NeuroMamba: A State-Space Foundation Model for Functional MRI

Jubin Choi
Interdisciplinary Program in Artificial Intelligence
Seoul National University
wnqlszoq123@snu.ac.kr

David Keetae Park
Artificial Intelligence Department
Brookhaven National Laboratory
dpark1@bnl.gov

Junbeom Kwon
Department of Psychology
The University of Texas at Austin
kjb961013@gmail.com

Shinjae Yoo
Artificial Intelligence Department
Brookhaven National Laboratory
sjyoo@bnl.gov

Jiook Cha *
Department of Psychology,
Interdisciplinary Program in Artificial Intelligence,
Department of Brain and Cognitive Sciences,
Graduate School of Data Science
Seoul National University
connectome@snu.ac.kr

Abstract

Foundation models for whole-brain fMRI analysis are dominated by two competing paradigms: Region-of-Interest (ROI) approaches that discard fine-grained spatial information, and hierarchical models with Transformer or Mamba backbones that retain it. However, the rigid, grid-based architecture of hierarchical models imposes critical limitations. They are forced to process vast, uninformative non-brain background regions (up to 60% of the data), creating significant computational inefficiency, and struggle with inter-subject anatomical variation. To overcome these challenges, we introduce **NeuroMamba**, a foundation model that enables *direct sequence modeling* of 4D whole-brain fMRI. NeuroMamba leverages Mamba backbone trained with an autoregressive objective given small spatiotemporal patches. Our approach facilitates the adaptive removal of nuisance background signals—a core inefficiency bottleneck for prior grid-based methods. To address anatomical variability, the model incorporates explicit frequency-based positional encodings, enhancing robustness across subjects. While direct sequence modeling is computationally demanding, the linear-time efficiency of Mamba, coupled with adaptive background suppression, makes whole-brain pre-training tractable. We demonstrate the first successful autoregressive pre-training on full-resolution fMRI, achieving state-of-the-art accuracy for sex classification and establishing a scalable paradigm for neuroimaging analysis.

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body.

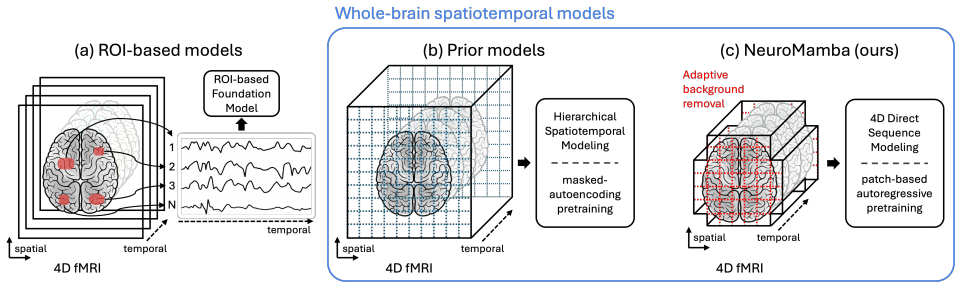

Figure 1: **Schematic comparison against prior approaches**. While (a) region-of-interest (ROI) based foundation models incorporate spatially averaged out data and (b) hierarchical approaches employ the whole brain, including the empty background, (c) our model employs patch-based autoregressive training, which enables adaptive removal of uninformative background using patch-based Mamba, supporting efficiency and training stability.

1 Introduction

Functional MRI (fMRI) measures spatiotemporal neural activities, providing unprecedented access to whole-brain dynamics underlying human cognition and behavior [1, 2]. This non-invasive imaging modality has revolutionized our ability to decode complex mental processes, from reconstructing visual imagery during perception and imagination [3, 4] to identifying neural signatures of psychiatric disorders and cognitive decline [5]. However, despite these remarkable capabilities, the complexity and heterogeneity of fMRI data present fundamental challenges. Individual studies typically capture limited behavioral contexts with small sample sizes, creating a fragmented landscape where models trained on specific tasks or populations fail to generalize across the diverse spectrum of human brain function [6]. Foundation models offer a transformative solution by leveraging vast, unlabeled fMRI data to learn universal representations of brain activity patterns, thereby mitigating label scarcity, improving cross-dataset generalization, and enhancing downstream performance in cognitive decoding and clinical prediction tasks [3, 7].

