

A RELATED WORK

Due to the space constraints in the main text, we present an extensive survey of related work on task-specific and self-supervised learning models for fMRI, and continuous-time dynamical modeling.

Task-specific Models for fMRI. Leveraging fMRI data for applications such as disease diagnosis and demographic trait inference has driven the development of a variety of supervised deep learning architectures. BrainNetCNN introduces convolutional kernels that operate directly on functional connectivity matrices, capturing both edge-to-node and node-to-graph interactions (Kawahara et al., 2017). BrainGNN builds on this by employing ROI-aware graph neural networks with a hierarchical pooling mechanism to focus on diagnostically salient regions (Li et al., 2021). More recently, BrainNetTF applies transformer encoders to functional connectivity matrices of ROI time-series and incorporates a cluster-sensitive readout layer for functional-module detection (Kan et al., 2022). While task-specific models achieve strong performance on their supervised tasks, they rely on labeled data and do not learn generalizable representations that transfer across diverse downstream applications.

This limitation motivates our focus on SSL, a paradigm designed to overcome the reliance on labeled data. By exploiting the abundance of unlabeled fMRI scans, SSL aims to distill rich, task-agnostic representations that generalize across a wide spectrum of downstream fMRI analyses.

Self-Supervised Learning for fMRI. SSL has emerged as a powerful paradigm for extracting rich, generalizable representations from large-scale unlabeled data in both vision and language domains, through approaches such as latent representation alignment-based (e.g., MoCo (He et al., 2020), BYOL (Grill et al., 2020)) and reconstruction-based methods (e.g., MAE (He et al., 2022)). More recently, JEPA has demonstrated the effectiveness of predicting latent representations across different views (Assran et al., 2023; Bardes et al., 2024). Existing SSL models for fMRI generally follow one of two routes as described in Figure 2: image-based and graph-based approaches.

First, image-based approaches, which patchify fMRI time-series into spatiotemporal tokens, represent each scan as a fixed grid of spatiotemporal tokens. BrainLM pioneered the image-style paradigm and used a MAE objective to reconstruct masked patches (Caro et al., 2024), and Brain-JEPA builds on this by replacing reconstruction with a JEPA that learns to predict the latent embedding of one view from another (Dong et al., 2024). Beyond the loss of fine-grained temporal dynamics during patchification and the difficulty of aligning temporal resolution across datasets as pointed out in Section 1, a key limitation of this design is its rigid dependence on sequence length in the SSL scenario. For example, a model pre-trained on UKB data with 160 time points grouped into 10 non-overlapping patches of length 16 expects exactly the same 10-token layout during inference. However, datasets such as ABIDE contain scans with as few as 76 time points, which cannot be processed without upsampling. This length rigidity undermines the core premise of SSL, namely broad out-of-the-box transfer across heterogeneous unlabeled cohorts, and therefore poses a critical limitation for SSL in fMRI.

In contrast, graph-based approaches compress the fMRI time-series into a single static functional-connectivity (FC) graph, freeing them from constraints such as heterogeneous temporal resolution or scan-length variability; however, this coarse summarization inevitably discards fine-grained temporal dynamics in BOLD signals. BrainMass, for example, generates augmented static graphs by randomly dropping time points to produce pseudo-FC variants, and leverages latent representation alignment and masked-ROI prediction (Yang et al., 2024). ST-JEMA tries to restore temporal information by partitioning the scan into multiple sliding windows and building a dynamic graph sequence on which it performs JEPA-style SSL objective (Choi et al., 2024). While this design captures slow changes in connectivity, it still averages out intra-window dynamics and therefore inherits the same limitations as image-based patchification such as loss of high-frequency temporal details and the practical difficulty of matching window size to the diverse TRs across different datasets as pointed out in Section 1.

To address these issues, we adopt continuous-time latent dynamical modeling. By viewing each fMRI scan as noisy observations of a stochastic differential equation, the model naturally aligns data with different TRs on a shared time axis, handles variable sequence lengths without resampling, and retains the fine-grained temporal dynamics that patch-based or static-graph approaches inevitably discard. Moreover, we integrate two complementary self-supervised objectives, MAE reconstruction and JEPA alignment, into a unified stochastic optimal control framework, yielding a single training objective that preserves fine-grained temporal structure and prevents representation collapse.

