NATURAL IMAGE AUTOENCODER-BASED FMRI REPRESENTATIONS FOR TRAIT AND STATE PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation models pre-trained on large-scale fMRI datasets have shown strong downstream performance, but at substantial data and computation cost. To investigate how much fMRI-specific pre-training is actually needed for such performance, we introduce FReD, which derives fMRI representations from a frozen Deep Compression AutoEncoder (DCAE) pre-trained exclusively on natural images and pairs them with a task specific readout. For trait prediction, FReD summarizes frame-wise representations by their temporal mean and log-standard deviation and applies linear probing, with late fusion across two normalization schemes. For state prediction, it represents each frame as a single token and models temporal dependencies with a shallow Transformer. Across four resting-state datasets spanning six trait-prediction targets, linear probes on frozen DCAE features generally outperform those on fMRI foundation model representations and remain competitive with fully fine-tuned fMRI foundation models. On three task-fMRI state-prediction tasks, a temporal readout on DCAE features performs comparably to the strongest foundation models evaluated. A Gaussian injection analysis further shows that localized signal changes are recovered more accurately from the frozen DCAE features than from the evaluated foundation-model representations. Together, these results show that strong performance on current fMRI benchmarks is possible without fMRI-specific representation pre-training, making frozen natural-image features as a useful baseline for assessing its added value.

1 INTRODUCTION

Deep learning-based approaches to fMRI signal processing have been widely explored. More recently, foundation models pre-trained on large-scale fMRI datasets have demonstrated strong performance across a broad range of downstream tasks (Kim et al., 2023; Dong et al., 2024; Xia et al., 2026; Lane et al., 2026; Wang et al., 2026a;b; Yang et al., 2024). Developing such models, however, typically relies on large-scale neuroimaging datasets that are costly to collect and preprocess, while the pre-training itself may also require substantial computational resources.

One alternative is to reuse representations learned outside the fMRI domain. TABLeT (Kim et al., 2026), for example, showed that Deep Compression AutoEncoder (DCAE) (Chen et al., 2025), pre-trained exclusively on natural images, can be applied to fMRI data and used for modeling long-range temporal dynamics, suggesting that an encoder trained without any fMRI data may already provide a useful starting point. How far such a frozen representation can go, however, remains unclear. TABLeT considers only volumetric fMRI and represents each frame as multiple tokens so that its Transformer (Vaswani et al., 2017) must model spatial and temporal relationships jointly. It also applies the same tokenization and temporal modeling to every downstream task; a single uniform recipe is appealing, but whether it is adequate has not been tested, since trait and state prediction differ in whether the target varies within a subject. Moreover, it is compared primarily with models trained from scratch, leaving open how a frozen off-the-shelf natural-image-based representation fares against fMRI foundation models.

In this work, we therefore ask: *Can competitive fMRI-based prediction be achieved without any fMRI-specific pre-training?* To answer this, we present **FReD**, which builds fMRI representations from a DCAE pre-trained exclusively on natural images and kept frozen throughout, and adapts only the readout to the structure of the downstream task. Our study spans both trait and state pre-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

diction, which differ in whether the target is constant or varies within a subject, respectively. We first consider resting-state fMRI-based trait prediction. Because the target is constant over time, we summarize each fMRI sequence using the temporal mean and log-standard deviation of its frame-level DCAE representations and fit a linear predictor. Despite its simplicity, this approach generally outperforms existing fMRI foundation models under linear probing. Fusing the predictions obtained under two complementary normalization schemes goes further, remaining competitive on most tasks even when those are fully fine-tuned. These results show that strong trait prediction can be achieved without fMRI-specific pre-training or end-to-end encoder adaptation.

We next consider task-fMRI-based state prediction, where the target changes within a subject. Here, the same order-invariant aggregation performs substantially worse, indicating that temporal information is important for state decoding. We therefore represent each frame as a single token and model their temporal dependencies with a shallow Transformer while keeping the DCAE frozen. On cognitive task-state decoding (Van Essen et al., 2013) and object-category decoding (Allen et al., 2022), this simple temporal readout perform comparably to the strongest fMRI foundation models. It also outperforms TABLeT while using a 27-fold shorter attention sequence (one token per frame) and 40% fewer trainable parameters.

We further examine why such a frozen off-the-shelf representation supports downstream tasks. Our analyses show that a substantial fraction of fMRI variation is captured by the principal directions of the DCAE latent space, and that localized perturbations injected into resting-state fMRI are recovered more accurately from the frozen DCAE features than from the evaluated foundation-model representations. Together, these results show that the performance of the fMRI foundation models on current popular benchmarks can be matched by a frozen natural-image autoencoder, paired with a task-appropriate readout. We therefore argue that frozen natural-image representations should serve as a baseline for assessing the added value of fMRI-specific pre-training.

2 PRELIMINARIES

2.1 fMRI DATA TYPES & NORMALIZATION SCHEMES

Existing fMRI foundation models consider three fMRI data types: *volume*, *cortical flat map*, and *parcellation*, which we denote by the superscript $r \in \{V, F, P\}$. A volumetric fMRI sequence is denoted by $\mathbf{X}^V \in \mathbb{R}^{T\times 1\times D\times H\times W}$, where T is the number of temporal frames and D, H, and W are the spatial dimensions of the volume, and a cortical flat-map sequence by $\mathbf{X}^F \in \mathbb{R}^{T\times 1\times H_F\times W_F}$, where H_F and W_F are the spatial dimensions of the flat map. Both retain spatial structure without aggregation over predefined regions, whereas a parcellation $\mathbf{X}^P \in \mathbb{R}^{T\times N_P}$ averages voxel signals within N_P predefined regions of interest. Our FReD operates on $\mathbf{X}^V$ and $\mathbf{X}^F$, since both retain the spatial structure required by a 2D image encoder; we consider $\mathbf{X}^P$ only through the baselines.

For $\mathbf{X}^V$ and $\mathbf{X}^F$, we introduce three normalization schemes. *Global normalization* computes a shared set of statistics over the entire fMRI sequence and applies the same transformation to all temporal frames and spatial coordinates, as commonly adopted by existing fMRI foundation models (Kim et al., 2023; Wang et al., 2026a;b). *Coordinate-wise normalization* independently standardizes the temporal signal at each spatial coordinate, whereas *frame-wise normalization* independently standardizes the spatial values within each temporal frame. In our experiments, we consider two normalization schemes: global normalization and *coordinate–frame normalization*, where the latter sequentially applies coordinate-wise normalization followed by the frame-wise normalization. Throughout this paper, we use normalization subscripts only for $\mathbf{X}^V$ and $\mathbf{X}^F$. We index this choice by $n \in \{g, cf\}$ and write $\mathbf{X}^r_n$ for $r \in \{V, F\}$, where $\mathbf{X}^r_g$ is globally normalized and $\mathbf{X}^r_{cf}$ coordinate–frame-normalized; for $\mathbf{X}^P$ we follow each baseline model's own convention.

2.2 fMRI FOUNDATION MODELS & DOWNSTREAM TASKS

fMRI foundation models. Following the data types in Sec. 2.1, existing fMRI foundation models can be grouped into volume-, cortical flat-map-, and parcellation-based approaches. Volume-based models such as SwiFT (Kim et al., 2023), NeuroSTORM (Wang et al., 2026a), and Omni-fMRI (Wang et al., 2026b) directly process volumetric fMRI. CortexMAE (Lane et al., 2026) instead operates on cortical flat maps. Parcellation-based models, including BrainMASS (Yang et al., 2024),

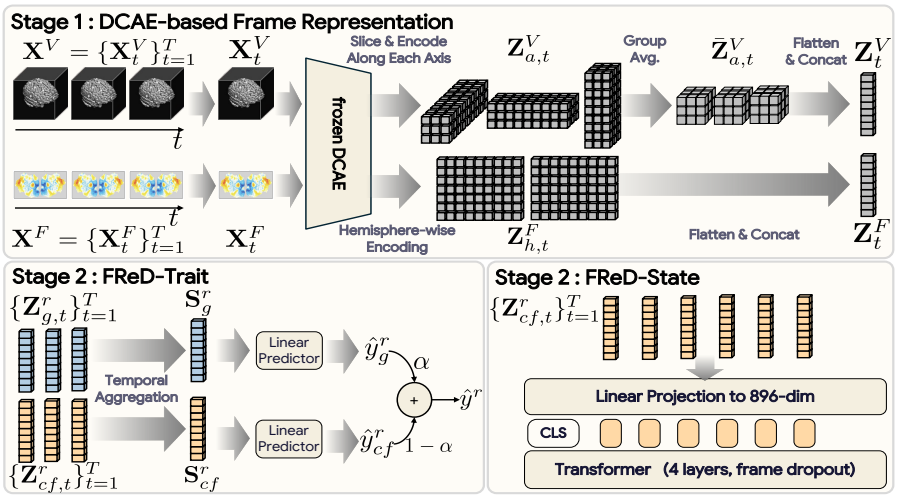


Figure 1: Overview of FReD. A frozen DCAE pre-trained on natural images constructs a frame representation for each fMRI frame, for both $\mathbf{X}^V$ and $\mathbf{X}^F$. For trait prediction, the frame representations are summarized by their temporal mean and log-standard deviation and read out by a linear predictor, and the predictions obtained under the two normalization schemes $n \in \{g, cf\}$ are combined by late fusion. For state prediction, the frame representations obtained under coordinate–frame normalization are projected into a common embedding space and processed by a Transformer.

Brain-JEPA (Dong et al., 2024), BrainHarmonix-F (Dong et al., 2025), and Brain-DiT (Xia et al., 2026), operate on ROI-level representations. These models are generally pre-trained on large-scale neuroimaging datasets with self-supervised objectives.

Downstream tasks. We broadly categorize downstream tasks into *trait prediction* and *state prediction*, following Lane et al. (2026), depending on whether the target varies within a subject. Trait prediction targets subject-level attributes that remain fixed within a subject, such as age or sex. Since the target does not vary over time, such tasks may be addressed without explicitly modeling temporal dependencies, and representations that capture stable differences between subjects are expected to be informative. State prediction, on the contrary, targets cognitive or stimulus-driven states, such as labels for cognitive tasks (Van Essen et al., 2013) or viewed stimuli (Allen et al., 2022). Here the label changes within a subject, so modeling temporal variations is likely to be important.

3 FReD

We present **FReD**, a framework for constructing and modeling **f**MRI **Re**presentations based on a natural-image-pre-trained **D**CAE. FReD operates on two fMRI data types, $\mathbf{X}^V$ and $\mathbf{X}^F$, and consists of two stages, as illustrated in Fig. 1: DCAE-based frame representation and task-dependent readout.

Let $\mathbf{X}^r_n = \{\mathbf{X}^r_{n,t}\}_{t=1}^T, r \in \{V, F\}, n \in \{g, cf\}$, denote an fMRI sequence of representation type r normalized by scheme n. A frozen DCAE encoder first constructs a representation for each temporal frame $\mathbf{X}^r_{n,t}$. FReD then processes the resulting frame sequence according to the downstream task. For trait prediction, it summarizes the temporal distribution of the frame representations and fits a linear predictor. For state prediction, it collapses each frame to a single token before the Transformer, so that spatial mixing happens in the projection and the Transformer models only temporal structure. This contrasts with TABLeT (Kim et al., 2026), which retains 27 tokens per frame and relies on self-attention to model both spatial and temporal relationships.

3.1 DCAE-based Frame Representation

The DCAE (Chen et al., 2025) is a convolutional autoencoder designed to provide highly compressed latent spaces for high-resolution image generation. We use the publicly released `dc-ae-f32c32-in-1.0` checkpoint without any modification, which downsamples an input image by a factor of 32 along each spatial dimension into a 32-channel latent map. The checkpoint is trained only on ImageNet (Deng et al., 2009), and we keep it frozen throughout.

Frame-wise encoding. We independently encode each fMRI frame with the DCAE. The same encoder is shared across data types, normalization schemes, and temporal frames. Since the encoding procedure is identical for both normalization schemes, we omit the subscript n throughout this subsection and write $\mathbf{X}_t^r$ for a single frame.

$\mathbf{X}^V$ case. We follow TABLeT in slicing the volume along the three anatomical axes and encoding each slice independently with the frozen DCAE, and depart from it in how the resulting latents are assembled into a frame representation. A volumetric fMRI frame is represented as $\mathbf{X}_t^V \in \mathbb{R}^{1\times D\times H\times W}$, where $D = H = W = 96$ in our experiments. We slice the volume along the three anatomical axes $a \in \mathcal{A} = \{\mathrm{sag}, \mathrm{cor}, \mathrm{axi}\}$, which yields $N_{\mathrm{sag}} = D$, $N_{\mathrm{cor}} = H$, and $N_{\mathrm{axi}} = W$ two-dimensional slices, respectively. Each axis thus produces 96 slices of size 96×96 in our implementation. Each slice is replicated across three channels and independently encoded by the frozen DCAE, yielding $\mathbf{Z}_{a,t}^V \in \mathbb{R}^{N_a\times C_z\times H_z\times W_z}$, where $C_z \times H_z \times W_z$ denotes the DCAE latent size per slice, which is $32 \times 3 \times 3$.

