SegMaST: Mamba-based Spatio-Temporal Modeling to Improve Longitudinal Disease Detection and Segmentation

Aswathi Varma[1,2,3] ASWATHI@TUM.DE
Jonas Weidner[1,3] J.WEIDNER@TUM.DE
Laurin Lux[2] LAURIN.LUX@TUM.DE
Cosmin Bercea[4,5] COSMIN.BERCEA@TUM.DE
Mark Mühlau[6] MARK.MUEHLAU@TUM.DE
Jan Kirschke[7] JAN.KIRSCHKE@TUM.DE
Benedikt Wiestler[*1,3] B.WIESTLER@TUM.DE
Daniel Rueckert[*2,3,8] DANIEL.RUECKERT@TUM.DE

[1] *AI for Image-Guided Diagnosis and Therapy, TUM, Munich, Germany*
[2] *Chair for AI in Healthcare and Medicine, Technical University of Munich (TUM) and TUM University Hospital, Munich, Germany*
[3] *Munich Center for Machine Learning, Munich, Germany*
[4] *Technical University of Munich, Munich, Germany*
[5] *Helmholtz AI and Helmholtz Center Munich, Munich, Germany*
[6] *Department of Neurology, TUM University Hospital, Munich, Germany*
[7] *Department of Neuroradiology, TUM University Hospital, Munich, Germany*
[8] *Department of Computing, Imperial College London, London, UK*

Editors: Under Review for MIDL 2026

Abstract

Longitudinal medical image segmentation is fundamental for quantifying disease progression and evaluating treatment efficacy. However, two critical challenges persist: First, methods that jointly segment baseline and follow-up images remain underexplored, often missing the contextual benefits of simultaneous assessment and lacking longitudinal consistency. Second, real-world datasets typically exhibit severe class imbalance, as scans showing actual disease progression are far rarer than those showing stable anatomy, an issue frequently neglected by existing models. To address these limitations, we propose *SegMaST*, a novel *Mamba*-based spatio-temporal framework. Unlike conventional approaches that treat timepoints in isolation, *SegMaST* leverages cross-temporal information and spatial correspondences to jointly segment the initial baseline mask and explicitly localize new or progressive pathologies in follow-up scans. Additionally, we introduce an imbalance-aware loss accumulation strategy to enhance robustness in realistic clinical settings. On longitudinal cohorts of patients with Multiple Sclerosis (MS) and glioma, *SegMaST* outperforms established CNN- and attention-based baselines for follow-up segmentation (mean follow-up Dice MS *in-house* 0.536, *MSSEG-2* 0.620, and glioma 0.631) and lesion detection (F1 *in-house* 0.688, *MSSEG-2* 0.723), while maintaining state-of-the-art accuracy in baseline segmentation (Dice: 0.617 MS, 0.844 glioma).

Keywords: Longitudinal Segmentation, State Space Models, Mamba, Imbalance Aware Loss

[*] Contributed equally as senior authors

1. Introduction

Longitudinal Magnetic Resonance Imaging (MRI) is the cornerstone of disease assessment for neurological pathologies. Unlike single-timepoint analysis, longitudinal imaging captures disease progression and therapy response over time (1). This is particularly critical in *Multiple Sclerosis* (MS), where the detection of new and evolving lesions serves as the primary biomarker for modifying treatment plans. Similarly, in *diffuse glioma*, precise follow-up segmentation is required to distinguish tumor recurrence from treatment effects. While manual volumetry remains the gold standard (2; 3), it is labor-intensive and prone to inter- and intra-variability. Deep learning offers a scalable solution to this bottleneck; however, popular architectures like *UNet* (4) and *Swin-UNETR* (5) typically process timepoints independently, ignoring the rich temporal correlations in serial imaging. This risks *longitudinal inconsistency*, where lesions appear and disappear between scans due to noise, whereas deformable registration introduces alignment errors that obscure progression.

To ensure longitudinal consistency, recent architectures explicitly model temporal dependencies. Simple input concatenation, as in *Neuropoly* (6), often fails to separate static anatomy from true temporal change. In contrast, *SNAC* (7) uses parallel encoders to compare features across multiple resolutions, offering clearer temporal cues. Advanced methods explicitly model scan differences to reduce longitudinal inconsistency: *CoActSeg* (10) uses voxel-wise difference maps, and the *Temporal Difference Weighting (TDW)* block (11) performs subtraction on latent features to amplify evolving regions.

Architectural design is often shaped by the available annotation scheme. One common strategy fully annotates both time points, enabling temporally consistent learning (12; 13) but at the cost of extensive manual labelling. Another approach reduces this burden by annotating only new or progressing abnormalities at follow-up (14), though models trained under this setting (6; 7) often lack the anatomical context needed to interpret changes. A third annotation configuration, followed in our work, strikes a balance by providing a full baseline segmentation while annotating only new or enlarging regions in the follow-up scans. However, prior efforts often suffer from significant limitations. Some rely on cumbersome auxiliary inputs (8), while others require labor-intensive, full-mask supervision for every longitudinal scan (9). Consequently, the hybrid setting remains underexplored.

Furthermore, typically a strong class imbalance of stable *vs.* progressive cases exists in real-world longitudinal data. This imbalance drives models toward two extremes: majority-class bias (missing subtle progression) or over-sensitivity (falsely flagging stable patients) - both challenge the safe clinical use of such models. To mitigate this issue, strategies range from augmenting training sets with synthetic abnormalities (15) to leveraging abundant cross-sectional data to compensate for limited longitudinal pairs (10). However, ensuring robustness across both progressing and non-progressing cases remains an open problem. Additionally, enhancing baseline segmentation alongside progression detection ensures more reliable longitudinal analysis. To this end, our contributions are as follows:

1. We introduce *SegMaST*, a *Mamba*-based spatio-temporal framework that exploits longitudinal dependencies through efficient state-space modeling to simultaneously refine baseline anatomical segmentation and precisely localize new or enlarging abnormalities in follow-up scans.

