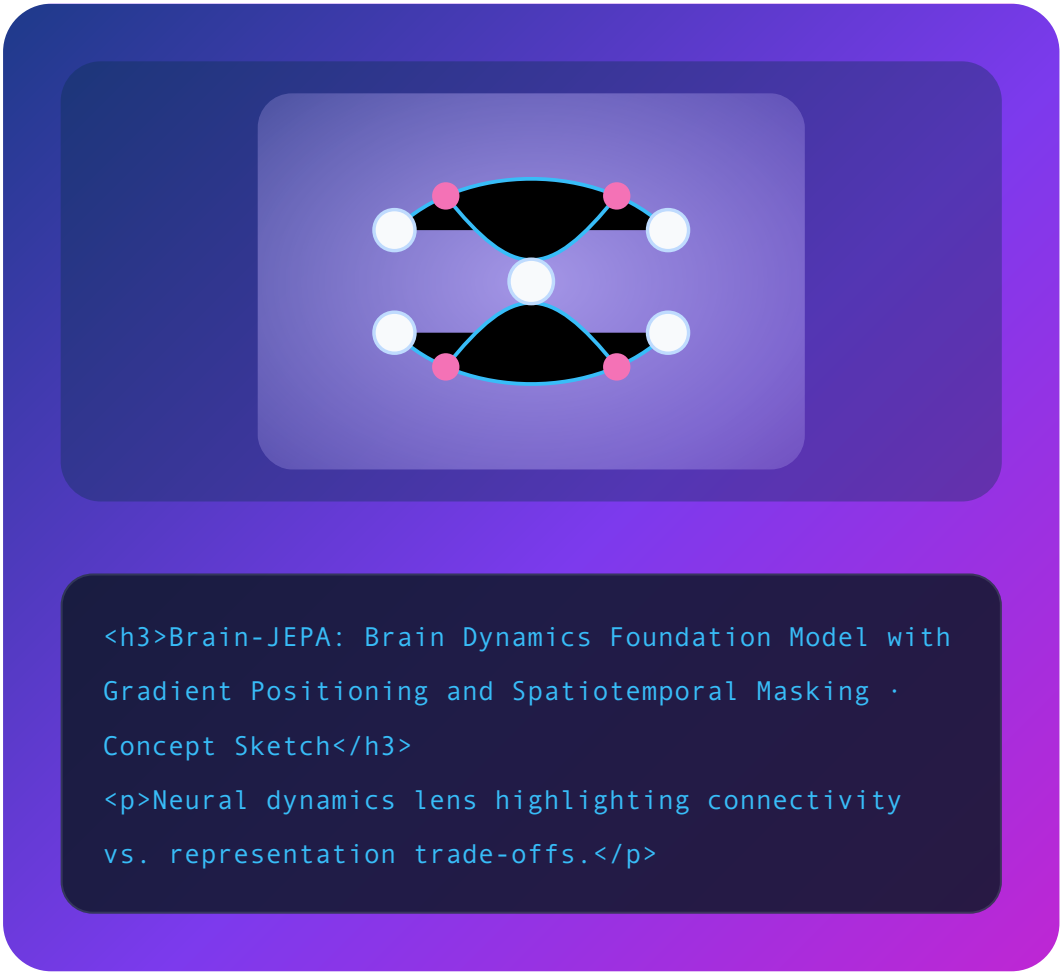

Brain-JEPA: Brain Dynamics Foundation Model with Gradient Positioning and Spatiotemporal Masking

Authors: 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

Year: 2024 (approx., based on arXiv:2409.19407 and NeurIPS 2024)

Venue: NeurIPS 2024 (Brain-JEPA: Brain Dynamics Foundation Model)

1. Classification

- **Domain Category:**
  - **Brain FM.** The paper develops a large self-supervised foundation model for fMRI time-series (resting-state fMRI across multiple cohorts) to support many downstream neuro-related tasks (demographics, traits, and disease prediction). It focuses entirely on brain activity recordings and their functional organization rather than genomics or generic multimodal integration.
- **FM Usage Type:**
  - **Core FM development.** Brain-JEPA is a new brain dynamics foundation model that adapts the Joint-Embedding Predictive Architecture (JEPA) framework to fMRI, introduces a novel functional positional encoding (Brain Gradient Positioning), and a customized pretraining mask (Spatiotemporal Masking). The paper primarily proposes this new architecture and training scheme, then applies it to a broad suite of downstream tasks.
- **Key Modalities:**
  - Resting-state fMRI BOLD time series (450 ROI-level time series, including cortical and subcortical regions).
  - Associated clinical / behavioral labels for downstream tasks (age, sex, cognitive scores, disease status), but the core model itself is pretrained purely on fMRI time series.

