BRAIN-SEMANTOKS: LEARNING SEMANTIC TOKENS OF BRAIN DYNAMICS WITH A SELF-DISTILLED FOUNDATION MODEL

Anonymous authors

ABSTRACT

The development of foundation models for functional magnetic resonance imaging (fMRI) time series holds significant promise for predicting phenotypes related to disease and cognition. Current models, however, are often trained using a mask-and-reconstruct objective on small brain regions. This focus on low-level information leads to representations that are sensitive to noise and temporal fluctuations, necessitating extensive fine-tuning for downstream tasks. We introduce Brain-Semantoks, a self-supervised framework designed specifically to learn abstract representations of brain dynamics. Its architecture is built on two core innovations: a semantic tokenizer that aggregates noisy regional signals into robust tokens representing functional networks, and a self-distillation objective that enforces representational stability across time. We show that this objective is stabilized through a novel training curriculum, ensuring the model robustly learns meaningful features from low signal-to-noise time series. We demonstrate that learned representations enable strong performance on a variety of downstream tasks even when only using a linear probe. Furthermore, we provide comprehensive scaling analyses indicating more unlabeled data reliably results in out-of-distribution performance gains without domain adaptation.

1 INTRODUCTION

The investigation of brain dynamics has been a cornerstone of neuroscience, progressing our understanding of human cognition, disease, and aging. Functional magnetic resonance imaging (fMRI) has been an instrumental modality in this endeavor; its blood-oxygen-level-dependent (BOLD) measurement relates to local changes in brain activity, and its non-invasive nature has made it a primary tool across numerous research fields (Ogawa et al., 1990; Logothetis, 2008). Despite being extremely high-dimensional, fMRI data is often collected in limited samples, which can severely constrain the potential insights (Button et al., 2013; Poldrack et al., 2017). This challenge motivates a shift towards data-driven representation learning, where progress in self-supervised learning (SSL) can enable the training of highly capable 'foundation' models from large quantities of unlabeled data, promising a new way forward for neuroimaging analysis.

Current fMRI foundation models, however, adapt reconstruction-centric paradigms from NLP and vision, such as masked signal prediction (Caro et al., 2023; Wang et al., 2025). While latent-space reconstruction (e.g., JEPA; Dong et al. (2024)) avoids modeling the substantial noise in the BOLD signal, these models still focus on low-level, regional information. We argue this objective is misaligned with predicting stable, high-level phenotypes, as the resulting representations require extensive fine-tuning for such tasks, reducing the utility of a foundation model. This dependency is particularly problematic for fMRI, where transfer is challenged by significant variations across datasets in participant cohorts, hardware, and acquisition protocols.

To address these issues, we hypothesize that effectively predicting stable phenotypes requires a shift from reconstruction to abstraction. The goal should not be to perfectly encode the BOLD signal, but to abstract away from it to find the underlying phenotypic signature. We propose Brain-Semantoks, a foundation model built on a strong neuroscientific inductive bias to learn such abstract representations. Our approach starts at the input level, recognizing that self-attention mechanisms, central to modern transformers, perform best on sequences of low-noise, semantic tokens, akin to words in natural language. Time series of individual, small regions are poor tokens in this regard as they are noisy and lack high-level meaning. We therefore introduce a semantic tokenizer, which aggregates information from regions within a functional brain network (e.g., default mode network) into a single, robust token. This creates a shorter, more computationally efficient, and semantically meaningful sequence for the transformer to operate on.

With these semantic tokens, we then shift the learning objective itself. Instead of focusing on the reconstruction of masked signals, Brain-Semantoks is trained using a self-distillation objective to produce a stable, summary representation across different temporal views of the same scan (Grill et al., 2020; Caron et al., 2021). This explicitly trains the model to capture a stable, high-level representation of an individual's brain dynamics, which we expect to transfer

better across data distributions. However, while conceptually highly suitable, we found that applying this objective to low signal-to-noise fMRI data can lead to training instability, where the model converges on a poor, simple solution. To solve this, we introduce a Teacher-guided Temporal Regularizer (TTR), a novel training curriculum active only at the start of training. This regularizer guides the model to first learn the time-averaged signature of each network before modeling more complex temporal variations, ensuring robust and meaningful pretraining convergence. The resulting representations are particularly powerful for linear probing, indicating they are well-disentangled and broadly useful without task-specific fine-tuning.

Contributions. Our contributions are threefold. First, we propose a new pre-training approach that prioritizes abstract representations over signal reconstruction, enabled by a novel semantic tokenizer and a Teacher-guided Temporal Regularizer (TTR) to stabilize training. Second, we introduce Brain-Semantoks, a foundation model trained with this method that achieves state-of-the-art performance on diverse downstream tasks under a rigorous linear probing protocol. Finally, we provide the first detailed scaling analysis for fMRI foundation models, showing consistent out-of-distribution performance gains without domain adaptation.

2 Related Work

Self-Supervised Learning for MRI. Much early self-supervised learning work focused on reconstruction using auto-encoders (Han et al., 2019; Pinaya et al., 2019; Kim et al., 2021), using relatively limited data. The first effort to build an fMRI foundation model similarly adapted reconstruction-based objectives popular in other domains. BrainLM (Caro et al., 2023) employs a masked modeling objective to reconstruct the BOLD signal in input space. While effective, this approach risks modeling the substantial noise inherent in fMRI data. More recent work like Brain-JEPA (Dong et al., 2024) mitigates this by predicting masked representations in a latent space, thereby learning to ignore noise. Concurrent preprint NeuroSTORM operates on 4D voxel data, performing spatio-temporal reconstruction (Wang et al., 2025). However, all these methods remain fundamentally focused on predicting low-level information. In contrast, models like BrainMass learn from static functional connectivity matrices (Yang et al., 2024), ignoring the rich temporal dynamics central to our work.

Self-Distillation Learning. Self-distillation has proven highly effective for learning semantic features, improving upon contrastive learning methods (Chen et al., 2020). Seminal works include MoCo (He et al., 2020), BYOL (Grill et al., 2020), and DINO (Caron et al., 2021; Oquab et al., 2023; Siméoni et al., 2025) demonstrated that a student network can learn powerful, linearly separable representations by matching the output of a teacher network (a momentum-updated version of itself) across different views of a sample. This approach avoids "representational collapse" without requiring negative samples or a reconstruction loss. The iBOT framework (Zhou et al., 2021) further advanced this by integrating a masked-token prediction objective within the distillation framework, enabling the model to learn both a global summary representation as well as rich, context-aware local features. Recent work by Wu et al. (2025) makes important progress in understanding what is necessary to prevent collapse and significantly simplifying the approach, reducing the number of hyperparameters.