We then partition the N_a slices along each axis into G contiguous, equally sized groups and average the DCAE latents within each group, which gives $\bar{\mathbf{Z}}_{a,t}^V \in \mathbb{R}^{G\times C_z\times H_z\times W_z}$. Without this step a frame representation would have $3N_aC_zH_zW_z = 82{,}944$ dimensions, which is impractical both for a linear probe and for a linear projection into the Transformer; grouping reduces this to $3GC_zH_zW_z$ while retaining spatial granularity along each axis. Finally, we flatten and concatenate the grouped latents from the three axes to construct the frame representation:

$$\mathbf{Z}_t^V = \mathrm{Concat}_{a\in\mathcal{A}}\left(\mathrm{Flatten}\left(\bar{\mathbf{Z}}_{a,t}^V\right)\right) \in \mathbb{R}^{3GC_zH_zW_z}. \tag{1}$$

TABLeT instead keeps the slice and latent-grid positions as 27 separate tokens per frame; grouping and flattening here reduce this to one, which is the change we ablate in Tab. 5. We use $G = 3$ for trait prediction and $G = 24$ for state prediction, supported by the sensitivity analysis in Fig. 5.

$\mathbf{X}^F$ case. A cortical flat-map frame is represented as $\mathbf{X}_t^F \in \mathbb{R}^{1\times H_F\times W_F}$, where $(H_F, W_F) = (224, 576)$ in our experiments. We split the frame along its width into the left- and right-hemisphere maps $h \in \mathcal{H} = \{\mathrm{lh}, \mathrm{rh}\}$, each of size $H_F \times (W_F/2)$. Each hemisphere map is replicated across three channels and independently encoded by the frozen DCAE, yielding $\mathbf{Z}_{h,t}^F \in \mathbb{R}^{C_z\times H_z\times W_z}$, where each hemisphere map has size $H_F \times (W_F/2) = 224 \times 288$, and the DCAE downsamples each spatial dimension by 32, giving $C_z \times H_z \times W_z = 32 \times 7 \times 9$. We then flatten and concatenate the latents of the two hemispheres to construct the frame representation:

$$\mathbf{Z}_t^F = \mathrm{Concat}_{h\in\mathcal{H}}\left(\mathrm{Flatten}\left(\mathbf{Z}_{h,t}^F\right)\right) \in \mathbb{R}^{2C_zH_zW_z}, \tag{2}$$

which has dimension $4{,}032$ in our implementation. Unlike the volumetric case, there is no slice axis to aggregate, and its dimensionality is moderate compared with that of X^V. The same flat-map representation is therefore used for both trait and state prediction.

3.2 Task-dependent Readout

We hereafter write $\mathbf{Z}_{n,t}^r$ for the frame representation obtained from $\mathbf{X}_{n,t}^r$. Starting from the frame representations $\mathbf{Z}_{n,t}^r$, FReD uses statistical aggregation with linear probing for trait prediction (FReD-Trait) and Transformer-based temporal modeling for state prediction (FReD-State). The two differ in how they model temporal information and which normalization schemes they use.

FReD-Trait. Since trait labels do not vary over time, we compress the temporal dynamics into the element-wise mean and log-standard deviation:

$$\mathbf{S}_n^r = \mathrm{Concat}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{Z}_{n,t}^r,\ \log\left(\mathrm{Std}_t(\mathbf{Z}_{n,t}^r) + \epsilon\right)\right), \tag{3}$$

where Std_t denotes the element-wise standard deviation over the T frames and ϵ is a small constant for numerical stability. A linear predictor is then fitted on $\mathbf{S}_n^r$. As the two normalization schemes are expected to retain complementary information, we apply this pipeline independently to $\mathbf{X}_g^r$ and $\mathbf{X}_{cf}^r$, and combine the resulting predictions $\hat{y}_g^r$ and $\hat{y}_{cf}^r$ by prediction-level late fusion:

$$\hat{y}^r = \alpha\,\hat{y}_g^r + (1-\alpha)\,\hat{y}_{cf}^r, \qquad \alpha \in [0,1], \tag{4}$$

where α is selected based on the validation set, detailed in Sec. B.2.

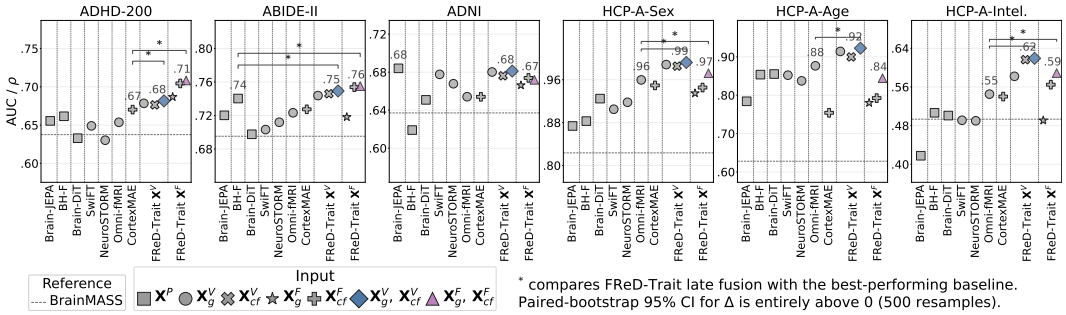


Figure 2: Linear-probing results on trait prediction tasks over 500 random splits, reported as AUC/ρ.

FReD-State. For state prediction, we explicitly model the temporal dependencies among frame representations. Each $\mathbf{Z}^{r}_{cf,t}$ is projected into an 896-dimensional embedding by a learnable linear layer, and we prepend a learnable `[CLS]` token to the resulting sequence. The sequence is processed by a four-layer self-attention Transformer with 1D-RoPE (Su et al., 2024). The backbone follows modern practices, including GQA (Ainslie et al., 2023) and Q/K normalization (Henry et al., 2020) with soft-capped attention logits (Gemma Team et al., 2024). We give the full specification in Tab. 11. The output at the `[CLS]` token is used for prediction. We additionally apply frame dropout during training as regularization. In contrast to FReD-Trait, we use only coordinate–frame-normalized inputs; we examine this choice empirically in Fig. 3.

4 Experimental Results

4.1 Experimental Settings

Datasets and tasks. We evaluate trait prediction on four resting-state fMRI datasets: ADHD classification using ADHD-200 (Bellec et al., 2017), autism spectrum disorder (ASD) classification using ABIDE-II (Di Martino et al., 2017), mild cognitive impairment (MCI) classification using ADNI (Jack Jr et al., 2008), and prediction of sex, age, and intelligence using HCP-A (Bookheimer et al., 2019). For state prediction, we consider three task-fMRI datasets: viewed-movie binary classification on HBN (Alexander et al., 2017), 21-class cognitive task-state decoding on HCP-Task (Van Essen et al., 2013), and decoding of 24 MS-COCO object categories from NSD (Allen et al., 2022). All splits are constructed at the subject level, and we use the fixed splits provided by BrainMarks (Lane et al., 2026) for HCP-Task and NSD. More details are provided in Sec. B.1.

Evaluation protocol. The linear-probing experiments in Fig. 2, Tab. 3, and Fig. 6 use 500 independently sampled 85/15 train–test splits. All other experiments use 70/15/15 train/validation/test splits, 12 splits for the clinical diagnosis tasks, and four splits for HCP-A and HBN; HCP-Task and NSD instead use three random seeds on their fixed splits. We report accuracy (Acc.), AUROC (AUC), and F1 for classification, and mean absolute error (MAE) and Pearson's correlation (ρ) for regression, with MAE expressed in units of the standardized target.

Baselines. We compare FReD against eight fMRI foundation models and TABLeT (Kim et al., 2026): BrainMASS (Yang et al., 2024), Brain-JEPA (Dong et al., 2024), BrainHarmonix-F (BH-F) (Dong et al., 2025), and Brain-DiT (Xia et al., 2026) operate on parcellations; SwiFT (Kim et al., 2023), NeuroSTORM (Wang et al., 2026a), and Omni-fMRI (Wang et al., 2026b) on volumes; and CortexMAE (Lane et al., 2026) on cortical flat maps. Since BrainMASS operates on functional connectivity rather than on the fMRI time series, we report it as a reference line in Figs. 2 and 3 indicating the performance attainable from a basic connectivity-based representation. TABLeT is the most closely related baseline to ours, as it also encodes 2D fMRI slices using a DCAE pre-trained on natural images, but it uses a single data type and modeling scheme across tasks.

4.2 Trait Prediction

Linear probing. We first compare the downstream utility of frozen representations. We standardize the input features and perform five-fold cross-validation, using `LogisticRegressionCV` from scikit-learn (Pedregosa et al., 2011) for classification and `RidgeCV` for regression.

Table 1: Comparison of trait prediction performance. All foundation models are fully fine-tuned, TABLeT is trained from scratch, and FReD-Trait uses linear probing. Results are reported as the mean and standard deviation across data splits. **Bold** and underlined values indicate the best and second-best performance for each metric, respectively.

Method	Input	ADHD-200		ABIDE-II		ADNI		HCP-A					
		Diagnosis		Diagnosis		Diagnosis		Sex		Age		Intelligence	
		AUC ↑	F1 ↑	AUC ↑	F1 ↑	AUC ↑	F1 ↑	AUC ↑	F1 ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑
Brain-JEPA	$\mathbf{X}^P$	$.660_{\pm.00}$	$.606_{\pm.01}$	$.720_{\pm.05}$	$.622_{\pm.04}$	$\underline{.693}_{\pm.07}$	$.597_{\pm.05}$	$.917_{\pm.00}$	$.829_{\pm.00}$	$.418_{\pm.00}$	$.849_{\pm.00}$	$.681_{\pm.01}$	$.487_{\pm.01}$
BH-F	$\mathbf{X}^P$	$.631_{\pm.05}$	$.573_{\pm.06}$	$.697_{\pm.03}$	$.592_{\pm.05}$	$.601_{\pm.06}$	$.452_{\pm.12}$	$.867_{\pm.01}$	$.729_{\pm.10}$	$.435_{\pm.02}$	$.838_{\pm.02}$	$.679_{\pm.04}$	$.498_{\pm.06}$
Brain-DiT	$\mathbf{X}^P$	$.648_{\pm.05}$	$.568_{\pm.04}$	$.712_{\pm.04}$	$.535_{\pm.08}$	$.659_{\pm.05}$	$.484_{\pm.14}$	$.923_{\pm.02}$	$.833_{\pm.03}$	$.386_{\pm.01}$	$.877_{\pm.00}$	$.674_{\pm.01}$	$.546_{\pm.01}$
SwiFT	$\mathbf{X}_g^V$	$.664_{\pm.05}$	$\mathbf{.633}_{\pm.05}$	$.711_{\pm.03}$	$.606_{\pm.04}$	$.688_{\pm.07}$	$.582_{\pm.06}$	$.946_{\pm.01}$	$.853_{\pm.02}$	$.350_{\pm.03}$	$.897_{\pm.02}$	$.678_{\pm.03}$	$.507_{\pm.02}$
NeuroSTORM	$\mathbf{X}_g^V$	$.658_{\pm.04}$	$.572_{\pm.07}$	$.716_{\pm.03}$	$.606_{\pm.02}$	$.673_{\pm.04}$	$\mathbf{.607}_{\pm.02}$	$.974_{\pm.01}$	$.888_{\pm.02}$	$\underline{.343}_{\pm.02}$	$\underline{.906}_{\pm.01}$	$.670_{\pm.04}$	$.525_{\pm.05}$
Omni-fMRI	$\mathbf{X}_g^V$	$.676_{\pm.05}$	$.598_{\pm.05}$	$.712_{\pm.03}$	$.586_{\pm.03}$	$.681_{\pm.06}$	$.584_{\pm.06}$	$\underline{.975}_{\pm.01}$	$\underline{.910}_{\pm.01}$	$.351_{\pm.01}$	$.894_{\pm.01}$	$.657_{\pm.04}$	$.546_{\pm.05}$
CortexMAE	$\mathbf{X}_{cf}^F$	$\underline{.690}_{\pm.05}$	$.605_{\pm.06}$	$.735_{\pm.03}$	$\mathbf{.639}_{\pm.04}$	$.638_{\pm.05}$	$.540_{\pm.06}$	$.944_{\pm.02}$	$.859_{\pm.02}$	$.508_{\pm.02}$	$.774_{\pm.02}$	$.661_{\pm.05}$	$.539_{\pm.06}$
TABLeT	$\mathbf{X}_g^V$	$.639_{\pm.04}$	$.583_{\pm.06}$	$.713_{\pm.03}$	$.624_{\pm.04}$	$\mathbf{.702}_{\pm.05}$	$\underline{.602}_{\pm.05}$	$.963_{\pm.02}$	$.872_{\pm.05}$	$.362_{\pm.02}$	$.891_{\pm.01}$	$.660_{\pm.04}$	$.549_{\pm.03}$
TABLeT	$\mathbf{X}_{cf}^V$	$.624_{\pm.05}$	$.592_{\pm.06}$	$.705_{\pm.03}$	$.588_{\pm.04}$	$.620_{\pm.07}$	$.564_{\pm.06}$	$.946_{\pm.02}$	$.860_{\pm.04}$	$.398_{\pm.01}$	$.871_{\pm.01}$	$.667_{\pm.05}$	$.513_{\pm.04}$
FReD-Trait	$\mathbf{X}_g^V, \mathbf{X}_{cf}^V$	$.686_{\pm.05}$	$.604_{\pm.05}$	$\underline{.743}_{\pm.03}$	$\underline{.629}_{\pm.04}$	$.690_{\pm.04}$	$.591_{\pm.05}$	$\mathbf{.995}_{\pm.00}$	$\mathbf{.956}_{\pm.03}$	$\mathbf{.320}_{\pm.01}$	$\mathbf{.917}_{\pm.01}$	$\mathbf{.633}_{\pm.06}$	$\mathbf{.607}_{\pm.05}$
FReD-Trait	$\mathbf{X}_g^F, \mathbf{X}_{cf}^F$	$\mathbf{.717}_{\pm.03}$	$\underline{.607}_{\pm.05}$	$\mathbf{.747}_{\pm.03}$	$.615_{\pm.05}$	$.679_{\pm.05}$	$.541_{\pm.04}$	$.969_{\pm.00}$	$.884_{\pm.00}$	$.476_{\pm.02}$	$.828_{\pm.01}$	$\underline{.655}_{\pm.03}$	$\underline{.576}_{\pm.04}$

