Can Natural Image Autoencoders Compactly Tokenize fMRI Volumes for Long-Range Dynamics Modeling?

Appendix

A. Implementation Details

For the 2D DCAE, we brought the unmodified `dc-ae-f32c32-in-1.0` checkpoint provided by Chen et al. [2] for all 2D natural image DCAE experiments.

All experiments were conducted on the NVIDIA A100-40GB and RTX A6000 GPUs. We used `fp16` mixed precision for the training of all models except for TFF due to NaN error during training.

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, TFF, SwiFT, and TABLeT, training was performed by randomly sampling consecutive 3D volumes. For evaluation, following Kim et al. [3], 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 UKB-Sex and HCP-Sex tasks, we searched the hyperparameter based on the validation AUROC for each model. For the UKB-Age, HCP-Age, and HCP-Intelligence tasks, we searched based on the validation MAE. For 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, meanMLP, Brain-JEPA, and TFF 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

Brain-JEPA We trained Brain-JEPA from scratch for fair comparison with the following setup.

- `Learning rate`: Chose between 1×10^{-5} and 7×10^{-4}.
- `Batch size`: 16
- `Epochs`: 50 epochs of training

TFF We trained TFF with the following setup:

- Phase 1
  - `Learning rate`: 3×10^{-3} for UKB, ADHD, and 7×10^{-4} for HCP
  - `Batch size`: 4
  - `Epochs`: 100 epochs of training
- Phase 2
  - `Learning rate`: 1×10^{-5} for UKB, ADHD, and chosen between 1×10^{-5} and 1×10^{-6}
  - `Batch size`: 2
  - `Epochs`: 50 epochs of training
- Fine-tuning
  - `Learning rate`: Chosen between 1×10^{-5} and 1×10^{-6} for UKB and ADHD, chosen between 3×10^{-7} and 1×10^{-6} for HCP,
  - `Batch size`: 4
  - `Epochs`: 10 epochs of training for UKB-Sex, 20 epochs of training for HCP, UKB-Age, and 30 epochs of training for ADHD.

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 UKB, HCP, 30 epochs for ADHD.

TABLeT We trained TABLeT with the following setup:

- `Learning rate`: Chosen between 3×10^{-7} and 5×10^{-5}
- `Batch size`: 4
- `Epochs`: 50 epochs of training for HCP-Sex, HCP-Intelligence, ADHD, 30 epochs for age regression, 15 epochs for UKB-Sex.

B. Training Details of 3D fMRI-trained DCAE

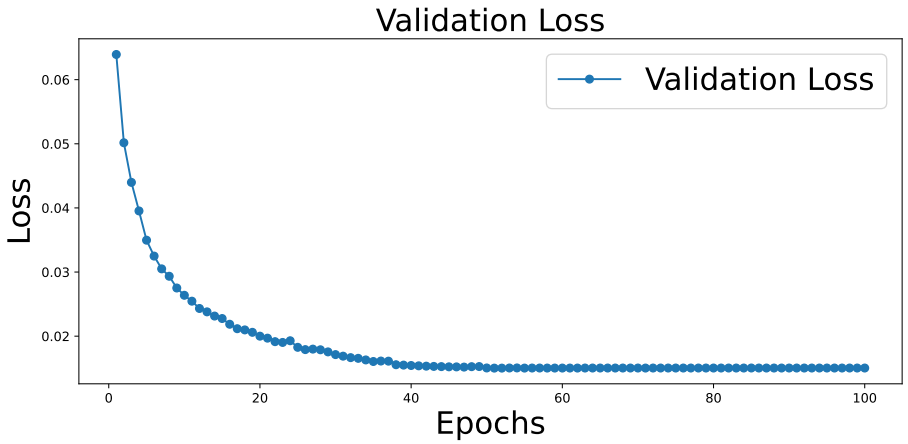

Figure 1. Validation Loss Curve for Training of 3D DCAE.

We developed 3D DCAE by adapting the architecture of 2D DCAE [2] 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]`, to make the same compression ratio as the 2D DCAE.

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}.

As the training curve in Fig. 1 shows, we made every effort to train the 3D DCAE model to achieve the best performance and ensure full convergence, for fair comparison.

C. Additional Experimental Results

C.1. Experiments with Matched T

We report the performance of TABLeT with $T = 20, 50$ in Tab. A1 on HCP and ADHD-200. The results demonstrate that TABLeT shows comparable performance to SwiFT. Therefore, it proves that TABLeT is able to maintain competitive performance even with a reduced number T, highlighting that the performance gain in Tab. 1 is not solely from the increased T, but rather from the effectiveness of the proposed method itself.

Table A1. Results of experiments with matched T on HCP and ADHD-200.

