Processing fMRI Brain Signals Using Latents from Natural Image Autoencoders

Juhyeon Park[*]
IPAI, Seoul National University
parkjh9229@snu.ac.kr

Peter Yongho Kim[*]
ECE, Seoul National University
peterkim98@snu.ac.kr

Jungwoo Park[*]
ECE, Seoul National University
lawjwpark@snu.ac.kr

Jubin Choi
IPAI, Seoul National University
wnqlszoq123@snu.ac.kr

Jungwoo Seo
BCS, Seoul National University
jungwoo.seo95@gmail.com

Jiook Cha
Psychology, Seoul National University
connectome@snu.ac.kr

Taesup Moon[†]
ECE/ASRI/INMC/AIIS, Seoul National University
tsmoon@snu.ac.kr

Abstract

Modeling long-range spatiotemporal dynamics in functional Magnetic Resonance Imaging (fMRI) remains a key challenge due to the high dimensionality of the four-dimensional signals. Prior voxel-based models, although demonstrating excellent performance and interpretation capabilities, are constrained by prohibitive memory demands and thus can only capture limited temporal windows. To address this, we propose TABLeT (Two-dimensionally Autoencoded Brain Latent Transformer), a novel approach that tokenizes fMRI volumes using a pre-trained 2D natural image autoencoder. Each 3D fMRI volume is compressed into a compact set of continuous tokens, enabling efficient long-sequence modeling with a simple transformer encoder. Across large-scale benchmarks including the Human Connectome Project (HCP) and ADHD-200 datasets, TABLeT consistently outperforms existing models in multiple tasks, while demonstrating substantial gains in computational and memory efficiency over the state-of-the-art voxel-based method. Our findings highlight a new paradigm for scalable spatiotemporal modeling of brain activity.

1 Introduction

The human brain is a spatiotemporal dynamic system whose activity can be non-invasively measured using functional magnetic resonance imaging (fMRI). A large body of work has leveraged fMRI to investigate functional connectivity patterns for tasks such as neurological disorder diagnosis or demographic attribute prediction [13, 12, 19, 15, 14, 4, 7]. Existing approaches can be broadly divided into two categories: *ROI-based methods* and *voxel-based methods*.

[*] Equal contributions.

[†] Corresponding author.

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

ROI-based methods first define a set of regions of interest (ROIs) based on anatomical segmentation [20], extract their corresponding time-series signals, and then compute functional connectivity matrices as model inputs. Although this approach is computationally efficient for managing the high dimensionality of fMRI data, it has several limitations: performance strongly depends on the choice of ROIs, fine-grained 3D spatial structures may be lost, and aggressive compression can discard informative signals. To overcome these limitations, voxel-based methods such as TFF [15] and SwiFT [14] have been proposed. These methods directly process raw 4D fMRI data, thereby preserving spatial and temporal information, while also allowing detailed interpretation as they directly operate on the given image. However, due to the massive scale of fMRI volumes, the temporal length that could be simultaneously processed by the model is severely restricted (e.g., SwiFT uses only 20 timesteps at once), potentially missing informative long-range temporal dynamics, and limiting use for tasks that require longer-range interactions, such as the infraslow BOLD–LFP coupling and global arousal waves that unfold over tens of seconds [17, 22].

In this work, we address this challenge by proposing to *tokenize* fMRI volumes into a compact set of continuous tokens, thereby enabling Transformers to model substantially longer temporal sequences. To this end, we paid attention to the remarkable perceptual information preservation capability of the Deep Compression Autoencoder (DCAE) [5] and aimed to leverage it, as it effectively tokenizes a 256×256 2D natural image into *just* 64 continuous tokens (a compression ratio of 32). Motivated by this, we ask *whether a high-performing **2D** autoencoder trained on **natural images** can serve as an effective tokenizer for **4D fMRI** data.*

Our findings reveal that such an autoencoder can indeed be applied to tokenize fMRI volumes. By rearranging the tokens extracted from each 2D slice of a 3D fMRI volume, we compress an entire volume into *only* 27 continuous tokens, thereby dramatically reducing the input size and enabling efficient long-sequence modeling with a simple Transformer encoder-based architecture. We dub our method TABLeT, **T**wo-dimensionally **A**utoencoded **B**rain **L**at**e**nt **T**ransformer, which achieves superior performance compared to both ROI-based and voxel-based baselines on demographic attribute prediction and attention-deficit hyperactivity disorder (ADHD) diagnosis tasks, while drastically saving memory and computation costs compared to the voxel-based baseline.