---

2. Executive Summary

This paper introduces **Brain-JEPA**, a large-scale **brain dynamics foundation model** that learns from resting-state fMRI time series using a **Joint-Embedding Predictive Architecture (JEPA)** instead of reconstructing raw signals. The authors argue that fMRI BOLD signals are noisy and sparsely informative, making direct reconstruction (as in the earlier BrainLM model) suboptimal, especially for off-the-shelf evaluations

like linear probing. Brain-JEPA instead predicts latent representations of masked fMRI patches using a Vision Transformer (ViT) encoder, a JEPA-style predictor, and an exponential moving average (EMA) target encoder. Two core innovations tailor JEPA to brain data: **Brain Gradient Positioning**, which uses functional connectivity gradients to define a **functional coordinate system** for regions of interest (ROIs), and **Spatiotemporal Masking**, which structures the self-supervised prediction task across ROIs and timesteps to provide a strong inductive bias. Pretrained on large-scale UK Biobank fMRI, Brain-JEPA achieves **state-of-the-art performance** across a wide range of downstream tasks (demographics, cognitive traits, and disease classification) and generalizes well across ethnic groups and external datasets. For a new grad student, this paper is important as a blueprint for how to construct and pretrain **foundation models for brain time series**, and as a demonstration that latent prediction architectures plus principled positional encodings can yield robust, generalizable brain representations.

---

3. Problem Setup and Motivation

- **Scientific / practical problem:**
  - Build a **foundation model for brain dynamics** that learns general-purpose representations of resting-state fMRI time series, which can be adapted to many downstream tasks: demographic prediction (age, sex), cognitive and personality traits (e.g., Neuroticism, Flanker score), and disease diagnosis/prognosis (e.g., Alzheimer's, mild cognitive impairment, amyloid positivity).
  - Move beyond **task-specific fMRI models** (e.g., BrainNetCNN, BrainGNN, Brain Network Transformer, SwiFT) that must be trained separately for each dataset/task and often fail to exploit large unlabeled fMRI repositories.

    - Improve on **BrainLM**, the first fMRI foundation model that uses a masked autoencoder (MAE) to reconstruct masked BOLD time series patches, which performs well under heavy fine-tuning but has weaker off-the-shelf representations and limited evaluation across ethnicities.

- **Why this is hard:**
    - **Noisy, low-SNR data:** fMRI BOLD signals are an indirect proxy for neural activity, influenced by physiological noise and scanner artifacts, with relatively low signal-to-noise ratio. Directly reconstructing all masked voxels or ROI time series can encourage modeling noise rather than meaningful structure.
    - **Sparse, complex spatiotemporal structure:** Unlike images (with local edges and dense spatial information), fMRI signals are **sparse and distributed** across ROIs and time, without clear local edges; reconstructive MAE losses can struggle to learn subtle patterns in such data.
    - **Lack of natural spatial ordering:** Transformers rely on positional embeddings, but there is no simple 1D order for ROIs across the 3D brain. Anatomical coordinates do not necessarily correspond to functional organization; spatially adjacent ROIs can have very different activity profiles.
    - **Heterogeneous time-series patches:** fMRI patches differ both in space (ROI location in functional networks) and time (brain states, tasks, intrinsic fluctuations). Masking and prediction strategies must reflect this heterogeneity to avoid trivial shortcuts and encourage robust learning.
    - **Generalization across cohorts and ethnicities:** Foundation models for clinical use must generalize beyond the pretraining cohort (e.g., UK Biobank, mostly Caucasian) to external datasets and different ethnic groups, which is challenging given domain shift in acquisition protocols and demographics.