3 Method: Brain-Semantoks

Our proposed framework, Brain-Semantoks, learns abstract and temporally stable representations from fMRI time series. The methodology is built upon three core innovations designed to address the unique challenges of fMRI data. First, we introduce a paradigm performing self-distillation across time, that explicitly trains for high-level representations suitable for transfer learning. Second, we develop a semantic tokenizer with a strong neuroscientific inductive bias to create a robust and meaningful input space for our encoder model. Finally, we introduce a training curriculum that stabilizes the learning objective, ensuring convergence on low signal-to-noise data. The framework uses a student-teacher architecture, as depicted in Figure 1.

3.1 A Self-Distillation Framework for Semantic Representations

The primary goal of a foundation model is to learn representations that are broadly applicable without requiring task-specific fine-tuning. To achieve this with fMRI, a model must learn to capture the stable, underlying phenotypic signature of a subject, abstracting away from transient noise and acquisition-specific artifacts.

Input Data and Augmentation: We represent a subject's fMRI time series as matrix $X \in \mathbb{R}^{C \times T}$, where C is the number of brain regions of interest (ROIs) and T is the number of time points. To generate different views of the same underlying brain dynamics we create two long temporal segments of length $T_{crop} < T$ resulting in two views, $X^{(1)}$ and $X^{(2)}$. Unlike computer vision, a large set of intuitive augmentations are not available for fMRI. We therefore

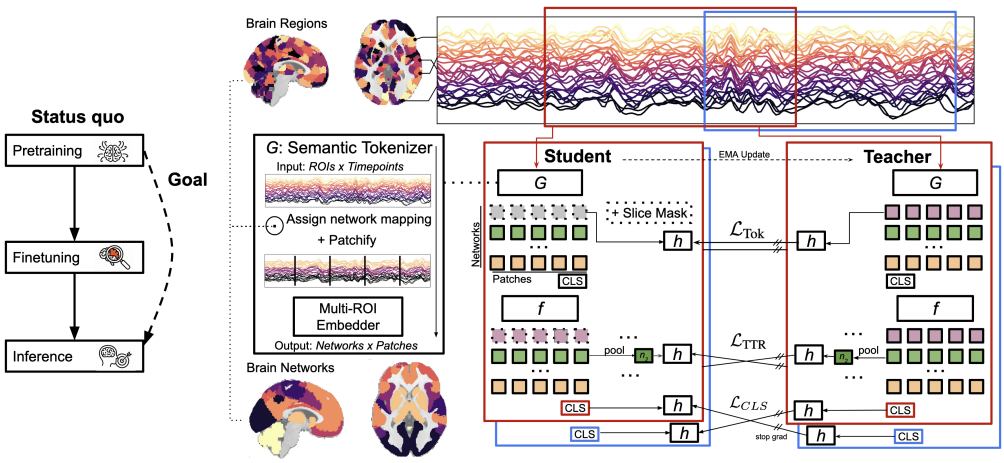

Figure 1: **Brain-Semantoks.** A student-teacher architecture is used to learn stable brain dynamics representations across time by aligning long temporal views. The semantic tokenizer (G) is used to produce robust tokens of functional brain networks, which serve as input to a transformer encoder (f). Three losses are used during pretraining: a temporary regularisation loss for stability ($\mathcal{L}_{TTR}$), a within-view, latent space prediction loss of masked tokens ($\mathcal{L}_{Tok}$), and a global cross-view loss to learn a high-level, semantic representation ($\mathcal{L}_{CLS}$).

mainly rely on self-distillation across time and only lightly further augment the views with corrupting transformations: we randomly select a fraction of channels (τ_c) and contiguous timepoints (τ_t) and set them to zero, add gaussian noise sampled with $\mu = 0$ and $\sigma = \tau_\sigma$, and finally scale the time series amplitude $X\tau_s$.

Student-Teacher Framework: We use a student-teacher architecture to enforce representational consistency across these two views. The student network $f_s(\theta_s)$, is trained to match the output of the teacher network $f_t(\theta_t)$. The teacher provides a stable regression target as its weights are an exponential moving average (EMA) of the student's weights:

$$\theta_t \leftarrow \alpha\theta_t + (1-\alpha)\theta_s \tag{1}$$

where the momentum coefficient α gradually increases during training. This forces the student to learn high-level representations which are stable across time.

3.2 Semantic Tokenizer

We posit that standard tokenization such as a direct linear projection of ROI signals is suboptimal for fMRI data. This approach creates overly long sequences of noisy, low-level tokens that hinder a transformer's ability to learn meaningful long-range dependencies. Our innovation is a semantic tokenizer, $G(\Phi)$, that addresses this by creating a compact and robust input space grounded in a core neuroscientific prior: the brain's organization into functional networks. For a given fMRI scan $\mathbf{X}^{(v)}$, our tokenizer uses N parallel, network-specific modules, g_n, each operating on the time series of a single network, $M_n \odot \mathbf{X}^{(v)}$, to produce a sequence of P semantically rich, D-dimensional tokens, $\mathbf{Z}_n^{(v)}$.

To model the BOLD signal's complex temporal structure, each module g_n first divides its input time series into P relatively large patches, x_p, inspired by the temporal stability of macro-scale brain dynamics (Allen et al., 2014; Vidaurre et al., 2017). Within each patch, a multi-scale convolutional filter bank, composed of a standard convolutional branch, $\text{Conv}_{\text{std}}(\cdot)$, and a structured convolutional branch, $\text{Conv}_{\text{str}}(\cdot)$ (Li et al., 2022), captures hierarchical temporal patterns. The Conv_{std} branch uses short kernels for local, cross-ROI dependencies, while the parameter-efficient Conv_{str} branch uses longer, decaying kernels to enforce a temporal inductive bias absent in standard ViT architectures.

A final D-dimensional token embedding for each patch is generated via the transformation: $\text{token} = \text{LayerNorm}\left(\text{AvgPool}\left(\text{GELU}\left([\text{Conv}_{\text{std}}(x_p); \text{Conv}_{\text{str}}(x_p)]\right)\right)\right)$. The complete token tensor is then formed by concatenating the outputs from all network modules, $\mathbf{Z}^{(v)} = [\mathbf{Z}_1^{(v)}; \ldots; \mathbf{Z}_N^{(v)}]$, resulting in a final tensor of shape $\mathbb{R}^{N\times P\times D}$. This approach yields a compact and semantically rich input that provides the transformer encoder with a better starting point for learning, which is crucial for the stability of the self-distillation framework.