As shown in Fig. 2, FReD-Trait outperforms the evaluated foundation-model baselines on five of the six tasks under linear evaluation, with ADNI being the only case where the methods perform comparably. These results show that frozen DCAE features, summarized by simple temporal statistics under a single normalization scheme, can support strong trait prediction without fMRI-specific pre-training. Late fusion of the predictions obtained under the two normalization schemes (blue diamond and pink triangle) further improves performance generally, indicating that $\mathbf{X}_g^r$ and $\mathbf{X}_{cf}^r$ retain complementary information relevant to trait prediction. The selected fusion weights vary across tasks, suggesting that the preferred normalization scheme is itself task-dependent; we report their distribution in Sec. C.2. One may argue that this comparison is confounded by the dimensionality of the representations, which is larger for FReD-Trait than for the foundation models. To address this, we adopt dimension reduction and find its performance to remain robust, as detailed in Sec. C.1.

Can fine-tuning yield better performance than FReD-Trait? We next investigate whether fine-tuning existing fMRI foundation models on individual downstream tasks improves their performance. For each foundation model, we follow its original fine-tuning protocol and search at least eight learning rates per dataset, with the full search ranges reported in Sec. B. In addition, we evaluate partial fine-tuning, in which only the final few layers of the pre-trained encoder are updated.

The full fine-tuning results are reported in Tab. 1, and the partial fine-tuning results in Sec. C.3. Despite adapting the entire pre-trained encoder to each downstream task, the fully fine-tuned foundation models remain broadly competitive with FReD-Trait rather than consistently surpassing it. Moreover, we additionally observe that fine-tuning itself does not yield consistent improvements over linear probing on the matched splits, as detailed in Sec. C.4. These results show that a frozen off-the-shelf representation paired with simple temporal aggregation can provide a competitive alternative to task-specific adaptation of fMRI foundation models.

4.3 State Prediction

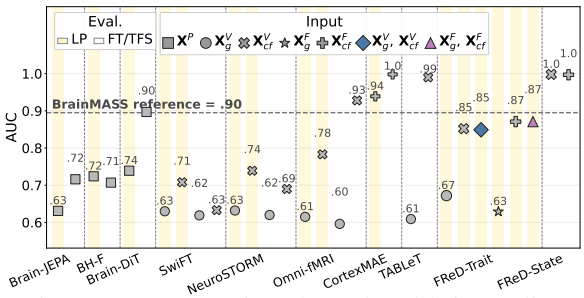
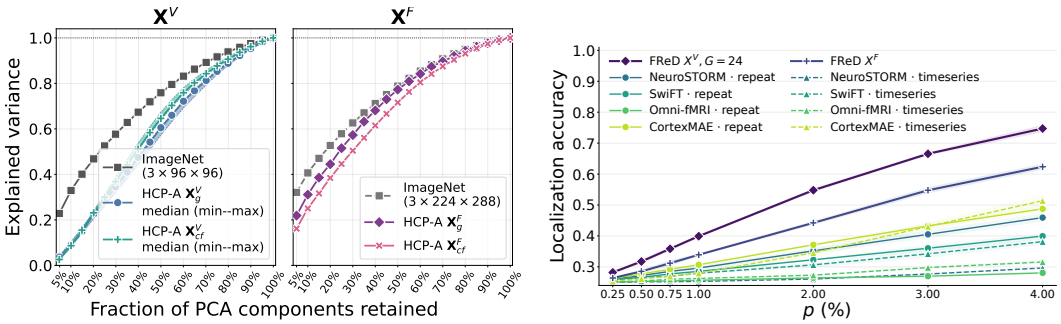


Figure 3: HBN-Movie task results with four splits.

Table 2: Results on state prediction tasks. **Bold** and underlined values indicate the best and second-best performance, respectively. FT: Full fine-tuning, TFS: Train from scratch.

Method	Input	Eval	HBN-Movie		HCP-Task	NSD
			AUC ↑	F1 ↑	Acc. ↑	Acc. ↑
Omni-fMRI	$\mathbf{X}_{cf}^V$	FT	$.928_{\pm.03}$	$.838_{\pm.02}$	$94.5_{\pm0.2}$	$21.5_{\pm0.9}$
CortexMAE	$\mathbf{X}_{cf}^F$	FT	$\underline{.997}_{\pm.00}$	$.959_{\pm.01}$	$\mathbf{99.2}_{\pm0.1}$	$\underline{29.8}_{\pm0.5}$
TABLeT	$\mathbf{X}_{cf}^V$	TFS	$.990_{\pm.01}$	$.948_{\pm.03}$	$94.2_{\pm0.2}$	$23.9_{\pm1.0}$
FReD-State	$\mathbf{X}_{cf}^V$	TFS	$\mathbf{.998}_{\pm.00}$	$\mathbf{.973}_{\pm.01}$	$97.4_{\pm0.1}$	$29.5_{\pm0.2}$
FReD-State	$\mathbf{X}_{cf}^F$	TFS	$\underline{.997}_{\pm.00}$	$\underline{.971}_{\pm.01}$	$\underline{98.8}_{\pm0.2}$	$\mathbf{30.0}_{\pm0.6}$

Insights from the HBN-Movie task. We first consider a relatively simple state prediction task, where the model predicts which of two movies a participant was watching from the corresponding fMRI sequence. As shown in Fig. 3, the volume-based foundation models, SwiFT, NeuroSTORM, and Omni-fMRI, perform only slightly above chance on $\mathbf{X}_g^V$ under both linear probing and full



(a) (b)

Figure 4: **(a)** Fraction of variance explained by the leading ImageNet principal directions, as a function of the fraction of components retained, shown separately for $\mathbf{X}^V$ (left) and $\mathbf{X}^F$ (right). Each panel includes an ImageNet reference computed on held-out ImageNet latents at the corresponding DCAE input resolution; a curve close to the ImageNet reference indicates a similar concentration of variance along the leading ImageNet principal directions. For $\mathbf{X}^V$, the curves are medians over the three anatomical slicing axes and the bands span their minimum and maximum. **(b)** Four-way localization accuracy as a function of the injected percent signal change, averaged over the three spatial scales σ; chance accuracy is 25%.

fine-tuning, whereas linear probing on $\mathbf{X}^V_{cf}$ consistently yields better performance. We further find that SwiFT and NeuroSTORM do not substantially improve even when fully fine-tuned on $\mathbf{X}^V_{cf}$, and that among the volume-based models only Omni-fMRI benefits from doing so. The effect is even more pronounced for the DCAE-based models, TABLeT and FReD-Trait, whose performance improves substantially on coordinate-frame-normalized inputs. Coordinate-frame normalization is thus critical for state prediction, which motivates its exclusive use in FReD-State.

Although FReD-Trait performs strongly on trait prediction, it underperforms TABLeT on HBN-Movie. We attribute this to the fact that FReD-Trait compresses an entire fMRI sequence into its temporal mean and log-standard deviation and thus discards the temporal ordering of the frame representations, which may help distinguishing stimulus-induced states in this task. Statistical temporal aggregation with linear probing is therefore insufficient to capture within-subject variability, which motivates the temporal modeling in FReD-State. By explicitly modeling the temporal interactions among frame representations, FReD-State outperforms FReD-Trait by a large margin and becomes competitive with the strongest foundation models, indicating that DCAE latents retain sufficient information for state prediction once paired with an appropriate temporal architecture.

More challenging tasks. We next consider two harder state prediction tasks, 21-class cognitive task-state decoding on HCP-Task and 24-class object-category decoding on NSD. For these evaluations, we include Omni-fMRI and CortexMAE that clearly outperformed BrainMASS (Yang et al., 2024) on HBN-Movie, along with TABLeT, the most closely related DCAE-based baseline. As shown in Tab. 2, FReD-State outperforms Omni-fMRI and TABLeT by clear margins on both tasks, while performing comparably to CortexMAE. Between the two data types, $\mathbf{X}^F$ performs better on HCP-Task and comparably on NSD. With $\mathbf{X}^F$, FReD-State matches the strongest flat-map baseline.

5 Analyses

5.1 Does the DCAE Latent Space Accommodate fMRI Variation?

We first examine how much fMRI variation is captured by the principal directions of DCAE representations learned from natural images. We fit a PCA basis on DCAE latents from ImageNet test images and measure the fraction of fMRI variance explained by the leading k ImageNet principal directions, $\|\mathbf{H}\mathbf{V}_{1:k}\|_F^2/\|\mathbf{H}\|_F^2$, where $\mathbf{V}_{1:k}$ contains the leading k ImageNet principal directions and $\mathbf{H}$ denotes the standardized latents of the domain being evaluated, centered by its own mean. Since the DCAE latent size depends on the input resolution, we compare each fMRI representation against an ImageNet reference computed at the matching resolution: $3 \times 96 \times 96$ for volumetric slices and $3 \times 224 \times 288$ for hemisphere maps. The detailed setup can be found in Sec. B.3.

As shown in Fig. 4(a), every fMRI representation falls below its ImageNet reference when only a small fraction of principal directions is retained, indicating that the directions explaining the largest variation in natural images account for a smaller fraction of fMRI variation. The gap gradually narrows as more components are included, and by 60% of the basis the retained ImageNet directions explain roughly 0.72–0.85 of the variance across the fMRI representations. Flat-map representations remain closer to their corresponding ImageNet reference than volumetric slices. Overall, these results show that a substantial fraction of fMRI variation is recoverable from the ImageNet-derived basis, which supports using DCAE pre-trained with natural images to obtain fMRI representations.

5.2 Are Localized Signal Changes Accessible from DCAE Representations?

Downstream tasks may depend on small, spatially localized changes in the fMRI signal, and whether such changes survive a $32\times$ spatial compression is not obvious. We therefore inject controlled Gaussian perturbations of varying amplitude into resting-state HCP-A fMRI and evaluate whether their locations can be recovered from each frozen representation using a linear probe. At one of four fixed cortical centers, we add an isotropic Gaussian profile of width σ, restricted to the cortical mask, modulated by a canonical double-gamma hemodynamic response and scaled so that the added signal at the center equals a requested percent signal change p of the local baseline. We sweep $p \in [0.25, 4]\%$ and $\sigma \in \{2, 3, 4\}$. After injection we match the global statistics of each sequence to those of its clean counterpart, so that no probe can exploit a shift in intensity or variance. A four-way linear probe is then fitted on the frozen representations over four different splits and asked which center was perturbed. Each representation is given the frame at which the response peaks: FReD encodes it directly, while the foundation models, which require a temporal input, receive the same volume repeated across their window (`repeat`) or, as an additional condition, the full injected sequence (`timeseries`). The DCAE features are reduced to 768 dimensions by PCA, matching the largest foundation-model feature dimension. Full details are given in Sec. B.4.

As shown in Fig. 4(b), all representations remain near chance at the weakest perturbation and improve monotonically as the injected signal becomes stronger. Both FReD variants stay above every foundation-model representation at every amplitude, with $\mathbf{X}^V$ the strongest throughout and $\mathbf{X}^F$ second; CortexMAE is the closest baseline, and neither input condition consistently outperforms the other across the foundation models. Information about localized signal changes therefore remains linearly accessible from the frozen DCAE features, even though the encoder has never been trained on fMRI. However, we note that this spatial-localization probe only measures the accessibility of perturbation, not the general representation quality.

5.3 Task-dependent Normalization and Readout

Normalization. Since HBN is a task-fMRI dataset, one may ask whether the sensitivity to normalization observed in Sec. 4.3 stems from the dataset itself rather than from the type of prediction task. We therefore add a trait prediction task on the same dataset, sex prediction of each subject, which let us compare the two task types under an identical data distribution. As shown in Tab. 3, HBN-Movie is far more sensitive to the choice of normalization than HBN-Sex; we assert the sensitivity therefore follows the prediction task rather than the dataset.

