Brain-JEPA: Brain Dynamics Foundation Model with Gradient Positioning and Spatiotemporal Masking

Zijian Dong[*], **Ruilin Li**[*], **Yilei Wu**, **Thuan Tinh Nguyen**, **Joanna Su Xian Chong**, **Fang Ji**, **Nathanael Ren Jie Tong**, **Christopher Li Hsian Chen**, **Juan Helen Zhou**[†]

zijian.dong@u.nus.edu, {li.rl, helen.zhou}@nus.edu.sg

National University of Singapore

Abstract

We introduce *Brain-JEPA*, a brain dynamics foundation model with the Joint-Embedding Predictive Architecture (JEPA). This pioneering model achieves state-of-the-art performance in demographic prediction, disease diagnosis/prognosis, and trait prediction through fine-tuning. Furthermore, it excels in off-the-shelf evaluations (*e.g.*, linear probing) and demonstrates superior generalizability across different ethnic groups, surpassing the previous large model for brain activity significantly. Brain-JEPA incorporates two innovative techniques: **Brain Gradient Positioning** and **Spatiotemporal Masking**. Brain Gradient Positioning introduces a functional coordinate system for brain functional parcellation, enhancing the positional encoding of different Regions of Interest (ROIs). Spatiotemporal Masking, tailored to the unique characteristics of fMRI data, addresses the challenge of heterogeneous time-series patches. These methodologies enhance model performance and advance our understanding of the neural circuits underlying cognition. Overall, Brain-JEPA is paving the way to address pivotal questions of building brain functional coordinate system and masking brain activity at the AI-neuroscience interface, and setting a potentially new paradigm in brain activity analysis through downstream adaptation. *Code is available at: https://github.com/Eric-LRL/Brain-JEPA.*

1 Introduction

Understanding large-scale brain activity data is crucial for deciphering the complex mechanisms underlying cognitive processes and human behavior. Functional magnetic resonance imaging (fMRI) captures blood-oxygen-level dependent (BOLD) signals that reflect regional brain activity. It emerges as an indispensable tool in neuroscience for identifying the neural bases of cognitive processes [1, 2, 3]. Deep learning approaches have been developed for fMRI analysis, improving brain disease diagnosis and deepening insights into cognition and behavior [4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. Despite notable advances, these task-specific models suffer from limited generalizability and adaptability to other tasks. In addition, they fail to leverage the vast amounts of unlabeled fMRI data available [14, 15].

Artificial intelligence (AI) is experiencing a paradigm shift from task-specific training to building foundation models that are trained on extensive data using self-supervision at scale [16]. Unlike the models with singular functions, foundation models can be adapted to a diverse array of downstream tasks. Large language models such as GPT [17] and LLaMA [18] have shown significant potential in natural language processing, with expansive applications in healthcare, biomedicine, and beyond [16].

[*] *Equal contribution*

[†] *Corresponding author*

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-2732

In the field of fMRI time series analysis, brain language model (brainLM) is one of the most representative foundation models [14]. It is a masked autoencoder (MAE) [19] trained to reconstruct masked fMRI time series. However, as an indirect measure of neuronal activity, the BOLD signal has a relatively low signal-to-noise ratio (SNR), which is influenced by a mixture of factors and distorted by non-neuronal fluctuations [20]. Filling every bit of the fMRI time series as in brainLM can hinder the model's ability to distinguish between noise and actual signals. This can result in either amplifying noise or missing critical subtle variations in brain activity. Unlike natural images, which have high information density with structures such as edges and colors, fMRI data has spatiotemporally sparser signals distributed across brain volumes without clear boundaries, making it difficult to accurately reconstruct signal of masked regions of interest (ROIs). Furthermore, it has been widely shown that masked pretraining in generative architectures such as MAE leads to suboptimal performance in off-the-shelf evaluations (*e.g.*, linear probing) [21]. Because of that, BrainLM requires computationally intensive end-to-end finetuning, with a three-layer MLP attached to the pretrained encoder, to achieve optimal performance. Furthermore, the absence of comparisons with state-of-the-art methods for downstream task performance and the focus only on Caucasian cohorts limit BrainLM's applicability in clinical settings.

Therefore, rather than focusing on the original brain activity time series, the inherent noise and sparse information density of fMRI lead us to explore the latent space of fMRI time series extracted from a strong encoder (*e.g.*, Vision Transformer (ViT) [22]). It potentially offers a higher SNR after "compression", achieving a greater level of abstraction that captures subtle yet crucial patterns [23]. Recently, Imaged-based Joint-Embedding Predictive Architecture (I-JEPA) has been proposed as a non-generative architecture for self-supervised learning from images [21]. It predicts the representations of various target blocks rather than reconstructing the masked input like MAE during pretraining. By predicting representations in the latent space, I-JEPA enhances the semantic quality of learned representations and boosts scalability and efficiency.

Training a brain dynamics foundation model using a JEPA-like architecture might offer advantages over the MAE approach. However, the distinct spatiotemporal characteristics of fMRI data make direct application of the JEPA architecture suboptimal: ❶ Positional embeddings in transformer play a crucial role by incorporating information about the order or position of tokens in the input data (*e.g.*, the order of different words in a sentence or the locations of pixels in an image) [24]. However, there is no such natural "order" for different ROIs across the 3D brain volume in fMRI. BrainLM utilizes anatomical positions to label each ROI [14], yet it does not account for brain functional parcellation, where nearby anatomical ROIs might exhibit rather different brain activation patterns represented by a lack of local coherence in fMRI data [25]. ❷ I-JEPA employs a random multi-block selection of context and target. However, unlike images, fMRI presents complex patterns across both spatial and temporal domains. Given the smaller sample size and sparser information density in fMRI datasets compared to datasets like ImageNet [26], learning in fMRI requires a stronger inductive bias. This would enhance the efficiency of training models by better capturing the underlying patterns specific to brain activity. Given the unique challenges presented by fMRI data, there is a pressing need to develop a functinal coordinate system and a tailored masking strategy for large-scale pretraining on fMRI data.

These neglected yet crucial questions of developing a functional coordinate system and a masking strategy for large-scale pretraining with fMRI data, lie at the intersection of AI and neuroscience, highlighting important interdisciplinary challenges.

To address these gaps, here we introduce *Brain-JEPA*, a brain dynamics foundation model with the Joint-Embedding Predictive Architecture (JEPA). Instead of reconstructing masked inputs during pretraining, Brain-JEPA predicts abstract representations of sampled targets from the observation. We propose two innovative techniques to enhance model performance and address key questions in AI for neuroscience: First, **Brain Gradient Positioning** provides a brain functional coordinate system for positional embedding of brain functional parcellation (Section 3.1). Second, **Spatiotemporal Masking** offers a tailored masking strategy for the heterogeneous time-series patches inherent in fMRI (Section 3.2). Moreover, in downstream experiments, our proposed Brain-JEPA achieves state-of-the-art results in demographic prediction, disease diagnosis/prognosis, and trait prediction through fine-tuning. It also excels in off-the-shelf evaluations (*e.g.*, linear probing), and shows superior generalizability across different ethnic groups. Brain-JEPA enhances brain activity analysis and deepens our understanding of critical AI-neuroscience questions related to constructing functional coordinate systems and developing spatiotemporal masking strategies.