3.3 TRANSFORMER ENCODER AND MASKING

The sequence of network-tokens $\mathbf{Z}^{(v)}$ is flattened into a sequence of length $N \times P$. We add sinusoidal positional embeddings to encode the temporal order of the patches and a learnable network-specific embedding for each of the N networks to encode their identity. Finally, a learnable `[CLS]` token is prepended to the sequence, yielding a sequence of length $N \times P + 1$.

The student and teacher models each consist of two core components: a transformer encoder backbone, $f(\theta)$, and a projection head, $h(\psi)$. The backbone, f, processes the input token sequence to produce network and `[CLS]` embeddings. These embeddings are then processed by the projector, h, into the space where the distillation loss is minimized. The projector is used only during pre-training and is discarded thereafter, while the transferable representations are the outputs from the teacher's encoder backbone, $f_t(\theta_t)$.

The student network, $f_s(\theta_s)$, receives a masked version of the token sequence. Masking is determined by a binary mask $\mathbf{B}^{(v)} \in \{0,1\}^{N \times P}$, which replaces tokens with a learnable mask embedding. The teacher network always receives the full, unmasked sequence. To reduce the degree to which the model can rely on simple, interpolative relationships to predict masked tokens, we perform 'slice masking'. Specifically, we treat the input network-tokens as a 2D matrix of size $N \times P$ (networks by temporal patches) and mask about entire 'slices' rather than random individual tokens. We randomly select one of two strategies on a sample-level to increase data diversity. The first, network slicing, masks entire rows, hiding all temporal data for one or more selected networks. Second, temporal slicing masks a contiguous block of entire columns, hiding information from all networks for a specific period. By masking large, contiguous parts of data, we force the model to learn more complex relationships between networks and across time.

3.4 PRETRAINING OBJECTIVE

Finally, we propose a multi-component objective function that includes a curriculum to ensure stable training. All loss components are computed using the outputs of the projection heads (h_s and h_t), which are denoted as $\mathbf{z_s}$ and $\mathbf{z_t}$. Following recent work simplifying DINO (Wu et al., 2025), we regularize with a coding rate term to prevent representation collapse.

Global Loss To learn a high-level, stable representation of the brain time series, we enforce consistency between the `[CLS]` tokens across two views. The loss is bidirectional and regularized:

$$\mathcal{L}_{CLS} = \mathbb{E}\left[d(z_{s,\text{CLS}(1)}, z_{t,\text{CLS}(2)}) + d(z_{s,\text{CLS}(2)}, z_{t,\text{CLS}(1)})\right] - \gamma_{\text{CLS}} R_\epsilon\left(\text{Cov}[z_{s,\text{CLS}}]\right) \tag{2}$$

where d is the squared Euclidean distance, the expectation $\mathbb{E}$ is taken over the data, R_ϵ is the total coding rate regularizer, and γ is a hyperparameter governing the regularization strength. R_ϵ is a differentiable measure related to the determinant of the feature covariance matrix. Minimizing its negative forces the covariance matrix to have a large determinant, which prevents the learned features $\mathbf{z}$ from collapsing into a lower-dimensional subspace.

Network Token Loss To promote the model learning rich, temporally-sensitive representations, we apply an auxiliary distillation loss on the network tokens that were masked in the student's input, guided by the 2D mask matrix $\mathbf{B}^{(v)}$ (Zhou et al., 2021). This loss is computed within each view:

$$\mathcal{L}_{\text{Tok}} = \mathbb{E}_{v \in \{1,2\}}\left[\sum_{n=1}^{N}\sum_{p=1}^{P} B_{n,p}^{(v)} \cdot d(\mathbf{z}_{s,(n-1)P+p}^{(v)}, \mathbf{z}_{t,(n-1)P+p}^{(v)})\right] \tag{3}$$

Due to the semantic tokenizer, this task is performed on a more semantic network-level, rather than on a noisier, lower region-level.

Teacher-guided Temporal Regularizer

Although our tokenization strategy significantly improves pretraining effectiveness, we found that a direct application of these distillation objectives can still lead to poor solutions with noisy fMRI time series. We therefore develop a principled stabilizing curriculum based on this observation.

Specifically, as we observe that a more compact token sequence aided convergence and yielded representations with better predictive performance, we guide the student network to first learn the time-averaged representation of each network. Conceptually, this constrains the token space $N \times P + 1$ towards $N + 1$, which helps find a good initial representation which can thereafter be refined with temporal variability. This can be directly adopted in the distillation

framework by using the teacher model to provide the network-specific targets. The summary token for each network n is computed by averaging its P patch embeddings from the transformer output:

$$\bar{\mathbf{z}}_{t,n}^{(v)} = \frac{1}{P} \sum_{p=1}^{P} \mathbf{z}_{t,(n-1)P+p}^{(v)} \tag{4}$$

The regularisation loss is then applied across views to these N summary tokens:

$$\mathcal{L}_{\text{TTR}} = \mathbb{E}\left[\sum_{n=1}^{N} \left(d(\bar{\mathbf{z}}_{s,n}^{(1)}, \bar{\mathbf{z}}_{t,n}^{(2)}) + d(\bar{\mathbf{z}}_{s,n}^{(2)}, \bar{\mathbf{z}}_{t,n}^{(1)}) \right)\right] - \gamma_{\text{TTR}} \sum_{n=1}^{N} R_{\epsilon}(\text{Cov}[\bar{\mathbf{z}}_{s,n}]) \tag{5}$$

Total Loss Function

The final training objective is a weighted sum of the three hierarchical loss components:

$$\mathcal{L}_{Total} = \mathcal{L}_{CLS} + \lambda_{Tok}\mathcal{L}_{Tok} + \lambda_{TTR}\mathcal{L}_{TTR} \tag{6}$$

where λ_{Tok} and λ_{TTR} are scalar hyperparameters balancing the contributions of different levels of self-supervision. Following training, the student weights are discarded while the teacher weights are used for downstream evaluation.

4 Experimental Setup

4.1 Datasets

Pretraining Data

We leveraged the largest 3T resting-state fMRI corpus available for unlabeled pretraining. Specifically, we use 39139 preprocessed recordings from the UKBioBank as well as the participant age and sex variables (UKB; Miller et al. (2016); application number [withheld for anonymous review], We held out 1625 recordings for downstream evaluation.