4. Data and Modalities

- **Datasets used:**
  - **UK Biobank (UKB):**
    - Large-scale public dataset (resting-state fMRI) with **40,162** participants aged 44–83.
    - Used for **self-supervised pretraining** and internal downstream tasks (age and sex prediction); 80% for pretraining, 20% held-out for evaluation.
  - **HCP-Aging (Human Connectome Project – Aging):**
    - 656 healthy elderly participants with resting-state fMRI.
    - Used for **external evaluation** of demographics (age, sex) and behavioral traits (Neuroticism, Flanker score).
  - **ADNI (Alzheimer's Disease Neuroimaging Initiative):**
    - Resting-state fMRI for 189 participants (NC vs. MCI classification) and 100 cognitively normal participants (amyloid positive vs. negative classification).
  - **MACC (Memory, Ageing and Cognition Centre; Asian cohort):**
    - Resting-state fMRI for 539 participants, used for NC vs. MCI classification in Asian participants to test cross-ethnic generalization.
  - **Additional external datasets (Appendix):**
    - **OASIS-3:** AD conversion prediction in MCI participants.
    - **CamCAN:** Depression diagnosis.
- **Modalities:**
  - **Primary modality:** Resting-state fMRI BOLD time series.
  - Labels or metadata include age, sex, cognitive and personality scores (e.g., Neuroticism, Flanker), and disease/prognosis labels (NC vs. MCI, amyloid status, AD conversion, depression).
- **Preprocessing / representation:**
  - **Parcellation:** All fMRI data is parcellated into **n = 450 ROIs**, using:
    - Schaefer-400 atlas for cortical regions.

        - Tian-Scale III atlas for subcortical regions.
    - **Scaling:** Robust scaling per ROI (subtract median, divide by interquartile range) across participants.
    - **Temporal resolution alignment:**
        - UKB and HCP-Aging: multiband acquisition (TR ≈ 0.735s).
        - ADNI and MACC: single-band acquisition (TR ≈ 2s).
        - Multi-band data are **downsampled (stride 3)** to align all datasets to TR ≈ 2s.
    - **Input size:** Default model input is **450 × 160** (ROIs × timesteps).
    - **Patchify and shuffle:** For JEPA pretraining, time series are patchified into temporal patches per ROI (e.g., 10 patches per ROI), with ROI shuffling and spatiotemporal partitioning into observation and target regions.

If any fine-grained preprocessing details are missing in the main text, they are specified in Appendix B and table summaries, but the above captures the key design choices.

---

5. Model / Foundation Model

- **Model Type:**
    - **Vision Transformer (ViT)-based JEPA model for fMRI time series**, using a **latent predictive architecture** rather than reconstruction.
    - Core components:
        - **Observation encoder**: ViT (ViT-Small, ViT-Base, or ViT-Large).
        - **Target encoder**: ViT with EMA-updated parameters (JEPA-style).
        - **Predictor network**: a narrower ViT that maps the observation representation to predicted target representations.

- Architecturally, this is analogous to I-JEPA for images, but adapted to fMRI time series with specific choices for positional encoding and masking.

- **Is it a new FM or an existing one?**
  - **New foundation model.** Brain-JEPA is not simply an application of an existing FM; it adapts the JEPA framework to brain dynamics and introduces **Brain Gradient Positioning** and **Spatiotemporal Masking** as domain-specific innovations.

- **Key components and innovations:**

Component	Description
Observation encoder (f_)	ViT backbone (ViT-S, ViT-B, or ViT-L), processes an **observation block** (subset of ROIs × timesteps).
Target encoder (f_) (EMA)	Same architecture as observation encoder, parameters updated via **Exponential Moving Average** of (f_).
Predictor (g_)	Narrower ViT that takes observation representations and positional embeddings to predict target block embeddings.
Brain Gradient Positioning	Functional connectivity gradient-based **spatial positional embedding** that defines a functional coordinate system for ROIs.
Temporal Positioning	Standard sine–cosine positional encoding applied along the time dimension for each ROI.

Component	Description
Spatiotemporal Masking	Partitioning of the input into **Cross-ROI (α)**, **Cross-Time (β)**, and **Double-Cross (γ)** regions, plus overlapped sampling to control masking ratios.