2. We address the clinically relevant issue that many scans show no disease activity at follow-up, by employing an *imbalance-aware loss accumulation*. This ensures robust performance on real-world clinical data where disease activity is intermittent.
3. We extensively validate *SegMaST* on two distinct pathologies (MS and glioma), demonstrating that it significantly reduces longitudinal inconsistency compared to both cross-sectional and longitudinal baselines, effectively distinguishing true disease progression from noise in stable and active cases.

2. Methodology

The above contributions motivate a methodological design that combines longitudinal feature modeling with stable learning in non-progressing cases, along with a joint prediction strategy for segmenting baseline regions and identifying follow-up abnormalities. Following this design, *SegMaST* processes baseline and follow-up scans within one coherent pipeline to generate both masks (see Figure 1-(a)). To balance performance and efficiency, we adopt a 2.5D architecture rather than a resource-intensive full-3D model, preserving contextual information while significantly lowering resource demands.

To effectively capture nuanced spatio-temporal interactions within this framework, our pipeline leverages *Mamba*-based state-space modeling. This approach addresses the computational bottlenecks inherent in standard spatial self-attention, which suffers from quadratic complexity ($O(N^2)$) in terms of sequence length. By strictly utilizing a selective state-space model (SSM) with linear complexity ($O(N)$) (18), we overcome the challenges that arise when flattening spatial patches results in sequence lengths that are prohibitive for explicit attention matrices. Consequently, we achieve global receptive fields comparable to *SegFormer* (16) while maintaining high throughput. We realize this design through a modular pipeline consisting of a hierarchical encoder, where Spatio-Temporal (ST) blocks capture longitudinal disease activity at multiple scales, and a dual-head decoder, detailed as follows.

Hierarchical Encoder. To capture features ranging from fine-grained lesion boundaries to global semantic context, *SegMaST* employs a four-layer hierarchical encoder. For an input pair with spatial dimensions $H \times W$ and channels C, we define the input tensor as $X \in \mathbb{R}^{N \times C \times H \times W}$, where $N = 2$ represents the timepoints. We first apply a convolutional patch embedding layer to tokenize the input while preserving local spatial continuity, yielding $X_1 \in \mathbb{R}^{N \times C_1 \times \frac{H}{4} \times \frac{W}{4}}$. In subsequent stages, we progressively increase the receptive field, such that the feature tensor at layer i follows $X_i \in \mathbb{R}^{N \times C_i \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}}$.

ST Block and MaST Module. We employ ST blocks to capture joint dynamics from the encoder features. Central to this block is the *Mamba* Spatio-Temporal (*MaST*) module, which processes the input by layer normalizing and reshaping it into two complementary sequences: (1) **Temporal-first** ($X_{i,t} \in \mathbb{R}^{C_i \times N(H/2^{i+1}W/2^{i+1})}$), where spatial patches are unfolded and concatenated along the time axis; (2) **Spatial-first** ($X_{i,s} \in \mathbb{R}^{C_i \times (H/2^{i+1}W/2^{i+1})N}$), where patches are stacked to preserve spatial correspondence across time. These flattened sequences, which create contexts too long for standard attention, are efficiently processed by the *Mamba* SSM (18). By compressing context into a hidden state rather than calculating pairwise interactions, the SSM enables the effective learning of spatio-temporal dependencies with linear complexity. We repeat this ST block $M = 4$ times before applying overlapped patch merging to downsample features for the next stage.

Decoder. We utilize a *CNN*-based decoder to generate segmentation masks, adopting the lightweight design of (17; 16). First, an *MLP* unifies the multi-level features from the encoder along the channel dimension. These unified representations are upsampled to a common spatial resolution and concatenated. A subsequent *MLP* fuses this aggregated representation, projecting the concatenated tensor $(N, 4C, H/4, W/4)$ to a lower embedding dimension C. The resulting features are then reshaped to recover the temporal structure, separating the latent representations of the two time points. Finally, a dual-head prediction module applies a 1×1 convolution to the respective time steps, producing the baseline segmentation $\hat{Y}_1$ from the first time point and the progression segmentation $\hat{Y}_P$ from the second time point.

Imbalance-Aware Loss Accumulation. In clinical settings, follow-up datasets are dominated by stable cases. These cases have *empty* progression masks. Progressive cases, which contain *non-empty* masks, are comparatively rare. This imbalance skews model predictions and reduces sensitivity to disease evolution (19). We address this through a combination of an *imbalance-aware* loss accumulation that filters some empty masks and a sampling strategy that maintains a balanced progression-to-no-progression ratio per batch. This prevents gradients from being dominated by empty masks while preserving useful learning signals.

Given a batch of segmentation outputs, let $Y_1, \hat{Y}_1$ denote the ground truth and predicted baseline masks, and $Y_P, \hat{Y}_P$ denote the ground truth and predicted progression masks. We use the *Dice-Focal* loss, computed separately for the baseline and progression heads, and sum them to obtain the final loss. To prevent trivial cases from dominating training, we apply a filtering mechanism for progression head loss, as many samples have no progression $(Y_P = 0)$.