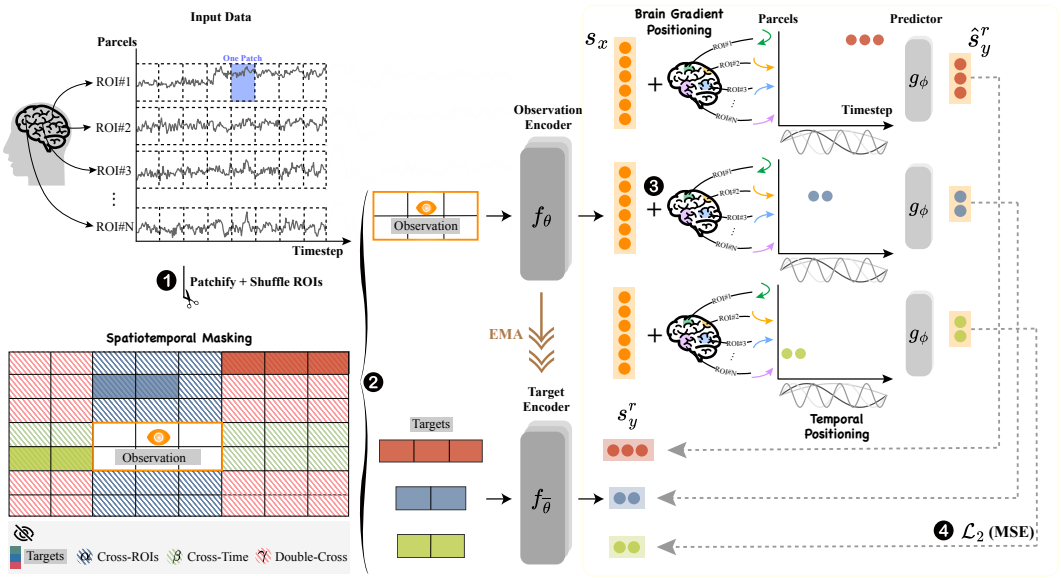

Figure 1: **Brain-JEPA.** With a Vision Transformer (ViT) as the observation encoder f_θ, Brain-JEPA employs a single observation block to predict the representations of target blocks. **(1)** The input fMRI data is initially segmented into patches for subsequent processing. **(2)** Through Spatiotemporal Masking, the input data—excluding the observation block—is divided into three distinct regions: Cross-ROI (α), Cross-Time (β), and Double-Cross (γ). The target blocks are sampled from different regions separately. **(3)** A narrower ViT, serving as the predictor g_ϕ, takes the output $\boldsymbol{s}_x$ from f_θ. It predicts the representations of a target block $\hat{\boldsymbol{s}}_y^r$ conditioned on positional embedding (brain gradient positioning for ROI locations and sine and cosine functions for temporal positioning). **(4)** These predicted representations align with those $\boldsymbol{s}_y^r$ from the target encoder $f_{\bar{\theta}}$, whose parameters are incrementally updated through an Exponential Moving Average (EMA) of the observation encoder's parameters.

2 Related Work

Task-specific Models for fMRI (state-of-the-art). SVR and MLP have been used in fMRI analysis, utilizing Pearson correlation matrices derived from fMRI time series as input [8, 9]. Deep learning models have substantially advanced fMRI analysis in recent years. BrainNetCNN [6] introduces a convolutional neural network (CNN) with specialized convolutional filters tailored for brain network. BrainGNN [5] utilizes ROI-aware graph neural networks (GNNs) to effectively harness functional brain network information, incorporating a pooling operator to highlight key ROIs. More recently, Brain network transformer (BNT) [4] employs transformer encoders to generate embeddings for ROIs based on Pearson correlation matrices, alongside a readout layer designed to identify clusters within the brain. Swift [27] applies Swin Transformer architecture [28] to process brain functional data. As noted in Section 1, these task-specific models have limited generalizability and adaptability across different tasks, and fail to utilize extensive unlabeled fMRI data.

The fMRI Foundation Model. BrainLM [14] stands out as the first fMRI foundation model, employing MAE for self-supervised pretraining of fMRI data. In this approach, fMRI time series are treated as images and patchified. The training goal is to reconstruct the masked patches of the time series. As outlined in Section 1, BrainLM exhibits several limitations: 1) Direct reconstruction of masked input may not be suitable for inherently noisy data with low information density, such as fMRI. It complicates the differentiation between noise and signal, making it difficult to capture underlying patterns. 2) Generative architectures like MAE result in suboptimal performance in linear probing, a critical method for evaluating learned representations. 3) The absence of comparisons with state-of-the-art models and evaluations limited to Caucasian cohorts restricts its broader applicability. BrainMass [29], a concurrent work in large-scale self-supervised learning for neuroimaging, focuses on brain network analysis rather than brain dynamics, distinguishing it from our research.

https://doi.org/10.52202/079017-2732

3 Method

In this section, we outline the methodology of Brain-JEPA. Instead of reconstructing masked patches of fMRI time series, Brain-JEPA operates in the latent space, as depicted in Figure 1. With the observation block excluded, the input data is divided into three non-overlapping regions: Cross-Time (α), Cross-ROI (β), and Double-Cross (γ). This division forces the model to engage in forecasting time series, generalizing across unseen ROIs, and predicting time series for unseen ROIs. Section 3.1 details the **Brain Gradient Positioning** we proposed, which encodes the functional relationships among different ROIs, serving as a brain functional coordinate system in the brain's functional organization. In Section 3.2, we introduce **Spatiotemporal Masking**, which injects a strong inductive bias during the masking process, leading to faster convergence during pretraining and superior performance in downstream tasks.

3.1 Brain Gradient Positioning

We propose Brain Gradient Positioning, which provides a brain functional coordinates system based on the functional connectivity gradient. Positional embeddings are crucial in transformer architectures, as they encode information about the positions of tokens in a sequence. These embeddings can be implemented using fixed sine and cosine functions across various frequencies [24] or through learnable embeddings that adapt during training [30]. However, the integration of positional information into fMRI time series has long been neglected. FMRI data, incorporating complex spatiotemporal information, requires separate consideration of its temporal and spatial domains. The temporal domain, representing timesteps during scanning, is well-suited for conventional sine and cosine positional embeddings, as the time series in each ROI is sequentially ordered by time. However, this method is not appropriate for the spatial domain, where ROIs across brain volumes lack a simple, inherent order, making sine and cosine embeddings unsuitable for capturing spatial relationships. Anatomical locations of ROIs offer an alternative to sine and cosine functions [14] but fall short in capturing functional parcellation. Spatially adjacent ROIs can exhibit significantly different brain activation patterns, reflecting the inherent lack of local coherence in fMRI data [25].

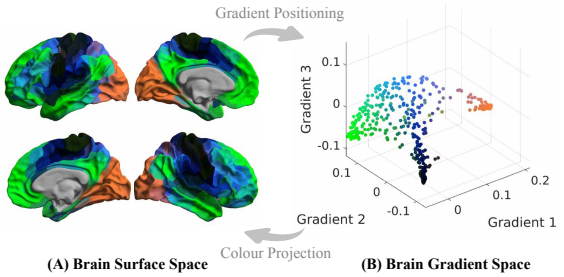

Figure 2: **Brain gradient positioning.** Brain cortical regions are situated in the top 3 gradient axes and colored based on their positions. These colors are then projected back into the brain surface.

The functional connectivity gradient is a continuous measure that captures the functional relations among different ROIs. Each attribute in the gradient represents an axis in the latent space of brain regions and networks. The relative distance between different ROIs indicates the similarity in their connectivity (*i.e.*, shorter distance means higher similarity in connectivity). The concept of a spatial gradient as conceptualized by Mesulam in 1998 entailed a synaptic hierarchy that supports cognitive processes [31]. Recent studies have built upon this concept, revealing that brain networks in adult humans and macaques exhibit linear distributions across different gradient axes [32]. Using this methodology, it has been shown that these gradients reflect the functional changes related to age [33, 34, 35, 36, 37], cognition [37, 38] and brain diseases [35, 39, 40]. These gradients together provide a framework to assess the relationship between brain regions based on their relative positioning across different gradient axes.

Before deriving the gradients, we first calculate a non-negative affinity matrix $\boldsymbol{A}(i,j)$ (a graph Laplacian) as follows:

$$\boldsymbol{A}(i,j) = 1 - \frac{1}{\pi}\cos^{-1}\left(\frac{\mathbf{c}_i \mathbf{c}_j^T}{\|\mathbf{c}_i\|\|\mathbf{c}_j\|}\right) \tag{1}$$

where $\boldsymbol{c}_i$ and $\boldsymbol{c}_j$ represents the features (functional connectivity) across the ROI i and j, respectively. Gradients are then derived using diffusion map [41, 42], a nonlinear dimension reduction method used to identify the underlying manifold structure of the data. We can obtain the diffusion matrix $\boldsymbol{L}_\delta$ and the diffusion operator $\boldsymbol{M}_\delta$ from $\boldsymbol{A}$ as follows:

$$\boldsymbol{M}_\delta = \boldsymbol{D}^{-1}\boldsymbol{L}_\delta, \boldsymbol{L}_\delta = \boldsymbol{D}^{-\frac{1}{\delta}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{\delta}} \tag{2}$$

where $\boldsymbol{D}$ is the degree matrix of $\boldsymbol{A}$. Here δ is set to 0.5 to maintain the global relations between ROIs in the embedding space.

Finally, we can compute the eigenvectors and eigenvalues of $\boldsymbol{M}_\delta$, and stack the column vectors to formulate the diffusion map $\boldsymbol{\Phi}_t \in \mathbb{R}^{n\times m}$ (n ROIs in total, with m gradients for each) at time t and the gradient matrix $\boldsymbol{G}$ with the same dimension:

$$\boldsymbol{\Phi}_t = [\lambda_1^t\psi_1, \lambda_2^t\psi_2, ..., \lambda_m^t\psi_m], \boldsymbol{G} = [\psi_1, \psi_2, ..., \psi_m] \tag{3}$$

where λ_k are the eigenvalues and ψ_k are the corresponding eigenvectors (gradients) of the graph Laplacian. The parameter t represents the diffusion time, which controls the scale of the diffusion process. Here we estimated the eigenvalues λ_k at time t by dividing it by $1-\lambda_k$ to enhance robustness against noisy eigenvalues.