- **Brain Gradient Positioning (spatial positional encoding):**
  - Compute a **non-negative affinity matrix** (A(i,j)) from functional connectivity features (c_i, c_j) across ROIs using a cosine-based similarity.
  - Use **diffusion maps** to compute gradients (eigenvectors) that capture macroscale functional organization; each gradient is a dimension of a latent manifold.
  - Stack eigenvectors into a gradient matrix (G ^{n m}) (n ROIs, m gradient components; default m = 30).
  - Map (G) through a trainable linear layer to (^{n d/2}), where (d) is the ViT embedding dimension.
  - Combine () with temporal positional encoding (T ^{n d/2}) to get final positional embeddings (P = [T,] ^{n d}).
  - This yields a **functional coordinate system** where distances reflect connectivity similarity, enabling the model to respect functional rather than purely anatomical proximity.
- **Spatiotemporal Masking and targets:**
  - The fMRI input (after patchifying and ROI shuffling) is divided into an **observation block** and three **non-overlapping regions** for targets:
    - **Cross-ROI (α):** Same time range as observation but different ROIs → forces spatial generalization across ROIs.
    - **Cross-Time (β):** Same ROIs but different timestep patches → forces temporal forecasting/generalization.

        - **Double-Cross (γ):** Different ROIs and timesteps → most challenging region, requiring generalization across space and time.
    - Sample K target blocks from each region (K = 1 in experiments), giving multiple prediction sub-tasks per sample.
    - Use **overlapped sampling** to flexibly adjust observation-to-input ratio and encourage diverse masking patterns.
    - Remove overlaps between observation and α/β targets to ensure non-trivial prediction.
    - Loss is mean squared error in **latent space** between predicted target embeddings (^r_y) and target encoder outputs (s^r_y), averaged over regions and blocks.

- **Training setup (as far as available):**
    - **Pretraining objective:**
        - **Joint-Embedding Predictive Architecture (JEPA)-style latent prediction:**
            - Given observation encoding (s_x) and positional embeddings (P), predictor (g_) outputs (^r_y = g_(s_x | P)) for each target region r; minimize (| ^r_y - s^r_y |_2^2) averaged over r and targets.
            - Crucially, the model **predicts representations**, not raw fMRI signals, which helps ignore noise and focus on more abstract features.
    - **Model scale:**
        - ViT-S (~22M parameters), ViT-B (~86M), ViT-L (~307M) for the observation encoder.
        - Predictors have matching architectures but shallower depth and smaller embedding dimensions (e.g., 6 layers with dim 192/384, etc.).
        - No [CLS] token in pretraining; for downstream evaluation, they use the **target encoder** and average pooling over patches to obtain global fMRI embeddings.

- **Optimization and hyperparameters (pretraining):**
  - Optimizer: **AdamW** with carefully tuned weight decay and cosine scheduling.
  - Learning rate: warmup cosine schedule, peak LR around 1e-3 with warmup and final LR 1e-6; weight decay schedule from 0.04 to 0.4.
  - EMA momentum increases from 0.996 to 1.0.
  - Batch size: effective 4 GPUs × 8 gradient accumulation steps × 16 batch size.
  - Training epochs: **300** epochs for main ViT-B model.
  - Patch size p = 16 time points; gradient vector dimension m = 30.
- **Downstream evaluation:**
  - Fine-tuning and linear probing use AdamW or LARS with cosine decay, 50 training epochs, and dataset-specific settings.
  - For linear probing, they add a **BatchNorm layer** before the linear head, following MAE practice, to stabilize learning.

---

6. Multimodal / Integration Aspects (If Applicable)

- **Is the paper multimodal?**
  - **Not in the sense of integrating multiple different data modalities.** Brain-JEPA operates on a single modality—resting-state fMRI time series. The "multimodality" here is spatiotemporal within the fMRI itself (ROIs × time), not multiple distinct biological modalities (like EEG + MRI or fMRI + genetics).
  - The paper does, however, span **multiple datasets and cohorts** (UKB, HCP-Aging, ADNI, MACC, OASIS-3, CamCAN), which could be viewed as multi-site integration, but there is no explicit multimodal fusion mechanism.

- **Relation to integration baseline plan:**
  - The **Integration Baseline Plan** emphasizes **late fusion** of modality-specific features, careful covariate adjustment, CCA-based cross-modality exploration, and robustness-focused evaluation (e.g., standardized preprocessing, consistent CV folds, significance testing).
  - Brain-JEPA's design is conceptually aligned with the plan's priority to **preserve modality-specific signals** (here, fMRI) by learning a strong fMRI foundation model before any cross-modal integration. The model produces high-quality, compressed fMRI embeddings that could be **plugged into late fusion schemes** (e.g., concatenation with genomic or clinical features followed by logistic regression or GBDT).
  - The evaluation practices—multiple independent runs, reporting means and standard deviations, and significance markers (*)—mirror the **robustness and evaluation discipline** recommended in the plan (e.g., repeated runs, clear metrics, statistical testing).
  - Although the paper does not explicitly use **CCA, partial correlations, or complex multimodal fusion**, it provides the **fMRI tower** that future work could combine with genomic, behavioral, or structural MRI towers using late fusion or more advanced contrastive frameworks, directly aligning with the "modality sequencing" and "escalation" steps of the plan.