Filtering Mechanism. The total loss consists of two terms: the baseline head loss $(\mathcal{L}_1)$ and a subset of the progression head loss $(\mathcal{L}_p)$. We define a binary indicator M, indicating whether a sample contains progression. Using this value, we separate the progression loss values into:

$$\mathcal{L}_p^+ = \mathcal{L}_p[M], \quad \mathcal{L}_p^0 = \mathcal{L}_p[\neg M]$$

Here $\mathcal{L}_p^+$ represents the loss for samples with progression $(M = 1)$, while $\mathcal{L}_p^0$ corresponds to the loss for no-progression samples $(M = 0)$. Since we want the model to learn from no-progression cases while preventing imbalance, we randomly sample a subset from $\mathcal{L}_p^0$, matching the number of progression cases in the batch. This selected subset is included in the final loss term, denoted as $\mathcal{L}_p^\emptyset$. The filtered progression loss is obtained by concatenating the two:

$$\mathcal{L}_p^{\text{filtered}} = \mathit{Concat}(\mathcal{L}_p^+, \mathcal{L}_p^\emptyset).$$

The final loss is computed as the sum of mean terms, as shown in Figure 1 - (b), where B is the original batch size and B' is the number of samples used in the progression head after filtering. There is a trade-off between excluding all zero masks, which could bias training towards detecting only positive cases, and retaining all zero masks, which could lead to the model ignoring changes in progression. Our approach ensures a proportionate distribution,

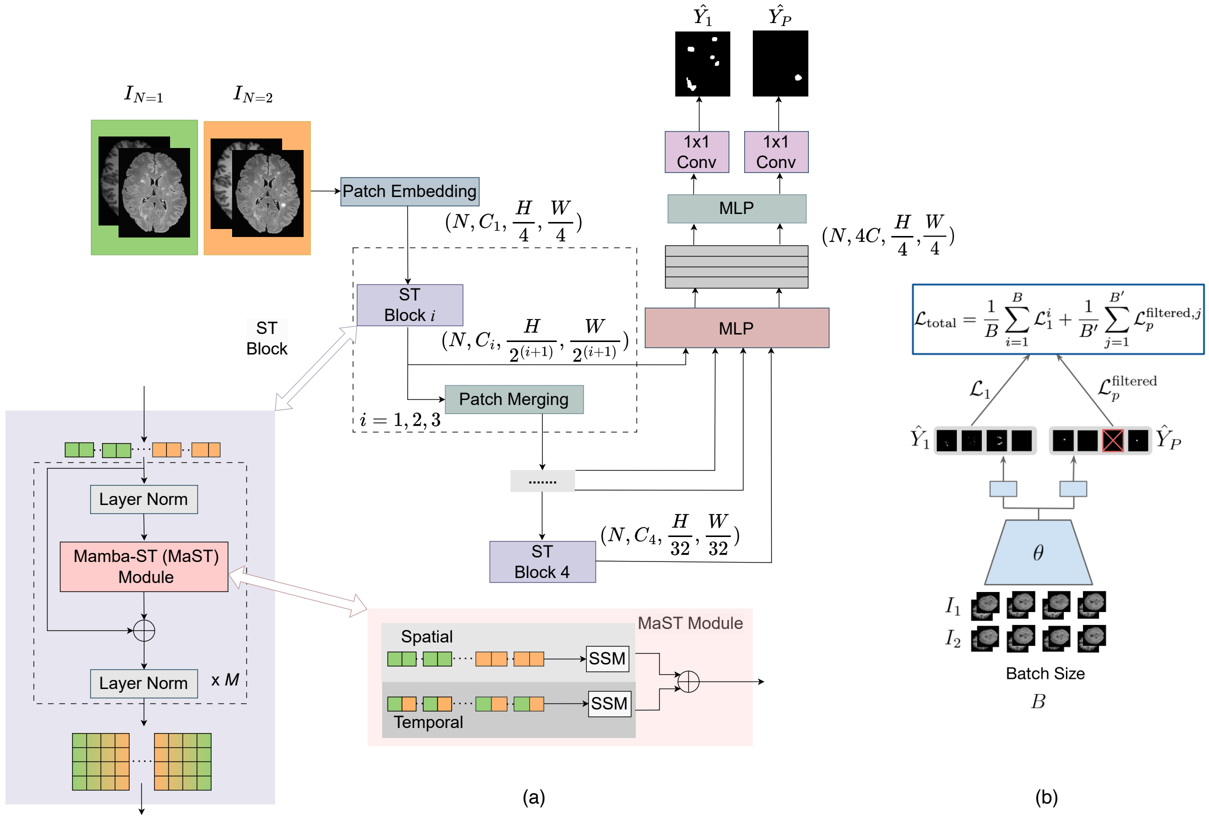
$I_{N=1}$
$I_{N=2}$
Patch Embedding
$(N, C_1, \frac{H}{4}, \frac{W}{4})$
ST Block i
$(N, C_i, \frac{H}{2^{(i+1)}}, \frac{W}{2^{(i+1)}})$
Patch Merging
$i = 1, 2, 3$
....
ST Block 4
$(N, C_4, \frac{H}{32}, \frac{W}{32})$
MLP
MLP
$(N, 4C, \frac{H}{4}, \frac{W}{4})$
1x1 Conv
1x1 Conv
$\hat{Y}_1$
$\hat{Y}_P$
ST Block
Layer Norm
Mamba-ST (MaST) Module
Layer Norm
x M
MaST Module
Spatial
SSM
SSM
Temporal
(a)
$\mathcal{L}_{\text{total}} = \frac{1}{B}\sum_{i=1}^{B}\mathcal{L}_1^i + \frac{1}{B'}\sum_{j=1}^{B'}\mathcal{L}_p^{\text{filtered},j}$
$\mathcal{L}_1$
$\mathcal{L}_p^{\text{filtered}}$
$\hat{Y}_1$
$\hat{Y}_P$
θ
I_1
I_2
Batch Size
B
(b)