We extract parcel-wise time series using the cortical Schaefer-400 atlas (Schaefer et al., 2018), subcortical Tian-III atlas (Tian et al., 2020), and cerebellar Buckner-7 atlas (Buckner et al., 2011), yielding 457 total ROIs. Data normalization is a crucial step to aid transfer learning. Whereas prior work has relied on robust scaling, which preserves ROI-specific 'DC' offsets resulting from the fMRI scanner, it impairs transferability to datasets which have less or no such offsets. We therefore applied z-scoring to each ROI per scan. As the UKB's temporal resolution (0.735s) is higher than most available datasets, we temporally downsample to 2s, resulting in 180 timepoints for the 6 minute recordings. Both normalization and resampling happen on the parcellated time series, meaning these are light operations that can be performed online during data loading and thereby ease transferability.

Downstream Tasks

For downstream prediction, we construct a varied set of tasks with differing sample sizes in order to evaluate consistency across contexts (Appendix Table 8). We transform continuous targets into multi-class targets to facilitate direct comparisons between linear probing and finetuning performance, as linear probes can be inadequate for regression problems. We perform internal evaluation using sex and age prediction on the UKB dataset. For age, we construct five age-ranges with equal sample sizes and perform five-class classification. We leverage multiple additional datasets for external evaluation. SRPBS is a Japanese cohort of patients with schizophrenia, major depressive disorder, and healthy controls (Tanaka et al., 2021). We perform binary classification versus healthy controls for each disorder. ABIDE includes participants with autism-spectrum disorder (ASD) and healthy participants, enabling binary prediction (Craddock et al., 2013; Di Martino et al., 2014). Healthy Brain Network (HBN) is a pediatric, clinical cohort for which we predict scores on language (measured by the CELF) and cognitive (WISC) scales using three bins with matched sample sizes (Alexander et al., 2017). We furthermore use this dataset to test out-of-distribution demographic prediction, as this dataset has no overlap in age with the UKB data. Next, LEMON is a German cohort of healthy participants; we predict scores on a personality scale (MDBF) and two cognitive tasks (CVLT, TMT) again following the construction of three bins with equal sample sizes (Babayan et al., 2019). For comparisons with baselines, we therefore use four demographic prediction tasks, three clinical diagnoses, three cognitive and language scores, and one personality score. To further contrast in- and out-of-distribution scaling performance, we additional use the demographic data from SRPBS and the pediatric ADHD200 dataset (Bellec et al., 2017). Critically, these datasets differ in terms of the participant cohort, acquisition hardware and protocols, as well as data processing. The only time series standardization are the aforementioned z-scoring and resampling (to 2s) transformations.

4.2 IMPLEMENTATION

Training and Model Architecture

We sample temporal views of length $T_{crop} = 100$ timepoints. At a sample level, we apply light augmentations by randomly zeroing out a fraction of channels $\tau_c \sim \mathcal{U}[0, 0.1]$ and contiguous timepoints $\tau_t \sim \mathcal{U}[0, 0.3]$, adding Gaussian noise with $\sigma = \tau_\sigma = 0.1$, and scaling the amplitude by $\tau_s \sim \mathcal{U}[0.8, 1.2]$. The semantic tokenizer G maps an input time series $\mathbf{X} \in \mathbb{R}^{C \times T_{crop}}$ to a token tensor $\mathbf{Z} \in \mathbb{R}^{N \times P \times D_f}$ with a patch length of 20. We define $N = 9$ functional networks, based on the Yeo 7-network parcellation for the cortex (Yeo et al., 2011), with two additional networks comprising all subcortical and cerebellar ROIs, respectively. Within each network-specific tokenizer g_n, the standard and structured convolutional branches each output features of dimension $D_f/2$, which are then concatenated. The standard convolution uses a kernel size of 3. The structured, depthwise-separable convolutions use a base kernel size of 4 across 3 scales with a decay rate of 0.5, creating a receptive field of 16 timepoints. A subsequent linear layer projects the network-specific ROI features to the target dimension for the structured convolutional branch.

The transformer encoder f uses a dimensionality $D_f = 768$ with 8 layers. The projection head h is an MLP with 2 hidden layers ($D_h = 1024$) and an output layer projecting down to 128 dimensions and applying ℓ_2-normalization. The head is shared across all three distillation objectives. We find it suffices to set $\lambda_{TTR} = 0.5$ (i.e., weighting of $\mathcal{L}_{TTR}$) and cosine-decay this weighting to zero over the first 5% of training steps. We use slice-masking with $\lambda_{Tok} = 0.5$ (weighting-factor) and a high masking ratio sampled from $\mathcal{U}[0.65, 0.85]$. Following Wu et al. (2025), we set $\gamma = (D_f + B)/(D_f B)$ where D_f is the feature dimension and B is the batch size. We provide the complete set of optimization hyperparameters in Appendix Table 9. The semantic tokenizer allows pretraining in under two hours on a single GPU with less than 20 GB of memory.

Evaluation

We evaluate representations in two settings: linear probing and full fine-tuning. For linear probing, we freeze the pretrained teacher encoder and train a single linear layer on top of its outputs to assess raw representation quality. For fine-tuning, the entire model is updated to provide a comparison with supervised baselines. All evaluations use a 10-fold cross-validation procedure while stratifying on the target. At test time, we average predictions of 8 equally-spaced temporal crops for all methods. For Brain-Semantoks evaluations and ablations, we pretrain using three random seeds and average their scores for each fold to improve reliability.

The linear probing setup is standardized across models (Appendix Table 9). For Brain-Semantoks, the input to the linear layer is the concatenation of the teacher's `[CLS]` token and the average of all network-patch tokens. For the linear probing comparisons, we omit the LEMON dataset as we only have access to filtered, preprocessed data. The ROI-specific offsets that BrainLM and Brain-JEPA were pretrained on are therefore not present, and we observe chance-level performance of these models on this dataset. We note that our normalization strategy is universally applicable however.

Baselines

The two main baselines we compare against are BrainLM (Caro et al., 2023) and Brain-JEPA (Dong et al., 2024), which are both fMRI time series foundation models using parcellated data. Importantly, both models are pretrained solely on the UKB, which is necessary for informative comparisons. We further compare against strong supervised approaches. The most widely-used method for prediction with fMRI is to compute pairwise Pearson correlations between ROI time series as a measure of 'functional connectivity' (FC) and use a support vector machine for prediction. We select the regularisation strength using the validation set ($C = \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$). Additionally, BolT is a Transformer-based architecture that classifies fMRI time series by using a novel fused window attention mechanism to hierarchically build representations from local to global temporal scales (Bedel et al., 2023). Finally, Brain Network Transformer (BNT) utilizes ROI connection profiles for positional encoding and introduces an Orthonormal Clustering Readout mechanism. This novel pooling function learns cluster-aware embeddings by grouping functionally similar brain regions to improve graph-level predictions (Kan et al., 2022).