In Brain-JEPA, we leverage $\boldsymbol{G}$ as the spatial positioning of ROIs. Specifically, the gradient $\boldsymbol{G} \in \mathbb{R}^{n\times m}$ is transformed into $\hat{\boldsymbol{G}} \in \mathbb{R}^{n\times d/2}$ through a trainable linear layer, where d represents the embedding dimension of the ViT backbone. The predefined temporal positioning $\boldsymbol{T} \in \mathbb{R}^{n\times d/2}$ is obtained using sine and cosine functions [43]. The final positional embedding can then be formulated as $\boldsymbol{P} = [\boldsymbol{T}, \hat{\boldsymbol{G}}] \in \mathbb{R}^{n\times d}$. Figure 2 provides a visualization of the top 3 gradients in Euclidean space with each ROI color coded by their locations. As shown, the brain gradient positioning reflects functional network architecture, such as the **somatomotor**, **default mode** and **visual** networks, consistent with previous literature [32, 37].

3.2 Spatiotemporal Masking

Observation. Brain-JEPA aims to predict representations of multiple target blocks based on the representation of a single observation block. For an input fMRI time series, the temporal signal for each parcel is divided into patches after shuffling ROIs, each containing p time points (dash boxes in Figure 1). The observation block $\boldsymbol{x}$ is obtained by randomly sampling a block within the range $\{\eta_R^o, \eta_T^o\}$. η_R^o specifies the range ratio along the ROI dimension, and η_T^o pertains to the timestep patches (10 in total). Subsequently, $\boldsymbol{x}$ is fed through the observation encoder f_θ, generating a corresponding patch-level representation $\boldsymbol{s}_x$:

$$\boldsymbol{s}_x = \{\boldsymbol{s}_{x_j}\}_{j\in\mathcal{B}_x} \tag{4}$$

where $\mathcal{B}_x$ represents the mask associated with the observation block $\boldsymbol{x}$, $\boldsymbol{s}_{x_j}$ is the representation of the j^{th} patch.

Targets. Given a single observation, the model is trained to predict other parts of the fMRI within the latent space. Random sampling of targets like MAE [19] might allow the model to learn shortcuts (*e.g.*, interpolation of time series) or rely heavily on simpler, more frequent patterns in the data, which could limit its generalizability. It is crucial to recognize that patches in fMRI vary spatially depending on the positions in their brain functional organization, and temporally regarding brain states and task conditions. The nonlinear relationship among brain networks further complicates the interactions between different brain patches.

As shown in Figure 1, we categorize the remaining parts (with the observation excluded) into three distinct and non-overlapping regions: Cross-ROI (α), Cross-Time (β), and Double-Cross (γ). For targets in the α and β regions, the model should generalize the observation across different ROIs spatially or timesteps temporally. For targets in the γ regions, which are the most challenging, the model should generalize to unseen ROIs at unencountered timesteps. We randomly sample K blocks from each of the three types of regions as targets, forcing the model to handle a variety of prediction tasks with a stronger inductive bias. We denote the mask corresponding of the region r ($r \in \{\alpha, \beta, \gamma\}$) as $\mathcal{B}_y^r$.

Overlapped sampling. It has been shown in [21] that a sufficiently large dynamic range of masking ratio could benefit pretraining. To effectively adjust the observation-to-input ratio during pretraining, we implement an overlapped sampling strategy that allows for a flexible, rather than fixed, ratio. When sampling the target block $\boldsymbol{s}_y^r$ from region r, for $r = \alpha$ or β, we sample the target from the union of the observation mask and region r mask; while for $r = \gamma$, we directly sample the target from the γ region mask. Formally, the overlapped sampling strategy is defined as:

$$\boldsymbol{s}_y^\alpha \sim \mathcal{B}_x \cup \mathcal{B}_y^\alpha, \boldsymbol{s}_y^\beta \sim \mathcal{B}_x \cup \mathcal{B}_y^\beta, \boldsymbol{s}_y^\gamma \sim \mathcal{B}_y^\gamma \tag{5}$$

https://doi.org/10.52202/079017-2732

Afterwards, part of the observation region might overlap with some α and β targets. We remove any ROIs in the observation that overlap with the α targets. Additionally, we eliminate all timesteps for ROIs that show overlap with the β targets. Refer to Table 6 for the block sizes.

Training. Given the output $\boldsymbol{s}_x$ from the observation encoder f_θ, the predictor g_ϕ is trained to predict the three kinds of targets $\boldsymbol{s}_y^r$ conditioned on the positional embedding $\boldsymbol{P}$ (Figure 1). The training loss $\mathcal{L}$ is the average L_2 distance between $\boldsymbol{s}_y^r$ and its corresponding prediction:

$$\mathcal{L} = \frac{1}{3K} \sum_r \left\| \hat{\boldsymbol{s}}_y^r - \boldsymbol{s}_y^r \right\|_2^2, \hat{\boldsymbol{s}}_y^r = g_\phi(\boldsymbol{s}_x | \boldsymbol{P}) \tag{6}$$

4 Experiments

4.1 Datasets

We leveraged the large-scale public dataset - UK Biobank (UKB) [44, 45] for the self-supervised pretraining of Brain-JEPA. It includes resting-state fMRI recordings with medical records from 40,162 participants aged 44 to 83. Multi-site recordings were acquired with the temporal resolution of 0.735s. We allocated 80% of this dataset for pretraining (of which we calculated the group-level gradients as well), with the 20% held-out for downstream evaluation (internal tasks of age and sex prediction).

We used three datasets for external evaluation: HCP-Aging, as a segment of the public Human Connectome Project (HCP) [46], includes resting-state fMRI from 656 healthy elderly participants. It was used to predict traits (Neuroticism and Flanker score) and demographics (age and sex). The Alzheimer's Disease Neuroimaging Initiative (ADNI) [47] was used for the early diagnosis and prognosis of neurodegenerative diseases, with fMRI from 189 participants for normal control (NC) *v.s.* mild cognitive impairment (MCI) classification, and 100 cognitively normal participants for amyloid positive v.s. negative classification. Moreover, to assess generalizability across different ethnic groups and real-world clinical applications, we included resting-state fMRI of Asian participants recruited by Memory, Ageing and Cognition Centre (MACC), with 539 participants for NC v.s. MCI classification. More details of the downstream tasks performed can be found in the Appendix A.

All fMRI data was parcellated into $n = 450$ ROIs, using Schaefer-400 [48] for cortical regions and Tian-Scale III [49] for subcortical regions. Robust scaling was implemented by subtracting the median and dividing by the interquartile range, calculated across participants for each ROI [14]. Our default input size is 160 timesteps for each of the 450 ROIs (*i.e.*, 450×160). UKB and HCP-Aging used multi-band acquisition with a high temporal resolution (TR ≈ 0.7 seconds), while ADNI and MACC used single-band acquisition with a lower resolution (TR ≈ 2 seconds). To ensure consistency across datasets, we standardized the temporal resolution by downsampling the multi-band data using a temporal stride of 3, aligning the TR of all datasets to approximately 2 seconds. During the fine-tuning and linear probing stage, all the downstream datasets were divided into a 6:2:2 ratio for training, validation, and testing.

4.2 Implementation details

For Brain-JEPA pretraining, we utilized ViT architectures for the observation encoder, target encoder, and predictor. We employed FlashAttention [50, 51] in our self-attention implementation to improve computational efficiency and reduce memory usage. Balancing the trade-off between data quantity and the model complexity, we experimented with ViT-Small (ViT-S) (22M), ViT-Base (ViT-B) (86M), and ViT-Large (ViT-L) (307M) for the observation encoder. For predictor, it is designed as a lightweight (narrow) ViT. Specifically, the predictor has the same architecture as the corresponding observation encoder, differing only in embedding dimension and depth. For the ViT-S and ViT-B observation encoders, the predictor has a depth of 6 and embedding dimensions of 192 and 384, respectively. The ViT-L observation encoder uses a predictor with a depth of 12 and an embedding dimension of 384. Brain-JEPA is pretrained without a [cls] token. For evaluation, we used the target encoder and average pooled its output to generate a global fMRI representation. The main results in Section 4.3, along with the analysis in Section 4.5, 4.6 and 4.7 were all based on ViT-B pre-trained for 300 epochs. Refer to Appendix B for optimization and masking details.

Table 1: Internal tasks of age and sex prediction on UKB 20% held-out. The mean (standard deviation) of Mean Squared Error (MSE), Pearson Correlation (ρ), and/or Accuracy (ACC), F1 score across 5 independent runs is reported. ↑: the higher, the better; ↓: the lower, the better. The best results are in **bold**, with * denoting significant improvement over previous approaches ($p < 0.05$).