Model	HCP									ADHD-200		
	Sex			Age			Intelligence			Diagnosis		
	ACC	AUC	F1	MSE	MAE	ρ	MSE	MAE	ρ	ACC	AUC	F1
SwiFT ($T = 20$)	**93.1**	0.978	**0.937**	**0.776**	0.719	0.450	0.940	0.782	0.297	63.3	0.693	**0.623**
TABLeT ($T = 20$)	91.6	**0.980**	0.923	0.784	**0.710**	**0.460**	**0.866**	**0.763**	**0.340**	**64.4**	**0.717**	0.621
SwiFT ($T = 50$)	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 ($T = 50$)	**93.4**	**0.986**	**0.940**	**0.763**	0.704	**0.473**	0.916	0.778	0.334	**64.3**	**0.710**	0.615

C.2. HBN-Movie Experiments

To test on a more temporally dynamic task, we experimented on Healthy Brain Network (HBN)-Movie [1], which includes 680 subjects and 1,360 fMRI scans where each subject watched two movies. We trained the models to predict which movie was being viewed. As shown in Tab. A2, TABLeT achieves performance comparable to SwiFT with the same temporal length T, and outperforms with longer T, which demonstrates that TABLeT can handle the temporal dynamics of fMRI data.

Table A2. Results on HBN-Movie.

Model	HBN-Movie		
	ACC	AUC	F1
SwiFT ($T = 50$)	$71.7_{\pm 4.29}$	$0.810_{\pm 0.040}$	$0.717_{\pm 0.040}$
TABLeT ($T = 50$)	$74.7_{\pm 2.38}$	$0.826_{\pm 0.033}$	$0.750_{\pm 0.024}$
TABLeT ($T = 250$)	$\mathbf{82.1}_{\pm 1.82}$	$\mathbf{0.976}_{\pm 0.014}$	$\mathbf{0.847}_{\pm 0.015}$

C.3. Axis Aggregation Scheme Variations

We varied the number of tokens and the latent dimensionality while keeping the total number of values fixed at $27 \times 3072 = 82{,}944$ and setting $T = 256$. In other words, we changed the number of tokens that are combined together in the token aggregation step, where combining a larger number of tokens together leads to a larger token dimensionality while reducing the number of total resulting tokens. Since the tokens are only concatenated and rearranged in this step, the total number of values are kept constant (82,944). The results are shown in Tab. A3. The results indicate that these adjustments do not lead to significant differences in downstream task performance.

Table A3. Results on HCP and ADHD-200 with varying axis aggregation schemes.

Model	# Token	Dim.	HCP			ADHD-200		
			Intelligence			Diagnosis		
			MSE	MAE	ρ	ACC	AUC	F1
TABLeT	27	3072	$0.835_{\pm 0.053}$	$0.741_{\pm 0.028}$	$0.392_{\pm 0.062}$	$\mathbf{65.8}_{\pm 3.50}$	$0.729_{\pm 0.029}$	$0.630_{\pm 0.038}$
TABLeT Alt.1	9	9216	$\mathbf{0.814}_{\pm 0.044}$	$\mathbf{0.733}_{\pm 0.026}$	$\mathbf{0.416}_{\pm 0.061}$	$65.7_{\pm 2.74}$	$0.710_{\pm 0.031}$	$0.632_{\pm 0.043}$
TABLeT Alt.2	3	27648	$0.840_{\pm 0.093}$	$0.736_{\pm 0.044}$	$0.401_{\pm 0.035}$	$65.1_{\pm 3.59}$	$\mathbf{0.690}_{\pm 0.032}$	$\mathbf{0.636}_{\pm 0.042}$

D. Detailed Experimental Results

We provide the results reported in the manuscript with the standard deviation in Tab. A4, Tab. A5, and Tab. A6.

Table A4. Experimental results with standard deviation on UKB.