Prior fMRI foundation models largely follow two design families. Region Of Interest (ROI)-based methods first reduce sequences of volumes to multivariate time series or connectivity matrix and then pre-train deep neural networks on these summarized features. Notable examples include BrainLM[7], Brain-JEPA [8], and fMRI-PTE [3]. These approaches scale efficiently but may discard voxel-level details and inherit atlas biases. In contrast, hierarchical models process 4-dimensional spatiotemporal volumes in an end-to-end manner via multiscale windowing or shifted windows. For example, SwiFT [9] applies a shifted-window attention over spatial and temporal dimensions and contrastive losses to capture individual differences in complex brain dynamics. Similarly, NeuroSTORM [10] introduces a shifted-window Mamba backbone pre-trained with masked image modeling. While these hierarchical designs preserve rich spatial structures, they are constrained by a rigid, grid-based architecture necessary for their windowing operations. This inflexibility prevents the adaptive removal of uninformative non-brain backgrounds, forcing these nuisance signals to be processed alongside genuine brain activity. Consequently, such models often require heavy downsampling, which compromises both computational efficiency and training stability.

We propose NeuroMamba, a model for the direct sequence modeling of whole-brain fMRI that resolves the trade-offs between prior methods (Figure 1c). NeuroMamba employs a Mamba state-space backbone [11] with two key innovations: 1) adaptive background removal to overcome the inefficiencies of hierarchical models, and 2) explicit frequency-based positional encodings to handle inter-subject anatomical variation [12]. This combination preserves the voxel-level detail lost by ROI methods while delivering superior efficiency and stability compared to rigid hierarchical architectures. Our contributions are threefold:

- **A Novel Autoregressive Pre-training Framework for fMRI**. We demonstrate the first successful application of a Mamba-based, autoregressive objective to model whole-brain fMRI sequences.

- **Efficient and Robust Modeling Techniques**. We introduce adaptive background suppression via patch-wise tokenization and improve inter-subject generalization with frequency-based positional encodings.
- **Strong Empirical Results**. We demonstrate stable pre-training and achieve state-of-the-art performance on sex classification, validating direct sequence modeling as a viable and powerful new direction for fMRI analysis.

2 Related Works

The primary goal of fMRI foundation models is to learn generalizable representations of brain activity transferable across tasks such as cognitive decoding, clinical diagnosis, personal trait prediction (e.g., sex or age), and cross-modal brain-to-image retrieval. Their development has followed two main trajectories: ROI-based approaches and whole-brain modeling approaches.

ROI-based fMRI Foundation Models. Early work focused on region-of-interest (ROI) based approaches. BrainLM [7] and Brain-JEPA [8] employed Vision Transformer (ViT) architectures with spatiotemporal masking strategies for self-supervised pre-training on ROI-level neural activity. These models demonstrated the feasibility of large-scale pre-training on neuroimaging data and strong transfer learning performance. However, their coarse spatial resolution ($\sim$400 ROIs) and reliance on predefined parcellation schemes limit their ability to capture fine-grained activity.

Whole-brain fMRI Foundation Models. Voxel-level modeling emerged to overcome these spatial limitations. TFF [13] used reconstruction-based self-supervised training with CNN-Transformer hybrids but suffered from instability and challenges coordinating two-stage learning. SwiFT [9] improved this with a hierarchical Swin Transformer for 4D fMRI, introducing spatiotemporal window attention for end-to-end voxel-level training. Building on this, NeuroSTORM [10] employed a Mamba-based architecture with shifted windows, optimized pre-training, and prompt tuning. Yet, their rigid grid-based designs remain inefficient, as they must also process noisy non-brain background.

