Token Mixer Placement Matters: A Systematic Encoder–Decoder Ablation Study of Mamba for Brain Tumour Segmentation on BraTS-Africa

Abdulbasit Oyetunji[1,4], Olamide Lawal[2,4], Kingsley Ibekwe[3,4], Grace Alele[4], Joy Abiodun[5,4], and Raymond Confidence[6]

[1]College of Medicine, University of Ibadan, Nigeria
[2]Obafemi Awolowo University, Nigeria
[3]University of Nigeria, Nsukka
[4]Machine Learning Collective
[5]Federal University of Agriculture Abeokuta, Nigeria
[6]McGill University, Canada
abdulbasitoyetunji88@gmail.com

Abstract. Accurate 3D brain tumour segmentation requires both fine-grained local feature extraction and effective modelling of long-range spatial dependencies. While convolutional neural networks excel at local feature extraction, they often struggle to capture global context. Conversely, Transformer-based models improve contextual modelling but incur prohibitive memory costs in 3D settings. State space models such as Mamba have emerged as a promising alternative by offering linear-complexity sequence modelling, yet their optimal placement within volumetric medical image segmentation networks remains insufficiently understood. Compounding this issue, many commonly used benchmarks do not adequately reflect the imaging and clinical heterogeneity encountered in real-world practice. We present a systematic encoder-decoder ablation of Mamba-based architectures for brain tumour segmentation on BraTS-Africa 2025, a 146-case multi-centre dataset representing diverse sub-Saharan African imaging environments. Using a unified MetaUNETR framework, we evaluate controlled CNN and Mamba configurations and perform encoder–decoder ablations to determine where Mamba contributes most effectively. We further benchmark the best-performing configurations against nnU-Net, TransUNet, and SwinUNETR. Results from this study indicate that incorporating Mamba within the encoder consistently yields the strongest performance among MetaUNETR variants, whereas Mamba-based decoding was associated with poorer boundary reconstruction, particularly for WT, within the evaluated configuration. These findings demonstrate that token mixer placement is critical, suggesting that Mamba-based encoders may provide a promising architectural alternative for developing brain tumour segmentation models suited for resource-constrained settings.

Keywords: Brain Tumour Segmentation · State Space Models · Mamba · MetaUNETR · BraTS-Africa · Token Mixing.

1 Introduction

A fundamental modelling trade-off constrains 3D brain tumour segmentation: models must simultaneously capture fine-grained local structure and long-range spatial dependencies within high-dimensional volumetric data. Historically, convolutional neural networks (CNNs) like U-Net [1] and nnU-Net [2] have long dominated this space due to their strong inductive biases and data efficiency. However, their inherent reliance on local receptive fields restricts their global context modelling. This limitation becomes critical when analyzing heterogeneous tumour structures spanning large spatial extents.

In recent years, Transformer-based architectures emerged as the primary solution to this global context deficit, via the use of self-attention, achieving strong empirical performance [3, 4]. However, this capability comes at a prohibitive quadratic memory cost in 3D settings. Consequently, these models remain impractical in many clinical environments with constrained hardware budgets. Recent advances in state space models introduce a fundamentally different scaling paradigm. Mamba [5] enables linear-time sequence modelling while preserving long-range dependencies, offering a highly scalable alternative to attention-based methods. However, its effectiveness in 3D medical image segmentation remains largely unexplored.

Beyond architectural efficiency, 3D segmentation faces a second major hurdle: evaluating true clinical robustness. Widely used benchmarks such as BraTS [6, 7] often fail to capture real-world variability in acquisition protocols and patient populations, producing models that degrade under distribution shift. To address this, BraTS-Africa introduces clinically relevant heterogeneity from sub-Saharan African sites, providing a rigorous testbed for robustness evaluation. We argue that next-generation segmentation models must be assessed not only on accuracy but also on their ability to deliver competitive performance under strict computational constraints and dataset shift.