Method	UKB					
	Sex			Age		
	ACC	AUC	F1	MSE	MAE	ρ
XGBoost	$84.1_{\pm 1.7}$	$0.916_{\pm 0.012}$	$0.830_{\pm 0.019}$	$0.698_{\pm 0.013}$	$0.686_{\pm 0.008}$	$0.553_{\pm 0.018}$
BrainNetCNN	$91.7_{\pm 0.9}$	$0.969_{\pm 0.007}$	$0.912_{\pm 0.009}$	$0.597_{\pm 0.017}$	$0.618_{\pm 0.007}$	$0.647_{\pm 0.012}$
BNT	$92.4_{\pm 0.9}$	$0.980_{\pm 0.003}$	$0.919_{\pm 0.009}$	$0.541_{\pm 0.016}$	$0.588_{\pm 0.011}$	$0.685_{\pm 0.011}$
meanMLP	$87.7_{\pm 1.8}$	$0.949_{\pm 0.009}$	$0.869_{\pm 0.020}$	$0.672_{\pm 0.031}$	$0.662_{\pm 0.016}$	$0.586_{\pm 0.027}$
Brain-JEPA	$86.8_{\pm 0.6}$	$0.943_{\pm 0.004}$	$0.862_{\pm 0.007}$	$0.688_{\pm 0.017}$	$0.669_{\pm 0.008}$	$0.574_{\pm 0.018}$
TFF ($T = 20$)	$98.3_{\pm 0.4}$	$0.998_{\pm 0.001}$	$0.982_{\pm 0.004}$	$0.440_{\pm 0.029}$	$0.525_{\pm 0.015}$	$0.760_{\pm 0.015}$
SwiFT ($T = 20$)	$97.4_{\pm 0.3}$	$0.998_{\pm 0.001}$	$0.972_{\pm 0.003}$	$0.366_{\pm 0.005}$	$0.480_{\pm 0.007}$	$0.800_{\pm 0.004}$
SwiFT ($T = 50$)	$98.1_{\pm 0.4}$	$0.999_{\pm 0.001}$	$0.980_{\pm 0.005}$	$0.364_{\pm 0.004}$	$0.477_{\pm 0.005}$	$0.802_{\pm 0.003}$
TABLeT ($T = 256$)	$97.6_{\pm 0.2}$	$0.998_{\pm 0.000}$	$0.975_{\pm 0.002}$	$0.340_{\pm 0.011}$	$0.466_{\pm 0.010}$	$0.814_{\pm 0.009}$

Table A5. Experimental results with standard deviation on HCP sex classification and age regression.

Method	HCP					
	Sex			Age		
	ACC	AUC	F1	MSE	MAE	ρ
XGBoost	$82.2_{\pm 2.5}$	$0.890_{\pm 0.028}$	$0.837_{\pm 0.025}$	$0.859_{\pm 0.074}$	$0.769_{\pm 0.033}$	$0.296_{\pm 0.112}$
BrainNetCNN	$86.3_{\pm 4.9}$	$0.937_{\pm 0.027}$	$0.866_{\pm 0.049}$	$0.847_{\pm 0.097}$	$0.749_{\pm 0.040}$	$0.372_{\pm 0.097}$
BNT	$86.3_{\pm 3.0}$	$0.935_{\pm 0.026}$	$0.872_{\pm 0.030}$	$0.794_{\pm 0.051}$	$0.719_{\pm 0.027}$	$0.444_{\pm 0.055}$
meanMLP	$84.5_{\pm 2.5}$	$0.915_{\pm 0.018}$	$0.855_{\pm 0.028}$	$0.846_{\pm 0.056}$	$0.751_{\pm 0.030}$	$0.370_{\pm 0.087}$
Brain-JEPA	$73.9_{\pm 3.2}$	$0.809_{\pm 0.018}$	$0.761_{\pm 0.043}$	$0.814_{\pm 0.037}$	$0.746_{\pm 0.009}$	$0.369_{\pm 0.046}$
TFF ($T = 20$)	$88.1_{\pm 5.0}$	$0.937_{\pm 0.055}$	$0.892_{\pm 0.042}$	$0.888_{\pm 0.062}$	$0.779_{\pm 0.036}$	$0.246_{\pm 0.061}$
SwiFT ($T = 20$)	$93.1_{\pm 0.5}$	$0.978_{\pm 0.008}$	$0.937_{\pm 0.004}$	$0.776_{\pm 0.043}$	$0.719_{\pm 0.015}$	$0.450_{\pm 0.031}$
SwiFT ($T = 50$)	$92.2_{\pm 1.1}$	$0.972_{\pm 0.014}$	$0.929_{\pm 0.010}$	$0.764_{\pm 0.092}$	$0.699_{\pm 0.047}$	$0.460_{\pm 0.071}$
TABLeT ($T = 20$)	$91.6_{\pm 1.5}$	$0.980_{\pm 0.007}$	$0.923_{\pm 0.014}$	$0.784_{\pm 0.120}$	$0.710_{\pm 0.046}$	$0.460_{\pm 0.087}$
TABLeT ($T = 50$)	$93.4_{\pm 1.3}$	$0.986_{\pm 0.004}$	$0.940_{\pm 0.011}$	$0.763_{\pm 0.069}$	$0.704_{\pm 0.031}$	$0.473_{\pm 0.051}$
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)	$92.2_{\pm 1.7}$	$0.973_{\pm 0.010}$	$0.929_{\pm 0.014}$	$0.767_{\pm 0.118}$	$0.693_{\pm 0.043}$	$0.475_{\pm 0.076}$
TABLeT (FT)	$95.3_{\pm 1.3}$	$0.986_{\pm 0.005}$	$0.958_{\pm 0.011}$	$0.650_{\pm 0.045}$	$0.655_{\pm 0.024}$	$0.552_{\pm 0.032}$

E. Detailed Data Description

We provide a detailed description of each dataset used in our study in Tab. A7.