---

7. Experiments and Results

- **Tasks / benchmarks:**
  - **Internal tasks on UKB (20% held-out):**
    - Age prediction (regression; MSE and Pearson correlation).
    - Sex prediction (binary classification; accuracy and F1).

- **External tasks on HCP-Aging:**
  - Age prediction.
  - Sex prediction.
  - **Neuroticism** score prediction (personality trait).
  - **Flanker** task performance prediction (attention / inhibitory control).
- **External tasks on ADNI and MACC:**
  - NC vs. MCI classification (ADNI; Caucasian cohort).
  - Amyloid-positive vs. negative classification (ADNI).
  - NC vs. MCI classification in Asian participants (MACC).
- **Additional tasks (Appendix C):**
  - AD conversion prediction (OASIS-3; MCI participants).
  - Depression diagnosis (CamCAN).

- **Baselines:**
  - **Task-specific fMRI models:**
    - BrainNetCNN (CNN-based on connectivity matrices).
    - BrainGNN (graph neural network on ROI-level graphs).
    - Brain Network Transformer (BNT; transformer-based on connectivity).
    - SwiFT (Swin Transformer-based model for raw fMRI).
  - **Non-deep baselines:**
    - SVM/SVR trained on connectivity features.
  - **fMRI foundation models and self-supervised baselines:**
    - BrainLM (MAE-based fMRI foundation model, ViT-B backbone).
    - BrainMass (graph-based large-scale self-supervised model for brain networks).
    - CSM (text-like representation model for brain data).
  - Variants and ablations of Brain-JEPA:
    - Using **anatomical locations** or sine–cosine spatial embeddings instead of Brain Gradient Positioning.

        - Using **standard multi-block sampling** instead of Spatiotemporal Masking.
        - Changing number of gradient components (3 vs. 30).
        - Removing the JEPA framework (i.e., MAE/BrainLM-like framework with contributions).

- **Key findings (high-level trends):**
    - **Overall performance:**
        - Brain-JEPA achieves **state-of-the-art** performance on almost all downstream tasks compared with task-specific models (BrainNetCNN, BrainGNN, BNT, SwiFT), non-deep baselines (SVM/SVR), and prior foundation models (BrainLM, BrainMass, CSM).
        - On UKB, Brain-JEPA substantially reduces age prediction error and improves correlation and sex classification metrics compared to BrainLM and others.
        - On HCP-Aging, Brain-JEPA shows **higher Pearson correlation** for age, **higher accuracy/F1** for sex, and strong improvements for Neuroticism and Flanker, especially in personality and cognition, where it surpasses BrainLM by a notable margin.
    - **Disease diagnosis and prognosis:**
        - On ADNI and MACC, Brain-JEPA achieves top or near-top performance in **NC vs. MCI**, **amyloid positivity**, and **NC vs. MCI (Asian)** tasks.
        - Particularly, Brain-JEPA shows **superior performance in NC vs. MCI classification for Asian participants**, even though it was **pretrained only on Caucasian UKB data**, indicating strong cross-ethnic generalization.
        - On OASIS-3 and CamCAN, Brain-JEPA again outperforms many baselines for AD conversion and depression diagnosis tasks.