Methods	Age		Sex	
	MSE ↓	ρ ↑	ACC(%) ↑	F1(%) ↑
BrainNetCNN [6]	0.985 (0.027)	0.225 (0.015)	77.86 (0.98)	78.17 (0.86)
BrainGNN [5]	0.931 (0.038)	0.332 (0.015)	77.31 (0.33)	79.23 (0.31)
BNT [4]	0.863 (0.031)	0.447 (0.017)	80.78 (0.40)	82.42 (0.36)
TFS†	0.812 (0.023)	0.487 (0.011)	82.60 (0.59)	83.00 (0.01)
BrainLM [14]	0.612 (0.041)	0.632 (0.020)	86.47 (0.74)	86.84 (0.43)
Brain-JEPA	**0.501*** (0.034)	**0.718*** (0.021)	**88.17*** (0.06)	**88.58*** (0.11)

† Trained-From-Scratched.

Table 2: External tasks of demographics and trait prediction on HCP-Aging.

Methods	Age		Sex		Neuroticism		Flanker	
	MSE ↓	ρ ↑	ACC (%) ↑	F1 (%) ↑	MSE ↓	ρ ↑	MSE ↓	ρ ↑
BrainNetCNN [6]	0.462 (.017)	0.611 (.023)	71.16 (0.88)	72.23 (0.92)	1.201 (.097)	0.096 (.006)	1.045 (.036)	0.201 (.018)
BrainGNN [5]	0.423 (.015)	0.672 (.024)	72.7 (0.54)	74.09 (0.67)	1.183(.096)	0.098 (.007)	0.982 (.043)	0.309 (.062)
BNT [4]	0.414 (.035)	0.731 (.057)	72.41 (1.09)	73.68 (1.11)	1.199 (.091)	0.101 (.005)	0.997 (.037)	0.307 (.026)
BrainLM [14]	0.331 (.018)	0.832 (.028)	74.39 (1.55)	77.51 (1.13)	0.942 (.082)	0.231 (.012)	**0.971 (.054)**	0.318 (.048)
Brain-JEPA	**0.298 (.017)**	**0.844 (.030)**	**81.52* (1.03)**	**84.26* (0.82)**	**0.897* (.055)**	**0.307* (.006)**	0.972 (.038)	**0.406* (.027)**

Table 3: External tasks of brain disease diagnosis and prognosis on ADNI and MACC.

Methods	NC/MCI		Amyloid $a\beta$+ve/−ve		NC/MCI (Asian)	
	ACC(%) ↑	F1(%) ↑	ACC(%) ↑	F1(%) ↑	ACC(%) ↑	F1(%) ↑
BrainNetCNN [6]	60.00 (3.51)	64.72 (3.18)	59.00 (2.00)	59.43 (1.14)	57.32 (4.45)	53.92 (4.25)
BrainGNN [5]	67.40 (2.93)	71.42 (2.87)	57.00 (4.00)	62.61 (3.48)	59.79 (2.35)	55.69 (2.29)
BNT [4]	**78.90** (4.12)	83.14 (3.58)	62.00 (2.45)	59.53 (0.58)	62.06 (3.88)	60.45 (4.52)
BrainLM [14]	75.79 (1.05)	85.66 (1.27)	67.00 (7.48)	68.82 (8.48)	61.65 (3.35)	60.26 (3.03)
Brain-JEPA	76.84 (1.05)	**86.32** (0.54)	**71.00*** (4.90)	**75.97*** (3.93)	**65.98*** (2.84)	**64.67*** (2.61)

4.3 Main results

Table 1, 2, and 3 compare Brain-JEPA with the existing deep learning models for fMRI analysis and foundation model BrainLM. We select the three deep learning baselines because they not only represent the previous state-of-the-art in fMRI analysis but also exemplify diverse model types: convolutional neural network (CNN)-based BrainNetCNN [6], graph neural network (GNN)-based BrainGNN [5], and transformer-based BNT [4]. For a fair comparison, both Brain-JEPA and BrainLM utilized a ViT-B backbone and were fine-tuned for downstream tasks (Section 4.4 will discuss performance scaling with different model sizes, and Section 4.5 will examine linear probing comparisons between the two models). BrainLM utilized [cls] token for downstream evaluation.

The results show that Brain-JEPA achieves state-of-the-art performance in various downstream tasks on both the unseen data from the same pretrained cohort and other independent datasets. Brain-JEPA effectively captures fundamental demographic information such as age and sex, cognitive/personality variance (Neuroticism and Flanker), and disease-related patterns for neurodegenerative diseases. Notably, Brain-JEPA demonstrates superior performance in classifying NC/MCI in Asian ethnic groups — one of the most challenging tasks for early diagnosis and prognosis of Alzheimer's Disease (AD) — even though it was trained exclusively on the Caucasian cohort. Please refer to C.1 for additional results on more datasets and comparisons with more baselines.

https://doi.org/10.52202/079017-2732

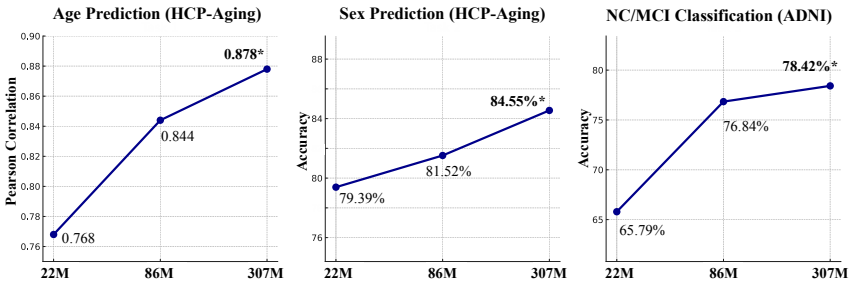

Figure 3: Performance scaling of the model sizes.

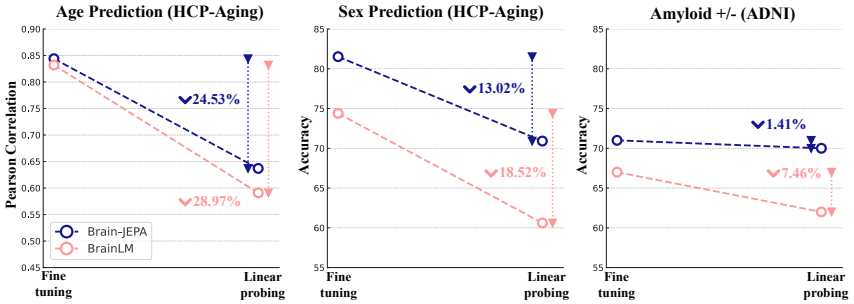

Figure 4: Fine-tuning *v.s.* linear probing.

4.4 Performance scaling

Figure 3 presents the performance of Brain-JEPA across various model sizes, using ViT-S, ViT-B, and ViT-L as backbones. The results demonstrate that the larger model configuration consistently achieves better performance. Specifically, there is a clear trend of increasing accuracy/correlation with larger models, with Brain-JEPA using ViT-L consistently achieving the best performance. We also studies the scaling property with respect to dataset size, please refer to C.2 for additional results.

4.5 Linear probing

BrainLM initially showcases its performance improvements through fine-tuning, complemented by an attached MLP [14]. However, to effectively assess the representations learned during pretraining, off-the-shelf evaluations such as linear probing are essential. As depicted in Figure 4, Brain-JEPA consistently outperforms BrainLM in linear probing and exhibits a smaller performance decline from fine-tuning to linear probing. This highlights the robustness and higher level of abstraction in the representations learned by Brain-JEPA.

4.6 Ablation study

We first compared Brain-JEPA with its ablated versions, employing sine and cosine functions [43] and anatomical locations [14] for ROI spatial positioning, as shown in Figure 5. Brain Gradient Positioning demonstrates superior performance over these two baseline methods. It indicates that Brain Gradient Positioning facilitates natural and accurate placement of brain functional parcellations, enhancing the learning of brain dynamics. Next, we assessed the effectiveness of our proposed Spatiotemporal Masking by comparing Brain-JEPA, pretrained over various numbers of epochs, to its ablated counterpart that utilizes standard multi-block sampling of targets [21]. This comparison, illustrated in Figure 6, highlights that not only does our proposed masking technique yield superior

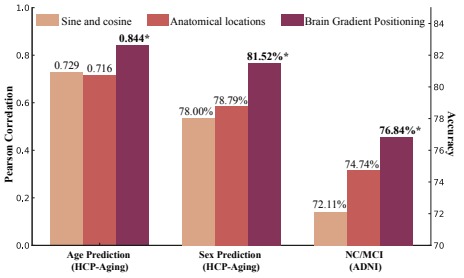

Figure 5: Comparisons of spatial positional embedding (For the first task, refer to the left y axis for the Pearson's Correlation, with the right y axis accuracy for the last two tasks).