Mamba for fMRI Foundation Models. Mamba has been explored for fMRI, though mainly with ROI-level or handcrafted features rather than end-to-end voxel modeling. Causal fMRI-Mamba [14] embeds a Mamba backbone in a causal learning framework for supervised task-fMRI decoding. FST-Mamba [15] applies hierarchical Mamba to model Dynamic Functional Network Connectivity (dFNC) matrices, a lower-dimensional proxy of brain dynamics. While validating Mamba's potential, these studies do not address self-supervised pre-training on raw voxel data. To our knowledge, NeuroMamba is the first to apply a direct-sequence Mamba architecture end-to-end on whole-brain voxels, treating fMRI as a large-scale spatiotemporal sequence modeling problem.

3 Methods

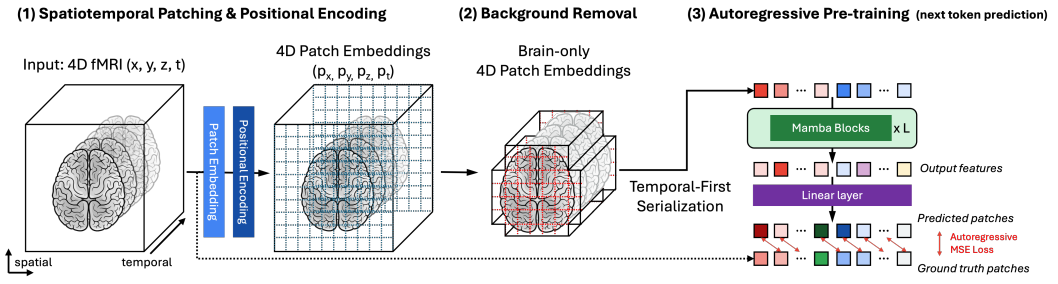

Figure 2: **Overview of NeuroMamba pre-training pipeline.** A 4D fMRI volume is converted into a sequence of patch embeddings with positional information. Non-brain background tokens are removed, and the model is trained with the autoregressive next-token prediction.

Our methodology employs a two-stage process: first, large-scale self-supervised pre-training, followed by downstream fine-tuning. This framework is built around our NeuroMamba architecture, designed to efficiently model whole-brain fMRI as a single, unified sequence.

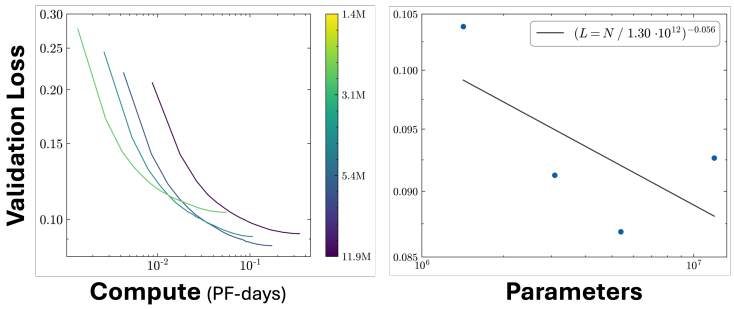

Figure 3: **Neural scaling laws of NeuroMamba.** (Left) Pre-training validation loss decreases with increased computational budget. (Right) Validation loss shows a power-law relationship with the number of model parameters up to 5.4M parameters, confirming the scalability of our approach.

NeuroMamba Architecture. As illustrated in Figure 2, an input fMRI scan (size of $96 \times 96 \times 96 \times 20$) is first divided into non-overlapping 4D patches of size $6 \times 6 \times 6 \times 2$. To preserve the origin of each patch, we add a Neural Radiance Fields (NeRF) [12] style positional encoding as suggested in [16]. A key aspect of our approach is the ability to handle variable-length inputs by identifying and discarding all tokens corresponding to non-brain background. The resulting sequence of brain-only tokens is then processed with Mamba2 [17] blocks (see Appendix B for further details and hyperparameters).

Autoregressive Pre-training. We pre-train NeuroMamba on a large amount of resting-state fMRI data curated from the UK Biobank [18], ABCD [19], and HCP datasets [20], totaling over 50,000 subjects (Appendix A). The objective is next-token prediction, where the model learns to predict the subsequent patch data. We use a temporal-first raster scan order $(t \rightarrow z \rightarrow y \rightarrow x)$.