Table 3: Linear probing results over 500 random splits on HBN-Movie/Sex.

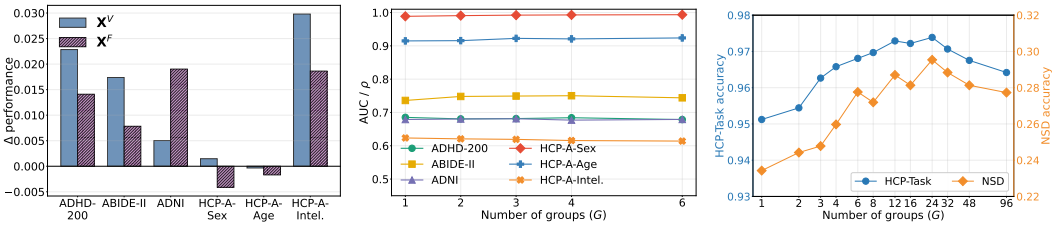
Method	Input	HBN-Movie		HBN-Sex	
		AUC ↑	F1 ↑	AUC ↑	F1 ↑
FReD-Trait	$\mathbf{X}_g^V$	$.667_{\pm.03}$	$.614_{\pm.04}$	$.882_{\pm.04}$	$.741_{\pm.06}$
FReD-Trait	$\mathbf{X}_{cf}^V$	$.863_{\pm.03}$	$.787_{\pm.03}$	$.822_{\pm.04}$	$.657_{\pm.06}$

Readout. We showed that FReD-Trait is ineffective for state prediction, which motivated the temporal modeling in FReD-State. This raises the converse question of whether FReD-State is better for trait prediction. As shown in Tab. 4, FReD-State does not improve over FReD-Trait for trait prediction tasks. Explicit temporal modeling therefore appears to matter specifically where temporal dynamics are relevant to the target, whereas simple temporal aggregation suffices for trait prediction.

Table 4: FReD-State performance on trait prediction tasks. **Bold** values indicate the best performance. LP: Linear probing, TFS: Train from scratch.

Method	Input	Eval	ADHD-200		HCP-A	
			Diagnosis		Intelligence	
			AUC ↑	F1 ↑	MAE ↓	ρ ↑
FReD-Trait	$\mathbf{X}_g^V, \mathbf{X}_{cf}^V$	LP	$.686_{\pm.05}$	$.604_{\pm.05}$	$\mathbf{.633}_{\pm.06}$	$\mathbf{.607}_{\pm.05}$
FReD-Trait	$\mathbf{X}_g^F, \mathbf{X}_{cf}^F$	LP	$\mathbf{.717}_{\pm.03}$	$.607_{\pm.05}$	$.655_{\pm.03}$	$.576_{\pm.04}$
FReD-State	$\mathbf{X}_g^V$	TFS	$.663_{\pm.05}$	$.584_{\pm.07}$	$.652_{\pm.03}$	$.564_{\pm.03}$
FReD-State	$\mathbf{X}_{cf}^V$	TFS	$.637_{\pm.05}$	$.570_{\pm.06}$	$.664_{\pm.06}$	$.556_{\pm.05}$
FReD-State	$\mathbf{X}_g^F$	TFS	$.669_{\pm.03}$	$.609_{\pm.05}$	$.689_{\pm.04}$	$.474_{\pm.08}$
FReD-State	$\mathbf{X}_{cf}^F$	TFS	$.687_{\pm.05}$	$\mathbf{.614}_{\pm.09}$	$.688_{\pm.01}$	$.505_{\pm.03}$

5.4 Component-wise Ablations



(a) (b) (c)

Figure 5: **(a)** Effect of the log-standard deviation feature on trait prediction, measured as the AUC or ρ difference against replacing it with a zero vector. **(b)** Trait prediction performance as a function of the number of groups G. **(c)** State prediction accuracy as a function of G.

Log-standard deviation features. We ablate the log-standard deviation feature by replacing it with a zero vector, which keeps the input dimensionality and parameter count unchanged. As shown in Fig. 5(a), the log-standard deviation feature generally improves performance, although the magnitude of the effect varies across tasks.

Number of groups G. We next vary the number of groups G, which controls how much spatial granularity along each slicing axis is retained in the frame representation. Trait prediction is largely insensitive to G and remains stable even with a single group per axis (Fig. 5(b)). State prediction, in contrast, degrades substantially when G is too small, but is stable across a moderate range (Fig. 5(c)).

From TABLeT to FReD-State. Finally, we identify where the improvement of FReD-State over TABLeT comes from. Since TABLeT operates on volumetric fMRI, we conduct this ablation with $\mathbf{X}^V$. The two models differ in their tokenization and in architectural and regularization choices. To isolate the former, we replace TABLeT's tokenization with ours, constructing a single token per frame from group-averaged DCAE latents, while keeping its temporal Transformer. As reported in Tab. 5, this intermediate model already improves performance over the original TABLeT. Moving from this intermediate model to FReD-State yields further gains while reducing the parameter count. Because this second step introduces several architectural and regularization changes simultaneously, we attribute the additional gain to these remaining changes jointly rather than to any single component, detailed in Sec. D.

Table 5: Comparison between TABLeT and FReD-State, trained from scratch.

Method	Input	Param.	HCP-Task	NSD
			ACC	ACC
TABLeT	$\mathbf{X}^V_f$	129.5M	$94.2_{\pm 0.2}$	$23.9_{\pm 1.0}$
TABLeT with $\bar{\mathbf{Z}}^V_{a,t}$	$\mathbf{X}^V_{cf}$	145.4M	$94.7_{\pm 0.3}$	$27.5_{\pm 0.4}$
FReD-State	$\mathbf{X}^V_{cf}$	78.2M	$\mathbf{97.4}_{\pm 0.1}$	$\mathbf{29.5}_{\pm 0.2}$

6 Conclusion and Limitations

We asked how much fMRI-specific representation pre-training is necessary for strong downstream prediction. We studied this question using a frozen DCAE pre-trained exclusively on natural images with task-matched readouts. For trait prediction, temporal mean and log-standard deviation followed by a linear predictor achieve performance competitive with fully fine-tuned fMRI foundation models. For state prediction, a shallow Transformer over one token per frame performs comparably to the strongest foundation models, while using a smaller readout and a shorter attention sequence than TABLeT. These results show that strong performance on current benchmarks can be achieved without fMRI-specific representation pre-training, and establish frozen natural-image representations as a baseline against which the benefits of large-scale fMRI pre-training should be measured.

However, several questions remain open. The Gaussian injection analysis is synthetic and spatially localized, so it speaks to linear accessibility rather than to downstream ranking. Relatedly, we do not resolve when each data type should be preferred, although we observe that $\mathbf{X}^F$ tends to be stronger on state prediction whereas $\mathbf{X}^V$ recovers injected perturbations far more accurately and leads on the HCP-A phenotype tasks. We also treat the encoder as fixed throughout, and do not explore whether adapting it to fMRI could further improve performance. Despite this, our results show that a frozen natural-image autoencoder can produce strong representations for fMRI, and suggest that the benefit of fMRI-specific pre-training should be demonstrated against such baselines rather than assumed.

AI USE STATEMENT

In this work, we used generative AI tools for providing feedback on research methodology or experiments, implementing methods, and assisting with translation. We have not used generative AI tools for proposing or refining hypotheses, supporting qualitative or thematic data analysis, and interpreting results. Generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and cleaning and reformat dataset are not applicable to this work.

Additionally, we used generative AI tools for modifying scientific figures or images, creating or editing software code, drafting parts of a research paper, editing a research paper to improve readability. All AI-assisted outputs were carefully reviewed by the authors. In particular, the authors created the initial versions of all scientific figures and images themselves and used LLM-based tools only to improve their visual clarity and presentation. In addition, the authors independently verified LLM-generated code and tested it for correctness. The authors also independently reviewed all LLM-generated drafts, using them only as starting points before revising and finalizing the text themselves. and write a final script by own. We take full responsibility for the final content of this work, including all text, claims, code, figures, and other artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work relies exclusively on publicly available, open-source fMRI datasets. We did not collect any new human-subject data or interact directly with study participants. Therefore, no additional ethical approval was required for this study.

REPRODUCIBILITY STATEMENT

We provide detailed experimental settings in Sec. B. We plan to release our code, preprocessing scripts, and data splits after the review process, subject to the licenses and access policies of the original datasets.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit 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 (EMNLP)*, pp. 4895–4901, 2023.

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.

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. Neurosci.*, 25(1):116–126, 2022.

Ulzee An, Moonseong Jeong, Simon Austin Lee, Aditya Gorla, Yuzhe Yang, and Sriram Sankararaman. Raptor: Scalable train-free embeddings for 3D medical volumes leveraging pretrained 2D foundation models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, pp. 1462–1482, 2025.

Pierre Bellec, Carlton Chu, Francois Chouinard-Decorte, Yassine Benhajali, Daniel S Margulies, and R Cameron Craddock. The neuro bureau adhd-200 preprocessed repository. *Neuroimage*, 144:275–286, 2017.

Susan Y Bookheimer, David H Salat, Melissa Terpstra, Beau M Ances, Deanna M Barch, Randy L Buckner, Gregory C Burgess, Sandra W Curtiss, Mirella Diaz-Santos, Jennifer Stine Elam, et al. The lifespan human connectome project in aging: an overview. *Neuroimage*, 185:335–348, 2019.

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 *International Conference on Learning Representations (ICLR)*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Adriana Di Martino, David O'connor, Bosi Chen, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Joshua H Balsters, Leslie Baxter, Anita Beggiato, Sylvie Bernaerts, et al. Enhancing studies of the connectome in autism using the autism brain imaging data exchange ii. *Sci. Data.*, 4(1):170010, 2017.

Zijian Dong, Ruilin Li, Yilei Wu, Thuan T Nguyen, Joanna S Chong, Fang Ji, Nathanael R Tong, Christopher L Chen, and Juan H Zhou. Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 86048–86073, 2024.

Zijian Dong, Ruilin Li, Joanna Chong, Niousha Dehestani, Yinghui Teng, Yi Lin, Zhizhou Li, Yichi Zhang, Yapei Xie, Leon Ooi, et al. Brain harmony: a multimodal foundation model unifying morphology and function into 1d tokens. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 122100–122127, 2025.

Oscar Esteban, Christopher J Markiewicz, Ross W Blair, Craig A Moodie, A Ilkay Isik, Asier Erramuzpe, James D Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, et al. fmriprep: a robust preprocessing pipeline for functional mri. *Nat. Methods.*, 16(1):111–116, 2019.

Oscar Esteban, Rastko Ciric, Karolina Finc, Ross W Blair, Christopher J Markiewicz, Craig A Moodie, James D Kent, Mathias Goncalves, Elizabeth DuPre, Daniel EP Gomez, et al. Analysis of task-based functional mri data preprocessed with fmriprep. *Nat. Protoc.*, 15(7):2186–2202, 2020.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253, 2020.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European Conference on Computer Vision (ECCV)*, pp. 646–661, 2016.

Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *J. Magn. Reson. Imaging.*, 27(4):685–691, 2008.

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 *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 42015–42037, 2023.

Peter Yongho Kim, Juhyeon Park, Jungwoo Park, Jubin Choi, Jungwoo Seo, Jiook Cha, and Taesup Moon. Can natural image autoencoders compactly tokenize fmri volumes for long-range dynamics modeling? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35321–35330, 2026.

Connor Lane, Mihir Tripathy, Leema Krishna Murali, Ratna Sagari Grandhi, Shamus Sim Zi Yang, Sam Gijsen, Debojyoti Das, Manish Ram, Utkarsh Kumar Singh, Cesar Kadir Torrico Villanueva, YUXIANG WEI, Will Beddow, Gianfranco Cortes, Suin Cho, Daniel Z Kaplan, Benjamin Warner, Tanishq Mathew Abraham, and Paul Steven Scotti. Scaling vision transformers for functional MRI with flat maps. In *Forty-third International Conference on Machine Learning (ICML)*, 2026.

Han Liu, Bogdan Georgescu, Yanbo Zhang, Youngjin Yoo, Michael Baumgartner, Riqiang Gao, Jianing Wang, Gengyan Zhao, Eli Gibson, Dorin Comaniciu, and Sasa Grbic. Revisiting 2d foundation models for scalable 3d medical image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30021–30031, 2026.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, NY, second edition, 2006. ISBN 978-0-387-30303-1. doi: 10.1007/978-0-387-40065-5.

Juhyeon Park, Peter Yongho Kim, Jungwoo Park, Jubin Choi, Jungwoo Seo, Jiook Cha, and Taesup Moon. Processing fMRI brain signals using latents from natural image autoencoders. In *NeurIPS 2025 Workshop on Foundation Models for the Brain and Body*, 2025.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems (NeurIPS)*, 2017.

Cheng Wang, Yu Jiang, Zhihao Peng, Chenxin Li, Chang-bae Bang, Lin Zhao, Wanyi Fu, Jinglei Lv, Jorge Sepulcre, Carl Yang, et al. Towards a general-purpose foundation model for functional mri analysis. *Nat. Biomed. Eng.*, pp. 1–12, 2026a.