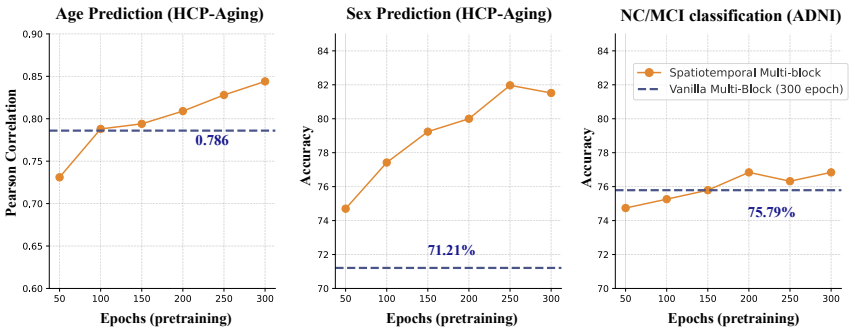

Figure 6: Comparisons of masking strategies.

performance, but it also introduces a stronger inductive bias leading to a more efficient pretraining. Notably, Brain-JEPA achieves or even surpasses the peak performance of the ablated version, which was pretrained for 300 epochs, with significantly fewer epochs—only 100, 200, and 50 respectively. For more ablation results regarding architectures and the number of gradient components, please refer to C.3.

4.7 Interpretation

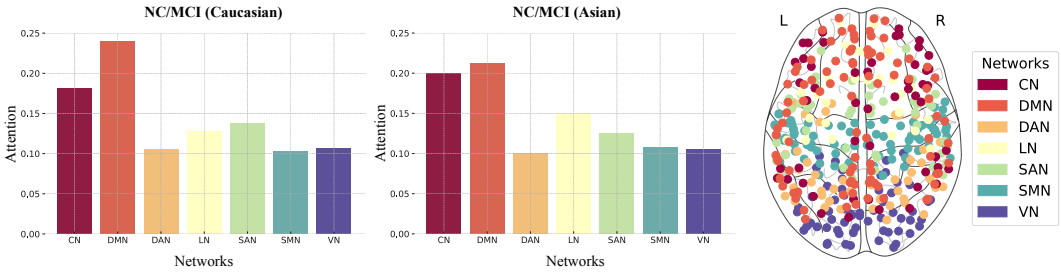

Figure 7: Attention across different brain networks for NC/MCI classification.

With the Schaefer functional atlas [48], the brain network is categorized into seven distinct sub-networks: the control network (CN), the default mode network (DMN), the dorsal attention network (DAN), the limbic network (LN), the salience ventral attention network (SAN), the somatomotor network (SMN), and the visual network (VN). To assess whether Brain-JEPA has captured the brain functional organization, we calculate the network-level attention for NC/MCI classification. For each ROI, we first average the self-attention across its 10 patches. Next, we average the values of the ROIs within each sub-network and normalize them to obtain the network-level attention distribution. As shown in Figure 7, we found consistent patterns across both Caucasian and Asian ethnic groups, with the model highlighting the critical roles of the DMN, CN, SAN, and LN in cognitive impairment, consistent with previous literature [52, 53, 54].

5 Conclusion

In this study, we developed Brain-JEPA, a brain dynamics foundation model based on the Joint-Embedding Predictive Architecture (JEPA). Brain-JEPA predicts abstract representations of sampled targets from observations during the pretraining stage. Utilizing Brain Gradient Positioning, Brain-JEPA encodes brain functional organization more naturally and accurately. With Spatiotemporal Masking, it effectively handles heterogeneous patches in fMRI time series. Brain-JEPA fosters generalizable and highly abstract representations of fMRI, achieving state-of-the-art performance across various tasks, including demographic prediction, trait prediction, and disease diagnosis and prognosis across different cohorts and ethnic groups. Our study provides new insights into applying large-scale self-supervised learning to brain activity modelling and contributes to addressing key questions in AI for neuroscience.

https://doi.org/10.52202/079017-2732

6 Limitation and future work

We acknowledge several limitations in our study, which also serve as inspirations for future research: 1) Larger models: Due to limited computing resources, we have not tested larger models like ViT-H. We expect that larger models could further improve performance. 2) More diverse datasets: A more diverse brain recording dataset for pretraining, including different ethnicity cohorts collected from various sites, scanning protocols, behavioral tasks, and disease groups, could enhance the generalizability and robustness of the representations learned by the model. 3) Fine-grained interpretation: More thorough interpretation can be achieved through the attention mechanism, such as comparing cortical and subcortical regions, identifying salient ROIs and critical timesteps. This would enable more nuanced and complex spatiotemporal interpretations. 4) Multi-modal integration: Brain-JEPA sets a potential foundation for integrating multimodal brain activity data such as MEG and EEG or even brain structure data like T1-weighted MRI. The integration could enhance our understanding of brain structure, function, and their links to human behavior and mental disorders. Please refer to Appendix D for the broader impact of Brain-JEPA.

7 Acknowledgements

This study was supported by the Singapore National Medical Research Council (NMRC/OFLCG19May-0035, NMRC/CIRG/1485/2018, NMRC/CSA-SI/0007/2016, NMRC/MOH-00707-01, NMRC/CG/435 M009/2017-NUH/NUHS, CIRG21nov-0007 and HLCA23Feb-0004), RIE2020 AME Programmatic Fund from A*STAR, Singapore (No. A20G8b0102), Ministry of Education (MOE-T2EP40120-0007 & T2EP2-0223-0025, MOE-T2EP20220-0001), and Yong Loo Lin School of Medicine Research Core Funding, National University of Singapore, Singapore.

References

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

[2] David J Heeger and David Ress. What does fmri tell us about neuronal activity? *Nature reviews neuroscience*, 3(2):142–151, 2002.

[3] Nikos K Logothetis, Jon Pauls, Mark Augath, Torsten Trinath, and Axel Oeltermann. Neurophysiological investigation of the basis of the fmri signal. *nature*, 412(6843):150–157, 2001.

[4] 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.

[5] Xiaoxiao Li, Yuan Zhou, Nicha Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H Staib, Pamela Ventola, and James S Duncan. Braingnn: Interpretable brain graph neural network for fmri analysis. *Medical Image Analysis*, 74:102233, 2021.

[6] Jeremy Kawahara, Colin J Brown, Steven P Miller, Brian G Booth, Vann Chau, Ruth E Grunau, Jill G Zwicker, and Ghassan Hamarneh. Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038–1049, 2017.

[7] Hejie Cui, Wei Dai, Yanqiao Zhu, Xuan Kan, Antonio Aodong Chen Gu, Joshua Lukemire, Liang Zhan, Lifang He, Ying Guo, and Carl Yang. Braingb: a benchmark for brain network analysis with graph neural networks. *IEEE transactions on medical imaging*, 42(2):493–506, 2022.

[8] Andrew T Drysdale, Logan Grosenick, Jonathan Downar, Katharine Dunlop, Farrokh Mansouri, Yue Meng, Robert N Fetcho, Benjamin Zebley, Desmond J Oathes, Amit Etkin, et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature medicine*, 23 (1):28–38, 2017.

[9] Tetsuya Iidaka. Resting state functional magnetic resonance imaging and neural network classified autism and control. *Cortex*, 63:55–67, 2015.

[10] Zijian Dong, Yilei Wu, Zijiao Chen, Yichi Zhang, Yueming Jin, and Juan Helen Zhou. Prompt your brain: Scaffold prompt tuning for efficient adaptation of fmri pre-trained model. *arXiv preprint arXiv:2408.10567*, 2024.

[11] Zijian Dong, Yilei Wu, Yu Xiao, Joanna Su Xian Chong, Yueming Jin, and Juan Helen Zhou. Beyond the snapshot: Brain tokenized graph transformer for longitudinal brain functional connectome embedding. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 348–357. Springer, 2023.

[12] Yilei Wu, Zijian Dong, Chongyao Chen, Wangchunshu Zhou, and Juan Helen Zhou. Mixup your own pairs. *arXiv preprint arXiv:2309.16633*, 2023.

[13] Zijian Dong, Joanna Su Xian Chong, Bing Cai Kok, and Juan Helen Zhou. Coop-dhgnn: a framework for joint classification and prediction of brain functional connectivity using sparse trajectory dataset with application to early dementia. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 4972–4978, 2022. doi: 10.1109/BigData55660.2022.10021043.

[14] Josue Ortega Caro, Antonio Henrique de Oliveira Fonseca, Christopher Averill, Syed A Rizvi, Matteo Rosati, James L Cross, Prateek Mittal, Emanuele Zappala, Rahul Madhav Dhodapkar, Chadi Abdallah, et al. Brainlm: A foundation model for brain activity recordings. In *The Twelfth International Conference on Learning Representations*, 2023.