2 Method

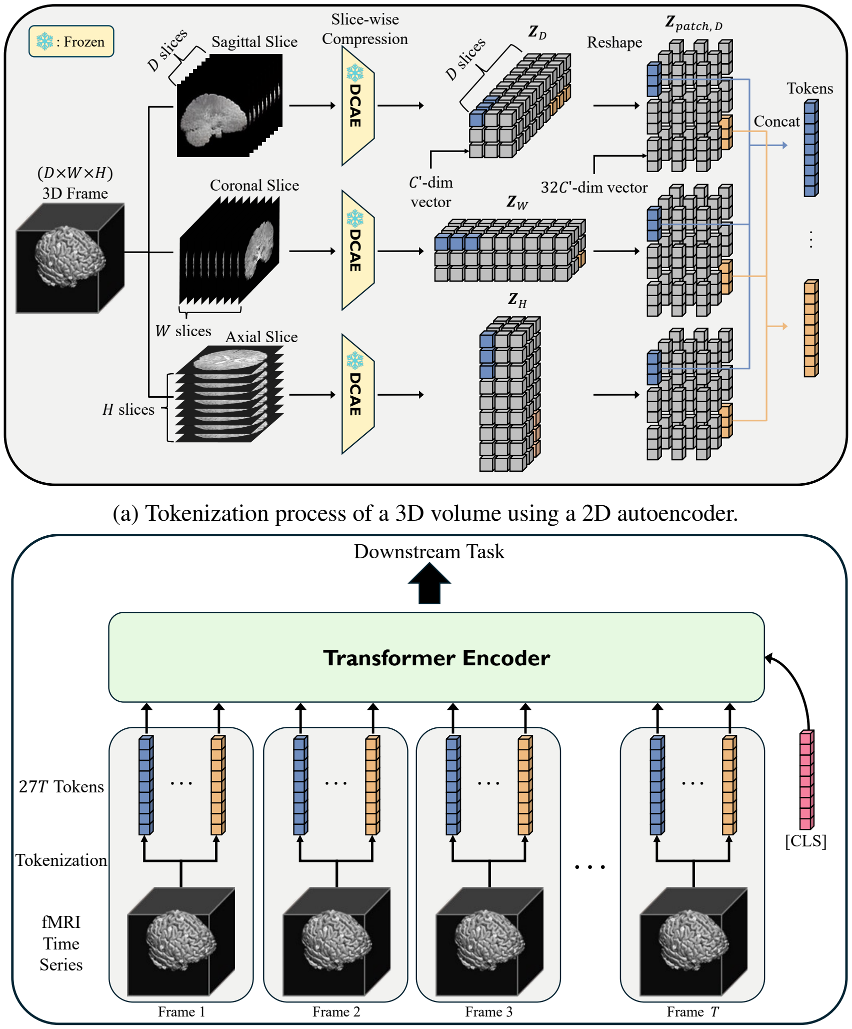
2.1 Tokenization of fMRI with 2D Natural Image Autoencoder

When tokenizing 4D fMRI data, the natural thing to do would be to train a tokenizer from scratch using the fMRI data itself. However, training such tokenizers is computationally expensive and may require a large number of samples to achieve reliable performance, which poses a significant bottleneck in domains where data are scarce, such as medical imaging. Moreover, their generalizability is questionable, as fMRI data can exhibit different characteristics depending on the scanning apparatus.

To bypass these problems, we sought a method to tokenize fMRI data without any additional training while faithfully preserving the original information. Based on recent findings showing that image autoencoders can effectively preserve the perceptual information of natural images, we hypothesized that these models could similarly preserve the perceptual information of fMRI data and thus be used for its tokenization. Among the available options, we selected DCAE [5], which has shown high compression capabilities while still preserving the details of the image. We use the unmodified *dc-ae-f32c32-in-1.0* DCAE provided by Chen et al. [5] for all 2D DCAE experiments.

One important problem is that our data is a timeseries of 3D images, while the autoencoder only operates with 2D images. Therefore, we slice the data into 2D images and independently feed them into the autoencoder.

We first experimented to see how well a pre-trained 2D DCAE reconstructs an fMRI brain image compared to a manually trained 3D DCAE, as detailed in Section 3.3. Surprisingly, the pre-trained 2D version was better at reconstructing the image even though it was not explicitly trained with it. Based on this finding, we propose to tokenize each single 3D volume at a time with the 2D DCAE, applying this across the entire fMRI sequence. *Remark*: for simplicity, we hereafter refer to the 2D natural image DCAE as 2D DCAE and the 3D fMRI-trained DCAE as 3D DCAE.



(a) Tokenization process of a 3D volume using a 2D autoencoder.

(b) Overview of TABLeT.

Figure 1: In TABLeT, each frame of the fMRI timeseries is tokenized by a 2D autoencoder, and the resulting tokens are processed by a Transformer.

Tokenization of a 3D Volume with Slicing. Formally, each fMRI frame is a 3D volume $\mathbf{X} \in \mathbb{R}^{1\times D\times H\times W}$. The single channel is first duplicated across three channels to simulate an RGB structure, giving $\mathbf{X} \in \mathbb{R}^{3\times D\times H\times W}$. One spatial dimension is then chosen as the slicing axis, so the volume becomes a stack of 2D images with the other two dimensions as spatial extents. For example, if the depth axis is chosen, the volume is treated as D images of shape $\mathbb{R}^{3\times H\times W}$. Each image is compressed independently into a latent representation $\mathbf{Z} \in \mathbb{R}^{C'\times\frac{H}{32}\times\frac{W}{32}}$, where the factor of 32 is the spatial compression ratio.