Accordingly, we conduct a systematic study of Mamba-based architectures within a unified MetaUNETR framework [8], isolating the role of the token-mixing mechanism under fully controlled architectural and training conditions. We evaluate segmentation quality on the BraTS-Africa 2025 dataset using Dice scores and the 95% Hausdorff Distance (HD95) [9]. To our knowledge, this work presents the first controlled encoder and decoder ablation of Mamba-based architectures for 3D brain tumor segmentation on sub-Saharan African clinical data, identifying exactly where state space models contribute most effectively to the segmentation pipeline.

2 Related Work

Foundational architectures such as U-Net [1] and its optimized variants like nnU-Net [2] established the standard for 3D medical image segmentation. To address the inherent limitations of localized receptive fields in convolutional models, architectures incorporating self-attention, including UNETR [3], TransUNet [4],

and Swin UNETR [10], were introduced to recover global context. State space models such as Mamba [5] have emerged as an alternative approach for modeling long-range dependencies. Preliminary applications of Mamba in medical imaging demonstrate competitive segmentation performance [21]; however, these studies typically propose end-to-end architectures without isolating the specific locus of improvement. A gap remains in understanding whether Mamba's sequence modeling is most effective when deployed for feature extraction or spatial reconstruction. To address this, we leverage the MetaUNETR framework [8], which permits the direct substitution of token-mixing mechanisms, to systematically ablate Mamba's placement in the encoder versus the decoder.

Beyond architectural design, evaluating model robustness against dataset heterogeneity is critical for real-world deployment. While established benchmarks like the standard BraTS challenges [6, 7] have driven methodological progress, models trained on high-resource data frequently degrade under distribution shifts caused by differing scanners and acquisition protocols. The BraTS-Africa dataset was introduced specifically to capture this clinical heterogeneity using multi-site sub-Saharan African data [22]. Recent works utilizing BraTS-Africa, such as the MD-SA2 evaluation [23], highlight the necessity of testing architectures on diverse, lower-quality MRI inputs typical of resource-constrained settings. Building on these evaluations, our study specifically benchmarks Mamba-based architectural variants on the BraTS-Africa dataset to determine the most effective token mixer placement for heterogeneous clinical data.

3 Method

3.1 Dataset

ImageNet (2D Pretraining). Aligning with the resource-constrained focus of this workshop, the encoder backbone avoids computationally expensive 3D pretraining. Instead, it is pretrained on a 540,000-image subset of ImageNet [15] using a self-supervised denoising autoencoder [16] objective. Gaussian noise ($\sigma = 0.15$) corrupts the input; the model reconstructs the clean image under MSE loss, yielding transferable visual representations without requiring labeled data. Only encoder weights are retained after pretraining; 2D convolutional kernels are inflated to 3D following the frame-repetition inflation strategy [17], repeating along the depth axis and dividing by kernel depth k.

BraTS-Africa 2025 (Fine-tuning). The target dataset comprises 146 multiparametric MRI cases (95 Glioma, 51 Other Neoplasms) acquired across sub-Saharan African clinical sites, providing four co-registered modalities: T1n, T1c, T2w, and T2f/FLAIR. Ground truth follows standard BraTS encoding, yielding three evaluation subregions: Enhancing Tumour (ET = $\{4\}$), Tumour Core (TC = $\{1, 4\}$), and Whole Tumour (WT = $\{1, 2, 4\}$). Preprocessing applies per-modality z-score normalisation within the non-zero brain mask, symmetric crop/pad to 128^3 voxels, and random axis-aligned flips ($p = 0.5$). The dataset was split into 109 training, 22 validation, and 15 test cases (approximately 75/15/10)

with seed = 42. This identical split is used for every model reported in this paper, including all SOTA baselines, to ensure comparisons are not confounded by differing data partitions.