5 RESULTS

We evaluated Brain-Semantoks on a diverse set of downstream tasks, assessing its performance via linear probing against state-of-the-art foundation models and fully supervised methods. We then conducted extensive scaling and ablation studies to validate our architectural and training choices, and finally leveraged our model's unique properties for built-in interpretability.

Table 1: Comparison of mean balanced accuracy (%) for fMRI time series foundation models using linear probes. Values are presented as mean ± standard deviation across ten folds.

Model	ABIDE	HBN CELF	HBN WISC	HBN Age	HBN Sex	UKB Age	UKB Sex	SRPBS MDD	SRPBS SZ
BrainLM	53.84 ± 3.00	42.03 ± 3.41	38.26 ± 4.11	**43.89** ± 2.12	65.44 ± 1.72	30.16 ± 1.41	86.71 ± 0.63	57.61 ± 4.14	55.72 ± 6.62
Brain-JEPA	52.92 ± 3.53	41.50 ± 5.16	38.34 ± 3.42	39.81 ± 2.08	63.96 ± 1.45	30.60 ± 2.12	83.23 ± 1.26	52.72 ± 4.18	57.63 ± 3.75
Brain-Semantoks	**65.13** ± 2.14	**42.18** ± 2.80	**40.87** ± 2.43	39.16 ± 0.81	**69.52** ± 0.93	**31.15** ± 1.15	**87.52** ± 0.52	**62.60** ± 4.79	**69.26** ± 3.98

Table 2: Model Performance Comparison with fully supervised and finetuned baselines (Balanced Accuracy (%)).

Model	UKB-Age	UKB-Sex	HBN-Age	HBN-Sex	HBN-CELF	HBN-WISC
FC	27.04 ± 1.51	80.63 ± 0.89	41.81 ± 1.36	66.51 ± 1.42	42.41 ± 2.91	39.79 ± 2.91
BNT	20.48 ± 1.08	77.91 ± 3.37	22.59 ± 4.31	61.74 ± 9.69	42.40 ± 3.98	38.53 ± 2.94
Bolt	26.85 ± 1.59	80.30 ± 0.91	37.67 ± 1.41	65.22 ± 1.23	42.45 ± 1.78	39.53 ± 4.42
BrainLM	30.26 ± 1.66	85.75 ± 0.77	39.31 ± 2.02	64.37 ± 2.32	39.27 ± 4.46	35.34 ± 3.61
Brain-JEPA	30.60 ± 1.60	86.70 ± 1.20	**41.91** ± 2.00	65.57 ± 2.28	39.60 ± 3.50	35.20 ± 3.10
Brain-Semantoks	31.15 ± 1.15	**87.52** ± 0.52	39.16 ± 0.81	**69.52** ± 0.93	42.18 ± 2.80	**40.87** ± 2.43
+ Finetune	**33.91** ± 0.87	87.13 ± 0.57	39.41 ± 3.77	69.31 ± 0.71	**42.59** ± 1.34	40.82 ± 1.51

Model	LEMON-CVLT	LEMON-MDBF	LEMON-TMT	ABIDE	SRPBS-MDD	SRPBS-SZ	Avg
FC	39.49 ± 8.32	32.29 ± 6.09	41.14 ± 6.58	65.12 ± 2.98	60.30 ± 4.65	**71.59** ± 5.84	50.68
BNT	36.76 ± 4.90	37.90 ± 5.12	33.86 ± 6.13	58.38 ± 6.51	57.60 ± 4.57	66.59 ± 5.09	46.23
Bolt	39.54 ± 4.71	37.98 ± 5.82	40.30 ± 7.52	64.89 ± 4.08	59.50 ± 4.52	67.12 ± 6.23	50.11
BrainLM	37.81 ± 6.91	34.05 ± 1.84	35.33 ± 3.78	53.91 ± 2.23	54.29 ± 2.38	60.10 ± 5.79	47.48
Brain-JEPA	30.94 ± 6.20	32.26 ± 6.58	35.48 ± 9.58	52.20 ± 4.00	54.00 ± 4.00	60.50 ± 4.40	47.08
Brain-Semantoks	42.10 ± 4.72	**40.23** ± 5.74	**42.88** ± 4.05	65.13 ± 2.14	62.60 ± 4.79	69.26 ± 3.98	52.72
+ Finetune	**44.36** ± 3.36	38.58 ± 4.65	39.21 ± 1.91	**65.44** ± 1.16	**63.60** ± 2.56	71.05 ± 4.39	**52.95**

5.1 Downstream Performance

A primary goal for a foundation model is to produce representations that are directly useful for downstream tasks without extensive fine-tuning. We therefore prioritize evaluation using a rigorous linear probing protocol, where the pretrained model weights are frozen.

As shown in Table 1, Brain-Semantoks consistently and significantly outperforms existing fMRI foundation models, BrainLM and Brain-JEPA, which are based on reconstruction objectives. Our model achieves the highest mean balanced accuracy on eight of the nine tasks, often by a large margin. The improvements are particularly striking on challenging out-of-distribution clinical datasets, where Brain-Semantoks achieves large performance gains for predicting ASD, Schizophrenia, and MDD.

We next compared the linear probing performance of Brain-Semantoks to fully fine-tuned models and strong supervised baselines in Table 2. With only a linear probe, Brain-Semantoks outperforms all baselines on eight diverse tasks. The ability to surpass fully supervised models, which are trained end-to-end on task-specific data, highlights the utility of the representations learned by our pretraining objective.

5.2 Scaling Laws of Semantic FMRI Representations

We conducted the first detailed scaling analysis for fMRI foundation models under a linear probing protocol to understand how performance varies with pretraining data size. We trained Brain-Semantoks on subsets of the UKB dataset and evaluated on both in-distribution (UKB hold-out) and out-of-distribution tasks.

As shown in Figure 2, performance on nearly all tasks improves predictably with the logarithm of the pretraining data size, following a power-law relationship characteristic of foundation models in other domains. Critically, we observe strong scaling laws for out-of-distribution (OOD) generalization. For age and sex prediction ($n_{train,probe} = 500$), which enable comparisons for matched prediction tasks, we observe consistent performance increases with more pretraining data. Remarkably, we note strong and reliable scaling for HBN, which has an age gap of more than 20 years with the UKB (HBN: up to 22, UKB: from 44 years old). Across the majority of tasks, we observe no plateau in OOD scaling.

Yet, in contrast to fields such as language processing, downstream probing performance is not only affected by scaling, but also by the baseline performance. We found that this starting point is significantly increased due to the neuroscientific inductive biases in Brain-Semantoks: a randomly initialized Brain-Semantoks model significantly outperforms a randomly initialized baseline with a ROI-level projection layer by up to 12%.