Downstream Fine-tuning. After pre-training, the generative head is replaced with a task-specific head for evaluation on the HCP dataset. We evaluate two head architectures: a simple Linear head and a more complex Mamba head. To test the effect of background removal, we train only the heads with a frozen backbone, while we perform full end-to-end fine-tuning for our main performance comparisons. To ensure robust results, each evaluation is repeated three times with different random seeds. See Appendix C for full details.

4 Results

4.1 Computational Efficiency Gain from Non-brain Background Removal

A key contribution of our direct-sequence approach is its ability to enhance computational efficiency by removing non-brain tokens. We used the DeepSpeed profiler [21] to measure the total Floating-Point Operations (FLOPs) per epoch for a 5.4M-parameter model. The standard model, processing all tokens, required 6.96×10^{17} FLOPs per epoch. In contrast, our model trained only on brain tokens required just 3.72×10^{17} FLOPs. This represents a **46.5% reduction** in computational cost, significantly improving the efficiency of both large-scale pre-training and downstream fine-tuning.

4.2 Pre-training and Scaling Laws

We pre-trained NeuroMamba models with varying sizes, observing a consistent trend where validation loss decreases as both compute and model parameters increase (Figure 3). The trend follows a power-law relationship [22], indicating that NeuroMamba is a scalable architecture that benefits from larger model sizes and more extensive training. Notably, our largest 11.9M parameter model deviates from this trajectory, which we attribute to training instabilities that emerged at this scale. We discuss this limitation and our plans for future work in Section 5.

Table 1: **Evaluation of background removal on HCP sex classification.** Performance improves significantly when uninformative background tokens are removed before training.

Method		HCP Sex	
		AUC	ACC
Linear head	All tokens	0.7432±0.00	0.6882±0.01
	Only brain tokens	**0.7769±0.00**	**0.7156±0.01**
Mamba head	All tokens	0.8456±0.05	0.7767±0.07
	Only brain tokens	**0.8628±0.08**	**0.8104±0.09**

Table 2: **HCP sex classification performance of various pre-trained models.** Fine-tuning outperforms training from scratch. The 3.1M model with a Mamba head achieves the highest accuracy.

Model Size	Method		HCP Sex	
			AUC	ACC
1.4M	Linear head	From scratch	0.9767±0.01	0.9427±0.01
		Full fine-tuning	0.9840±0.00	0.9458±0.01
	Mamba head	From scratch	0.9446±0.01	0.8872±0.01
		Full fine-tuning	0.9535±0.01	0.8781±0.01
3.1M	Linear head	From scratch	0.9813±0.01	0.9232±0.02
		Full fine-tuning	**0.9885±0.01**	0.9455±0.02
	Mamba head	From scratch	0.9825±0.01	0.9396±0.00
		Full fine-tuning	0.9874±0.01	**0.9486±0.02**
5.4M	Linear head	From scratch	0.9865±0.00	0.9347±0.01
		Full fine-tuning	0.9717±0.00	0.9198±0.00
	Mamba head	From scratch	0.9729±0.02	0.9175±0.03
		Fine-tuning	0.9766±0.00	0.9307±0.01

Table 3: **Comparison with SOTA models on HCP sex classification.** NeuroMamba achieves state-of-the-art accuracy, despite differences in data splits.

Model (# Params)	Data Split (Train / Validation / Test)	AUC	ACC (%)
SwiFT [9] (4.6M)	70 / 15 / 15	98.0	92.9
NeuroSTORM [10] (5.0M)	70 / 15 / 15	97.6	93.3
NeuroMamba (Ours, 3.1M)	80 / 10 / 10	**98.9**	**94.9**

4.3 Downstream Performances

Effect of Non-brain Background Removal. We first assessed our background removal strategy by fine-tuning only task heads on a pre-trained backbone. Table 1 shows a significant performance improvement when training only on brain tokens. With the Mamba head, accuracy increased from 77.67% to **81.04%**, demonstrating that focusing the model's capacity on relevant signals is crucial for learning effective representations.