[15] Guangqi Wen, Peng Cao, Lingwen Liu, Jinzhu Yang, Xizhe Zhang, Fei Wang, and Osmar R Zaiane. Graph self-supervised learning with application to brain networks analysis. *IEEE Journal of Biomedical and Health Informatics*, 2023.

[16] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[17] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[18] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[19] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.01553. URL `https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01553`.

[20] César Caballero-Gaudes and Richard C Reynolds. Methods for cleaning the bold fmri signal. *Neuroimage*, 154:128–149, 2017.

[21] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[23] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1), 2022.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[25] Stephen M Smith, Christian F Beckmann, Jesper Andersson, Edward J Auerbach, Janine Bijsterbosch, Gwenaëlle Douaud, Eugene Duff, David A Feinberg, Ludovica Griffanti, Michael P Harms, et al. Resting-state fmri in the human connectome project. *Neuroimage*, 80:144–168, 2013.

 https://doi.org/10.52202/079017-2732

[26] 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*, pages 248–255. Ieee, 2009.

[27] 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 A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 42015–42037. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/8313b1920ee9c78d846c5798c1ce48be-Paper-Conference.pdf.

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[29] 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*, pages 1–1, 2024. doi: 10.1109/TMI.2024.3414476.

[30] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR, 2017.

[31] M-Marsel Mesulam. From sensation to cognition. *Brain: a journal of neurology*, 121(6): 1013–1052, 1998. ISSN 1460-2156.

[32] Daniel S. Margulies, Satrajit S. Ghosh, Alexandros Goulas, Marcel Falkiewicz, Julia M. Huntenburg, Georg Langs, Gleb Bezgin, Simon B. Eickhoff, F. Xavier Castellanos, Michael Petrides, Elizabeth Jefferies, and Jonathan Smallwood. Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences*, 113(44):12574–12579, 2016. doi: 10.1073/pnas.1608282113. URL https://www.pnas.org/content/pnas/113/44/12574.full.pdf.

[33] Z. X. Zhou and X. N. Zuo. Editorial: Lifespan connectome gradients for a road to mental health. *J Am Acad Child Adolesc Psychiatry*, 63(1):25–28, 2024. ISSN 0890-8567. doi: 10.1016/j.jaac.2023.08.006. 1527-5418 Zhou, Zi-Xuan Zuo, Xi-Nian Editorial United States 2023/09/02 J Am Acad Child Adolesc Psychiatry. 2024 Jan;63(1):25-28. doi: 10.1016/j.jaac.2023.08.006. Epub 2023 Aug 30.

[34] Sara Larivière, Reinder Vos de Wael, Seok-Jun Hong, Casey Paquola, Shahin Tavakol, Alexander J. Lowe, Dewi V. Schrader, and Boris C. Bernhardt. Multiscale structure–function gradients in the neonatal connectome. *Cerebral Cortex*, 30(1):47–58, 2020. ISSN 1047-3211. doi: 10.1093/cercor/bhz069. URL https://doi.org/10.1093/cercor/bhz069.

[35] Thuan Tinh Nguyen, Xing Qian, Eric Kwun Kei Ng, Marcus Qin Wen Ong, Zhen Ming Ngoh, Shayne S. P. Yeo, Jia Ming Lau, Ai Peng Tan, Birit F. P. Broekman, Evelyn C. Law, Peter D. Gluckman, Yap-Seng Chong, Samuele Cortese, Michael J. Meaney, and Juan Helen Zhou. Variations in cortical functional gradients relate to dimensions of psychopathology in preschool children. *Journal of the American Academy of Child & Adolescent Psychiatry*, 2023. ISSN 0890-8567. doi: https://doi.org/10.1016/j.jaac.2023.05.029. URL https://www.sciencedirect.com/science/article/pii/S0890856723003702.

[36] Hao-Ming Dong, Daniel S. Margulies, Xi-Nian Zuo, and Avram J. Holmes. Shifting gradients of macroscale cortical organization mark the transition from childhood to adolescence. *Proceedings of the National Academy of Sciences*, 118(28):e2024448118, 2021. doi: 10.1073/pnas.2024448118. URL https://doi.org/10.1073/pnas.2024448118. doi: 10.1073/pnas.2024448118.

[37] Richard A. I. Bethlehem, Casey Paquola, Jakob Seidlitz, Lisa Ronan, Boris Bernhardt, Cam-Can Consortium, and Kamen A. Tsvetanov. Dispersion of functional gradients across the adult lifespan. *NeuroImage*, 222:117299–117299, 2020. ISSN 1095-9572 1053-8119. doi: 10.1016/j.neuroimage.2020.117299. URL https://pubmed.ncbi.nlm.nih.gov/32828920https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7779368/.

[38] Xiuyi Wang, Daniel S. Margulies, Jonathan Smallwood, and Elizabeth Jefferies. A gradient from long-term memory to novel cognition: Transitions through default mode and executive cortex. *NeuroImage*, 220:117074, 2020. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2020.117074. URL https://www.sciencedirect.com/science/article/pii/S1053811920305607.

[39] Seok-Jun Hong, Reinder Vos de Wael, Richard A. I. Bethlehem, Sara Lariviere, Casey Paquola, Sofie L. Valk, Michael P. Milham, Adriana Di Martino, Daniel S. Margulies, Jonathan Smallwood, and Boris C. Bernhardt. Atypical functional connectome hierarchy in autism. *Nature Communications*, 10(1):1022, 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-08944-1. URL https://doi.org/10.1038/s41467-019-08944-1.

[40] Yirong He, Qiongling Li, Zhenrong Fu, Debin Zeng, Ying Han, and Shuyu Li. Functional gradients reveal altered functional segregation in patients with amnestic mild cognitive impairment and alzheimer's disease. *Cerebral Cortex*, 33(21):10836–10847, 2023. ISSN 1047-3211. doi: 10.1093/cercor/bhad328. URL https://doi.org/10.1093/cercor/bhad328.

[41] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences*, 102(21):7426–7431, 2005. doi: 10.1073/pnas.0500334102. URL https://doi.org/10.1073/pnas.0500334102. doi: 10.1073/pnas.0500334102.

[42] Reinder Vos de Wael, Oualid Benkarim, Casey Paquola, Sara Lariviere, Jessica Royer, Shahin Tavakol, Ting Xu, Seok-Jun Hong, Georg Langs, Sofie Valk, et al. Brainspace: a toolbox for the analysis of macroscale gradients in neuroimaging and connectomics datasets. *Communications biology*, 3(1):103, 2020.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[44] 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.

[45] Maria A Di Biase, Robert E Smith, Andrew Zalesky, Caio Seguin, et al. Connectomes for 40,000 uk biobank participants: a multi-modal, multi-scale brain network resource. *Neuroimage*, 283: 120407, 2023.

[46] Jennifer Stine Elam, Matthew F Glasser, Michael P Harms, Stamatios N Sotiropoulos, Jesper LR Andersson, Gregory C Burgess, Sandra W Curtiss, Robert Oostenveld, Linda J Larson-Prior, Jan-Mathijs Schoffelen, et al. The human connectome project: a retrospective. *NeuroImage*, 244: 118543, 2021.

[47] 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. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.

[48] 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.

[49] 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.

[50] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[51] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

[52] Puneet Talwar, Suman Kushwaha, Monali Chaturvedi, and Vidur Mahajan. Systematic review of different neuroimaging correlates in mild cognitive impairment and alzheimer's disease. *Clinical neuroradiology*, 31(4):953–967, 2021.

[53] Yvette I Sheline and Marcus E Raichle. Resting state functional connectivity in preclinical alzheimer's disease. *Biological psychiatry*, 74(5):340–347, 2013.

[54] Matthew R Brier, Jewell B Thomas, and Beau M Ances. Network dysfunction in alzheimer's disease: refining the disconnection hypothesis. *Brain connectivity*, 4(5):299–311, 2014.

[55] Edmund R Thompson. Development and validation of an international english big-five mini-markers. *Personality and individual differences*, 45(6):542–548, 2008.

[56] Barbara A Eriksen and Charles W Eriksen. Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & psychophysics*, 16(1):143–149, 1974.

[57] Ronald Carl Petersen, Paul S Aisen, Laurel A Beckett, Michael C Donohue, Anthony Collins Gamst, Danielle J Harvey, CR Jack Jr, William J Jagust, Leslie M Shaw, Arthur W Toga, et al. Alzheimer's disease neuroimaging initiative (adni) clinical characterization. *Neurology*, 74(3): 201–209, 2010.

[58] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1xMH1BtvB.

[59] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/ioffe15.html.

[60] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[61] Boris Ginsburg, Igor Gitman, and Yang You. Large batch training of convolutional networks with layer-wise adaptive rate scaling, 2018. URL https://openreview.net/forum?id=rJ4uaX2aW.