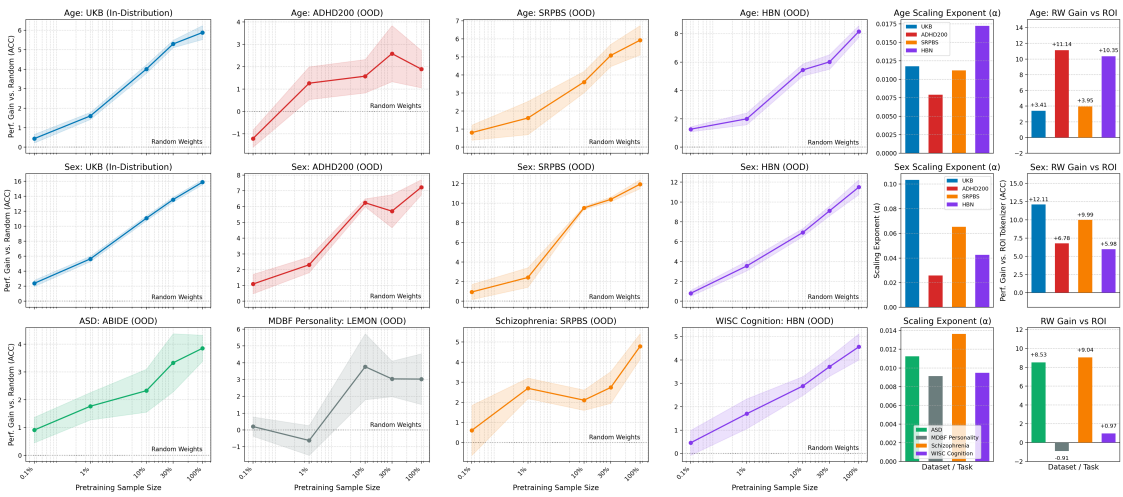

Figure 2: Scaling performance of linear probing following pretraining on increasing sample sizes. We compare within and out-of-distribution scaling. RW: Random weights. ROI: We compare to using a linear layer for projecting single-ROI timeseries instead of our semantic tokenizer.

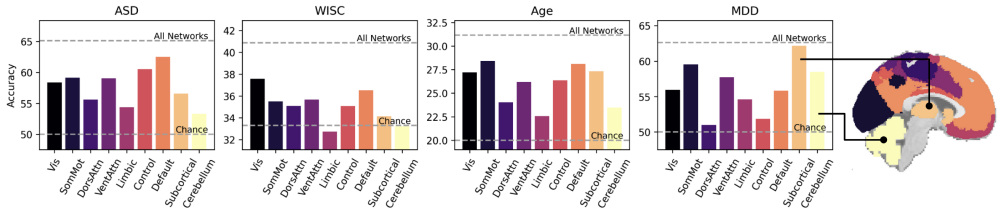

Figure 3: We investigate the predictive performance of individual network representations by performing all-but-one network masking using linear probing.

5.3 Interpretability

A common challenge for interpreting deep models is that post-hoc analyses, like input masking, shift the data out of the training distribution, potentially yielding unreliable results. Our distillation pretraining with slice-masking directly addresses this. Because the model is trained to predict global information from seeing only a subset of brain networks, we can probe its learned dependencies in an ”in-distribution” manner.

We assessed the importance of each of the 9 functional networks for various downstream tasks by masking all but one network and evaluating linear probing performance (Figure 3). This reveals which individual networks contain the most predictive information for a given phenotype. Multiple findings align well with neuroscientific research such as the importance of the default-mode network for ASD and subcortical regions for MDD (Ramasubbu et al., 2014). Interestingly, whereas the default-mode network has dominated MDD research, we find that cerebellar activity is more predictive, which is a more recent hypothesis (Wang et al., 2023).

6 Ablations

We perform ablation studies on core aspects of the Brain-Semantoks framework. We average linear probing performance across ten downstream task, omitting HBN-Age and HBN-Sex to reduce the influence of demographic prediction on the total score. First, we compare using a linear projection layer operating on single ROIs (as in Brain-JEPA), instead of our semantic tokenizer (Figure 4A). We observe unstable training dynamics, as the cosine similarity between the student’s reconstructed tokens and teacher tokens (i.e., the negative of the token-level loss) to quickly reach 0.95 and stabilize there (Figure 4C). This indicates partial collapse where the learned representations are simple and can be predicted at high accuracy immediately, which is associated with poor downstream performance. Indeed, we find large gains in downstream performance by adopting the semantic tokenizer, which also results in improved training dynamics, although instability still exists early in training. By including Teacher-guided Temporal Regulari-

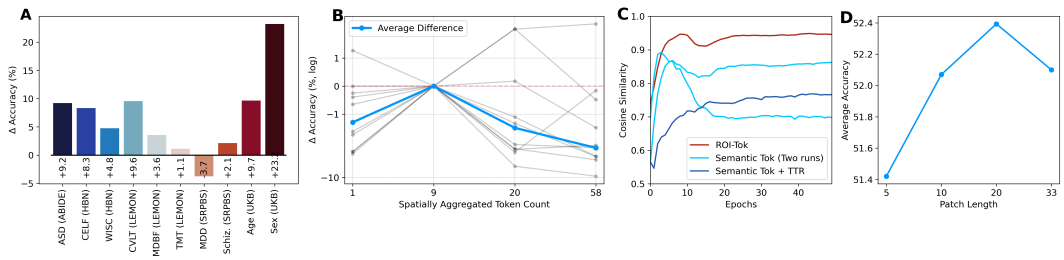

Figure 4: Tokenizer Ablations. A) Linear probing performance comparison after pretraining with our semantic tokenizer vs (minus) a linear projection layer applied to ROIs individually. B) We ablate aggregating into different numbers of 'networks' or spatial tokens. Grey lines indicate individual downstream tasks. C) Pretraining dynamics visualised by the cosine similarity between the teacher tokens and reconstructed student tokens. D) Temporal patch size ablation. TTR: Teacher-guided Temporal Regularizer

Table 3: Convolution type for tokenizer

Kernel	Score
Full kernel (20)	48.28
Short Kernel (3)	50.50
Structured Conv	52.19
Short + Struct	**52.39**

Table 4: TTR Duration

Duration	Score
0%	50.88
5%	**52.39**
100%	49.60

Table 5: Masking Type

Masking Type	Score
Random	51.03
Block	49.71
Network	52.09
Temporal Slice	51.50
Slice	**52.39**

Table 6: Loss Components

CLS	Mask w	Score
No	1.0	47.32
Yes	0.0	50.10
Yes	**0.5**	**52.39**
Yes	1.0	51.62

Table 7: Masking Ratio

Ratio	Score
$[0.1, 0.9]$	50.33
$[0.45, 0.55]$	51.23
$[0.5, 0.75]$	51.88
[0.65, 0.85]	**52.39**

Ablation studies using average balanced accuracy for linear probing across ten downstream tasks. w: Weight on masking loss. TTR Duration is noted as pretraining period during which TTR is active.

sation (TTR), which we decay to zero in the first 5% of training, we observe stable pretraining dynamics and strong downstream performance. We find using TTR for the entirety of pretraining to be overly restrictive (Table 4).