- **Scaling with model size and dataset size:**
  - Performance **increases consistently** as model size grows from ViT-S to ViT-B to ViT-L, mirroring scaling laws seen in vision and language FMs.
  - Likewise, using larger fractions of the UKB pretraining data (25% → 50% → 75% → 100%) yields monotonic improvements in downstream metrics, indicating that Brain-JEPA benefits from more data.
- **Fine-tuning vs. linear probing:**
  - Brain-JEPA achieves strong performance under **fine-tuning**, but also shows **excellent linear probing results**, outperforming BrainLM in both regimes.
  - The **performance drop** from fine-tuning to linear probing is **smaller for Brain-JEPA** than for BrainLM, suggesting that Brain-JEPA learns more robust and semantically rich representations during pretraining.
- **Ablation: Brain Gradient Positioning & Spatiotemporal Masking:**
  - Replacing Brain Gradient Positioning with sine–cosine or anatomical location embeddings **hurts performance** across age, sex, and NC vs. MCI tasks.
  - This indicates that gradient-based functional coordinates better capture the brain's functional architecture and benefit representation learning.
  - Replacing Spatiotemporal Masking with standard multi-block sampling leads to **slower pretraining and lower peak performance**; Brain-JEPA with Spatiotemporal Masking reaches or exceeds the ablated model's best performance in **fewer epochs**, emphasizing its efficient inductive bias.
- **Interpretability:**
  - Analysis of self-attention patterns across canonical brain networks (DMN, control network, salience/ventral attention, limbic, etc.) in NC vs. MCI tasks shows that Brain-JEPA

focuses on networks known to be implicated in cognitive impairment.

- Attention distributions are consistent across Caucasian and Asian cohorts, suggesting that the model has learned **robust, neurobiologically meaningful patterns**.

---

8. Strengths, Limitations, and Open Questions

- **Strengths:**
  - **Tailored JEPA architecture for brain dynamics:** The model combines JEPA-style latent prediction with fMRI-specific positional encoding and masking, addressing the limitations of MAE-style reconstruction for noisy, low-SNR BOLD signals.
  - **Principled functional positional encoding:** Brain Gradient Positioning offers a compelling way to incorporate functional connectivity gradients into transformer positional embeddings, grounding the model in known macroscale brain organization.
  - **Strong and broad empirical performance:** Brain-JEPA achieves state-of-the-art results across many tasks, datasets, and ethnic groups, demonstrating both **versatility** (demographics, traits, multiple clinical tasks) and **generalization**.
  - **Improved off-the-shelf representations:** Superior linear probing performance and smaller fine-tuning–to–linear-probe gaps suggest that Brain-JEPA learns more **semantic and robust representations** than prior MAE-based fMRI FMs.
  - **Interpretable attention patterns:** Network-level attention aligns with known neurobiological findings regarding cognitive impairment, providing a bridge between deep models and neuroscientific insights.
- **Limitations:**
  - **Model and compute scale:** The largest model used is ViT-L (~307M parameters); larger models (e.g., ViT-H) are not explored

due to compute limits, leaving open how far scaling could push performance.

- **Dataset diversity for pretraining:** Pretraining is primarily on UKB (mostly Caucasian), with external datasets used only for downstream evaluation. A more diverse, multi-ethnic, multi-site pretraining corpus might further improve robustness and fairness.
- **Single-modality focus:** While the paper hints at future multimodal integration (e.g., MEG, EEG, structural MRI), the current model only handles resting-state fMRI, so it does not directly address multimodal fusion challenges.
- **Complexity of gradients and masking:** Gradient computation, diffusion maps, and spatiotemporal masking introduce additional complexity and hyperparameters that may be non-trivial to implement and tune for new labs.
- **Interpretability depth:** Although some attention analyses are provided, rich causal or mechanistic interpretation of what the model has learned (e.g., specific ROIs and pathways) remains limited and could be extended.

- **Open Questions and Future Directions:**
  - How does **scaling up** the model size (e.g., ViT-H or mixture-of-experts) and pretraining data diversity (multi-site, multi-ethnic, multi-task) affect generalization and fairness across populations?
  - Can Brain-JEPA embeddings serve as the fMRI tower in a **multimodal FM** that integrates genetics, structural MRI, behavior, or clinical data via late fusion, contrastive objectives, or cross-attention, aligning with the integration baseline plan?
  - How robust are Brain Gradient Positioning and Spatiotemporal Masking to changes in parcellation schemes (e.g., different atlases, voxel-level modeling) or to task-based fMRI instead of resting-state data?
  - Can we design **more interpretable JEPA objectives** or probing methods that link latent representations to specific cognitive processes or disease mechanisms, beyond network-level attention

summaries?

- What are the implications of using JEPA-style latent prediction (rather than generative reconstruction) for identifying causal relationships or performing counterfactual reasoning in brain data?