Impact of Pre-training and Model Scale. Next, we compared the performance of fine-tuned pre-trained models against models trained from scratch across different sizes (Table 2). Pre-training consistently provides a substantial boost in performance. For the 3.1M parameter model with a Mamba head, fine-tuning achieves an accuracy of **94.86%**, a significant improvement over the 93.96% accuracy from scratch. This result also represents the state-of-the-art for this task. Interestingly, while larger models perform better, the benefits of pre-training are evident across all model scales, confirming the value of the learned representations. The 3.1M model appears to hit a sweet spot, achieving the highest performance on this task.

Comparison against State-of-the-Art. We compare NeuroMamba's performance against leading hierarchical models, SwiFT [9] and NeuroSTORM [10]. As shown in Table 3, NeuroMamba demonstrates state-of-the-art performance, outperforming the reported accuracy of previous models.

5 Discussion & Conclusion

Our experiments validate that autoregressive pre-training on fMRI is a feasible and highly effective strategy. However, as this research is a work-in-progress, there are clear avenues for improvement. A key challenge we encountered was training instability with our largest 11.9M parameter model, causing it to deviate from the scaling laws observed with smaller models. This is a common obstacle in scaling foundation models, and we are actively exploring several strategies to mitigate it, such as refining the learning rate schedule and implementing more advanced optimizers. We frame these efforts as immediate future work and are confident that with additional compute and tuning, our architecture will demonstrate consistent scaling.

The significant performance boost from fine-tuning confirms that NeuroMamba learns meaningful and transferable representations, and our results on the HCP sex classification benchmark are highly encouraging. Nonetheless, we must acknowledge a limitation in our comparison to prior work in Table 3. A direct comparison is challenging because our 80/10/10 data split differs from the 70/15/15 split used by other models, meaning the subject memberships in the test sets are not identical. While we argue that the large sample size of the dataset (N=1,084) likely mitigates major sampling bias, we plan to conduct a more rigorous benchmark by re-implementing baselines on our exact data splits in future work to ensure a perfectly fair comparison.

To more rigorously test NeuroMamba's capabilities as a foundation model, we also plan to extend our evaluation to a wider range of downstream tasks, such as predicting cognitive scores and clinical diagnoses. We will also assess its generalization by fine-tuning and evaluating the pre-trained model on external, unseen datasets to ensure its robustness and applicability across different data acquisition settings.

Acknowledgments and Disclosure of Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1C1C1006503, RS-2023-00266787, RS-2023-00265406, RS-2024-00421268, RS-2024-00342301, RS-2024-00435727, NRF-2021M3E5D2A01022515, and NRF-2021S1A3A2A02090597), by the Creative-Pioneering Researchers Program through Seoul National University (No. 200-20240057, 200-20240135). Additional support was provided by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [No. RS-2021-II211343, Artificial Intelligence Graduate School Program, Seoul National University] and by the Global Research Support Program in the Digital Field (RS-2024-00421268). This work was also supported by the Artificial Intelligence Industrial Convergence Cluster Development Project funded by the Ministry of Science and ICT and Gwangju Metropolitan City, by the Korea Brain Research Institute (KBRI) basic research program (25-BR-05-01), by the Korea Health Industry Development Institute (KHIDI) and the Ministry of Health and Welfare, Republic of Korea (HR22C1605), and by the Korea Basic Science Institute (National Research Facilities and Equipment Center) grant funded by the Ministry of Education (RS-2024-00435727). We acknowledge the National Supercomputing Center for providing supercomputing resources and technical support (KSC-2023-CRE-0568). An award for computer time was provided by the U.S. Department of Energy's (DOE) ASCR Leadership Computing Challenge (ALCC). This research used resources of the National Energy Research Scientific Computing Center (NERSC), a DOE Office of Science User Facility, under ALCC award m4750-2024, and supporting resources at the Argonne and Oak Ridge Leadership Computing Facilities, U.S. DOE Office of Science user facilities at Argonne National Laboratory and Oak Ridge National Laboratory. The BNL team was supported by the U.S. Department of Energy (DOE), Office of Science (SC), Advanced Scientific Computing Research program under award DE-SC-0012704 and used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility using NERSC award DDR-ERCAP0033558.