We furthermore ablate the choice of nine functional networks for the semantic tokenizer and we compare to spatially aggregating more aggressively (into 1 spatial token per temporal patch) or less so (20 or 58 spatial tokens; Figure 4B; see appendix Table 10 for a detailed description). We observe a significant bias for most downstream tasks towards fewer spatial tokens, with the best overall performance for the nine network solution. We also ablate the temporal patch size and convolutional filter bank for the tokenizer, finding that relatively long patches and structured convolutions are important (Table 3).

Finally, we provide ablations on masking. We find that the influence of the mask loss should not be too high (Table 6), masking types which reduce interpolation learning are most effective (Table 5), combined with a high masking ratio (Table 7).

7 Discussion

This paper presents Brain-Semantoks, a novel foundation model for fMRI that marks a significant shift to learning abstract, semantic representations of brain dynamics. By introducing a neuroscientifically-grounded semantic tokenizer and employing a self-distillation objective, the model effectively learns high-level phenotypic signatures. The results demonstrate the strength of this approach, achieving state-of-the-art performance under a rigorous linear probing protocol and often surpassing supervised methods on diverse tasks. This indicates the learned representations are broadly applicable without domain adaptation.

Future work may benefit from including task-based fMRI data, as our approach here only relied on resting-state data. Furthermore, while we find that using neuroscience-based functional networks is effective for many downstream tasks, follow-up research will explore learning how to group ROIs from data rather than having them fixed. Finally, investigations to understand which distribution shifts between pretraining and downstream data are particularly harmful may help explain why Brain-Semantoks performs better or worse on some tasks and provide insight in how to best address them.

Ethics Statement

This research utilized publicly available, pre-existing neuroimaging datasets, including the UK Biobank, ABIDE, HBN, SRPBS, LEMON, and ADHD200. No new data were collected for this study. All data were acquired in accordance with the data use agreements specific to each dataset, such as the UK Biobank application process. The original data collection procedures for these cohorts were conducted under the approval of their respective institutional review boards (IRBs) and ethics committees, with all participants providing informed consent. We have complied with all data privacy and sharing agreements, ensuring that sensitive information is stored securely and handled according to the established protocols.

Reproducibility Statement

To ensure the reproducibility of our findings, we will release all source code. Furthermore, the weights of the pre-trained Brain-Semantoks models will be made publicly available upon publication. A comprehensive description of our methodology, including model architecture, training procedures, and evaluation protocols, is provided in the Experimental Setup (Section 3). Specific hyperparameter configurations for all stages of our experiments are detailed in Appendix Table 9. All datasets used in this work are publicly accessible, and relevant citations are provided to guide their acquisition. These resources are intended to allow for the full verification of our results and facilitate future research building upon our work.

LLM Usage Statement

The authors utilized large language models (LLMs) during the preparation of this manuscript. Specifically, Google's Gemini was used for proofreading, correcting grammatical errors, and improving sentence structure for clarity. All LLM-generated suggestions were manually reviewed and edited by the authors to ensure the final text accurately reflects our research and claims. The core scientific contributions and claims were solely authored by the human authors.

References

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. *Scientific data*, 4(1):1–26, 2017.

Elena A Allen, Eswar Damaraju, Sergey M Plis, Erik B Erhardt, Tom Eichele, and Vince D Calhoun. Tracking whole-brain connectivity dynamics in the resting state. *Cerebral cortex*, 24(3):663–676, 2014.

Anahit Babayan, Miray Erbey, Deniz Kumral, Janis D Reinelt, Andrea MF Reiter, Josefin Röbbig, H Lina Schaare, Marie Uhlig, Alfred Anwander, Pierre-Louis Bazin, et al. A mind-brain-body dataset of mri, eeg, cognition, emotion, and peripheral physiology in young and old adults. *Scientific data*, 6(1):1–21, 2019.

Hasan A Bedel, Irmak Sivgin, Onat Dalmaz, Salman UH Dar, and Tolga Çukur. Bolt: Fused window transformers for fmri time series analysis. *Medical image analysis*, 88:102841, 2023.

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.

Randy L Buckner, Fenna M Krienen, Angela Castellanos, Julio C Diaz, and BT Thomas Yeo. The organization of the human cerebellum estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 106(5):2322–2345, 2011.

Katherine S Button, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience*, 14(5):365–376, 2013.

Josue Ortega Caro, Antonio H de O Fonseca, Christopher Averill, Syed A Rizvi, Matteo Rosati, James L Cross, Prateek Mittal, Emanuele Zappala, Daniel Levine, Rahul M Dhodapkar, et al. Brainlm: A foundation model for brain activity recordings. *bioRxiv*, pp. 2023–09, 2023.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Cameron Craddock, Yassine Benhajali, Carlton Chu, Francois Chouinard, Alan Evans, András Jakab, Budhachandra Singh Khundrakpam, John David Lewis, Qingyang Li, Michael Milham, et al. The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Frontiers in Neuroinformatics*, 7(27):5, 2013.

Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667, 2014.

Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Chong, Fang Ji, Nathanael Tong, Christopher Chen, and Juan Helen Zhou. Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. *Advances in Neural Information Processing Systems*, 37:86048–86073, 2024.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Kuan Han, Haiguang Wen, Junxing Shi, Kun-Han Lu, Yizhen Zhang, Di Fu, and Zhongming Liu. Variational autoencoder: An unsupervised model for encoding and decoding fmri activity in visual cortex. *NeuroImage*, 198:125–136, 2019.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Xuan Kan, Wei Dai, Hejie Cui, Zilong Zhang, Ying Guo, and Carl Yang. Brain network transformer. *Advances in Neural Information Processing Systems*, 35:25586–25599, 2022.

Jung-Hoon Kim, Yizhen Zhang, Kuan Han, Zheyu Wen, Minkyu Choi, and Zhongming Liu. Representation learning of resting state fmri with variational autoencoder. *NeuroImage*, 241:118423, 2021.