3.2 Architecture and Experimental Stages

Stage 1 – 2D Pretraining. A 2D MetaUNETR-inspired encoder was pretrained on ImageNet using a denoising autoencoder objective. The model processes 224×224 RGB images through a four-stage architecture with depths $(2, 2, 2, 2)$, feature dimension $C = 48$, MLP ratio 4.0, drop path 0.1, and Mamba parameters $d_{\text{state}} = 16$, $d_{\text{conv}} = 4$, and expansion factor of 2. Gaussian noise ($\sigma = 0.15$) was added to inputs, and the network was trained to reconstruct clean images. Training used AdamW [19] (learning rate 10^{-3}, weight decay 0.05), cosine decay to 10^{-6}, one warm-up epoch, batch size of 64, automatic mixed precision, and gradient clipping of 1.0 for three epochs. A 5% validation split was used for monitoring. Following pretraining, only the encoder weights are retained and inflated from 2D to 3D using the frame-repetition inflation strategy [17] along the depth dimension, providing transferable visual representations for downstream brain tumour segmentation on BraTS-Africa.

Stage 2 – MetaUNETR Baseline. We adopt a lightweight MetaUNETR architecture, informed by CKA network similarity [14] with an initial feature dimension of $C = 48$, four hierarchical stages with depths $(2, 2, 2, 2)$, and feature channels $\{48, 96, 192, 384\}$. The MetaUNETR framework supports six token-mixing mechanisms: Conv (depthwise convolution), GF (Global Filter) [13], MLP (Vision Permutator-style) [12], Recur (bidirectional LSTM), Attn (Swin Transformer) [11], and Mamba (selective state-space model) [5].However, for the present encoder–decoder ablation, we focus on the CNN and Mamba configurations to isolate the effect of Mamba placement. All variants share an identical MONAI UnetrUpBlock CNN decoder, isolating the encoder mixer as the only varying factor.

Fine-tuning is performed in two stages. During Phase I, the pretrained encoder is frozen while the decoder is trained for 20 epochs using a learning rate of 3×10^{-4}. During Phase II, the entire network is unfrozen and optimized end-to-end for an additional 30 epochs using discriminative learning rates [18]: 3×10^{-5} for the encoder and 3×10^{-4} for the decoder. Training employs AdamW [19] with a weight decay of 0.05, cosine learning-rate scheduling with a minimum learning rate of 10^{-6}, three warm-up epochs, and gradient clipping at 1.0. Models are trained using automatic mixed precision on a single NVIDIA Tesla P100 GPU with a batch size of one. Validation and testing follow fixed splits of 15% and 10%, respectively, while sliding-window inference is performed with a patch size of $96 \times 96 \times 96$, an overlap of 0.5, and a batch size of two.

Stage 3 – Mamba Architectural Ablations. Two conditions isolate where Mamba's contribution is concentrated. Mod A (Mamba Encoder + CNN Decoder) replaces all encoder token mixers with Mamba cross-scan SSM blocks while the

decoder is held fixed.Note that the Stage 2 MetaUNETR Baseline restricts the Mamba token mixer exclusively to the bottleneck, whereas Mod A integrates it across all hierarchical encoder stages. Mod B (CNN Encoder + Mamba Decoder) retains the convolutional encoder and instead replaces the CNN residual decoder blocks with Mamba3DCore cross-scan blocks. Both configurations are initialised from the same ImageNet checkpoint and trained under the identical protocol described in Stage 2, isolating decoder versus encoder placement as the sole architectural variable between them.

Stage 4 – Comparison with State-of-the-Art Models. The MetaUNETR Baseline, Mod A (Mamba Encoder + CNN Decoder), and Mod B (CNN Encoder + Mamba Decoder) were evaluated against three widely used segmentation architectures: nnU-Net [2], TransUNet [4], and Swin UNETR [10, 11] using the same 75/15/10 train–validation–test split to ensure a fair comparison. nnU-Net was trained using its self-configuring full-resolution 3D pipeline and fine-tuned in two stages, with the encoder frozen for 20 epochs (Phase I) before end-to-end optimization for an additional 50 epochs (Phase II). TransUNet, initialized with ImageNet-21k pretrained ViT-B/16 weights and adapted for four-channel MRI input, was fine-tuned end-to-end for 50 epochs. Swin UNETR was trained for 300 epochs following its recommended configuration. For the two-stage protocols, learning rates of 10^{-3} and 10^{-4} were used during Phases I and II, respectively, with a weight decay of 10^{-5}. This experimental design enables a direct assessment of Mamba-based architectures against established CNN- and Transformer-based segmentation models under identical data conditions.