Figure 1: (a) *SegMaST* Architecture. Longitudinal image pairs are processed by a hierarchical encoder utilizing Spatio-Temporal (ST) blocks, where a *Mamba*-based module (MaST) efficiently captures spatial and temporal dependencies. Multi-scale features are aggregated by a *CNN*-based decoder, which feeds a dual-head module to predict baseline ($\hat{Y}_1$) and progression ($\hat{Y}_P$) masks. (b) *Imbalance-Aware Loss Accumulation.* The total loss combines the baseline term ($\mathcal{L}_1$) with a filtered progression term ($\mathcal{L}_p^{\text{filtered}}$). This mechanism selectively retains a balanced subset of zero-mask samples to prevent the high prevalence of non-progression from biasing the model.

striking a balance between these extremes. The impact of different proportions is analyzed in an ablation study (section 4.1).

3. Experimental Setup

Datasets. We evaluate *SegMaST* on three datasets spanning two distinct brain pathologies: MS and diffuse glioma. The MS dataset was derived from a well-characterized subgroup of the cohort of our *in-house* observational MS study, TUM-MS (20). This dataset contains MRI scans at baseline, 6, 12, and 24 months (3–4 time points per patient). Each subject has a baseline whole-lesion mask and new or enlarging lesion masks for follow-ups. We use 3D FLAIR and T2 volumes ($1 \times 1 \times 1$ mm^3) for binary segmentation of baseline lesions and progression, with 224 successive time-point pairs (125 without progression and 99 with progression). The second MS dataset is the public *MSSEG-2* dataset, which provides only new-lesion annotations of follow-up scans. We use the official training split, which consists of two-time-point 3D FLAIR scans (1×1×1 mm^3), comprising 40 timepoint pairs (11 without progression and 29 with progression). Follow-up scans were acquired 1–3 years after the baseline scan. Following the challenge protocol, the validation set excludes subjects without new lesions.

The third dataset, *UCSF-ALTPD* (21), consists of multimodal MRI scans from glioma patients with two consecutive follow-up time points. The preprocessing includes skull stripping (23), N4 bias correction (22), coregistration, and rigid SRI24 atlas registration (24). Each case provides FLAIR, T1, contrast-enhanced T1 (T1-ce), and T2 volumes (1×1×1 mm^3), along with tumor masks delineating enhancing tissue (ET), surrounding nonenhancing FLAIR hyperintensity (SNFH), nonenhancing tumor core (NETC), and resection cavity (RC). In our setup, the baseline head performs binary whole-tumor segmentation, while the progression head predicts a binary difference mask between the two time points. We randomly sample 200 patients for training and evaluation (88 without progression and 112 with progression).

Training and Evaluation. *SegMaST* employs a 2.5D training strategy processing 2D slices independently, while 3D inference averages probability maps from three orthogonal views. We include both 2.5D and 3D spatiotemporal baselines. For 2.5D, we adopt a *SegFormer*-based variant. For 3D, we evaluate *DynUNet* (MONAI's *nnUNet*), *SwinUNETR*, and *LongiUNet-DW*, extending a *UNet* with *TDW* blocks (11). For the *in-house* MS dataset, we additionally train two cross-sectional *DynUNet* models (*DynUNet (CS)*), estimating progression by subtraction. All baselines for *in-house* and *UCSF-ALTPD* datasets use dual convolutional heads to match *SegMaST*. For *MSSEG-2*, we also compare against *CoactSeg*, *Neuropoly*, and *SNAC* (10).

We employ the weighted patch-sampling strategy of (25) with crop sizes tailored to the datasets (*In-house*: 128; *MSSEG-2*: 80; *UCSF-ALTPD*: 160). Standard geometric augmentations, such as random flips and rotations, are applied. For all models, redicted components smaller than minimum volume thresholds (27 mm^3 for MS, 50 mm^3 for UCSF-ALTPD) are suppressed (26). Across datasets, 2.5D models remain lightweight (∼9M for MS, ∼16M for diffuse glioma) and are trained for up to 150–500 epochs with early stopping, using a batch size of 32. The 3D models are larger (∼15M for MS, ∼25M for diffuse glioma) and follow longer schedules of up to 1000 epochs, with a batch size of 8. All models are optimized with AdamW (27) (weight decay 0.05, learning rate 10^{-4}) using cosine decay (28) and a 40-epoch warm-up (29). The proposed *imbalance-aware loss accumulation* is applied throughout, and training is performed on an NVIDIA RTX A6000 GPU.

In accordance with recent recommendations for image analysis validation (30), we use the Dice score to assess voxel-wise segmentation quality. To capture lesion-level detection performance, we further report the lesion-wise F1 score. Following the methodology established by (31), we exclude tiny lesions (defined as fewer than 11 voxels) from the F1 calculation to filter out noise and focus on clinically significant observations. We apply the Dice score to both baseline and progression regions, with progression metrics further stratified into progression (P) and non-progression (NP) cases. For NP cases, we assign a strict Dice score of 1 when the predicted progression mask is empty and 0 otherwise. Additionally,

Table 1: Baseline (B), progression (P) and no-progression (NP) metrics for the *in-house* MS dataset. Values are mean (SD). **Bold** and underline denote best and second-best. * and ** indicate statistical differences to *SegMaST* using a paired t-test with significance values <0.05 and <0.01, respectively. Disease Activity Assessment (DAA) shows the percentage of correctly classified follow-up scans.