[62] Marc-Andre Schulz, B. T. Thomas Yeo, Joshua T. Vogelstein, Janaina Mourao-Miranada, Jakob N. Kather, Konrad Kording, Blake Richards, and Danilo Bzdok. Different scaling of linear models and deep learning in ukbiobank brain images versus machine-learning datasets. *Nature Communications*, 11(1):4238, Aug 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-18037-z. URL https://doi.org/10.1038/s41467-020-18037-z.

[63] Armin Thomas, Christopher Ré, and Russell Poldrack. Self-supervised learning of brain dynamics from broad neuroimaging data. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 21255–21269. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/8600a9df1a087a9a66900cc8c948c3f0-Paper-Conference.pdf.

A Task Details

A.1 Neuroticism

Neuroticism is a personality trait linked to negative emotions and is one of the Big Five personality traits. People who score high in neuroticism tend to experience negative feelings more frequently than others [55]. The HCP uses the 60-question version of the NEO-FFI Short Form (ages 16+) questionnaire, which provides a quick, reliable, and accurate assessment of the Big Five personality traits: neuroticism, extraversion, openness, agreeableness, and conscientiousness. More detailed information can be found in the Lifespan HCP 2.0 Data Release Appendix 2: Details and References for Behavioral & Clinical Instruments.

A.2 Flanker

The Flanker task is designed to assess both attention and inhibitory control in participants [56]. It involves the participant focusing on a central stimulus while ignoring adjacent stimuli, which are either fish for ages 3-7 or arrows for ages 8-85. Sometimes the central stimulus points in the same direction as the flanking stimuli (congruent) and sometimes in the opposite direction (incongruent). For participants aged 8-85, the task consists of twenty trials and takes about three minutes to complete. More detailed information can be found in the Lifespan HCP 2.0 Data Release Appendix 2: Details and References for Behavioral & Clinical Instruments.

A.3 NC/MCI

For the ADNI dataset [47], the criteria for NC was as follows: 1) No subjective memory complaints, 2) preserved activities of daily living and cognitive function, 3) Mini-mental state examination (MMSE) score of between 24 to 30 inclusive, 4) Clinical Dementia Rating (CDR) score of 0, and 5) education-adjusted score on delayed recall of one paragraph from Wechsler Memory Scale Logical Memory II of >=3 for 0-7 years of education, >= 5 for 8-15 years of education, and >= 9 for >=16 years of education. The criteria for MCI was as follows: 1) significant subjective memory complaints reported by the participant, clinician or informant, 2) not significantly impaired in other cognitively domains, 3) essentially preserved activities of daily living and does not meet criteria for diagnosis of dementia, 4) MMSE score of between 24 to 30 inclusive, 5) CDR score of 0.5, and 6) education-adjusted score on delayed recall of one paragraph from Wechsler Memory Scale Logical Memory II of 3-6 for 0-7 years of education, 5-9 for 8-15 years of education, and 9-11 for >=16 years of education [57].

For the Asian disease cohort, all participants completed a locally validated neuropsychological test battery, which assessed seven domains: executive function, attention, language, visuomotor speed, verbal memory, and visual memory. Impairment in a particular domain was defined as failing at least half of the individual tests in a domain, and failure in an individual test was determined using education-adjusted cut-offs of 1.5 standard deviations below established normal means. NC was defined as having no impairment in all cognitive domains on the neuropsychological test battery, while MCI was defined as having an impairment in at least one cognitive domain of the neuropsychological test battery. Detailed descriptions of the neuropsychological assessments and diagnostic criteria are described in previous work which will be added upon acceptance.

A.4 Amyloid +/-

Participants from the ADNI cohort were also classified as amyloid positive or amyloid negative, using a threshold of global [18F]-Florbetapir amyloid PET SUVR >= 1.11 to define amyloid positivity [57].

B Additional Implementation Details

Optimization for pre-training. The default settings are detailed in Table 4. We initialized all transformer blocks using the Xavier uniform method, as described in [19]. The pre-training process utilized four A100 GPUs, each with 40GB of memory.

Optimization for downstream tasks. The default settings for end-to-end fine-tuning and linear probing are detailed in Table 5. For fine-tuning, following [19], we applied layer-wise *lr* decay [58].

https://doi.org/10.52202/079017-2732

Table 4: Pre-training settings. GAS: Gradient accumulation steps; BS: Batch size

config	value
optimizer	AdamW [60]
optimizer momentum	β_1,β_2=0.9,0.999
learning rate schedule	warmup cosine schedule [21]
start learning rate	5×10^{-5}
learning rate	1×10^{-3}
final learning rate	1×10^{-6}
weight decay schedule	cosine weight decay schedule [21]
weight decay	0.04
final weight decay	0.4
EMA momentum schedule	linear [21]
EMA start momentum	0.996
EMA final momentum	1
total batch size	4 GPUs $\times$ 8 GAS $\times$ 16 BS
warmup epochs	40
patch size p	16
dimension of gradient vector m	30
training epochs	300

Table 5: Settings of end-to-end fine-tuning and linear probe.

config	value of FT	value of LP
optimizer	AdamW	LARS [61]
optimizer momentum	(0.9, 0.999)	0.9
learning rate schedule	cosine decay [19]	cosine decay
base learning rate	0.001	0.01
weight decay	0.05	N.A.
layer-wise lr decay	0.75	N.A.
batch size	16	64
warmup epochs	0	0
training epochs	50	50

Table 6: Hyperparameters for spatiotemporal masking.

Region	Mask ratio
observation block	$\{\eta_R^o, \eta_T^o\}$=$\{(0.84, 1.0), (0.84, 1.0)\}$
target α	$\{\eta_R^\alpha, \eta_T^\alpha\}$=$\{(0.45, 0.6), (0.2, 0.6)\}$
target β	$\{\eta_R^\beta, \eta_T^\beta\}$=$\{(0.15, 0.3), (0.0, 0.4)\}$
target γ	$\{\eta_R^\gamma, \eta_T^\gamma\}$=$\{(0.15, 0.3), (0.0, 0.4)\}$

For linear probing, we incorporated an additional BatchNorm layer [59] before the linear head, as per [19].

Masking. The range ratios for obtaining the observation block and three target blocks introduced in Section 3.2 are presented in Tables 6. η_b^a denotes the mask range along the b dimension for the a block. We set $K=1$ which is the number of randomly sampled blocks in three target regions.

C Additional Results

C.1 Results on additional baselines and datasets

We incorporated more baseline results for downstream tasks on external datasets in Tables 7-8, including commonly used SVM/SVR [62] and recent self-supervised learning methods. It is observed that Brain-JEPA outperforms these models on most tasks. We note that for the compared baselines, BrainMass

[29] is a concurrent work. Additionally, CSM [63] and SwiFT [27] are not time series models; CSM utilizes text-like representations, while SwiFT operates on raw fMRI data.

Table 7: Results of additional baselines on HCP-aging.

Methods	Age		Sex	
	MSE ↓	ρ ↑	ACC (%) ↑	F1 (%) ↑
SVM/SVR	0.586 (.019)	0.699 (.022)	76.67 (1.88)	80.82 (1.15)
BrainMass	0.396 (.002)	0.831 (.014)	74.09 (3.87)	75.78 (3.37)
CSM	0.409 (.012)	0.733 (.023)	74.85 (1.11)	76.23 (0.37)
SwiFT	0.341 (.007)	0.755 (.063)	73.48 (2.20)	74.65 (2.32)
Brain-JEPA	**0.298** (.017)	**0.844** (.030)	**81.52** (1.03)	**84.26** (0.82)

Table 8: Results of additional baselines on ADNI.

Methods	NC/MCI		Amyloid $a\beta$+ve/−ve	
	ACC (%) ↑	F1 (%) ↑	ACC (%) ↑	F1 (%) ↑
SVM/SVR	64.21 (5.16)	73.06 (4.71)	62.00 (4.00)	63.84 (5.44)
BrainMass	74.21 (5.10)	81.36 (3.56)	68.00 (7.48)	69.29 (8.96)
CSM	68.42 (4.99)	76.74 (4.54)	63.00 (9.80)	65.89 (9.79)
SwiFT	73.16 (5.31)	80.46 (4.16)	65.00 (6.32)	67.79 (6.38)
Brain-JEPA	**76.84** (1.05)	**86.32** (0.54)	**71.00** (4.90)	**75.97** (3.93)

To further demonstrate the diversity of our downstream applications, we conducted additional experiments using two aging-related public datasets: OASIS-3 and CamCAN, for AD conversion prediction in MCI participants and depression diagnosis, respectively. The results are shown in Table 9. By applying Brain-JEPA to five downstream datasets across eight distinct tasks totally, we have demonstrated its versatility in a wider range of applications compared to the existing models. Specifically, Brain-JEPA excels in demographic prediction, trait prediction, and disease diagnosis and prognosis. This stands in contrast to experiments done in BrainLM, which is limited to demographic and clinical score prediction, and BrainMass, which focuses solely on disease diagnosis and prognosis.