Aggregation of 3 Axes. This procedure is repeated for all three slicing axes, producing three latent volumes: $\mathbf{Z}_D \in \mathbb{R}^{D\times C'\times\frac{H}{32}\times\frac{W}{32}}$, $\mathbf{Z}_H \in \mathbb{R}^{H\times C'\times\frac{D}{32}\times\frac{W}{32}}$, $\mathbf{Z}_W \in \mathbb{R}^{W\times C'\times\frac{D}{32}\times\frac{H}{32}}$. For each latent volume, the uncompressed dimension (the slicing axis) is further partitioned into patches of size 32, reshaping it to $\mathbf{Z}_{\text{patch},D}, \mathbf{Z}_{\text{patch},H}, \mathbf{Z}_{\text{patch},W} \in \mathbb{R}^{32C'\times\frac{D}{32}\times\frac{H}{32}\times\frac{W}{32}}$. This yields $\frac{D}{32}\times\frac{H}{32}\times\frac{W}{32}$ tokens per axis, where each token corresponds to a position in the downsampled 3D grid and has hidden dimension $32C'$. The three axis-specific representations are then aligned to the same grid and concatenated along the feature dimension, resulting in $\frac{D}{32}\times\frac{H}{32}\times\frac{W}{32}$ tokens per frame with hidden dimension $96C'$. In our case, $H = W = D = 96$ and $C' = 32$. Thus, a 3D volume of shape $(1, 96, 96, 96)$ is tokenized into 27 tokens with an embedding dimension of 3072. Finally, we note that tokenization is performed only once, and the tokens are cached for later use. Consequently, its computational cost is negligible compared to the subsequent training process.

2.2 TABLeT Model Architecture

We model the spatiotemporal dynamics between the tokenized fMRI frames with a simple Transformer encoder [27], naming the pipeline **TABLeT** (Two-dimensionally Autoencoded Brain Latent Transformer). Our model is based on a Transformer encoder architecture and incorporates modern components widely used in large language models [21, 11]. In particular, we adopt grouped query attention [1] to efficiently handle long sequences, along with the increasingly popular rotary positional encoding [25]. In addition, we employ `F.scaled_dot_product_attention` from PyTorch [18] to achieve more memory-efficient and faster attention computations. Before being fed into the Transformer, fMRI tokens are normalized and projected into a lower-dimensional embedding space via a linear layer. A `[CLS]` token is prepended to the sequence, followed by an additional normalization step to enhance training stability. Unless stated otherwise, the model is composed of 12 Transformer layers with 14 attention heads and 2 key–value heads, and it processes sequences of tokens from 256 volumes at once. We randomly sampled 256 frames from the entire sequence of each subject at every training iteration, while for validation, we used all frames by partitioning the sequence from the beginning and averaging the outputs across partitions, following Kim et al. [14].

3 Experimental Results

3.1 Experimental Setting

Dataset & Task. We used resting-state fMRI data from 1,061 healthy young adults in the Human Connectome Project (HCP) [24] and from 533 children and adolescents, including both individuals diagnosed with ADHD and healthy controls, included in ADHD-200 [3].

For the HCP dataset, we used the preprocessed data provided by Smith et al. [24], which goes through the preprocessing pipeline including bias field reduction, skull-stripping, cross-modality registration, and spatial normalization to the MNI space [10]. For the ADHD-200 dataset, we downloaded the fMRIPrep [8, 9] processed data from Bellec et al. [3] and regressed out nuisance variables using cosine bases, six motion parameters, and aCompCor components. Following Kim et al. [14], we set each fMRI volume to the shape of $(96, 96, 96)$ by cropping out the background and padding appropriately. We then applied global min-max normalization and rescaled it to the $[-1, 1]$ range to match the input range of the DCAE.

We split the HCP dataset using stratified sampling based on age, gender, and intelligence scores. For the ADHD-200 dataset, we performed stratified sampling based on diagnosis labels and image acquisition sites, following Kan et al. [12]. For all of the splits, the training, validation, and test sets were assigned in a 0.7:0.15:0.15 ratio. We generated four different random stratified splits and used them for all experiments. For the ADHD-200 dataset, we used three random seeds for each split to ensure reliable results, given the relatively small size of the dataset.

We considered sex, age, and intelligence score (`CogTotalComp-AgeAdj`) for the prediction targets for the HCP dataset, and the diagnosis label for the ADHD-200 dataset. The continuous targets (age, intelligence) are z-normalized using statistics of the training set.

Baselines. We considered four ROI-based models as our baseline: XGBoost (eXtreme Gradient Boosting) [6], BrainNetCNN [13], Brain Network Transformer (BNT) [12], and meanMLP [19]. To preprocess the data, we first construct the functional connectivity (FC) matrix using a total of 450 ROIs, comprising 400 ROIs from the Schaefer-400 atlas [23] and 50 additional ROIs from the Tian-Scale III atlas [26]. The Fisher-transformed FC matrix serves as the input for the ROI-based models. For the XGBoost model, we used the upper-triangular part of the FC matrix as the input. The detailed description of each baseline is provided in Section A.

For the voxel-based baseline, we adopted SwiFT [14], the state-of-the-art voxel-based model. We reproduced the original model with 20 input time frames and then extended the temporal window size for longer input time frames. Specifically, we used a temporal window size of $T/5$ for T input time frames, as the window size was 4 for 20 input frames in the original implementation.

Table 1: Performance comparison to baselines on classification and regression tasks

Method	HCP									ADHD-200		
	Sex			Age			Intelligence			Diagnosis		
	ACC	AUC	F1	MSE	MAE	R^2	MSE	MAE	R^2	ACC	AUC	F1
XGBoost [6]	82.2	0.890	0.837	0.859	0.769	0.296	0.908	0.779	0.292	62.3	0.650	0.555
BrainNetCNN [13]	86.3	0.937	0.866	0.847	0.749	0.372	0.967	0.788	0.286	59.2	0.640	0.545
BNT [12]	86.3	0.935	0.872	0.794	0.719	0.444	0.920	0.778	0.318	63.6	0.677	0.624
meanMLP [19]	84.5	0.915	0.855	0.846	0.751	0.370	0.887	0.767	0.340	56.8	0.617	0.532
SwiFT ($T = 20$) [14]	93.1	0.978	0.937	0.776	0.719	0.450	0.940	0.782	0.297	63.3	0.693	0.623
SwiFT ($T = 50$) [14]	92.2	0.972	0.929	**0.764**	**0.699**	0.460	0.865	0.758	0.354	63.9	0.701	0.627
TABLeT ($T = 256$)	**93.8**	**0.987**	**0.943**	0.773	0.705	**0.473**	**0.835**	**0.741**	**0.392**	**65.8**	**0.728**	**0.631**
TABLeT (3D DCAE)	92.2	0.973	0.929	0.767	0.693	0.475	0.869	0.755	0.387	65.8	0.711	0.643