Mo Wang, Wenhao Ye, Junfeng Xia, Junxiang Zhang, Xuanye Pan, Minghao Xu, Haotian Deng, Hongkai Wen, and Quanying Liu. Omni-fMRI: A universal atlas-free fMRI foundation model. In *Forty-third International Conference on Machine Learning (ICML)*, 2026b.

Junfeng Xia, Wenhao Ye, Xuanye Pan, Xinke Shen, Mo Wang, and Quanying Liu. Brain-dit: A universal multi-state fmri foundation model with metadata-conditioned pretraining. *arXiv preprint arXiv:2604.12683*, 2026.

Yanwu Yang, Chenfei Ye, Guinan Su, Ziyao Zhang, Zhikai Chang, Hairui Chen, Piu Chan, Yue Yu, and Ting Ma. Brainmass: Advancing brain network analysis for diagnosis with large-scale self-supervised learning. *IEEE Trans. Med. Imaging.*, 43(11):4004–4016, 2024.

Natural Image Autoencoder-Based fMRI Representations for Trait and State Prediction

Appendix

A Related Work

A.1 Utilizing Natural Image Prior for Medical Imaging

The limited availability and diversity of medical imaging data have long posed challenges for learning generalizable representations directly from the medical domain. In contrast, foundation models pre-trained on large-scale natural image datasets benefit from substantially richer and more diverse visual supervision, yielding robust representations. Motivated by this discrepancy, recent studies have investigated transferring such pre-trained representations to medical imaging, rather than training modality-specific foundation models solely on medical data.

Natural-image priors have proven effective even for volumetric medical data despite considerable domain and dimensionality gaps. Raptor (An et al., 2025) leverages a frozen 2D foundation model pre-trained on natural images to extract representations from orthogonal slices of 3D medical volumes and compresses the resulting features through random projections. This training-free approach demonstrates that representations learned from natural images can effectively support diverse downstream tasks on 3D medical data. AnyMC3D (Liu et al., 2026) further extends this direction to 3D medical image classification, adapting a pre-trained 2D foundation model using lightweight task-specific LoRA modules and attention-based slice aggregation. Notably, it shows that properly adapted general-purpose models can match or even outperform medical-specific foundation models across diverse 3D classification tasks. More recently, TABLeT (Kim et al., 2026; Park et al., 2025) employs a pre-trained 2D natural-image autoencoder as an off-the-shelf tokenizer for fMRI volumes, transforming high-dimensional 3D brain images into compact tokens for modeling long-range temporal dynamics. Together, these studies highlight the potential of transferring rich visual priors learned from natural images to medical imaging, offering an alternative to costly domain-specific pre-training under limited medical data.

B Implementation Details

B.1 Datasets

Table 6: Summary of the datasets used for trait prediction (top) and state prediction (bottom). The numbers of subjects and samples represent the totals after preprocessing. Duration indicates the temporal duration of each sample. Trait prediction datasets contain one sample per subject, whereas state prediction datasets contain multiple samples per subject. Dx denotes diagnosis classification; Task21 denotes 21-class cognitive task-state decoding; and COCO24 denotes 24-class object-category decoding. The majority-class proportion is omitted for regression tasks.

Dataset	Target	Subjects	Samples	Duration	# Classes	Majority (%)
ADHD-200	ADHD Dx	430	430	300 s	2	56.3
ABIDE-II	ASD Dx	616	616	300 s	2	56.3
ADNI	MCI Dx	464	464	300 s	2	59.5
HCP-A	Sex	1,069	1,069	382.4 s	2	57.1
HCP-A	Age, Intelligence	1,069	1,069	382.4 s	–	–
HBN-Movie	Movie	428	856	200 s	2	50.0
HCP-Task	Task21	604	26,995	16 s	21	16.9
NSD	COCO24	8	47,589	16 s	24	6.8

The detailed description of each dataset is provided in Tab. 6.

Preprocessing. For ADHD-200 (Bellec et al., 2017), ABIDE-II (Di Martino et al., 2017), ADNI (Jack Jr et al., 2008), and HBN (Alexander et al., 2017), we perform minimal preprocessing using fMRIPrep (Esteban et al., 2019; 2020). For ADHD-200, ABIDE-II, and ADNI, we additionally regress out nuisance signals using cosine basis functions, six rigid-body motion parameters, and anatomical CompCor components. For HCP-A, we use the minimally processed data provided by Bookheimer et al. (2019), and for HCP-Task and NSD, the preprocessed data provided by BrainMarks (Lane et al., 2026). We then resample all fMRI data to match the TR each foundation model was pre-trained with.

Data splits. For ADHD-200, ABIDE-II, and ADNI, we perform stratified sampling to preserve the distributions of acquisition sites and diagnostic labels across the splits, and for HCP-A we preserve the distributions of sex, age, and intelligence. HBN is split randomly, whereas for HCP-Task and NSD we use the fixed splits provided by BrainMarks (Lane et al., 2026). All splits are constructed at the subject level, so that no subject appears in more than one split. For the linear-probing experiments in Fig. 2, Tab. 3, and Fig. 6, we evaluate each method over 500 independently sampled train–test splits, assigning 85% of the data to training and the remaining 15% to testing. All other experiments use a 70/15/15 train/validation/test ratio, repeated over 12 splits for the clinical diagnosis tasks and four splits for HCP-A and HBN-Movie; HCP-Task and NSD instead use three random seeds on their fixed splits.

B.2 Training and Evaluation

DCAE. We use the unmodified `dc-ae-f32c32-in-1.0` checkpoint provided by Chen et al. (2025) in all experiments, and keep it frozen throughout.

Windowing. For the $\mathbf{X}^V$- and $\mathbf{X}^F$-based models, training uses randomly sampled consecutive frames. At evaluation time, following Kim et al. (2023; 2026), we average the model outputs (logits) over all windows starting from the first frame. For linear probing, we follow the protocol of BrainMarks (Lane et al., 2026), which uses the window-averaged embedding. If inputs are shorter than a model's expected temporal sequence length, we pad the input with the per-coordinate mean, following Lane et al. (2026).

Late fusion. We select α without using the test set at any stage. For experiments with an 85/15 train/test split, α is selected using out-of-fold predictions obtained from five-fold cross-validation on the training set. For experiments with a 70/15/15 train/validation/test split, it is selected on the validation set.

Because the search is one-dimensional, we first evaluate the selection criterion on a uniform grid over $\alpha \in [0, 1]$ with a step size of 0.1. For smooth criteria, such as mean absolute error and weighted cross-entropy, the best-performing grid point is used to initialize an SLSQP (Nocedal & Wright, 2006) refinement, yielding a continuous value of α. For rank-based criteria such as AUROC, we use the best grid point directly.

Shared settings. For neural-network training, unless stated otherwise, all experiments use the following settings.

- `Target normalization`: For regression tasks, we z-normalize the target values using the mean and standard deviation computed from the training split.
- `Optimizer`: AdamW with a cosine decay learning rate schedule and weight decay 10^{-2}. We use `fp16` mixed precision, `BCEWithLogitsLoss` for binary classification, and `L1Loss` for regression. For ADHD-200, ABIDE-II, and ADNI, we enable the `pos-weight` option to account for class imbalance. For HCP-Task and NSD, we use `CrossEntropyLoss`.
- `Hyperparameter search`: We select hyperparameters per model on the validation set, using validation loss for ADHD-200, ABIDE-II, and ADNI (so that `pos-weight` is reflected), AUROC for HCP-A-Sex and HBN-Movie, accuracy for HCP-Task and NSD, and MAE for HCP-A-Age and HCP-A-Intelligence. For every model and dataset, we search at least eight learning rates on a log-uniform grid within the range listed below.

- `Early stopping`: We report the early-stopped model for all methods except SwiFT, for which we report the final epoch, as we observed its training to be stable.

Brain-JEPA

- `Learning rate`: Selected from $[1 \times 10^{-6}, 3 \times 10^{-4}]$.
- `Batch size`: 16.
- `Epochs`: 50.

BrainHarmonix-F

- `Learning rate`: Selected from $[3 \times 10^{-8}, 5 \times 10^{-3}]$.
- `Batch size`: 8.
- `Epochs`: 50.

Brain-DiT

- `Learning rate`: Selected from $[1 \times 10^{-6}, 3 \times 10^{-4}]$.
- `Batch size`: 2.
- `Epochs`: 20.

SwiFT

- `Learning rate`: Selected from $[3 \times 10^{-8}, 9 \times 10^{-5}]$.
- `Batch size`: 8.
- `Epochs`: 30.
- `Window size`: 20.
- `TR`: 0.72s

NeuroSTORM

- `Learning rate`: Selected from $[3 \times 10^{-8}, 9 \times 10^{-5}]$.
- `Batch size`: 8.
- `Epochs`: 30.
- `Window size`: 20.
- `TR`: 0.72s

Omni-fMRI

- `Learning rate`: Selected from $[3 \times 10^{-8}, 9 \times 10^{-5}]$ for trait prediction, $[3 \times 10^{-8}, 5 \times 10^{-4}]$ for HBN-Movie, and $[3 \times 10^{-7}, 5 \times 10^{-4}]$ for HCP-Task and NSD. The learning rate of the prediction head is set to $10\times$ the base learning rate.
- `Batch size`: 8 for trait prediction and HBN-Movie, and 16 for HCP-Task and NSD.
- `Epochs`: 30.
- `Window size`: 40.
- `TR`: 0.8s

CortexMAE

- `Learning rate`: Selected from $[3 \times 10^{-8}, 9 \times 10^{-5}]$ for trait prediction, $[3 \times 10^{-8}, 9 \times 10^{-6}]$ for HBN-Movie, and $[3 \times 10^{-8}, 9 \times 10^{-5}]$ for HCP-Task and NSD.
- `Batch size`: 8 for trait prediction and HBN-Movie, and 64 for HCP-Task and NSD.
- `Epochs`: 30.
- `Window size`: 16.
- `TR`: 1s.

TABLeT

- `Learning rate`: Selected from $[5 \times 10^{-9}, 5 \times 10^{-6}]$ for trait prediction and HBN-Movie, and $[7 \times 10^{-6}, 2 \times 10^{-4}]$ for HCP-Task and NSD.
- `Batch size`: 4 for trait prediction and HBN-Movie, and 16 for HCP-Task and NSD.

- `Epochs`: 30 for trait prediction and HBN-Movie, 50 for HCP-Task, and 15 for NSD.
- `Window size`: We use full frames.
- `TR`: 0.8s for trait prediction and HBN, 1s for HCP-Task and NSD.

FReD-Trait

- `Hyperparameters`: We use scikit-learn's default search space.
- `TR` : 0.8s for trait prediction and HBN.

FReD-State

- `Learning rate`: Selected from $[1\times10^{-6},\ 1\times10^{-4}]$ for HBN-Movie, and $[7\times10^{-6},\ 2\times10^{-4}]$ for HCP-Task and NSD. For the trait prediction tasks in Sec. 5.3, the learning rate is selected from $[5\times10^{-9},\ 5\times10^{-6}]$ matched with TABLeT's search space.
- `Batch size`: 8 for trait prediction and HBN-Movie, and 16 for HCP-Task and NSD.
- `Epochs`: 50 for trait prediction, HBN-Movie, and HCP-Task, and 15 for NSD.
- `Window size` : We use full frames.
- `TR` : 0.8s for trait prediction and HBN, 1s for HCP-Task and NSD.

B.3 PCA Analysis Details

We extract DCAE latents from 100k ImageNet test images and from HCP-A, using only the first temporal frame of each fMRI sequence to keep the computation tractable. For HCP-A we consider $\mathbf{X}_n^r$ with $r \in \{V, F\}$ and $n \in \{g, cf\}$: for $\mathbf{X}_n^V$ we take the center sagittal, coronal, and axial slices, evaluate each axis separately, and report their median and range in Fig. 4(a), and for $\mathbf{X}_n^F$ we treat the two hemispheres as independent samples. The ImageNet latents are randomly split into 80k samples for standardization and PCA fitting and 20k held-out samples for evaluation.

B.4 Gaussian Injection Experiments

The experiment uses all HCP-A subjects and a 16 s analysis window sampled at $\mathrm{TR} = 0.8\,\mathrm{s}$, giving $T = 20$ frames. Since the procedure is identical for every subject, we describe it for a single subject and omit the subject index.

Injection centers. The four centers are shared across every model, p, σ, and data split. They are selected from the intersection of the cohort-wide anatomical support and a cortical atlas mask C, subject to a minimum distance of 8 voxels from the outer support boundary and a minimum pairwise separation of 12 voxels. In native LAS-oriented MNI152 2-mm voxel indices they are $\mathbf{c}_1 = (60, 39, 34)$, $\mathbf{c}_2 = (20, 47, 33)$, $\mathbf{c}_3 = (31, 30, 34)$, and $\mathbf{c}_4 = (24, 38, 43)$, with a realized minimum pairwise distance of 13.93 voxels.