References

[1] Seiji Ogawa, David W. Tank, Ravi Menon, Jutta M. Ellermann, Seong-Gi Kim, Hellmut Merkle, and Kamil Ugurbil. Intrinsic signal changes accompanying sensory stimulation: Functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, 89(13):5951–5955, 1992.

[2] José M. Soares, Ricardo Magalhães, Pedro S. Moreira, Alexandre Sousa, Edward Ganz, Adriana Sampaio, Victor Alves, Paulo Marques, and Nuno Sousa. A hitchhiker's guide to functional magnetic resonance imaging. *Frontiers in Neuroscience*, 10:515, 2016.

[3] Xuelin Qian, Yun Wang, Jingyang Huo, Jianfeng Feng, and Yanwei Fu. fMRI-PTE: A large-scale fmri pretrained transformer encoder for multi-subject brain activity decoding. *arXiv preprint arXiv:2311.00342*, 2023.

[4] Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36:24705–24728, 2023.

[5] Stefan Frässle, Andre F Marquand, Lianne Schmaal, Richard Dinga, Dick J Veltman, Nic JA Van der Wee, Marie-José van Tol, Dario Schöbi, Brenda WJH Penninx, and Klaas E Stephan. Predicting individual clinical trajectories of depression with generative embedding. *NeuroImage: Clinical*, 26:102213, 2020.

[6] Xinliang Zhou, Chenyu Liu, Zhisheng Chen, Kun Wang, Yi Ding, Ziyu Jia, and Qingsong Wen. Brain foundation models: A survey on advancements in neural signal processing and brain discovery. *arXiv preprint arXiv:2503.00580*, 2025.

[7] Josue Ortega Caro, Antonio H. de O. Fonseca, Syed A. Rizvi, Matteo Rosati, and David van Dijk. Brainlm: A foundation model for brain activity recordings. In *International Conference on Learning Representations (ICLR)*, 2024.

[8] Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Su Xian Chong, Fang Ji, Nathanael Ren Jie Tong, Christopher Li Hsian Chen, and Juan Helen Zhou. Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[9] Peter Yongho Kim, Junbeom Kwon, Sunghwan Joo, Sangyoon Bae, Donggyu Lee, Yoonho Jung, Shinjae Yoo, Jiook Cha, and Taesup Moon. Swift: Swin 4d fmri transformer. *arXiv preprint arXiv:2307.05916*, 2023.

[10] Cheng Wang, Yu Jiang, Zhihao Peng, Chenxin Li, Changbae Bang, Lin Zhao, Jinglei Lv, Jorge Sepulcre, Carl Yang, Lifang He, Tianming Liu, Daniel Barron, Quanzheng Li, Randy Hirschtick, Byung-Hoon Kim, Xiang Li, and Yixuan Yuan. Towards a general-purpose foundation model for fmri analysis. *arXiv preprint arXiv:2506.11167*, 2025.

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

[12] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[13] Itzik Malkiel, Gony Rosenman, Lior Wolf, and Talma Hendler. Self-supervised transformers for fmri representation. In *International Conference on Medical Imaging with Deep Learning*, pages 895–913. PMLR, 2022.

[14] Weihao Deng, Fei Han, Qinghua Ling, Qing Liu, and Henry Han. Causal fmri-mamba: Causal state space model for neural decoding and brain task states recognition. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[15] Yuxiang Wei, Anees Abrol, and Vince D Calhoun. Hierarchical spatio-temporal state-space modeling for fmri analysis. In *International Conference on Research in Computational Molecular Biology*, pages 86–98. Springer, 2025.

[16] David Park, Shuhang Li, Yi Huang, Xihaier Luo, Haiwang Yu, Yeonju Go, Christopher Pinkenburg, Yuewei Lin, Shinjae Yoo, Joseph Osborn, et al. Fm4npp: A scaling foundation model for nuclear and particle physics. *arXiv preprint arXiv:2508.14087*, 2025.

[17] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

[18] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[19] Betty Jo Casey, Tariq Cannonier, May I Conley, Alexandra O Cohen, Deanna M Barch, Mary M Heitzeg, Mary E Soules, Theresa Teslovich, Danielle V Dellarco, Hugh Garavan, et al. The adolescent brain cognitive development (abcd) study: imaging acquisition across 21 sites. *Developmental cognitive neuroscience*, 32:43–54, 2018.