Model	**Dim**	**Dice ↑**			**Lesion-F1 ↑**		**PPV ↑**		**DAA (%) ↑**
		B	P	NP	B	P	B	P	All
DynUNet (CS)	3D	0.620 (0.162)	0.170** (0.166)	0.000** (0.000)	0.552 (0.158)	0.157** (0.192)	0.580** (0.179)	0.134** (0.127)	28.9%
LongiUNet-DW	3D	0.510** (0.175)	0.380** (0.256)	0.520 (0.500)	0.526 (0.186)	0.455 (0.368)	0.380** (0.172)	0.343* (0.277)	**64.4%**
Swin UNETR	3D	**0.657**** (0.149)	0.371** (0.317)	0.520 (0.500)	**0.621*** (0.132)	0.440* (0.402)	**0.800**** (0.122)	0.493 (0.413)	57.8%
SegFormer	2.5D	0.533** (0.180)	0.396 (0.290)	0.360** (0.480)	0.451** (0.151)	0.528 (0.355)	0.534** (0.217)	0.389 (0.314)	55.6%
SegMaST (ours)	2.5D	0.617 (0.148)	**0.513** (0.270)	**0.560** (0.496)	0.567 (0.148)	**0.688** (0.320)	0.648 (0.193)	**0.494** (0.286)	**64.4%**

Table 2: Comparison of segmentation models on the *MSSEG-2* dataset, with progression subjects only. Values are mean (SD). **Bold** and underline denote best and second-best.

Model	**Dim**	**Dice ↑**	**Lesion-F1 ↑**	**PPV ↑**
SNAC	3D	0.531	0.307	*N/A*
SNAC (VNet)	3D	0.568	0.576	*N/A*
Neuropoly	3D	0.563	0.175	*N/A*
CoactSeg	3D	**0.638**	0.620	0.636
DynUNet	3D	0.586 (0.238)	0.622 (0.276)	0.636 (0.325)
LongiUNet-DW	3D	0.616 (0.177)	0.684 (0.229)	0.569 (0.216)
SwinUNETR	3D	0.538 (0.280)	0.557 (0.330)	**0.642** (0.355)
SegFormer	2.5D	0.532 (0.226)	0.518 (0.308)	0.621 (0.327)
SegMaST (ours)	2.5D	0.620 (0.201)	**0.723** (0.235)	0.630 (0.281)

we compute the Disease Activity Assessment (DAA), a clinically relevant score indicating the proportion of follow-up scans that are correctly assessed. The *in-house* MS set has 45 patients (25 NP), while *MSSEG2* is progression-only. For MS datasets, we report Dice, F1, and *PPV*. *UCSF-ALTPD* comprises 25 patients (16 NP).

4. Results

For the *in-house* MS dataset, Table 1 reports the test-set performance for baseline and progression predictions. For the baseline head, 3D *Swin UNETR* achieves the highest Dice and F1, while *SegMaST* remains close at far lower computational cost. For progression, *SegMaST* is the best-performing model, achieving the highest P Dice (0.513) and lesion-wise F1 (0.688), markedly outperforming all 3D models and the next-best 2.5D baseline, *SegFormer*. In NP cases, *SegMaST* again attains the highest Dice (0.560), with *Swin UNETR*

Table 3: Comparison of segmentation models for baseline (B), progression (P), and no progression (NP) mask prediction in glioma cases. Values are mean (SD). **Bold** and underline denote best and second-best. * and ** indicate statistical differences to *SegMaST* using a paired t-test with significance values <0.05 and <0.01, respectively.

Model	Dim	Dice ↑		
		B	P	NP
DynUNet	3D	**0.846** (0.092)	**0.520** (0.096)	0.125 (0.331)**
LongiUNet-DW	3D	0.836 (0.096)	0.496 (0.141)	0.000 (0.000)**
Swin UNETR	3D	0.784 (0.153)*	0.449 (0.131)	0.063 (0.242)**
SegFormer	2.5D	0.791 (0.126)*	0.497 (0.128)	0.500 (0.500)*
SegMaST (ours)	2.5D	0.844 (0.075)	0.511 (0.132)	**0.750** (0.433)

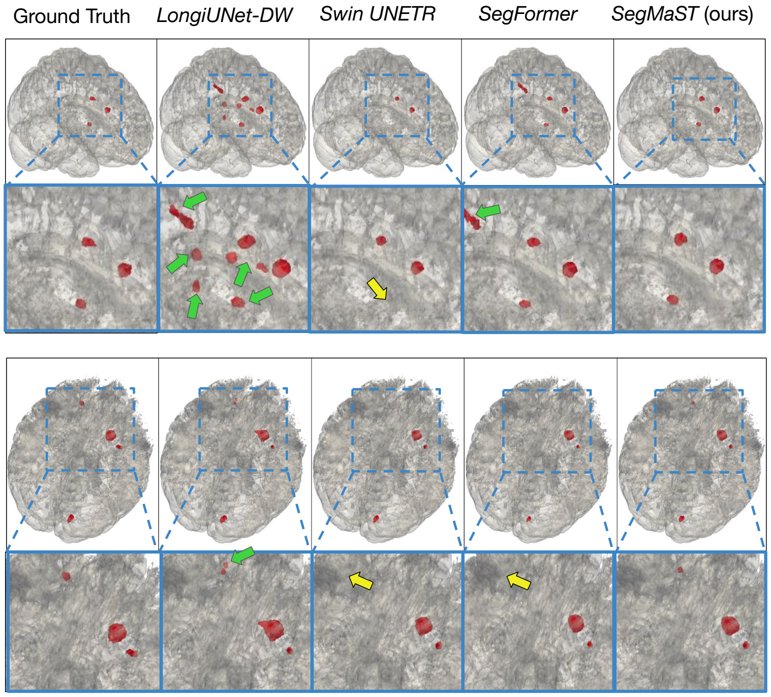

Figure 2: Exemplar results of *SegMaST* and other models for MS lesion progression segmentation. **Top:** *in-house* dataset. **Bottom:** *MSSEG-2* dataset. Green arrows indicate false positives, and yellow arrows indicate false negatives.