3.2 Main Results

The experimental results are presented in Table 1, and the second-order statistics are detailed in Section F. The findings demonstrate that TABLeT achieves superior performance compared to baseline methods, including both ROI-based and voxel-based approaches, across four tasks and two datasets, with only marginal improvement observed for the HCP-Age task. Furthermore, the results of SwiFT ($T = 20, 50$) and TABLeT indicate a clear association between temporal window length and performance in intelligence prediction and ADHD diagnosis, suggesting that modeling longer temporal variability may be particularly advantageous for these tasks.

3.3 Comparison of 2D Natural Image DCAE and 3D fMRI-trained DCAE

Reconstruction Quality. We evaluated the reconstruction quality of the 2D DCAE on fMRI. Specifically, we computed PSNR and SSIM for 3D volumes at each time step and then averaged the results across all time steps and subjects. As a baseline, we trained a 3D DCAE on 8,178 subjects from the UK Biobank [16, 2]; a detailed training procedure is provided in Section D. To assess generalizability, we deliberately excluded HCP and ADHD-200 from the training set. Remarkably, the 2D DCAE achieved higher reconstruction quality, compared to the 3D DCAE trained directly on fMRI data, despite not being fine-tuned on fMRI data.

We also attempted to fine-tune the 2D DCAE with fMRI data while freezing different parts of the autoencoder, but discovered that any fine-tuning consistently harmed the reconstruction quality.

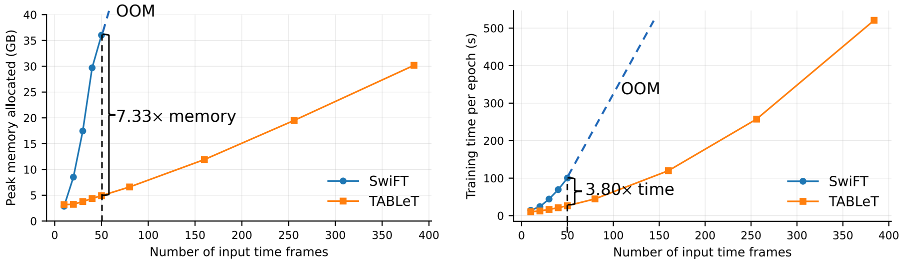
Table 2: Reconstruction quality of 2D DCAE and 3D DCAE

Model	HCP		ADHD-200	
	PSNR	SSIM	PSNR	SSIM
3D DCAE	$28.92_{\pm 0.93}$	$0.8440_{\pm 0.0111}$	$28.26_{\pm 2.42}$	$0.8324_{\pm 0.1167}$
2D DCAE, sagittal	$29.35_{\pm 0.96}$	$0.8448_{\pm 0.0110}$	$\mathbf{32.76}_{\pm 1.02}$	$\mathbf{0.9209}_{\pm 0.0078}$
2D DCAE, coronal	$\mathbf{29.54}_{\pm 0.97}$	$0.8427_{\pm 0.0102}$	$32.69_{\pm 1.01}$	$0.9152_{\pm 0.0078}$
2D DCAE, axial	$29.51_{\pm 0.97}$	$\mathbf{0.8462}_{\pm 0.0110}$	$32.43_{\pm 1.02}$	$0.9136_{\pm 0.0084}$

Training Performance. We also compared models trained with latents from the 3D DCAE and the 2D DCAE. As shown in Table 1, both models achieve competitive performance, with the 2D DCAE outperforming the 3D counterpart in most cases. These findings demonstrate that the 2D DCAE can reliably reconstruct and tokenize fMRI data, even though it was trained exclusively on natural images.

3.4 Memory and Computational Efficiency

As the development of TABLeT was motivated by the goal of making a fast and efficient voxel-based model, here we conduct a quantitative analysis to compare the memory and computational efficiency between TABLeT and SwiFT. To ensure a fair comparison, all tests were performed on a single GPU, and the batch size of both models was fixed to 4. We were only able to run SwiFT up to $T = 50$ (number of input frames) due to memory limitations. At $T = 50$, compared to SwiFT, TABLeT is 7.33 times more efficient in terms of memory, and 3.8 times faster in terms of training speed. With a similar memory budget (~30GB), the number of input frames can be extended nearly tenfold between SwiFT (40 frames) and TABLeT (384 frames).



(a) Peak memory allocation

(b) Training time per epoch

Figure 2: Comparison of (a) memory and (b) training time, between TABLeT and SwiFT.

4 Conclusion & Future Work

We presented TABLeT, a simple and efficient framework that leverages a 2D autoencoder trained on natural images to tokenize fMRI volumes. This approach enables long-range temporal modeling with Transformers while substantially reducing memory and computational costs. Experiments on HCP and ADHD-200 demonstrated that TABLeT achieves competitive or superior performance compared to both ROI-based and voxel-based baselines. Notably, the 2D DCAE consistently outperformed a brain fine-tuned 3D DCAE, highlighting the surprising transferability of vision models to neuroimaging.