Signal injection. Let $\mathbf{X}_t(\mathbf{v})$ denote the clean BOLD value at voxel $\mathbf{v}$ in frame t, in the native MNI volume and before any normalization. For center $\mathbf{c}_j$, spatial scale σ, and percent signal change p, the injected frame is

$$\mathbf{X}_t^{\mathrm{inj}}(\mathbf{v}) = \mathbf{X}_t(\mathbf{v}) + b_j \tfrac{p}{100}\, h(t)\, A_{j,\sigma}(\mathbf{v}), \tag{5}$$

which adds a fixed spatial profile $A_{j,\sigma}$, modulated by a fixed time course h and scaled by an amplitude $b_j \cdot p/100$. The clean condition leaves $\mathbf{X}_t$ unchanged. The three factors are defined below.

Spatial profile. $A_{j,\sigma}$ is an isotropic Gaussian centered at $\mathbf{c}_j$, restricted to the cortical atlas mask C and to the subject's anatomical support (foreground) M:

$$A_{j,\sigma}(\mathbf{v}) = \exp\!\left(-\frac{\|\mathbf{v} - \mathbf{c}_j\|_2^2}{2\sigma^2}\right) C(\mathbf{v})\, M(\mathbf{v}). \tag{6}$$

The perturbation is thus confined to cortex: both the centers and the Gaussian tails are restricted to C.

Amplitude and the definition of PSC. b_j is the local baseline intensity at $\mathbf{c}_j$, taken as the mean of the temporal-mean image $\overline{\mathbf{X}}(\mathbf{v}) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_t(\mathbf{v})$ over the nonzero supported voxels in the

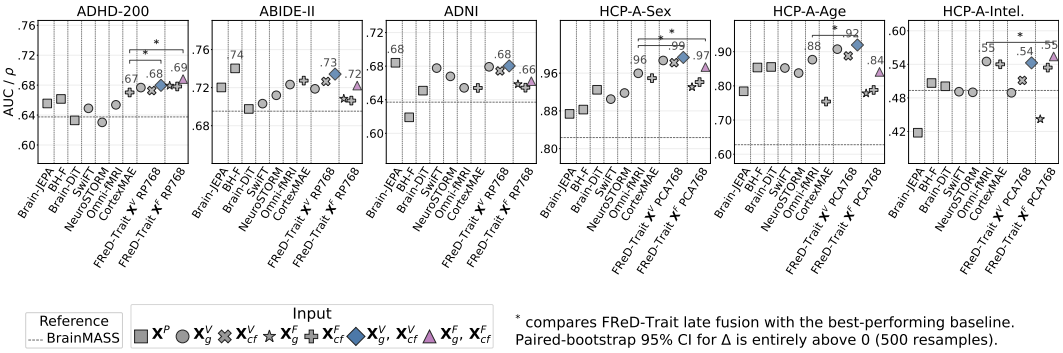


Figure 6: Linear-probing results over 500 random train-test splits after reducing the FReD-Trait features to 768 dimensions, reported as AUC for classification and ρ for regression. RP768 and PCA768 denote Gaussian random projection and PCA, respectively.

$7 \times 7 \times 7$ neighborhood $\mathcal{N}_j$ centered on it, $b_j = |\mathcal{N}_j|^{-1} \sum_{\mathbf{v} \in \mathcal{N}_j} \overline{\mathbf{X}}(\mathbf{v})$. Since $A_{j,\sigma}(\mathbf{c}_j) = 1$ and $\max_t h(t) = 1$, the added signal at the Gaussian center and at the frame where the response peaks is exactly p percent of b_j, before the global-statistics matching described below.

Time course. $h(t)$ is a canonical SPM hemodynamic response convolved with a 4 s boxcar $1_{[0,4)}$ and normalized to unit peak,

$$h(t) \propto 1_{[0,4)} * \left(\mathrm{GammaPDF}(\cdot\,; 6, 1) - \tfrac{1}{6}\,\mathrm{GammaPDF}(\cdot\,; 16, 1)\right), \qquad \max_t h(t) = 1. \tag{7}$$

Removing global-intensity shortcuts. After injection we match the mean and standard deviation of the modified sequence to those of its clean counterpart, computed jointly over all frames and all voxels in M:

$$\widehat{\mathbf{X}}_t^{\mathrm{inj}}(\mathbf{v}) = \frac{\mathbf{X}_t^{\mathrm{inj}}(\mathbf{v}) - \mu_{\mathrm{inj}}}{\sigma_{\mathrm{inj}} + \epsilon}\,(\sigma_{\mathrm{clean}} + \epsilon) + \mu_{\mathrm{clean}}, \qquad \mathbf{v} \in M, \tag{8}$$

with values outside M reset to zero afterwards. The matched sequence is then coordinate–frame normalized, yielding $\mathbf{X}_{cf}^V$.

Frozen representations. For NeuroSTORM, SwiFT, Omni-fMRI, and CortexMAE we extract two frozen representations. The `repeat` condition fills the model's temporal input with the frame at which $h(t)$ peaks, removing genuine temporal evolution while preserving the peak spatial perturbation, whereas the `timeseries` condition adapts the full injected sequence to each checkpoint's temporal sampling and input length.

The two FReD variants use the peak frame only, since the DCAE encodes frames independently and the repeat condition is its natural counterpart. For $\mathbf{X}_{cf}^V$ we compute $\mathbf{Z}_{cf,t}^V$ following Eq. (1) with $G = 24$, and for $\mathbf{X}_{cf}^F$ we project the matched volume to the fsLR surface and compute $\mathbf{Z}_{cf,t}^F$ following Eq. (2), in both cases at the peak frame. Both are reduced to 768 dimensions by PCA fitted on the training subjects only, matching the largest foundation-model feature dimension.

Probing protocol. Each subject contributes one injected sample at each of the four centers, giving 4,276 balanced samples scored by accuracy against a chance level of 25%. We use four predefined subject-disjoint 70/15/15 train/validation/test splits and fit a `LogisticRegression` probe from scikit-learn.

C Additional Experimental Results

C.1 Dimension-reduced Linear Probing

The frame representations of FReD-Trait are higher-dimensional than those of the foundation models, which could by itself account for its linear-probing performance in Sec. 4.2. To assess this effect, we reduce the FReD-Trait features to 768 dimensions, matching the dimensionality of CortexMAE, and repeat the same 500-split protocol. The reduction method depends on the sample size: with

85% of the data used for training, the clinical datasets provide only 365–523 training samples, so a 768-component PCA is not defined and we use Gaussian random projection instead. For HCP-A, where 908 training samples are available, we use PCA.

As shown in Fig. 6, the competitiveness of FReD-Trait is largely preserved at matched dimensionality. On ADHD-200, HCP-A sex, and HCP-A age, the reduced representations remain significantly better than the strongest baseline. Performance does drop on ABIDE-II and HCP-A intelligence, where the reduced features fall to the level of the best baselines, indicating that part of the margin on these tasks may come from the larger feature dimension. Overall, however, FReD-Trait remains competitive with fully fine-tuned foundation models after reducing its features to the largest foundation-model feature dimension, so its linear-probing performance cannot be attributed to dimensionality alone.

C.2 Late Fusion Weight Distribution

We analyze the fusion weight α selected for each task, where larger α favors the globally normalized branch and smaller α the coordinate–frame-normalized one. As shown in Fig. 7, the selected weights vary substantially across tasks. ASD classification and intelligence prediction consistently favor coordinate–frame normalization under both data types, whereas ADHD and ADNI diagnosis and the HCP-A sex and age tasks place at least as much weight on the globally normalized branch.

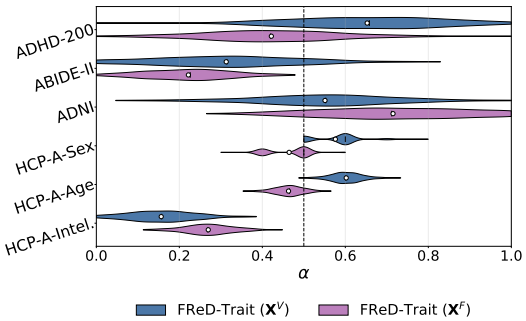


Figure 7: Distribution of the late-fusion weight α over the 500 train-test splits, shown separately for FReD-Trait with $\mathbf{X}^V$ and $\mathbf{X}^F$. Larger α places more weight on the globally normalized branch, and the dashed line marks equal weighting. White markers indicate the mean.

The distributions are broad for the clinical diagnosis tasks and much narrower for the HCP-A tasks, mirroring their split-to-split variability. Across all tasks, however, the mean weight rarely approaches 0 or 1, suggesting that both normalization schemes can contribute complementary information. These results highlight normalization as an important component of the downstream pipeline: the two schemes retain complementary information, and their relative utility varies across tasks.

C.3 Partial Fine-tuning

Table 7: Partial fine-tuning results on trait prediction tasks. Each foundation model is partially fine-tuned, while FReD-Trait uses linear probing on frozen features. Results are reported as the mean and standard deviation across 12 data splits for the clinical diagnosis tasks and four splits for HCP-A. **Bold** and underlined values indicate the best and second-best performance for each metric.

Method	Input	ADHD-200		ABIDE-II		ADNI		HCP-A					
		Diagnosis		Diagnosis		Diagnosis		Sex		Age		Intelligence	
		AUC ↑	F1 ↑	AUC ↑	F1 ↑	AUC ↑	F1 ↑	AUC ↑	F1 ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑
Brain-JEPA	$\mathbf{X}^P$	$.662_{\pm.04}$	$.595_{\pm.03}$	$.721_{\pm.03}$	$.572_{\pm.06}$	$.674_{\pm.04}$	$.527_{\pm.07}$	$.905_{\pm.02}$	$.817_{\pm.01}$	$.454_{\pm.01}$	$.830_{\pm.01}$	$.704_{\pm.04}$	$.429_{\pm.05}$
BH-F	$\mathbf{X}^P$	$.662_{\pm.05}$	$\underline{.617}_{\pm.03}$	$.710_{\pm.03}$	$.593_{\pm.05}$	$.614_{\pm.06}$	$.421_{\pm.10}$	$.896_{\pm.01}$	$.781_{\pm.01}$	$.430_{\pm.02}$	$.839_{\pm.01}$	$.677_{\pm.04}$	$.502_{\pm.05}$
Brain-DiT	$\mathbf{X}^P$	$.650_{\pm.03}$	$.604_{\pm.04}$	$.704_{\pm.04}$	$.558_{\pm.09}$	$.668_{\pm.05}$	$.562_{\pm.04}$	$.936_{\pm.03}$	$.856_{\pm.01}$	$.408_{\pm.02}$	$.863_{\pm.01}$	$.674_{\pm.02}$	$.517_{\pm.02}$
SwiFT	$\mathbf{X}_g^V$	$.654_{\pm.06}$	$\mathbf{.618}_{\pm.04}$	$.705_{\pm.03}$	$.588_{\pm.04}$	$\mathbf{.699}_{\pm.07}$	$\mathbf{.613}_{\pm.08}$	$.945_{\pm.02}$	$.839_{\pm.02}$	$\underline{.391}_{\pm.02}$	$\underline{.866}_{\pm.01}$	$.683_{\pm.03}$	$.500_{\pm.03}$
NeuroSTORM	$\mathbf{X}_g^V$	$.647_{\pm.04}$	$.583_{\pm.06}$	$.706_{\pm.03}$	$.599_{\pm.04}$	$.652_{\pm.05}$	$.575_{\pm.07}$	$.959_{\pm.01}$	$.868_{\pm.02}$	$.423_{\pm.02}$	$.847_{\pm.01}$	$.686_{\pm.05}$	$.500_{\pm.05}$
Omni-fMRI	$\mathbf{X}_g^V$	$.655_{\pm.05}$	$.591_{\pm.04}$	$.706_{\pm.03}$	$.598_{\pm.03}$	$.668_{\pm.06}$	$\underline{.591}_{\pm.07}$	$.958_{\pm.01}$	$.859_{\pm.02}$	$.392_{\pm.01}$	$.862_{\pm.01}$	$.676_{\pm.05}$	$.514_{\pm.02}$
CortexMAE	$\mathbf{X}_{cf}^F$	$.681_{\pm.04}$	$\mathbf{.618}_{\pm.05}$	$.732_{\pm.03}$	$\underline{.623}_{\pm.03}$	$.645_{\pm.06}$	$.555_{\pm.06}$	$.928_{\pm.02}$	$.811_{\pm.02}$	$.558_{\pm.05}$	$.720_{\pm.06}$	$\underline{.653}_{\pm.04}$	$.539_{\pm.05}$
FReD-Trait	$\mathbf{X}_g^V, \mathbf{X}_{cf}^V$	$\underline{.686}_{\pm.05}$	$.604_{\pm.05}$	$\underline{.743}_{\pm.03}$	$\mathbf{.629}_{\pm.04}$	$\underline{.690}_{\pm.04}$	$\underline{.591}_{\pm.05}$	$\mathbf{.995}_{\pm.00}$	$\mathbf{.956}_{\pm.03}$	$\mathbf{.320}_{\pm.01}$	$\mathbf{.917}_{\pm.01}$	$\mathbf{.633}_{\pm.06}$	$\mathbf{.607}_{\pm.05}$
FReD-Trait	$\mathbf{X}_g^F, \mathbf{X}_{cf}^F$	$\mathbf{.717}_{\pm.03}$	$.607_{\pm.05}$	$\mathbf{.747}_{\pm.03}$	$.615_{\pm.05}$	$.679_{\pm.05}$	$.541_{\pm.04}$	$\underline{.969}_{\pm.00}$	$\underline{.884}_{\pm.00}$	$.476_{\pm.02}$	$.828_{\pm.01}$	$.655_{\pm.03}$	$\underline{.576}_{\pm.04}$

To examine whether the trends in Tab. 1 also hold under partial fine-tuning, we fine-tune only the final layer (one last block) of each foundation model together with its prediction head, freezing all remaining layers, except for Brain-DiT. For Brain-DiT, we consider the last four blocks because

the downstream representation aggregates hidden states from the 13th, 15th, and 16th blocks of the backbone, with the 13th block defining the earliest layer involved in downstream feature extraction.

