ye, J., Dao, S.D., Wu, Y., George, Y., Nguyen-Duc, T., Schmidt, D.F., Shi, H., Chong, W., Cai, J.: New multiple sclerosis lesion segmentation via calibrated interpatch blending. In: MICCAI. pp. 365–375. Springer (2025)
35. Zha, Y., Wang, Y., Guo, H., Wang, J., Dai, T., Chen, B., Ouyang, Z., Yuerong, X., Chen, K., Xia, S.T.: PMA: Towards parameter-efficient point cloud understanding via point mamba adapter. In: CVPR. pp. 16976–16986 (2025)
36. Zhan, T., Qi, J., Zhang, J., Yu, X., Du, Q., Wu, Z.: Spatial-spectral feature–enhanced mamba and SAM-guided hyperspectral multi-class change detection. IEEE Trans. Geosci. Remote Sens. **63**, 1–13 (2025)
37. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 (2023)