4 Results

4.1 SOTA Comparison

Table 1 summarizes segmentation performance across all benchmarked architectures on the BraTS-Africa 2025 test split.

Table 1. Segmentation accuracy (Dice) and boundary error (HD95, mm) across all benchmarked architectures on the BraTS-Africa 2025 test split. Values are reported as mean ± standard deviation across the 15 held-out test cases.

Model	ET Dice ↑	TC Dice ↑	WT Dice ↑	ET HD95 ↓	TC HD95 ↓	WT HD95 ↓
MetaUNETR (Mamba)	0.717 ± 0.129	0.641 ± 0.186	0.805 ± 0.103	41.56 ± 28.12	44.22 ± 28.01	40.50 ± 22.47
Mod A (Mamba Enc + CNN Dec)	0.822 ± 0.071	0.756 ± 0.203	0.873 ± 0.106	16.49 ± 16.89	20.84 ± 19.32	$\mathbf{12.39 \pm 13.85}$
Mod B (CNN Enc + Mamba Dec)	0.766 ± 0.094	0.677 ± 0.202	0.826 ± 0.097	15.43 ± 15.87	24.15 ± 16.68	29.42 ± 21.08
TransUNet	0.868 ± 0.094	0.869 ± 0.029	0.814 ± 0.009	26.77 ± 32.43	25.00 ± 10.71	33.86 ± 1.26
nnU-Net	$\mathbf{0.898 \pm 0.051}$	$\mathbf{0.900 \pm 0.067}$	$\mathbf{0.908 \pm 0.061}$	$\mathbf{1.96 \pm 24.43}$	$\mathbf{14.03 \pm 24.08}$	20.99 ± 30.11
Swin UNETR	0.363 ± 0.194	0.335 ± 0.185	0.478 ± 0.154	129.87 ± 22.77	133.84 ± 17.35	131.89 ± 12.34

nnU-Net achieves the strongest Dice across all three subregions (WT 0.908, TC 0.900, ET 0.898), consistent with its established role as a highly optimized CNN baseline at full 3D resolution. TransUNet achieves the second-highest ET

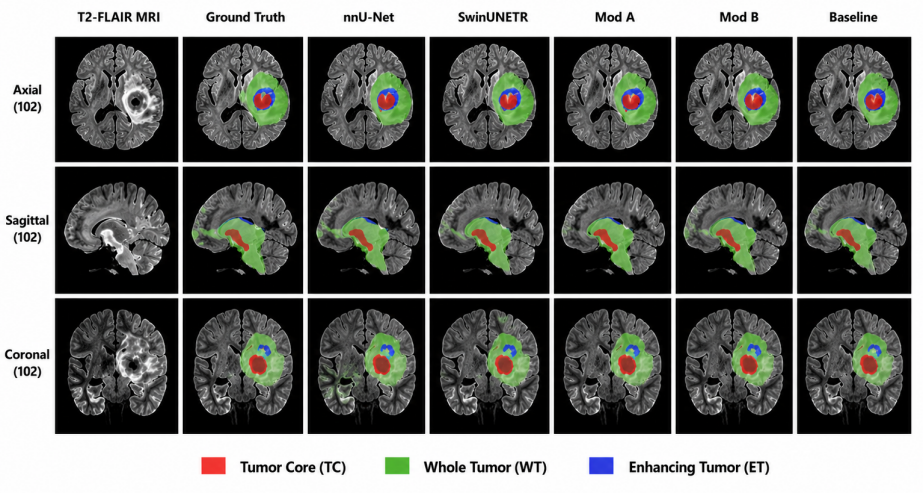

Fig. 1. Comparison of tumor sub-region segmentation on BraTS-Africa test cases across baseline architectures and proposed MetaUNETR variants.

and TC Dice (0.868, 0.869, respectively). Swin UNETR achieved the lowest mean Dice scores across all subregions, including a WT Dice of 0.478.