To follow up on this work, we plan to further apply TABLeT across diverse tasks and datasets, while exploring self-supervised masked image modeling strategies to enhance its downstream performance by pre-training on large-scale fMRI datasets.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C2007884, RS-2025-02263628, 2021R1C1C1006503, RS-2023-00266787, RS-2023-00265406, RS-2024-00421268, RS-2024-00342301, NRF-2021M3E5D2A01022515), by Creative-Pioneering Researchers Program through Seoul National University(No. 200-20240057, 200-20240135), by Semi-Supervised Learning Research Grant by SAMSUNG(No.A0342-20220009), by Identify the network of brain preparation steps for concentration Research Grant by LooxidLabs(No.339-20230001), by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II212068, RS-2022-II220113, RS-2022-II220959, RS-2021-II211343] by the MSIT(Ministry of Science, ICT), Korea, under the Global Research Support Program in the Digital Field program(RS-2024-00421268) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation), by the National Supercomputing Center with supercomputing resources including technical support (KSC-2023-CRE-0568), by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A3A2A02090597), by the Korea Health Industry Development Institute (KHIDI), by the Ministry of Health and Welfare, Republic of Korea (HR22C1605), by the Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT (MSIT, Korea) & Gwangju Metropolitan City and by KBRI basic research program through Korea Brain Research Institute funded by Ministry of Science and ICT(25-BR-05-01), by Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(Ministry of Education) (P0025681-G02P22450002201-10054408, Semiconductor-Specialized University), and by Hyundai Motor Chung Mong-Koo Foundation.

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.

This research has been conducted using data from UK Biobank, a major biomedical database (`https://www.ukbiobank.ac.uk`). Our project ID of the UK Biobank is 32575. Data were also provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

References

[1] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebron, and S. Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901, 2023.

[2] F. Alfaro-Almagro, M. Jenkinson, N. K. Bangerter, J. L. Andersson, L. Griffanti, G. Douaud, S. N. Sotiropoulos, S. Jbabdi, M. Hernandez-Fernandez, E. Vallee, et al. Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. *Neuroimage*, 166: 400–424, 2018.

[3] P. Bellec, C. Chu, F. Chouinard-Decorte, Y. Benhajali, D. S. Margulies, and R. C. Craddock. The neuro bureau adhd-200 preprocessed repository. *Neuroimage*, 144:275–286, 2017.

[4] J. O. Caro, A. H. de Oliveira Fonseca, S. A. Rizvi, M. Rosati, C. Averill, J. L. Cross, P. Mittal, E. Zappala, R. M. Dhodapkar, C. Abdallah, et al. Brainlm: A foundation model for brain activity recordings. In *The Twelfth International Conference on Learning Representations*, 2024.

[5] J. Chen, H. Cai, J. Chen, E. Xie, S. Yang, H. Tang, M. Li, and S. Han. Deep compression autoencoder for efficient high-resolution diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[6] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[7] Z. Dong, R. Li, Y. Wu, T. T. Nguyen, J. Chong, F. Ji, N. Tong, C. Chen, and J. H. Zhou. Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. *Advances in Neural Information Processing Systems*, 37:86048–86073, 2024.

[8] O. Esteban, C. J. Markiewicz, R. W. Blair, C. A. Moodie, A. I. Isik, A. Erramuzpe, J. D. Kent, M. Goncalves, E. DuPre, M. Snyder, et al. fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, 16(1):111–116, 2019.

[9] O. Esteban, R. Ciric, K. Finc, R. W. Blair, C. J. Markiewicz, C. A. Moodie, J. D. Kent, M. Goncalves, E. DuPre, D. E. Gomez, et al. Analysis of task-based functional mri data preprocessed with fmriprep. *Nature protocols*, 15(7):2186–2202, 2020.

[10] A. C. Evans, D. L. Collins, S. Mills, E. D. Brown, R. L. Kelly, and T. M. Peters. 3d statistical neuroanatomical models from 305 mri volumes. In *1993 IEEE conference record nuclear science symposium and medical imaging conference*, pages 1813–1817. IEEE, 1993.

[11] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[12] X. Kan, W. Dai, H. Cui, Z. Zhang, Y. Guo, and C. Yang. Brain network transformer. *Advances in Neural Information Processing Systems*, 35:25586–25599, 2022.

[13] J. Kawahara, C. J. Brown, S. P. Miller, B. G. Booth, V. Chau, R. E. Grunau, J. G. Zwicker, and G. Hamarneh. Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038–1049, 2017.

[14] P. Kim, J. Kwon, S. Joo, S. Bae, D. Lee, Y. Jung, S. Yoo, J. Cha, and T. Moon. Swift: Swin 4d fmri transformer. *Advances in Neural Information Processing Systems*, 36:42015–42037, 2023.

[15] I. Malkiel, G. Rosenman, L. Wolf, and T. Hendler. Self-supervised transformers for fmri representation. In *International Conference on Medical Imaging with Deep Learning*, pages 895–913, 2022.

[16] K. L. Miller, F. Alfaro-Almagro, N. K. Bangerter, D. L. Thomas, E. Yacoub, J. Xu, A. J. Bartsch, S. Jbabdi, S. N. Sotiropoulos, J. L. Andersson, et al. Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523–1536, 2016.