---

9. Context and Broader Impact

- **Position in the FM landscape:**
  - Brain-JEPA sits in the emerging class of **brain foundation models**, analogous to how GPT-like models function for text and ViT/MAE/Jepa-like models for images. It extends the **JEPA paradigm**—originally developed for 2D images—to **spatiotemporal fMRI dynamics**, emphasizing latent prediction over pixel/voxel reconstruction.
  - It can be seen as a counterpart to **BrainLM** (MAE-based fMRI FM), providing an alternative pretraining objective that prioritizes robust, abstract representations suitable for **off-the-shelf use** in diverse tasks.
- **Relation to well-known ideas:**
  - Conceptually, Brain-JEPA is like **"I-JEPA but for fMRI time series"**, with key adaptations for brain data. Instead of predicting missing pixels, it predicts representations of masked ROI-time patches.
  - Brain Gradient Positioning draws on the **functional gradient literature** in neuroscience, which uses diffusion maps to uncover macroscale cortical organization. This connects the model's positional encoding to a well-established neuro-scientific framework.
  - Spatiotemporal Masking is akin to **structured masking strategies** in self-supervised learning (e.g., masked patches in MAE), but tailored to enforce meaningful fMRI forecasting and generalization across ROIs and timesteps.

- **Why this paper is a useful reference:**
  - For a new grad student interested in **AI for neuroscience**, this paper is a strong example of how to adapt modern self-supervised architectures (JEPA, ViT) to the constraints and opportunities of brain data.
  - It provides a concrete recipe for building a **foundation model for fMRI**, including data preprocessing, model architecture, pretraining objective, and extensive evaluation across multiple datasets and tasks.
  - It also showcases how to connect **architectural choices to neuroscience concepts** (e.g., functional gradients, network-level attention), which is crucial for interdisciplinary work and for gaining acceptance in the neuroscience community.
  - Finally, it provides a robust fMRI representation that fits naturally into the **integration baseline plan** as the brain modality component to be later fused with other modalities using late fusion, CCA, or contrastive learning.

---

10. Key Takeaways (Bullet Summary)

- **Problem:**
  - Brain-JEPA aims to build a **general-purpose foundation model for brain dynamics** that can be adapted to many downstream tasks (demographics, traits, disease diagnosis/prognosis) using resting-state fMRI data.
  - Existing task-specific models and MAE-based fMRI foundation models (like BrainLM) struggle with noisy BOLD signals, limited generalizability, and weaker off-the-shelf representations.
- **Method / Model:**
  - Brain-JEPA adopts a **Joint-Embedding Predictive Architecture (JEPA)**: instead of reconstructing raw fMRI signals, it predicts **latent representations** of masked fMRI patches in a ViT-based

architecture.

- It introduces **Brain Gradient Positioning**, a functional embedding of ROIs derived from diffusion maps on functional connectivity, providing a brain-specific positional encoding that reflects macroscale functional organization.
- It proposes **Spatiotemporal Masking**, which partitions fMRI patches into cross-ROI, cross-time, and double-cross regions and uses overlapped sampling to encourage generalization across space and time with a strong inductive bias.
- The model is pretrained on large UK Biobank data using ViT-S/B/L backbones, with EMA target encoder and predictor networks, and evaluated via fine-tuning and linear probing on multiple external datasets.

- **Results:**
  - Across UKB, HCP-Aging, ADNI, MACC, OASIS-3, and CamCAN, Brain-JEPA achieves **state-of-the-art performance** on age, sex, trait prediction, and multiple disease diagnosis/prognosis tasks, outperforming task-specific models, SVM/SVR, and prior fMRI FMs like BrainLM and BrainMass.
  - It scales well with **model size and data size**, with larger ViT backbones and more pretraining data yielding better performance.
  - Brain-JEPA shows strong **linear probing performance** and smaller gaps between fine-tuning and linear probing than BrainLM, indicating more robust and transferable representations.
  - Ablation studies confirm that **Brain Gradient Positioning** and **Spatiotemporal Masking** are crucial for performance and training efficiency.
- **Why it matters:**
  - Brain-JEPA demonstrates that **latent prediction architectures** with neuroscience-informed positional encodings can produce powerful, generalizable representations of brain activity.
  - It provides a solid foundation for future **multimodal integration** where fMRI embeddings are combined with genetics, structural

imaging, or behavioral data using late fusion or contrastive methods, aligning well with broader integration plans.

- For students and researchers, Brain-JEPA is an influential reference on how to design, train, and evaluate **brain foundation models** at scale, bridging modern self-supervised ML techniques and contemporary neuroscience.