As shown in Tab. 7, partial fine-tuning also fails to consistently outperform FReD-Trait across tasks.

C.4 Linear Probing Results with Matched Splits for Foundation Models

Table 8: Comparison of trait prediction performance. Each foundation model is evaluated both by linear probing on its frozen features and by full fine-tuning, while FReD-Trait uses linear probing. Results are reported as the mean and standard deviation across data splits. **Bold** and underlined values indicate the best and second-best performance for each metric, respectively. LP and FT denote linear probing and full fine-tuning, respectively.

Method	Input	Eval	ADHD-200		ABIDE-II		ADNI		HCP-A					
			Diagnosis		Diagnosis		Diagnosis		Sex		Age		Intelligence	
			AUC ↑	F1 ↑	AUC ↑	F1 ↑	AUC ↑	F1 ↑	AUC ↑	F1 ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑
Brain-JEPA	$\mathbf{X}^P$	LP	$.660_{\pm.04}$	$.589_{\pm.04}$	$.728_{\pm.03}$	$.623_{\pm.04}$	$.693_{\pm.03}$	$.564_{\pm.05}$	$.892_{\pm.03}$	$.785_{\pm.04}$	$.516_{\pm.01}$	$.773_{\pm.01}$	$.715_{\pm.05}$	$.407_{\pm.05}$
Brain-JEPA	$\mathbf{X}^P$	FT	$.660_{\pm.00}$	$.606_{\pm.01}$	$.720_{\pm.05}$	$.622_{\pm.04}$	$.693_{\pm.07}$	$.597_{\pm.05}$	$.917_{\pm.00}$	$.829_{\pm.00}$	$.418_{\pm.00}$	$.849_{\pm.00}$	$.681_{\pm.01}$	$.487_{\pm.01}$
BH-F	$\mathbf{X}^P$	LP	$.669_{\pm.03}$	$\underline{.616}_{\pm.02}$	$.731_{\pm.03}$	$.608_{\pm.03}$	$.601_{\pm.05}$	$.432_{\pm.07}$	$.872_{\pm.01}$	$.759_{\pm.02}$	$.434_{\pm.00}$	$.840_{\pm.00}$	$.677_{\pm.04}$	$.520_{\pm.05}$
BH-F	$\mathbf{X}^P$	FT	$.631_{\pm.05}$	$.573_{\pm.06}$	$.697_{\pm.03}$	$.592_{\pm.05}$	$.601_{\pm.06}$	$.452_{\pm.12}$	$.867_{\pm.01}$	$.729_{\pm.10}$	$.435_{\pm.02}$	$.838_{\pm.02}$	$.679_{\pm.04}$	$.498_{\pm.06}$
Brain-DiT	$\mathbf{X}^P$	LP	$.624_{\pm.04}$	$.563_{\pm.04}$	$.675_{\pm.03}$	$.547_{\pm.03}$	$.644_{\pm.05}$	$.553_{\pm.06}$	$.925_{\pm.02}$	$.828_{\pm.03}$	$.423_{\pm.02}$	$.852_{\pm.01}$	$.675_{\pm.03}$	$.514_{\pm.05}$
Brain-DiT	$\mathbf{X}^P$	FT	$.648_{\pm.05}$	$.568_{\pm.04}$	$.712_{\pm.04}$	$.535_{\pm.08}$	$.659_{\pm.05}$	$.484_{\pm.14}$	$.923_{\pm.02}$	$.833_{\pm.03}$	$.386_{\pm.01}$	$.877_{\pm.00}$	$.674_{\pm.01}$	$.546_{\pm.01}$
SwiFT	$\mathbf{X}_g^V$	LP	$.651_{\pm.05}$	$.606_{\pm.05}$	$.705_{\pm.03}$	$.603_{\pm.04}$	$\mathbf{.708}_{\pm.05}$	$.598_{\pm.05}$	$.913_{\pm.02}$	$.801_{\pm.03}$	$.437_{\pm.02}$	$.837_{\pm.01}$	$.683_{\pm.03}$	$.485_{\pm.03}$
SwiFT	$\mathbf{X}_g^V$	FT	$.664_{\pm.05}$	$\mathbf{.633}_{\pm.05}$	$.711_{\pm.03}$	$.606_{\pm.04}$	$.688_{\pm.07}$	$.582_{\pm.06}$	$.946_{\pm.01}$	$.853_{\pm.02}$	$.350_{\pm.03}$	$.897_{\pm.02}$	$.678_{\pm.03}$	$.507_{\pm.02}$
NeuroSTORM	$\mathbf{X}_g^V$	LP	$.645_{\pm.05}$	$.599_{\pm.05}$	$.697_{\pm.02}$	$.595_{\pm.03}$	$\underline{.701}_{\pm.06}$	$\underline{.600}_{\pm.05}$	$.924_{\pm.02}$	$.828_{\pm.02}$	$.460_{\pm.02}$	$.822_{\pm.01}$	$.696_{\pm.05}$	$.473_{\pm.05}$
NeuroSTORM	$\mathbf{X}_g^V$	FT	$.658_{\pm.04}$	$.572_{\pm.07}$	$.716_{\pm.03}$	$.606_{\pm.02}$	$.673_{\pm.04}$	$\mathbf{.607}_{\pm.02}$	$.974_{\pm.01}$	$.888_{\pm.02}$	$\underline{.343}_{\pm.02}$	$\underline{.906}_{\pm.01}$	$.670_{\pm.04}$	$.525_{\pm.05}$
Omni-fMRI	$\mathbf{X}_g^V$	LP	$.657_{\pm.04}$	$.609_{\pm.03}$	$.702_{\pm.03}$	$.602_{\pm.04}$	$.674_{\pm.04}$	$.588_{\pm.04}$	$.969_{\pm.01}$	$.873_{\pm.04}$	$.388_{\pm.02}$	$.875_{\pm.01}$	$\underline{.645}_{\pm.04}$	$.552_{\pm.03}$
Omni-fMRI	$\mathbf{X}_g^V$	FT	$.676_{\pm.05}$	$.598_{\pm.05}$	$.712_{\pm.03}$	$.586_{\pm.03}$	$.681_{\pm.06}$	$.584_{\pm.06}$	$\underline{.975}_{\pm.01}$	$\underline{.910}_{\pm.01}$	$.351_{\pm.01}$	$.894_{\pm.01}$	$.657_{\pm.04}$	$.546_{\pm.05}$
CortexMAE	$\mathbf{X}_{cf}^F$	LP	$.677_{\pm.03}$	$.613_{\pm.04}$	$.730_{\pm.02}$	$.612_{\pm.03}$	$.652_{\pm.05}$	$.569_{\pm.06}$	$.947_{\pm.01}$	$.858_{\pm.02}$	$.536_{\pm.02}$	$.739_{\pm.02}$	$.655_{\pm.04}$	$.553_{\pm.05}$
CortexMAE	$\mathbf{X}_{cf}^F$	FT	$\underline{.690}_{\pm.05}$	$.605_{\pm.06}$	$.735_{\pm.03}$	$\mathbf{.639}_{\pm.04}$	$.638_{\pm.05}$	$.540_{\pm.06}$	$.944_{\pm.02}$	$.859_{\pm.02}$	$.508_{\pm.02}$	$.774_{\pm.02}$	$.661_{\pm.05}$	$.539_{\pm.06}$
FReD-Trait	$\mathbf{X}_g^V, \mathbf{X}_{cf}^V$	LP	$.686_{\pm.05}$	$.604_{\pm.05}$	$\underline{.743}_{\pm.03}$	$\underline{.629}_{\pm.04}$	$.690_{\pm.04}$	$.591_{\pm.05}$	$\mathbf{.995}_{\pm.00}$	$\mathbf{.956}_{\pm.03}$	$\mathbf{.320}_{\pm.01}$	$\mathbf{.917}_{\pm.01}$	$\mathbf{.633}_{\pm.06}$	$\mathbf{.607}_{\pm.05}$
FReD-Trait	$\mathbf{X}_g^F, \mathbf{X}_{cf}^F$	LP	$\mathbf{.717}_{\pm.03}$	$.607_{\pm.05}$	$\mathbf{.747}_{\pm.03}$	$.615_{\pm.05}$	$.679_{\pm.05}$	$.541_{\pm.04}$	$.969_{\pm.00}$	$.884_{\pm.00}$	$.476_{\pm.02}$	$.828_{\pm.01}$	$.655_{\pm.03}$	$\underline{.576}_{\pm.04}$

The linear-probing comparison in Fig. 2 uses 500 randomly sampled splits, whereas the fine-tuning results in Tab. 1 use the 12 or four splits described in Sec. 4.1, so the two are not directly comparable. To compare the two adaptation strategies on identical partitions, we additionally linear-probe every foundation model on the same splits used for fine-tuning.

From Tab. 8, it can be observed that fine-tuning is not uniformly better: linear probing is ahead on ADNI for SwiFT and NeuroSTORM, on ABIDE-II and ADHD-200 for BrainHarmonix-F, and on intelligence regression for Omni-fMRI and CortexMAE, while fine-tuning helps most consistently on the HCP-A sex and age tasks. Adapting the entire pre-trained encoder therefore does not reliably extract more from these representations than a linear readout does, and neither protocol brings the foundation models past FReD-Trait on most tasks.

C.5 Full Results of Fig. 2 and Fig. 3

We present full results of Fig. 2 and Fig. 3 in Tab. 9 and Tab. 10, respectively.

Table 9: Full results of Fig. 2. **Bold** and underlined values indicate the best and second-best performance for each metric, respectively.