4.2 Mamba Encoder/Decoder Ablation

Among the MetaUNETR variants, Mod A (Mamba Encoder + CNN Decoder) outperforms both the pure Mamba MetaUNETR baseline and Mod B (CNN Encoder + Mamba Decoder) across nearly all evaluated metrics. Relative to the MetaUNETR baseline, Mod A improves WT Dice by 0.068 (0.805 $\rightarrow$ 0.873) while reducing WT HD95 by more than two-thirds (40.50 mm $\rightarrow$ 12.39 mm). Mod B improves Dice over the baseline as well, but its WT HD95 (29.42 mm) remains substantially worse than Mod A's, despite Mod B's ET HD95 (15.43 mm) being marginally the best of any MetaUNETR variant.

This asymmetry indicates that Mamba's contribution to segmentation quality within this framework is concentrated in the encoder. Conversely, replacing CNN reconstruction blocks with Mamba cross-scan blocks in the decoder degrades spatial reconstruction for the largest, most boundary-irregular subregion (WT).

4.3 Comparative Performance Analysis

Mod A achieves a WT Dice within 0.035 of nnU-Net (0.873 vs. 0.908) on the held-out test set. While TransUNet demonstrated competitive Dice scores on the ET and TC subregions, its training dynamics showed a noticeable discrepancy between training and validation loss. This may indicate a degree of overfitting and suggests that its performance could be sensitive to the limited training cohort.

Further evaluation across multiple splits would be required to establish the generalizability of this finding. Taken together, these results indicate that, within the MetaUNETR framework, Mamba-encoder hybrids (Mod A) offer a favorable performance balance, achieving accuracy that approaches the nnU-Net baseline while avoiding the potential overfitting behaviour observed in the attention-heavy TransUNet baseline.

5 Discussion and Conclusion

Within the evaluated MetaUNETR framework, the Mamba-CNN hybrid (Mod A: Mamba Encoder + CNN Decoder) achieved the strongest results among all tested variants, yielding a WT Dice of 0.873 and an HD95 of 12.39 mm. This performance approaches nnU-Net's full 3D-resolution accuracy (WT Dice 0.908). The encoder-decoder ablation results indicate that Mod A consistently outperforms Mod B (CNN Encoder + Mamba Decoder) across every subregion. This asymmetry suggests that Mamba's sequence modeling is more effective when utilized in the encoder, whereas CNN-based decoding retains a crucial inductive-bias advantage for reconstructing irregular tumor boundaries.

Comparisons against established baselines further contextualize these findings on the 146-case BraTS-Africa dataset. While TransUNet achieved competitive Dice scores on the ET and TC subregions, its training dynamics showed a noticeable discrepancy between training and validation loss. This may indicate a degree of overfitting and suggests that its performance could be sensitive to the limited training cohort; however, further evaluation across multiple splits would be required to establish the generalizability of this finding. Furthermore, Swin UNETR showed substantially lower performance, with a WT Dice of 0.478, despite being trained using the specified training configuration. This result highlights the difficulty of achieving robust performance with attention-heavy architectures on relatively small and heterogeneous medical imaging datasets. Limited sample size may constrain the effective representation learning of such models, particularly under substantial acquisition variability [20].

Several limitations temper our conclusions. First, this study relies on a single train-validation-test split (75/15/10) of 146 cases; incorporating multi-fold cross-validation in future work would provide stronger evidence of robustness against dataset variations. Second, the lack of statistical significance testing limits definitive claims regarding the exact margins of improvement between architectures. Moreover, a further direction is to evaluate stage-wise Mamba placement within the encoder, including bottleneck-only and deeper-stage configurations, to determine whether the observed encoder preference is uniform across hierarchical feature levels. Finally, comprehensive computational profiling (e.g., FLOPs, latency, and peak VRAM) was not conducted, meaning the deployment efficiency of these specific models remains to be formally quantified.

Despite these limitations, our findings provide concrete evidence that token mixer placement is critical. Within the evaluated MetaUNETR framework

and experimental setting, the results favour encoder integration over decoder integration for Mamba-based 3D brain tumour segmentation.