Table 9: AD conversion prediction and depression diagnosis on OASIS-3 and CamCAN datasets.

Methods	OASIS-3		CamCAN	
	AD Conversion		Depression	
	ACC (%) ↑	F1 (%) ↑	ACC (%) ↑	F1 (%) ↑
SVM/SVR	56.00 (2.81)	52.05 (1.66)	63.64 (3.07)	56.79 (2.32)
BrainNetCNN	62.00 (2.45)	59.53 (0.58)	62.73 (4.45)	56.85 (4.47)
BrainGNN	59.00 (2.00)	56.53 (4.34)	63.64 (4.98)	56.68 (3.26)
BNT	68.00 (8.72)	64.73 (11.29)	65.45 (4.64)	55.32 (8.67)
BrainLM	65.00 (7.75)	62.67 (9.04)	70.00 (6.17)	64.18 (3.82)
BrainMass	67.00 (6.00)	66.53 (6.95)	70.91 (2.23)	63.56 (2.93)
CSM	61.00 (4.90)	61.97 (5.49)	64.55 (4.45)	56.08 (6.23)
SwiFT	65.00 (6.32)	66.80 (4.12)	69.09 (6.68)	61.78 (9.26)
Brain-JEPA	**69.00** (7.35)	**67.32** (7.92)	**72.73** (2.87)	**67.45** (1.57)

C.2 Scaling properties with respect to dataset size.

We compared the performance of Brain-JEPA trained with varying portions of the UKB pretraining dataset: 25%, 50%, 75%, and 100%. As shown in Table 10, the performance improves as the dataset size increases, highlighting the scalability of Brain-JEPA in relation to the pretraining dataset size.

https://doi.org/10.52202/079017-2732

Table 10: Ablation on different dataset size for pretraining.

Percentage	HCP-Aging		ADNI
	Age	Sex	NC/MCI
	$\rho\uparrow$	ACC (%) $\uparrow$	ACC (%) $\uparrow$
25%	0.659 (.043)	68.03 (1.21)	67.89 (9.18)
50%	0.768 (.012)	74.24 (1.36)	71.05 (3.86)
75%	0.813 (.015)	77.42 (2.00)	74.74 (4.88)
100%	**0.844** (.030)	**81.52** (1.03)	**76.84** (1.05)

C.3 Additional ablations

Architectures/Frameworks. To thoroughly compare the performance between JEPA with anatomical locations (AL) and BrainLM (MAE-based), we extended our comparison to include all the tasks except for the three in the main content, as well as two newly added datasets, OASIS-3 and CamCAN. The results shown in Table 11, demonstrates that JEPA with AL outperforms BrainLM in seven out of eleven tasks, demonstrating the superiority of prediction in latent space. For the tasks where BrainLM performs better, it is likely that JEPA requires gradient positioning for precise ROI placement to achieve optimal performance. In future work, we will further investigate the possible interactions between the self-supervised learning framework and brain gradient positioning.

Table 11: Ablation on position embedding.

Methods	UKB		HCP-Aging	
	Age	Sex	Neurotism	Flanker
	$\rho\uparrow$	ACC (%) $\uparrow$	$\rho\uparrow$	$\rho\uparrow$
BrainLM	0.632 (0.020)	**86.47** (0.74)	0.231 (.012)	0.318 (.048)
Brain-JEPA w AL	**0.686** (0.013)	84.11 (0.50)	**0.267** (.003)	**0.374** (.022)
Methods	ADNI	MACC	OASIS-3	CamCAN
	Amy+/-	NC/MCI	AD Conversion	Depression
	ACC (%) $\uparrow$	ACC (%) $\uparrow$	ACC (%) $\uparrow$	ACC (%) $\uparrow$
BrainLM	**67.00** (7.48)	61.65 (3.35)	65.00 (7.75)	70.00 (6.17)
Brain-JEPA *w* AL	65.00 (6.32)	**64.33** (1.80)	**67.00** (4.00)	**71.82** (6.03)

We further compared Brain-JEPA without JEPA architecture (*i.e.*, BrainLM with contributions) to Brain-JEPA. As shown in Table 12, Brain-JEPA (JEPA framework) outperforms BrainLM (MAE framework) with contributions consistently, demonstrating the superiority of JEPA framework.

Table 12: Comparisons of different frameworks.

Methods	HCP-Aging		ADNI
	Age	Sex	Amy+/-
	$\rho\uparrow$	ACC (%) $\uparrow$	ACC (%) $\uparrow$
BrainLM	0.832 (.028)	74.39 (1.55)	67.00 (7.48)
BrainLM w contributions	0.838 (.014)	76.36 (2.58)	70.00 (11.40)
JEPA w contributions	**0.844** (.030)	**81.52** (1.03)	**71.00** (4.90)

The number of gradient components. We compared the model performance between 3-dimensional (3-dim) and 30-dim brain gradient positioning, shown in Table 13. The 30-dim model consistently outperformed the 3-dim model by a large margin. This indicates that higher-dimensional brain gradients may encapsulate finer-grained information on brain network organization, which benefits the learning of brain dynamics.

Table 13: Comparison of different number of gradient components.

Methods	HCP-Aging		ADNI
	Age	Sex	Amy+/-
	$\rho\uparrow$	ACC (%) $\uparrow$	ACC (%) $\uparrow$
3-dim brain gradient	0.819 (.003)	76.96 (1.77)	67.00 (6.00)
30-dim brain gradient	**0.844** (.030)	**81.52** (1.03)	**71.00** (4.90)

D Broader Impact

The introduction of Brain-JEPA marks a significant advancement in the interdisciplinary field of AI and neuroscience, particularly in the brain activity analysis. An assessment of the broader impact of this model has across various dimensions:

D.1 Neuroscience and medical advancements

Brain-JEPA's capabilities in demographic prediction, disease diagnosis, and prognosis could revolutionize how neurological disorders are diagnosed and treated. This may lead to earlier detection and more personalized therapeutic interventions, potentially improving outcomes for patients with conditions like AD, schizophrenia, or autism spectrum disorders. Furthermore, the model's innovative techniques, including Brain Gradient Positioning and Spatiotemporal Masking, offer new ways to understand the brain's functional organization. This could lead to breakthroughs in identifying how various cognitive processes are mapped in the brain, aiding in both basic science and clinical applications. On the other hand, by effectively predicting various traits, Brain-JEPA can aid in the study of the genetic and environmental influences on behavior and cognitive functions. This can enhance our understanding of the neural underpinnings of psychological traits and disorders.

D.2 Technological impact

Brain-JEPA sets a new standard in AI's application to complex brain activity data with a novel brain functional coordinate system and masking strategy, which could spur further innovations and applications of AI across different sub-fields of neuroscience. Furthermore, the model's success in performing well across different ethnic groups indicate potential for broad applications in diverse global settings, which is crucial for building inclusive and unbiased AI systems.

D.3 Ethical and social considerations

Ensuring the confidentiality and integrity of patient data while using such advanced AI systems is paramount. While Brain-JEPA has shown superior performance across different tasks, continuous monitoring for potential biases is essential, especially as the model is scaled and deployed in varied clinical settings. Besides, the deployment of advanced technologies like Brain-JEPA could exacerbate existing disparities in healthcare access unless carefully managed. Ensuring that these technologies benefit all segments of the population equally is critical.

https://doi.org/10.52202/079017-2732

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Refer to Section 1 for our contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Refer to Section 6 for the discussion of limitation and future work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided detailed information about the datasets in Section 4.1 and Appendix A. The implementation details are presented in Section 4.2 and Appendix B. The model checkpoints and anonymized code are provided in supplementary material with clear instructions in the readme file.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/079017-2732

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have provided model checkpoints and code in supplementary material with a readme file. The links to the publicly available datasets are provided.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We have provided detailed information about the datasets in Section 4.1 and Appendix A. We have also provided the implementation details and hyperparameters in Section 4.2 and Appendix B.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All main results in Table 1, 2, and 3 contain mean and standard deviation from 5 independent runs. The statistical significance of the experiments has been shown with * denoting significant improvement with $p < 0.05$

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the computer resources in Section 4.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The proposed research conforms with the Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Refer to Section D for potential positive societal impacts and negative societal impacts of Brain-JEPA.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

https://doi.org/10.52202/079017-2732

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The original papers that produced the code package or dataset are all properly cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The details of the dataset/code/model have been provided in 4.1, 4.2 and Appendix B, with the codes in supplementary materials.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The datasets of human subjects used in this paper are either publicly available or from previous work which will be cited upon acceptance.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The datasets of human subjects used in this paper are either publicly available or from previous work which will be cited upon acceptance.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

https://doi.org/10.52202/079017-2732

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.