[17] W.-J. Pan, G. J. Thompson, M. E. Magnuson, D. Jaeger, and S. Keilholz. Infraslow lfp correlates to resting-state fmri bold signals. *Neuroimage*, 74:288–297, 2013.

[18] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[19] P. Popov, U. Mahmood, Z. Fu, C. Yang, V. Calhoun, and S. Plis. A simple but tough-to-beat baseline for fmri time-series classification. *NeuroImage*, 303:120909, 2024.

[20] J. D. Power, A. L. Cohen, S. M. Nelson, G. S. Wig, K. A. Barnes, J. A. Church, A. C. Vogel, T. O. Laumann, F. M. Miezin, B. L. Schlaggar, et al. Functional network organization of the human brain. *Neuron*, 72(4):665–678, 2011.

[21] Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu. Qwen2.5 technical report, 2025.

[22] R. V. Raut, A. Z. Snyder, A. Mitra, D. Yellin, N. Fujii, R. Malach, and M. E. Raichle. Global waves synchronize the brain's functional systems with fluctuating arousal. *Science advances*, 7 (30):eabf2709, 2021.

[23] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, and B. T. Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral cortex*, 28(9):3095–3114, 2018.

[24] S. M. Smith, C. F. Beckmann, J. Andersson, E. J. Auerbach, J. Bijsterbosch, G. Douaud, E. Duff, D. A. Feinberg, L. Griffanti, M. P. Harms, et al. Resting-state fmri in the human connectome project. *Neuroimage*, 80:144–168, 2013.

[25] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[26] Y. Tian, D. S. Margulies, M. Breakspear, and A. Zalesky. Topographic organization of the human subcortex unveiled with functional connectivity gradients. *Nature neuroscience*, 23(11): 1421–1432, 2020.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Processing fMRI Brain Signals Using Latents from Natural Image Autoencoders

Technical Appendix

A Related Works

ROI-Based Methods. ROI-based methods parcellate the brain into ROIs and average the BOLD signals within each. The signals are transformed into functional connectivity (FC) matrices by computing the correlation between the time series of the ROIs. BrainNetCNN [13] treats the FC matrix as a 2D image and uses edge-to-edge, edge-to-node, and node-to-graph convolutional filters to utilize topological locality in ROI-based networks. The approach is powerful but limited due to the fixed structure of standard CNNs. The Brain Network Transformer [12] adapts Transformer architecture to process FC matrices as graphs of ROIs. It uses hierarchical attention mechanisms to learn node identities and group-specific patterns, addressing distribution shifts across subpopulations. meanMLP [19] is a lightweight MLP-based model that applies the same MLP repeatedly across parcellated fMRI time-series and averages the resulting embeddings across time before a final classification layer. It outperforms complex models like Transformers on clinical predictions.

Voxel-Based Methods. Voxel-based methods process 4D fMRI volumes, enabling end-to-end learning of spatiotemporal features without ROI aggregation. SwiFT [14] extends the Swin Transformer to 4D fMRI volumes with a 4D window multi-head self-attention mechanism and absolute positional embeddings. It learns brain dynamics end-to-end, outperforming baselines on large-scale datasets (HCP, ABCD, UKB) for sex, age, and intelligence prediction.

B Limitations

Tokenization. In TABLeT, each frame of the fMRI timeseries is independently tokenized, which may or may not harm the subtle temporal dynamics of the brain signal during the tokenization process, and there could potentially be a method that could explicitly consider the temporal dynamics and consistency during the tokenization process. Applying TABLeT to more temporally dynamic tasks could be an interesting follow-up to validate this point.

Model Architecture. TABLeT uses a refined Transformer encoder architecture, meaning it processes the input tokens all at once, without any special consideration of their underlying spatial position or temporal nature. A specialized model that explicitly considers the spatial and temporal alignment between the tokens might better capture the spatiotemporal dynamics.

C Implementation Details

All experiments are conducted on the NVIDIA A100-40GB, and RTX A6000 GPUs. We used `fp16` mixed precision for the training of all models.

We used `BCEWithLogitsLoss` for the classification task, and used `pos-weight` option for the ADHD task to account for class imbalance. We used `L1Loss` for the regression tasks.

For the voxel-based models, SwiFT and TABLeT, training was performed by randomly sampling consecutive 3D volumes. For evaluation, following [14], we computed the final prediction by averaging the model outputs over all possible windows starting from the first frame.

Shared Settings We used the following strategy for all of the experiments, unless explicitly stated.

- `Optimizer`: AdamW using a cosine decay learning rate scheduler, with weight decay of 10^{-2}.
- `Hyperparameter Search`: For the HCP-sex task, we searched the hyperparameter based on the validation AUROC for each model. For the HCP-age, and HCP-intelligence tasks, we

searched based on the validation MAE. For the ADHD, we searched based on the validation loss to consider the `pos-weight` for the class imbalance.
- `Early Stopping`: We chose the early-stopped model for the BrainNetCNN, BNT, and meanMLP by default. As we observed that SwiFT and TABLeT are more stable during training, we report results from the final epoch for all tasks.

XGBoost We grid searched for hyperparameter tuning of XGBoost for the following.