and *LongiUNet-DW* following (0.520). The subtractive *DynUNet (CS)* baseline performs poorly on progression metrics. Overall, *SegMaST* delivers the most consistent and accurate longitudinal segmentation performance, as visualized in Figure 2. Clinically, *SegMaST* also provides the most reliable DAA, achieving the highest progression *PPV*. On *MSSEG-2*, *SegMaST* achieves the highest lesion-wise F1 (0.723), surpassing all prior methods. It also attains a competitive Dice score of 0.620, on par with top 3D models such as *CoactSeg*.

The mean test-set metrics for baseline whole-tumor segmentation and progression/no-progression prediction on the *UCSF-ALTPD* dataset are shown in Table 3. For the baseline task, *DynUNet* achieves the highest Dice score (0.846), closely followed by *SegMaST* (0.844). For the progression task, *SegMaST* provides the best overall performance, achieving the highest Dice for NP cases (0.750) while obtaining P Dice that are comparable to the strongest baselines. Notably, the 3D baselines struggle with NP cases, underscoring the difficulty of accurately modeling stable follow-up scans in these architectures.

4.1. Ablation Study - Contribution of No Progression Cases

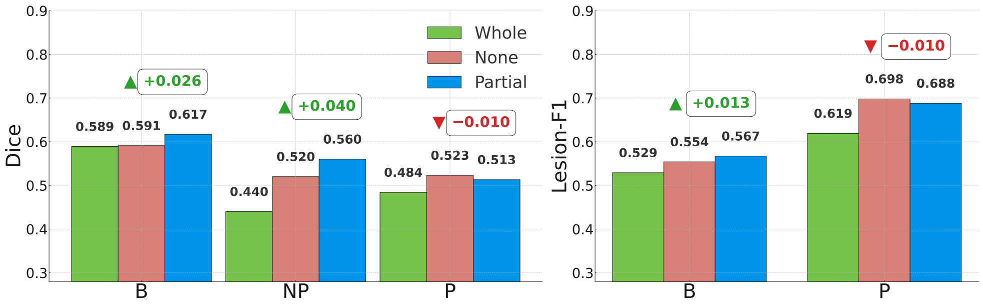

Figure 3: Comparison of three strategies for handling the imbalance between progression (P) and no-progression (NP) cases. *Whole* uses all NP cases, *None* excludes them, and the proposed *Partial* strategy includes a balanced subset. Deviations indicate differences between the *Partial* and *None* settings.

To counter the over-dominance of non-progression cases (which essentially are empty segmentation masks), we compare three distinct loss accumulation strategies: (1) **Whole:** all NP cases included, (2) **None:** all NP cases excluded, and (3) Our proposed **Partial:** or *imbalance-aware loss accumulation*, where a subset of NP cases is included, maintaining an equal number of NP and P cases per batch. As shown in Figure 3, the Whole strategy yields the lowest P Dice (0.484) and P F1 (0.619) due to the dominance of NP samples. The None setting achieves the highest P Dice (0.523) and P F1 (0.698) but performs poorly on B and NP. Our Partial strategy offers the best overall balance, obtaining the highest B Dice (0.617), NP Dice (0.560), and B F1 (0.567), while maintaining strong P F1 (0.688). This demonstrates that *imbalance-aware loss accumulation* effectively balances B, P, and NP performance, an important aspect for clinical use.

4.2. Ablation Study - Lesion Size-Wise Analysis

To evaluate the robustness of *SegMaST* across varying lesion volumes, we perform a size-wise analysis on the *in-house* MS dataset, reporting results for baseline lesions and progression lesions (Table 4). For the baseline segmentation (Left), *SegMaST* demonstrates superior performance across medium and large lesions, achieving the highest Lesion-F1 and Dice scores in both the Medium (Lesion-F1: 0.746, Dice: 0.424) and Large (Lesion-F1: 0.915,

Table 4: Size-wise lesion analysis on the private MS dataset. **Left:** baseline lesion analysis. Lesion bins are defined using global tertile-based volume thresholds across the dataset ($1\,\text{mm}^3$ isotropic voxels): Small (≤ 84) mm^3, Medium (84–208) mm^3, Large (> 208) mm^3. **Right:** analysis of new lesions in follow-up scans. Lesion bins use volume thresholds: Small (≤ 43) mm^3, Medium (43–113) mm^3 and Large (> 113) mm^3.