Method	Input	ADHD-200		ABIDE-II		ADNI		HCP-A					
		Diagnosis		Diagnosis		Diagnosis		Sex		Age		Intelligence	
		AUC ↑	F1 ↑	AUC ↑	F1 ↑	AUC ↑	F1 ↑	AUC ↑	F1 ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑
BrainMASS	$\mathbf{X}^P$	$.638_{\pm.06}$	$.552_{\pm.07}$	$.695_{\pm.04}$	$.596_{\pm.05}$	$.637_{\pm.07}$	$.533_{\pm.07}$	$.823_{\pm.03}$	$.699_{\pm.04}$	$.635_{\pm.03}$	$.628_{\pm.04}$	$.679_{\pm.04}$	$.493_{\pm.06}$
Brain-JEPA	$\mathbf{X}^P$	$.655_{\pm.04}$	$.589_{\pm.04}$	$.720_{\pm.03}$	$.619_{\pm.04}$	$\mathbf{.684}_{\pm.06}$	$.561_{\pm.07}$	$.874_{\pm.03}$	$.763_{\pm.04}$	$.502_{\pm.03}$	$.784_{\pm.02}$	$.712_{\pm.04}$	$.418_{\pm.06}$
BH-F	$\mathbf{X}^P$	$.662_{\pm.03}$	$\mathbf{.631}_{\pm.02}$	$.740_{\pm.03}$	$.621_{\pm.04}$	$.619_{\pm.07}$	$.453_{\pm.09}$	$.883_{\pm.03}$	$.770_{\pm.04}$	$.420_{\pm.02}$	$.854_{\pm.02}$	$.682_{\pm.04}$	$.507_{\pm.05}$
Brain-DiT	$\mathbf{X}^P$	$.633_{\pm.04}$	$.586_{\pm.03}$	$.698_{\pm.03}$	$.577_{\pm.04}$	$.651_{\pm.06}$	$.542_{\pm.07}$	$.925_{\pm.02}$	$.823_{\pm.03}$	$.419_{\pm.02}$	$.855_{\pm.02}$	$.680_{\pm.04}$	$.501_{\pm.06}$
SwiFT	$\mathbf{X}_g^V$	$.649_{\pm.04}$	$.600_{\pm.04}$	$.703_{\pm.03}$	$.600_{\pm.04}$	$.678_{\pm.06}$	$\mathbf{.577}_{\pm.07}$	$.905_{\pm.02}$	$.797_{\pm.03}$	$.424_{\pm.02}$	$.852_{\pm.02}$	$.679_{\pm.04}$	$.491_{\pm.05}$
NeuroSTORM	$\mathbf{X}_g^V$	$.630_{\pm.04}$	$.587_{\pm.04}$	$.712_{\pm.03}$	$.602_{\pm.04}$	$.668_{\pm.06}$	$.562_{\pm.07}$	$.918_{\pm.02}$	$.805_{\pm.03}$	$.445_{\pm.02}$	$.837_{\pm.02}$	$.687_{\pm.04}$	$.490_{\pm.05}$
Omni-fMRI	$\mathbf{X}_g^V$	$.654_{\pm.04}$	$.598_{\pm.05}$	$.723_{\pm.03}$	$.612_{\pm.04}$	$.654_{\pm.06}$	$.551_{\pm.07}$	$.960_{\pm.01}$	$.875_{\pm.03}$	$.386_{\pm.02}$	$.877_{\pm.01}$	$.655_{\pm.04}$	$.545_{\pm.05}$
CortexMAE	$\mathbf{X}_{cf}^F$	$.670_{\pm.04}$	$\underline{.618}_{\pm.05}$	$.727_{\pm.04}$	$.595_{\pm.05}$	$.654_{\pm.06}$	$.568_{\pm.06}$	$.949_{\pm.02}$	$.856_{\pm.03}$	$.529_{\pm.03}$	$.754_{\pm.03}$	$.665_{\pm.04}$	$.540_{\pm.05}$
FReD-Trait	$\mathbf{X}_g^V$	$.678_{\pm.04}$	$.603_{\pm.05}$	$.744_{\pm.03}$	$\mathbf{.630}_{\pm.05}$	$.680_{\pm.06}$	$\underline{.571}_{\pm.07}$	$\underline{.988}_{\pm.01}$	$\underline{.948}_{\pm.02}$	$\underline{.328}_{\pm.02}$	$\underline{.914}_{\pm.01}$	$.644_{\pm.04}$	$.582_{\pm.04}$
FReD-Trait	$\mathbf{X}_{cf}^V$	$.676_{\pm.04}$	$.612_{\pm.05}$	$.746_{\pm.04}$	$.625_{\pm.05}$	$.676_{\pm.06}$	$.571_{\pm.07}$	$.985_{\pm.01}$	$.931_{\pm.02}$	$.348_{\pm.02}$	$.900_{\pm.01}$	$\underline{.616}_{\pm.04}$	$\underline{.616}_{\pm.04}$
FReD-Trait	$\mathbf{X}_g^V, \mathbf{X}_{cf}^V$	$.682_{\pm.04}$	$.609_{\pm.05}$	$.749_{\pm.03}$	$\underline{.628}_{\pm.05}$	$\underline{.681}_{\pm.06}$	$.563_{\pm.07}$	$\mathbf{.992}_{\pm.01}$	$\mathbf{.961}_{\pm.02}$	$\mathbf{.312}_{\pm.02}$	$\mathbf{.922}_{\pm.01}$	$\mathbf{.616}_{\pm.04}$	$\mathbf{.619}_{\pm.04}$
FReD-Trait	$\mathbf{X}_g^F$	$.687_{\pm.04}$	$.617_{\pm.05}$	$.718_{\pm.03}$	$.618_{\pm.04}$	$.666_{\pm.06}$	$.551_{\pm.07}$	$.935_{\pm.02}$	$.833_{\pm.03}$	$.509_{\pm.03}$	$.780_{\pm.02}$	$.695_{\pm.04}$	$.491_{\pm.05}$
FReD-Trait	$\mathbf{X}_{cf}^F$	$\underline{.704}_{\pm.04}$	$.593_{\pm.07}$	$\underline{.754}_{\pm.03}$	$.618_{\pm.05}$	$.674_{\pm.06}$	$.560_{\pm.08}$	$.945_{\pm.02}$	$.855_{\pm.03}$	$.491_{\pm.03}$	$.792_{\pm.02}$	$.648_{\pm.04}$	$.565_{\pm.05}$
FReD-Trait	$\mathbf{X}_g^F, \mathbf{X}_{cf}^F$	$\mathbf{.708}_{\pm.04}$	$.612_{\pm.06}$	$\mathbf{.755}_{\pm.03}$	$.617_{\pm.04}$	$.672_{\pm.06}$	$.555_{\pm.07}$	$.972_{\pm.01}$	$.896_{\pm.03}$	$.448_{\pm.02}$	$.844_{\pm.02}$	$.641_{\pm.04}$	$.589_{\pm.05}$

Table 10: Full results of Fig. 3. LP rows are highlighted in yellow; FT and TFS denote fine-tuning and training from scratch, respectively. **Bold** and underlined values indicate the best and second-best distinct performance for each metric, respectively.

Method	Eval.	Input	HBN-Movie	
			AUC ↑	F1 ↑
BrainMASS	LP	$\mathbf{X}^P$	$.895_{\pm.02}$	$.793_{\pm.02}$
Brain-JEPA	LP	$\mathbf{X}^P$	$.631_{\pm.02}$	$.610_{\pm.04}$
Brain-JEPA	FT	$\mathbf{X}^P$	$.716_{\pm.10}$	$.656_{\pm.08}$
BH-F	LP	$\mathbf{X}^P$	$.724_{\pm.04}$	$.647_{\pm.05}$
BH-F	FT	$\mathbf{X}^P$	$.707_{\pm.18}$	$.673_{\pm.16}$
Brain-DiT	LP	$\mathbf{X}^P$	$.739_{\pm.03}$	$.675_{\pm.04}$
Brain-DiT	FT	$\mathbf{X}^P$	$.897_{\pm.13}$	$.691_{\pm.13}$
SwiFT	LP	$\mathbf{X}_g^V$	$.630_{\pm.04}$	$.599_{\pm.02}$
SwiFT	LP	$\mathbf{X}_{cf}^V$	$.708_{\pm.02}$	$.663_{\pm.00}$
SwiFT	FT	$\mathbf{X}_g^V$	$.619_{\pm.05}$	$.590_{\pm.06}$
SwiFT	FT	$\mathbf{X}_{cf}^V$	$.633_{\pm.06}$	$.602_{\pm.04}$
NeuroSTORM	LP	$\mathbf{X}_g^V$	$.632_{\pm.04}$	$.597_{\pm.04}$
NeuroSTORM	LP	$\mathbf{X}_{cf}^V$	$.739_{\pm.02}$	$.692_{\pm.05}$
NeuroSTORM	FT	$\mathbf{X}_g^V$	$.620_{\pm.06}$	$.538_{\pm.08}$
NeuroSTORM	FT	$\mathbf{X}_{cf}^V$	$.690_{\pm.09}$	$.659_{\pm.08}$
Omni-fMRI	LP	$\mathbf{X}_g^V$	$.615_{\pm.03}$	$.591_{\pm.03}$
Omni-fMRI	LP	$\mathbf{X}_{cf}^V$	$.783_{\pm.03}$	$.713_{\pm.02}$
Omni-fMRI	FT	$\mathbf{X}_g^V$	$.596_{\pm.06}$	$.561_{\pm.10}$
Omni-fMRI	FT	$\mathbf{X}_{cf}^V$	$.928_{\pm.03}$	$.838_{\pm.02}$
CortexMAE	LP	$\mathbf{X}_{cf}^F$	$.939_{\pm.02}$	$.852_{\pm.06}$
CortexMAE	FT	$\mathbf{X}_{cf}^F$	$\underline{.997}_{\pm.00}$	$.959_{\pm.01}$
TABLeT	TFS	$\mathbf{X}_g^V$	$.609_{\pm.05}$	$.499_{\pm.19}$
TABLeT	TFS	$\mathbf{X}_{cf}^V$	$.990_{\pm.01}$	$.948_{\pm.03}$
FReD-Trait	LP	$\mathbf{X}_g^V, \mathbf{X}_{cf}^V$	$.849_{\pm.03}$	$.780_{\pm.03}$
FReD-Trait	LP	$\mathbf{X}_g^V$	$.672_{\pm.03}$	$.623_{\pm.04}$
FReD-Trait	LP	$\mathbf{X}_{cf}^V$	$.852_{\pm.03}$	$.784_{\pm.03}$
FReD-Trait	LP	$\mathbf{X}_g^F, \mathbf{X}_{cf}^F$	$.871_{\pm.02}$	$.808_{\pm.01}$
FReD-Trait	LP	$\mathbf{X}_g^F$	$.629_{\pm.03}$	$.580_{\pm.04}$
FReD-Trait	LP	$\mathbf{X}_{cf}^F$	$.871_{\pm.02}$	$.802_{\pm.01}$
FReD-State	TFS	$\mathbf{X}_{cf}^V$	$\mathbf{.998}_{\pm.00}$	$\mathbf{.973}_{\pm.01}$
FReD-State	TFS	$\mathbf{X}_{cf}^F$	$\underline{.997}_{\pm.00}$	$\underline{.971}_{\pm.01}$

D Architectural Comparison between TABLeT and FReD-State

Tab. 11 lists the differences between TABLeT and FReD-State with $\mathbf{X}^V$ in full. Both models take the same DCAE latent tensor and the same downstream loss, so every difference below concerns how those latents are turned into a sequence and how that sequence is modeled. The ablation in Tab. 5 isolates the contribution of tokenization, while the remaining architectural and regularization changes are evaluated jointly.

Tokenization. The two models differ in whether the slice axis is aggregated before or after the temporal Transformer. TABLeT keeps three slice blocks and the 3×3 latent grid as separate tokens, giving 27 tokens per frame and a sequence of length $27T+1$, so the Transformer must resolve spatial and temporal structure jointly. FReD-State instead averages within slice groups and flattens all axes, groups, channels, and spatial positions into a single vector per frame, which a linear layer projects to the model width. The sequence is therefore $T + 1$ tokens long, and spatial mixing happens in the projection rather than in self-attention. This reduces the number of self-attention token pairs significantly and lets the Transformer allocate its capacity to temporal structure alone.

Transformer backbone. FReD-State uses a third of the depth, four blocks instead of twelve, and adopts several standard modernizations: RMSNorm in place of LayerNorm, SwiGLU (Shazeer, 2020) feed-forward blocks in place of a ReLU MLP, and Q/K normalization (Henry et al., 2020)

Table 11: Architectural comparison between TABLeT and FReD-State with $\mathbf{X}^V$. Both models receive the same DCAE latent tensor and use the same downstream loss; T denotes the number of input fMRI frames.

Component	**TABLeT**	**FReD-State with $\mathbf{X}^V$**
Input representation and tokenization		
DCAE latent input	$\mathbf{x} \in \mathbb{R}^{T\times 3\times 96\times 32\times 3\times 3}$	Same input tensor
Intra-frame aggregation	No slice averaging. Three contiguous 32-slice blocks are formed per axis, and corresponding features from the three anatomical axes are concatenated at each latent-grid position.	Contiguous slice-group averaging with $G = 24$ groups per axis ($96 \to 24$), followed by concatenation of all axes, groups, channels, and spatial locations into one frame vector.
Tokens per frame	3 slice blocks $\times$ 3 $\times$ 3 spatial positions $= 27$	1 whole-frame token
Pre-projection feature width	3,072	20,736
Frame projection	$\mathrm{LN}(3072) \to \mathrm{Linear}(3072, 896) \to \mathrm{LN}(896)$	$\mathrm{LN}(20736) \to \mathrm{Linear}(20736, 896) \to \mathrm{LN}(896)$
Sequence length with CLS	$27T + 1$	$T + 1$
Temporal Transformer backbone		
Transformer depth	12 blocks	4 blocks
Model and FFN widths	$d_{\mathrm{model}} = 896$, $d_{\mathrm{ff}} = 4864$	$d_{\mathrm{model}} = 896$, $d_{\mathrm{ff}} = 4864$
Grouped-query attention	14 query heads, 2 key/value heads; head dimension 64	16 query heads, 2 key/value heads; head dimension 56
Attention formulation	Scaled dot-product GQA	Q/K RMS normalization and soft-capped attention logits, $10\tanh(\ell/10)$, before softmax
Positional encoding	One-dimensional RoPE	One-dimensional RoPE
Transformer normalization	Pre-norm LayerNorm	Pre-norm RMSNorm
Feed-forward network	Vanilla two-layer MLP with ReLU: $896 \to 4864 \to 896$	SwiGLU with separate gate and value projections and a SiLU gate: $896 \to 4864 \to 896$
Readout	Learned `[CLS]` token followed by a linear task head	Learned `[CLS]` token followed by a linear task head
Regularization and computational scale		
Frame dropout	None	Training-only frame-token dropout with $p = 0.5$
Stochastic depth	None	Drop-path probability increases linearly from 0 to 0.1 across the four blocks
Trainable parameters	129,533,333 (129.53M)	78,225,493 (78.23M; 39.6% fewer)

with soft-capped attention logits (Gemma Team et al., 2024). Model and feed-forward widths are unchanged, and both models use a learned `[CLS]` token with rotary position embeddings (Su et al., 2024).

Regularization. Since a frame is now a single token, dropping frames during training is a natural regularizer, and we drop frame tokens with $p = 0.5$ while never dropping the `[CLS]` token. We additionally use stochastic depth (Huang et al., 2016) with a drop-path probability increasing linearly from 0 to 0.1 across the four blocks.

Together these changes reduce the trainable parameter count from 129.5M to 78.2M, a 39.6% reduction, while improving accuracy on both HCP-Task and NSD (Tab. 5).