- `Maximum depth`: Chosen between 3 and 5
- `Minimal child weight`: Chosen between 1 and 7
- `Gamma`: Chosen between 0.0 and 0.4
- `Learning rate`: Chosen between 0.05 and 0.3
- `Colsample by tree`: Chosen between 0.6 and 0.9

BrainNetCNN We trained BrainNetCNN with the following setup:

- `Learning rate`: Chosen between 1×10^{-6} and 2×10^{-4}
- `Batch size`: 64
- `Epochs`: 100 epochs of training

Brain Network Transformer We trained Brain Network Transformer with the following setup:

- `Learning rate`: Chosen between 1×10^{-6} and 2×10^{-4}
- `Batch size`: 64
- `Epochs`: 100 epochs of training

meanMLP We trained meanMLP with the following setup:

- `Learning rate`: Chosen between 1×10^{-4} and 1×10^{-2}
- `Batch size`: 32
- `Epochs`: 100 epochs of training

SwiFT We trained SwiFT with the following setup:

- `Learning rate`: Chosen between 1×10^{-6} and 5×10^{-5}
- `Batch size`: 4
- `Epochs`: 25 epochs of training for HCP, 30 epochs for ADHD.

TABLeT We trained TABLeT with the following setup:

- `Learning rate`: Chosen between 5×10^{-7} and 5×10^{-5}
- `Batch size`: 4
- `Epochs`: 50 epochs of training for HCP-sex, HCP-intelligence, ADHD, and 30 epochs for HCP-age.

D Training Details of 3D fMRI-trained DCAE

We developed 3D DCAE by adapting the architecture of 2D DCAE [5] to handle 3D volume inputs. To achieve this, we replaced 2D convolutional layers with 3D convolutional layers and adjusted components such as RMS normalization, batch normalization, `PixelUnshuffle`, and `PixelShuffle` to process 3D data effectively. The model was configured with 1 input channel, 1024 latent channels, encoder-decoder width of `[16, 64, 256, 256, 1024, 1024]`, and encoder-decoder depth of `[0, 2, 2, 5, 5, 5]`.

For training, we utilized a dataset of 8,178 subjects from the UK Biobank, splitting it into training and validation sets with a 9:1 ratio and stratification based on sex and age. The model was trained for 100 epochs with an initial learning rate of 4×10^{-5}, which was gradually reduced using `ReduceLROnPlateau` scheduler. During each epoch, we randomly selected a single fMRI frame from the full set of frames for each subject to train the model. The training process used $\mathcal{L}_2$ reconstruction loss and the AdamW optimizer with a weight decay of 1×10^{-4}.

Table 3: Performance comparison with respect to slicing axis on HCP sex classification and age regression.

Axis	HCP					
	Sex			Age		
	ACC	AUC	F1	MSE	MAE	R^2
Sagittal	91.3$_{\pm 3.6}$	0.972$_{\pm 0.017}$	0.920$_{\pm 0.033}$	0.783$_{\pm 0.111}$	0.721$_{\pm 0.041}$	0.458$_{\pm 0.076}$
Coronal	93.6$_{\pm 1.7}$	0.981$_{\pm 0.007}$	0.941$_{\pm 0.015}$	0.855$_{\pm 0.053}$	0.745$_{\pm 0.023}$	0.376$_{\pm 0.048}$
Axial	92.3$_{\pm 3.0}$	0.979$_{\pm 0.008}$	0.930$_{\pm 0.028}$	**0.748**$_{\pm 0.056}$	0.711$_{\pm 0.015}$	0.470$_{\pm 0.040}$
All	**93.8**$_{\pm 0.9}$	**0.987**$_{\pm 0.003}$	**0.943**$_{\pm 0.008}$	0.773$_{\pm 0.077}$	**0.705**$_{\pm 0.038}$	**0.473**$_{\pm 0.053}$

Table 4: Performance comparison with respect to slicing axis on HCP intelligence regression and ADHD diagnosis.

Axis	HCP			ADHD-200		
	Intelligence			Diagnosis		
	MSE	MAE	R^2	ACC	AUC	F1
Sagittal	0.842$_{\pm 0.058}$	0.744$_{\pm 0.028}$	**0.401**$_{\pm 0.060}$	**65.9**$_{\pm 2.7}$	0.712$_{\pm 0.026}$	**0.633**$_{\pm 0.038}$
Coronal	0.850$_{\pm 0.057}$	0.749$_{\pm 0.029}$	0.381$_{\pm 0.065}$	63.5$_{\pm 3.1}$	0.707$_{\pm 0.036}$	0.621$_{\pm 0.040}$
Axial	0.896$_{\pm 0.070}$	0.773$_{\pm 0.033}$	0.309$_{\pm 0.072}$	64.3$_{\pm 2.5}$	0.713$_{\pm 0.022}$	0.622$_{\pm 0.034}$
All	**0.835**$_{\pm 0.053}$	**0.741**$_{\pm 0.028}$	0.392$_{\pm 0.062}$	65.8$_{\pm 3.3}$	**0.728**$_{\pm 0.028}$	0.631$_{\pm 0.033}$

E Additional Ablation Studies

E.1 Effect of Aggregation of Three Axes

We compared the performance of two models: one trained with fMRI tokens derived from a single axis and the other with aggregated tokens. As shown in Table 3 and Table 4, the performance of TABLeT varies depending on the chosen axis for tokenization and training. In contrast, our aggregation scheme consistently achieves strong performance across diverse tasks, eliminating the dependence on any particular slicing axis.

F Detailed Experimental Results

We provide the results reported in the manuscript (Table 1) with the standard deviation in Tables 5 and 6.

Table 5: Main experimental results with standard deviation on HCP sex classification and age regression.