Model	**Lesion-F1 ↑**	**Dice ↑**
Small ($N = 394$)		
DynUNet	0.176	0.076
LongiUNet-DW	0.186	**0.216**
Swin UNETR	**0.298**	0.110
SegFormer	0.139	0.087
SegMaST (ours)	0.292	0.196
Medium ($N = 388$)		
DynUNet	0.525	0.272
LongiUNet-DW	0.609	0.382
Swin UNETR	0.699	0.345
SegFormer	0.625	0.302
SegMaST (ours)	**0.746**	**0.424**
Large ($N = 392$)		
DynUNet	0.840	0.525
LongiUNet-DW	0.850	0.523
Swin UNETR	0.891	0.601
SegFormer	0.816	0.530
SegMaST (ours)	**0.915**	**0.617**

Model	**Lesion-F1 ↑**	**Dice ↑**
Small ($N = 20$)		
DynUNet	0.034	0.010
LongiNet-DW	0.136	0.035
Swin UNETR	0.000	0.000
SegFormer	0.229	**0.185**
SegMaST (ours)	**0.235**	0.097
Medium ($N = 21$)		
DynUNet	0.081	0.172
LongiNet-DW	0.421	0.450
Swin UNETR	0.376	0.309
SegFormer	0.415	0.460
SegMaST (ours)	**0.703**	**0.598**
Large ($N = 21$)		
DynUNet	0.293	0.175
LongiNet-DW	0.871	0.612
Swin UNETR	0.761	0.634
SegFormer	0.746	0.683
SegMaST (ours)	**0.889**	**0.720**

Dice: 0.617) bins. In the challenging progression lesion analysis (Right), *SegMaST* exhibits a dominant trend, achieving the highest Lesion-F1 scores across all three bins: Small (0.235), Medium (0.703), and Large (0.889). It also achieves the best Dice scores for both Medium (0.598) and Large (0.720) progression lesions. This strong performance, particularly in the difficult small-to-medium progression bins, validates the ability of our *Mamba*-based spatio-temporal modeling to effectively capture subtle changes indicative of disease progression across different lesion scales.

5. Discussion & Conclusion

Longitudinal segmentation plays a crucial role in monitoring disease progression, yet existing methods often face challenges related to imbalanced datasets or foregoing the rich spatio-temporal information. To address these issues, we introduce *SegMaST*, a novel approach that jointly segments baseline and follow-up images while incorporating an *imbalance-aware loss accumulation* strategy to effectively manage the dominance of non-progression cases. Our results demonstrate that *SegMaST* outperforms established baseline methods, including both convolutional and attention-based models, by leveraging the rich spatio-temporal synergies inherent in medical imaging data. Notably, this performance improvement is consistent across two distinct and challenging clinical scenarios: multiple small, newly appearing lesions, as seen in MS, and the continuous growth patterns characteristic of gliomas. These findings underscore the versatility and robustness of *SegMaST* in advancing longitudinal medical image analysis. We make our codes for *SegMaST* and the *imbalance-aware loss accumulation* publicly available at https://github.com/Aswathi-Varma/SegMaST.

Acknowledgments

Aswathi Varma, Mark Mühlau, Benedikt Wiestler, and Daniel Rueckert are supported by the DFG as part of the SPP *Radiomics* (project number 428223038).

References

[1] Martí-Juan, Gerard, Gerard Sanroma-Guell, and Gemma Piella. "A survey on machine and statistical learning for longitudinal analysis of neuroimaging data in Alzheimer's disease." Computer Methods and Programs in Biomedicine 189 (2020): 105348.

[2] Barkhof, Frederik, et al. "2024 MAGNIMS–CMSC–NAIMS consensus recommendations on the use of MRI for the diagnosis of multiple sclerosis." The Lancet Neurology 24.10 (2025): 866-879.

[3] Wen, Patrick Y., et al. "RANO 2.0: update to the response assessment in neuro-oncology criteria for high-and low-grade gliomas in adults." Journal of Clinical Oncology 41.33 (2023): 5187-5199.

[4] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation." MICCAI, pp. 234–241, 2015.

[5] Tang, Y., Chen, D. F., Yan, Y., et al. "Swin-UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images." WACV, 2022.

[6] Macar, Uzay, et al. "Team neuropoly: description of the pipelines for the MICCAI 2021 MS new lesions segmentation challenge." arXiv preprint arXiv:2109.05409 (2021).

[7] Cabezas, Mariano, et al. "Estimating lesion activity through feature similarity: A dual path Unet approach for the MSSEG2 MICCAI challenge." MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure 107 (2021).

[8] Basaran, Berke Doga, Paul M. Matthews, and Wenjia Bai. "SegHeD+: Segmentation of Heterogeneous Data for Multiple Sclerosis Lesions with Anatomical Constraints and Lesion-aware Augmentation." arXiv preprint arXiv:2412.10946 (2024).

[9] Denner, Stefan, et al. "Spatio-temporal learning from longitudinal data for multiple sclerosis lesion segmentation." International MICCAI Brainlesion Workshop. Cham: Springer International Publishing, 2020.

[10] Wu, Yicheng, et al. "Coactseg: Learning from heterogeneous data for new multiple sclerosis lesion segmentation." International conference on medical image computing and computer-assisted intervention. Cham: Springer Nature Switzerland, 2023.

[11] Rokuss, Maximilian R., et al. "Longitudinal segmentation of MS lesions via temporal difference weighting." MICCAI, 2024.

[12] Carass, A., Roy, S., Jog, A., et al. "Longitudinal multiple sclerosis lesion segmentation: Resource and challenge." NeuroImage 148 (2017): 77–102.

[13] Wei, Jie, et al. ”Consistent segmentation of longitudinal brain MR images with spatio-temporal constrained networks.” MICCAI 2021.

[14] Commowick, O., Cervenansky, F., Cotton, F., Dojat, M. ”Multiple sclerosis new lesions segmentation challenge.” MSSEG-2 Challenge, MICCAI 2021.

[15] Tahghighi, Peyman, et al. ”Enhancing New Multiple Sclerosis Lesion Segmentation via Self-supervised Pre-training and Synthetic Lesion Integration.” International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2024.

[16] Xie, Enze, et al. ”SegFormer: Simple and efficient design for semantic segmentation with transformers.” NeurIPS 2021.