References

[1] Lindsay M Alexander, Jasmine Escalera, Lei Ai, Charissa Andreotti, Karina Febre, Alexander Mangone, Natan Vega-Potler, Nicolas Langer, Alexis Alexander, Meagan Kovacs, et al. An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Sci. Data.*, 4(1):170181, 2017. 3

[2] Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. 1, 3

Table A6. Main experimental results with standard deviation on HCP intelligence regression and ADHD diagnosis.

Method	HCP			ADHD-200		
	Intelligence			Diagnosis		
	MSE	MAE	ρ	ACC	AUC	F1
XGBoost	$0.908_{\pm 0.054}$	$0.779_{\pm 0.023}$	$0.292_{\pm 0.099}$	$62.3_{\pm 2.5}$	$0.650_{\pm 0.036}$	$0.555_{\pm 0.031}$
BrainNetCNN	$0.967_{\pm 0.119}$	$0.788_{\pm 0.044}$	$0.286_{\pm 0.112}$	$59.2_{\pm 10.7}$	$0.640_{\pm 0.095}$	$0.545_{\pm 0.118}$
BNT	$0.920_{\pm 0.092}$	$0.778_{\pm 0.054}$	$0.318_{\pm 0.083}$	$63.6_{\pm 5.4}$	$0.677_{\pm 0.062}$	$0.624_{\pm 0.057}$
meanMLP	$0.887_{\pm 0.076}$	$0.767_{\pm 0.028}$	$0.340_{\pm 0.045}$	$56.8_{\pm 6.8}$	$0.617_{\pm 0.067}$	$0.532_{\pm 0.095}$
Brain-JEPA	$0.959_{\pm 0.091}$	$0.799_{\pm 0.033}$	$0.171_{\pm 0.051}$	–	–	–
TFF ($T = 20$)	$0.898_{\pm 0.022}$	$0.767_{\pm 0.018}$	$0.312_{\pm 0.088}$	$63.3_{\pm 2.3}$	$0.700_{\pm 0.028}$	$0.608_{\pm 0.030}$
SwiFT ($T = 20$)	$0.940_{\pm 0.111}$	$0.782_{\pm 0.044}$	$0.297_{\pm 0.080}$	$63.3_{\pm 3.7}$	$0.693_{\pm 0.030}$	$0.623_{\pm 0.033}$
SwiFT ($T = 50$)	$0.865_{\pm 0.093}$	$0.758_{\pm 0.046}$	$0.354_{\pm 0.070}$	$63.9_{\pm 3.2}$	$0.701_{\pm 0.032}$	$0.627_{\pm 0.030}$
TABLeT ($T = 20$)	$0.866_{\pm 0.074}$	$0.763_{\pm 0.042}$	$0.340_{\pm 0.045}$	$64.4_{\pm 3.0}$	$0.717_{\pm 0.020}$	$0.621_{\pm 0.039}$
TABLeT ($T = 50$)	$0.916_{\pm 0.155}$	$0.778_{\pm 0.063}$	$0.334_{\pm 0.102}$	$64.3_{\pm 2.8}$	$0.710_{\pm 0.021}$	$0.615_{\pm 0.037}$
TABLeT ($T = 256$)	$0.835_{\pm 0.053}$	$0.741_{\pm 0.028}$	$0.392_{\pm 0.062}$	$65.8_{\pm 3.5}$	$0.729_{\pm 0.029}$	$0.630_{\pm 0.038}$
TABLeT (3D DCAE)	$0.869_{\pm 0.077}$	$0.755_{\pm 0.032}$	$0.387_{\pm 0.078}$	$65.8_{\pm 1.7}$	$0.711_{\pm 0.026}$	$0.644_{\pm 0.022}$
TABLeT (FT)	$0.796_{\pm 0.051}$	$0.732_{\pm 0.028}$	$0.435_{\pm 0.046}$	-	-	-

Table A7. Demographic information of the datasets used in our study

Category	UKB	HCP	ADHD -200
Number of subjects	8,178	1,061	533
Sex			
Male, n (%)	4,295 (52.5%)	488 (46.0%)	207 (38.8%)
Female, n (%)	3,883 (47.5%)	573 (54.0%)	325 (61.0%)
N/A, n (%)	–	–	1 (0.2%)
Age (years)	$54.98_{\pm 7.53}$	$28.79_{\pm 3.70}$	$11.94_{\pm 3.40}$
Intelligence	–	$113.32_{\pm 20.50}$	–
Diagnosed, n (%)	–	–	236 (44.3%)

[3] Peter Kim, Junbeom Kwon, Sunghwan Joo, Sangyoon Bae, Donggyu Lee, Yoonho Jung, Shinjae Yoo, Jiook Cha, and Taesup Moon. Swift: Swin 4d fmri transformer. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 42015–42037, 2023. 1