Yuhong Li, Tianle Cai, Yi Zhang, Deming Chen, and Debadeepta Dey. What makes convolutional models great on long sequence modeling? *arXiv preprint arXiv:2210.09298*, 2022.

Nikos K Logothetis. What we can do and what we cannot do with fmri. *Nature*, 453(7197):869–878, 2008.

Karla L Miller, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, Saad Jbabdi, Stamatios N Sotiropoulos, Jesper LR Andersson, et al. Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523–1536, 2016.

Seiji Ogawa, Tso-Ming Lee, Alan R Kay, and David W Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *proceedings of the National Academy of Sciences*, 87(24):9868–9872, 1990.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Walter HL Pinaya, Andrea Mechelli, and João R Sato. Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: A large-scale multi-sample study. *Human brain mapping*, 40(3):944–954, 2019.

Russell A Poldrack, Chris I Baker, Joke Durnez, Krzysztof J Gorgolewski, Paul M Matthews, Marcus R Munafò, Thomas E Nichols, Jean-Baptiste Poline, Edward Vul, and Tal Yarkoni. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature reviews neuroscience*, 18(2):115–126, 2017.

Rajamannar Ramasubbu, Nithya Konduru, Filomeno Cortese, Signe Bray, Ismael Gaxiola-Valdez, and Bradley Goodyear. Reduced intrinsic connectivity of amygdala in adults with major depressive disorder. *Frontiers in psychiatry*, 5:17, 2014.

Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and BT Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral cortex*, 28(9):3095–3114, 2018.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Saori C Tanaka, Ayumu Yamashita, Noriaki Yahata, Takashi Itahashi, Giuseppe Lisi, Takashi Yamada, Naho Ichikawa, Masahiro Takamura, Yujiro Yoshihara, Akira Kunimatsu, et al. A multi-site, multi-disorder resting-state magnetic resonance image database. *Scientific data*, 8(1):227, 2021.

Ye Tian, Daniel S Margulies, Michael Breakspear, and Andrew Zalesky. Topographic organization of the human subcortex unveiled with functional connectivity gradients. *Nature neuroscience*, 23(11):1421–1432, 2020.

Diego Vidaurre, Stephen M Smith, and Mark W Woolrich. Brain network dynamics are hierarchically organized in time. *Proceedings of the National Academy of Sciences*, 114(48):12827–12832, 2017.

Cheng Wang, Yu Jiang, Zhihao Peng, Chenxin Li, Changbae Bang, Lin Zhao, Jinglei Lv, Jorge Sepulcre, Carl Yang, Lifang He, et al. Towards a general-purpose foundation model for fmri analysis. *arXiv preprint arXiv:2506.11167*, 2025.

Xiang Wang, Jie Xia, Weiyan Wang, Jingjie Lu, Qian Liu, Jie Fan, Tamini Soondrum, Quanhao Yu, Changlian Tan, and Xiongzhao Zhu. Disrupted functional connectivity of the cerebellum with default mode and frontoparietal networks in young adults with major depressive disorder. *Psychiatry research*, 324:115192, 2023.

Ziyang Wu, Jingyuan Zhang, Druv Pai, XuDong Wang, Chandan Singh, Jianwei Yang, Jianfeng Gao, and Yi Ma. Simplifying dino via coding rate regularization. *arXiv preprint arXiv:2502.10385*, 2025.

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 transactions on medical imaging*, 43(11):4004–4016, 2024.

BT Thomas Yeo, Fenna M Krienen, Jorge Sepulcre, Mert R Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L Roffman, Jordan W Smoller, Lilla Zöllei, Jonathan R Polimeni, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 2011.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

A Appendix

Table 8: Summary of datasets, tasks, and sample sizes for downstream evaluations. "Tr/V/T" stands for Train/Validation/Test.

Dataset	Task	Classes	Total Size	Split Ratio (Tr/V/T)	Train Size
UKB	Sex	2	1625	0.31 / 0.08 / 0.61	500
UKB	Age	5	1625	0.31 / 0.08 / 0.61	500
HBN	Sex	2	1870	0.27 / 0.27 / 0.46	500
HBN	Age	5	1870	0.11 / 0.11 / 0.78	200
HBN	WISC FSIQ	3	884	0.60 / 0.20 / 0.20	530
HBN	CELF Total	3	1005	0.60 / 0.20 / 0.20	603
LEMON	CVLT	3	212	0.60 / 0.20 / 0.20	127
LEMON	TMT B-A	3	212	0.60 / 0.20 / 0.20	127
LEMON	MDBF	3	213	0.60 / 0.20 / 0.20	127
SRPBS	Schizophrenia	2	291	0.60 / 0.20 / 0.20	174
SRPBS	MDD	2	499	0.60 / 0.20 / 0.20	299
ABIDE	Autism	2	974	0.60 / 0.20 / 0.20	584

Table 9: Hyperparameter settings for all experimental stages.

(a) Pre-training

Hyperparameter	Value
Optimizer	AdamW
Base LR	0.0007
Epochs	100
Patch Size	20
Crop Length	100
Teacher Momentum	0.99
Weight Decay	$0.05 \rightarrow 0.3$
Batch Size	512
Warmup Ratio	3% (Linear)
LR Schedule	Cosine Decay
Layer Scale Init	0.1

(b) Fine-tuning

Hyperparameter	Value
Head Type	Linear Layer
Optimizer	AdamW
Base LR	0.0001
Epochs	50
LR Schedule	Cosine Decay
Warmup	None
Batch Size	16
Weight Decay	0.05
LR Decay Rate	0.9

(c) Linear Probing

Hyperparameter	Value
Head Type	BN + Linear
Optimizer	SGD
Momentum	0.9
Learning Rate	Fixed (best sel.)[a]
LR Schedule	None
Epochs	50
Batch Size	$\min(256, n/8)$

[a] We fit a linear layer for each of the following learning rates in parallel and choose the best one based on the validation data for test set evaluation: {0.03, 0.01, 0.003, 0.001, 0.0003, 0.0001}

Table 10: Ablation of Spatial Token Aggregation Strategies

Spatial Tokens	Aggregation Strategy
1	All 457 ROIs are aggregated into a single group.
9	7 Yeo networks + 1 subcortical group + 1 cerebellum group.
20	17 Yeo networks + 2 manually split subcortical groups + 1 cerebellum group.
58	17 Yeo networks (each split 3 times) + 6 manually split subcortical groups + 1 cerebellum group.*

*Note: The cerebellum was not split as the atlas contains only 7 ROIs for this region.