[17] Yang, Yijun, et al. ”ViViM: A Video Vision Mamba for Medical Video Segmentation.” arXiv:2401.14168, 2024.

[18] Gu, Albert, and Tri Dao. ”Mamba: Linear-time sequence modeling with selective state spaces.” arXiv:2312.00752, 2023.

[19] Karimian-Jazi, Kianush, et al. ”Gd contrast administration is dispensable in patients with MS without new T2 lesions on follow-up MRI.” Neurology: Neuroimmunology & Neuroinflammation 5.5 (2018): e480.

[20] Bayas, Antonios, et al. ”Prospective study validating a multidimensional treatment decision score predicting the 24-month outcome in untreated patients with clinically isolated syndrome and early relapsing–remitting multiple sclerosis, the ProVal-MS study.” Neurological Research and Practice 6.1 (2024): 15.

[21] Fields, Brandon KK, et al. ”The UCSF Adult Longitudinal Post-Treatment Diffuse Glioma MRI Dataset.” Radiology: AI 6.4 (2024).

[22] Tustison, Nicholas J., et al. “N4ITK: Improved N3 Bias Correction.” IEEE Transactions on Medical Imaging, vol. 29, no. 6, 2010, pp. 1310–1320, https://doi.org/10.1109/TMI.2010.2046908 .

[23] Isensee, Fabian, et al. “Automated Brain Extraction of Multisequence MRI Using Artificial Neural Networks.” Human Brain Mapping, 2019, https://doi.org/10.1002/hbm.24750 .

[24] Rohlfing, Torsten, et al. “The SRI24 Multimodal Neuroimaging Reference Annotations.” NeuroImage, vol. 51, no. 2, 2010, pp. 752–762, https://doi.org/10.1016/j.neuroimage.2010.02.106 .

[25] Zhang, Hang, et al. ”QSMRim-Net: Imbalance-aware learning for identification of chronic active multiple sclerosis lesions on quantitative susceptibility maps.” NeuroImage: Clinical 34 (2022): 102979.

[26] R. Saluja, BraTS-2023-Metrics, GitHub, 2023. [Online]. Available: https://github.com/rachitsaluja/BraTS-2023-Metrics

[27] Loshchilov, Ilya and Frank Hutter. ”Decoupled weight decay regularization.” arXiv:1711.05101, 2017.

[28] Loshchilov, Ilya and Frank Hutter. ”SGDR: Stochastic gradient descent with warm restarts.” arXiv:1608.03983, 2016.

[29] Goyal, Priya, et al. ”Accurate, large minibatch SGD: Training ImageNet in 1 hour.” arXiv:1706.02677, 2017.

[30] Maier-Hein, Lena, and Bjoern Menze. ”Metrics reloaded: Pitfalls and recommendations for image analysis validation.” arXiv. org 2206.01653 (2022).

[31] Commowick, Olivier, et al. ”Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure.” Scientific reports 8.1 (2018): 13650.

6. Appendix

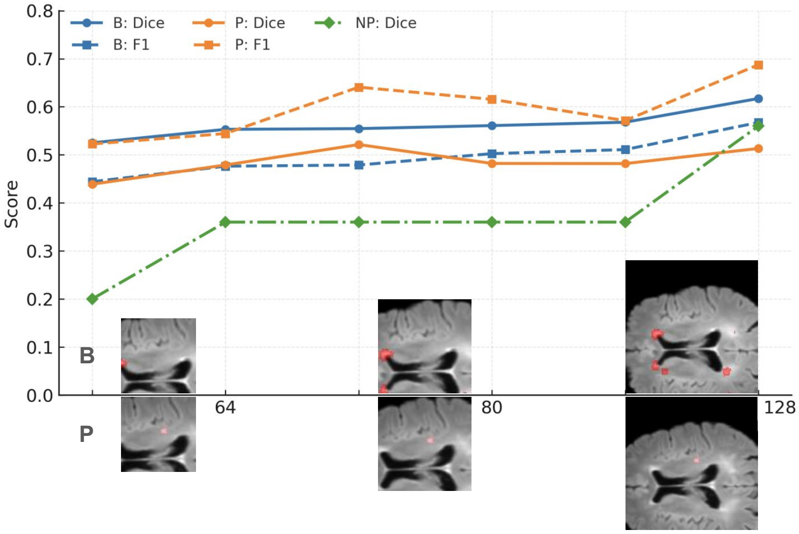

Figure 4: Ablation on patch size under weighted cropping. Larger patches (40^2 to 128^2) provide more global context and improve Dice and F1 performance on the *in-house* dataset.

Effect of Weighted Cropping Patch Size. When applying weighted cropping on the in-house dataset, the results show that larger patch sizes consistently outperform smaller ones in terms of both baseline Dice and P F1 (Figure 4). Weighted cropping makes small patches (e.g., 40^2) highly local and biased toward lesion-centered regions, limiting global contextual cues and resulting in lower baseline Dice (0.525) and weaker NP F1 (0.523). As patch size increases, the model benefits from more anatomical context, which improves its ability to distinguish true progression from stable regions. This is reflected in the steady rise in baseline Dice, reaching 0.617 at 128^2, and in the pronounced improvement in **non-empty progression F1, which peaks at 0.688 at the same size. Progression Dice follows a similar upward trend, increasing from 0.439 at 40^2 to above 0.51 at 128^2. Overall, combining weighted cropping with larger patches (128^2) provides the optimal balance: weighted sampling focuses the model on lesion-relevant areas, while the larger patch size supplies the necessary global context, yielding the strongest Dice and F1 performance for progression detection.