6 Impact in Resource-Constrained Settings

The relevant resource constraint for clinical and research adoption in low- to middle-income countries (LMICs) is typically not the cloud infrastructure used during initial model development but the local hardware required to fine-tune and run inference on the trained model afterward. While attention-based models struggle with quadratic memory scaling in 3D, and purely convolutional networks rely on localized receptive fields, Mamba's linear-complexity sequence modeling offers a theoretically scalable alternative for global context aggregation.

By demonstrating that Mamba's modeling capabilities are most effectively utilized within the encoder rather than the decoder, this study provides a crucial architectural blueprint. This ablation guides the design of future hybrid models that maximize segmentation accuracy without defaulting to computationally prohibitive attention mechanisms. Such optimized hybrid designs may facilitate local fine-tuning and deployment in resource-constrained settings, although their practical computational advantages require formal profiling in future work. Ultimately, we position the central contribution of this work not as a finalized software product, but as empirical guidance on token-mixer placement, accelerating the development of highly accurate and practically deployable medical imaging solutions for resource-constrained healthcare systems.

Acknowledgement. We thank the ML Collective community for providing generous computational resources through the Modal Academic Compute Grant as well as for their insightful weekly feedback, which helped shape the direction of this work. We also acknowledge the computational infrastructure support provided by the Digital Research Alliance of Canada through the Consortium for Advancement of MRI Education and Research in Africa (CAMERA). In addition, this work benefited from compute resource grants provided by the Nigeria AI Collectives.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: *MICCAI. LNCS*, vol. 9351, pp. 234–241. Springer (2015)
2. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* 18(2), 203–211 (2021)
3. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: UNETR: Transformers for 3D medical image segmentation. In: *WACV*. pp. 574–584 (2022)

4. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
5. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
6. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (BraTS). *IEEE TMI* 34(10), 1993–2024 (2015)
7. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data* 4, 170117 (2017)
8. Lyu, P., Zhang, J., Zhang, L., Liu, W., Wang, C., Zhu, J.: MetaUNETR: Rethinking token mixer encoding for efficient multi-organ segmentation. In: *MICCAI. LNCS*, vol. 15009, pp. 443–453. Springer (2024)
9. Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J.: Comparing images using the Hausdorff distance. *IEEE TPAMI* 15(9), 850–863 (1993)
10. 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)
11. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: *ICCV*. pp. 10012–10022 (2021)
12. Hou, Q., Jiang, Z., Yuan, L., Cheng, M.M., Yan, S., Feng, J.: Vision permutator: A permutable MLP-like architecture for visual recognition. *IEEE TPAMI* 45(1), 1328–1334 (2022)
13. Rao, Y., Zhao, W., Zhu, Z., Lu, J., Zhou, J.: Global filter networks for image classification. In: *NeurIPS*. vol. 34, pp. 980–993 (2021)
14. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited. In: *ICML*. pp. 3519–3529 (2019)
15. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. *IJCV* 115(3), 211–252 (2015)
16. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: *ICML*. pp. 1096–1103 (2008)
17. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the Kinetics dataset. In: *CVPR*. pp. 6299–6308 (2017)
18. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: *ACL*. pp. 328–339 (2018)
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: *ICLR* (2019)
20. Zhu, H., Chen, B., Yang, C.: Understanding why vit trains badly on small datasets: An intuitive perspective. arXiv preprint arXiv:2302.03751 (2023)
21. Ma, J., Li, F., Wang, B.: U-Mamba: Enhancing long-range dependency for biomedical image segmentation. *IEEE TMI* (2024)
22. Adewole, M., et al.: The BraTS-Africa challenge: A multi-national sub-Saharan African brain tumor dataset. arXiv preprint arXiv:2305.19369 (2023)
23. Li, B., et al.: MD-SA2: Optimizing Segment Anything 2 for multimodal, depth-aware brain tumor segmentation in sub-Saharan populations. *Journal of Medical Imaging* 12(2), 024007 (2025)