[20] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

[21] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506, 2020.

[22] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

A Dataset and Experimental Details

Pre-training Datasets. Our pre-training dataset was curated from three sources of resting-state fMRI: the UK Biobank [18] (N=40,647), the Adolescent Brain Cognitive Development (ABCD) study [19] (N=9,139), and the Human Connectome Project (HCP) Young Adult study [20] (N=1,084). All data were preprocessed using standard pipelines, including bias field correction, skull stripping, alignment to the structural image, and spatial normalization to MNI space. The preprocessed fMRI data have a spatial shape of $91 \times 109 \times 91$. We cropped and padded this data to have a spatial shape of $96 \times 96 \times 96$.

Data Splitting. We used a fixed 80% (training), 10% (validation), and 10% (testing) split across all subjects from all datasets. To ensure a fair and rigorous evaluation of generalization, the subjects comprising the HCP test set used for all downstream task evaluations were completely held out and were not included in any part of the pre-training data (neither the training nor validation sets).

B Model Architecture and Hyperparameters

Mamba2 Backbone. The NeuroMamba backbone consists of a stack of 12 Mamba2 blocks. We experimented with four model sizes by varying the embedding dimension ('embed_dim'), as detailed in Table 4.

Table 4: NeuroMamba Model Specifications.

Model Version	Parameters	Embedding Dim ('embed_dim')
NeuroMamba-Small	1.4M	128
NeuroMamba-Base	3.1M	192
NeuroMamba-Medium	5.4M	256
NeuroMamba-Large	11.9M	384

Positional Encoding. We use Neural Radiance Fields (NeRF) [12] positional encoding to encode the continuous (x, y, z, t) coordinate of each patch's centroid. This allows the model to learn high-frequency functions of the input coordinates, making it robust to inter-subject anatomical variability.

Training Environment. All models were trained on the ALCF Aurora supercomputer. We utilized a full precision training. To manage memory and scale efficiently, we employed the `DeepSpeed` library [21] with the FusedAdam optimizer and the ZeRO stage-1 or stage-3 strategy, as detailed in Table 5.

Hyperparameters. For pre-training, we used an effective batch size of 1,152, training for 40 epochs. For fine-tuning, the batch size was adjusted based on model size, and we trained for 30 epochs. The learning rate for all experiments was set to 1×10^{-4} with a weight decay of 1×10^{-2}. We used a `CosineAnnealingWarmUpRestarts` learning rate scheduler with a 5% warmup period, annealing down to a minimum learning rate of 1% of the peak value. Specific hyperparameters for each model size are summarized in Table 5.

C Downstream Fine-tuning Protocol

Downstream Head Architectures. We evaluated two types of heads for downstream tasks:

- **Linear Head:** An 'nn.AdaptiveAvgPool1d(1)' layer performs mean pooling across the output token sequence, followed by a single 'nn.Linear' layer for final prediction.
- **Mamba Head:** A stack of 3 Mamba2 blocks (with 'd_state' = 'embed_dim'/16) processes the output token sequence from the backbone. The hidden state of the final token in the sequence is then passed to a 'nn.Linear' layer for prediction.

Table 5: Pre-training and Fine-tuning Hyperparameters for NeuroMamba Models.

Hyperparameter	**-Small**	**-Base**	**-Medium**	**-Large**
Model Size	1.4M	3.1M	5.4M	11.9M
Pre-training Configuration				
GPU Nodes (12 ranks/node)	32	32	32	48
DeepSpeed ZeRO Stage	1	1	1	3
Micro Batch Size / GPU	3	3	3	2
Fine-tuning Configuration				
GPU Nodes (12 ranks/node)	2			
DeepSpeed ZeRO Stage	1			
Micro Batch Size / GPU	2	2	2	-

Evaluation Repetitions. To account for stochasticity in training, each downstream fine-tuning experiment was repeated three times with different random seeds. The mean and standard deviation of the primary performance metrics are reported in all result tables.