Method	HCP					
	Sex			Age		
	ACC	AUC	F1	MSE	MAE	R^2
XGBoost [6]	82.2$_{\pm 2.5}$	0.890$_{\pm 0.028}$	0.837$_{\pm 0.025}$	0.859$_{\pm 0.074}$	0.769$_{\pm 0.033}$	0.296$_{\pm 0.112}$
BrainNetCNN [13]	86.3$_{\pm 4.9}$	0.937$_{\pm 0.027}$	0.866$_{\pm 0.049}$	0.847$_{\pm 0.097}$	0.749$_{\pm 0.040}$	0.372$_{\pm 0.097}$
BNT [12]	86.3$_{\pm 3.0}$	0.935$_{\pm 0.026}$	0.872$_{\pm 0.030}$	0.794$_{\pm 0.051}$	0.719$_{\pm 0.027}$	0.444$_{\pm 0.055}$
meanMLP [19]	84.5$_{\pm 2.5}$	0.915$_{\pm 0.018}$	0.855$_{\pm 0.028}$	0.846$_{\pm 0.056}$	0.751$_{\pm 0.030}$	0.370$_{\pm 0.087}$
SwiFT (T = 20) [14]	93.1$_{\pm 0.5}$	0.978$_{\pm 0.008}$	0.937$_{\pm 0.004}$	0.776$_{\pm 0.043}$	0.719$_{\pm 0.015}$	0.450$_{\pm 0.031}$
SwiFT (T = 50) [14]	92.2$_{\pm 1.1}$	0.972$_{\pm 0.014}$	0.929$_{\pm 0.010}$	**0.764**$_{\pm 0.092}$	**0.699**$_{\pm 0.047}$	0.460$_{\pm 0.071}$
TABLeT (T = 256)	**93.8**$_{\pm 0.9}$	**0.987**$_{\pm 0.003}$	**0.943**$_{\pm 0.008}$	0.773$_{\pm 0.077}$	0.705$_{\pm 0.038}$	**0.473**$_{\pm 0.053}$
TABLeT (3D DCAE)	92.2$_{\pm 1.7}$	0.973$_{\pm 0.010}$	0.929$_{\pm 0.014}$	0.767$_{\pm 0.118}$	0.693$_{\pm 0.043}$	0.475$_{\pm 0.076}$

G Detailed Data Description

We provide a detailed description of each dataset used in our study in Table 7.

H Licenses

- BrainNetCNN, Brain Network Transformer [13, 12] : https://github.com/Wayfear/BrainNetworkTransformer, MIT License.

Table 6: Main experimental results with standard deviation on HCP intelligence regression and ADHD diagnosis.

Method	HCP			ADHD-200		
	Intelligence			Diagnosis		
	MSE	MAE	R^2	ACC	AUC	F1
XGBoost [6]	$0.908_{\pm0.054}$	$0.779_{\pm0.023}$	$0.292_{\pm0.099}$	$62.3_{\pm2.5}$	$0.650_{\pm0.036}$	$0.555_{\pm0.031}$
BrainNetCNN [13]	$0.967_{\pm0.119}$	$0.788_{\pm0.044}$	$0.286_{\pm0.112}$	$59.2_{\pm10.7}$	$0.640_{\pm0.095}$	$0.545_{\pm0.118}$
BNT [12]	$0.920_{\pm0.092}$	$0.778_{\pm0.054}$	$0.318_{\pm0.083}$	$63.6_{\pm5.4}$	$0.677_{\pm0.062}$	$0.624_{\pm0.057}$
meanMLP [19]	$0.887_{\pm0.076}$	$0.767_{\pm0.028}$	$0.340_{\pm0.045}$	$56.8_{\pm6.8}$	$0.617_{\pm0.067}$	$0.532_{\pm0.095}$
SwiFT ($T = 20$) [14]	$0.940_{\pm0.111}$	$0.782_{\pm0.044}$	$0.297_{\pm0.080}$	$63.3_{\pm3.7}$	$0.693_{\pm0.030}$	$0.623_{\pm0.033}$
SwiFT ($T = 50$) [14]	$0.865_{\pm0.093}$	$0.758_{\pm0.046}$	$0.354_{\pm0.070}$	$63.9_{\pm3.2}$	$0.701_{\pm0.032}$	$0.627_{\pm0.030}$
TABLeT ($T = 256$)	$\mathbf{0.835}_{\pm0.053}$	$\mathbf{0.741}_{\pm0.028}$	$\mathbf{0.392}_{\pm0.062}$	$\mathbf{65.8}_{\pm3.3}$	$\mathbf{0.728}_{\pm0.028}$	$\mathbf{0.631}_{\pm0.033}$
TABLeT (3D DCAE)	$0.869_{\pm0.077}$	$0.755_{\pm0.032}$	$0.387_{\pm0.078}$	$65.8_{\pm1.7}$	$0.711_{\pm0.025}$	$0.643_{\pm0.021}$

Table 7: Demographic information of the datasets used in our study

Category	HCP	ADHD-200
Number of subjects	1061	533
Sex		
Male, n (%)	488 (46.0%)	207 (38.8%)
Female, n (%)	573 (54.0%)	325 (61.0%)
N/A, n (%)	–	1 (0.2%)
Age (years)	$28.79_{\pm3.70}$	$11.94_{\pm3.40}$
Intelligence	$113.32_{\pm20.50}$	–
Diagnosed, n (%)	–	236 (44.3%)

- meanMLP [19] : https://github.com/neuroneural/meanMLP, MIT License.
- SwiFT [14]: https://github.com/Transconnectome/SwiFT, Apache-2.0 License.