Continuous Dynamical Models. Continuous-time dynamical models have been developed to capture irregular time-series dynamics and uncertainty. Neural ODEs parameterize a smooth vector field via a neural network and solve an initial-value problem to fit observations (Chen et al., 2018), while Latent ODEs extend this by learning a global latent initial state and replacing standard RNN encoders with ODE-based encoders (Rubanova et al., 2019). GRU-ODE-B further incorporates a Bayesian update rule into a GRU-driven ODE for online uncertainty quantification (De Brouwer et al., 2019), and on the stochastic front, Latent-SDE and Latent-SDEH introduce variational inference schemes for SDE-driven latent trajectories (Li et al., 2020; Zeng et al., 2023). Classical continuous–discrete state-space models like CD-SSM generalize discrete transitions to SDE-governed updates (Jazwinski, 2007), inspiring neural CD-SSM variants that leverage locally linear approximations for irregular real-world series (Schirmer et al., 2022; Ansari et al., 2023).

Although these frameworks excel at modeling temporal dynamics and noise, they are designed as task-specific predictors and rely on costly numerical solvers for integration and moment inference, making them impractical for large-scale, high-dimensional fMRI pretraining. Recently, (Park et al., 2024) proposed an efficient task-specific method for amortized inference and simulation-free latent dynamics in CD-SSMs, altogether enabling scalable modeling of irregularly sampled time-series without expensive numerical solvers. Inspired by this work, we introduce a simulation-free SSL framework that defines a closed-form latent dynamics objective to capture both temporal evolution and uncertainty in fMRI signals, designed to enable highly efficient and scalable pretraining.

B PROOFS AND DERIVATIONS

B.1 PROOF OF PROPOSITION 2.1

Proof. Bayesian filtering and smoothing techniques Särkkä (2013) are foundational tools for estimating latent states in probabilistic dynamical systems. The goal is to recover the posterior distribution of the latent trajectories given the observations:

$$p(\mathbf{X}_{[0,T]} | \mathcal{Y}) = \frac{1}{Z(\mathcal{Y})} \prod_{t \in \mathcal{T}} g_\psi(\mathbf{y}_{t_i} | \mathbf{X}_{t_i}) p(\mathbf{X}_{[0,T]}), \quad (12)$$

where $Z(\mathcal{Y})$ is a normalization constant:

$$Z(\mathcal{Y}) = \mathbb{E}_{\mathbf{X} \sim (1)} \left[\prod_{t \in \mathcal{T}} g_\psi(\mathbf{y}_{t_i} | \mathbf{X}_{t_i}) \right]. \quad (13)$$

Expanding this expectation, we have

$$\log Z(\mathcal{Y}) = \log \mathbb{E}_{\mathbf{X} \sim (1)} [p(\mathcal{Y} | \mathbf{X}_{[0,T]})] \quad (14)$$

$$= \log \mathbb{E}_{\mathbf{X} \sim (1)} \left[p(\mathcal{Y} | \mathbf{X}_{[0,T]}^\alpha) \frac{p(\mathbf{X}_{[0,T]})}{p(\mathbf{X}_{[0,T]}^\alpha)} \right] \quad (15)$$

$$\stackrel{(i)}{\geq} \mathbb{E}_{\mathbf{X} \sim (10)} \left[\log p(\mathcal{Y} | \mathbf{X}_{[0,T]}^\alpha) + \log \frac{p(\mathbf{X}_{[0,T]})}{p(\mathbf{X}_{[0,T]}^\alpha)} \right] \quad (16)$$

$$= \mathbb{E}_{\mathbf{X} \sim (10)} \left[\sum_{t \in \mathcal{T}} g(\mathbf{y}_t | \mathbf{X}_t^\alpha) + \log \frac{p(\mathbf{X}_{[0,T]})}{p(\mathbf{X}_{[0,T]}^\alpha)} \right] \quad (17)$$

$$\stackrel{(ii)}{=} \mathbb{E}_{\mathbf{X} \sim (10)} \left[\sum_{t \in \mathcal{T}} g(\mathbf{y}_t | \mathbf{X}_t^\alpha) - \frac{1}{2} \int_0^T \|\alpha_t\|^2 dt + \int_0^1 \alpha_t d\mathbf{W}_s \right] \quad (18)$$

$$\stackrel{(iii)}{=} \mathbb{E}_{\mathbf{X} \sim (10)} \left[\sum_{t \in \mathcal{T}} g(\mathbf{y}_t | \mathbf{X}_t^\alpha) - \frac{1}{2} \int_0^T \|\alpha_t\|^2 dt \right] \quad (19)$$

$$= -\mathcal{J}(\alpha, \mathcal{Y}), \quad (20)$$

where (i) results from Jensen’s inequality, (ii) follows by applying Girsanov’s theorem (Baldi, 2017, Theorem 12.1), and in the final equality, (iii) holds because \mathbf{W}_t is a martingale process.

□

B.2 PROOF OF THEOREM 2.2

Proof. Since each SPD matrix \mathbf{D}_t for $t \in \mathcal{T}$ admits an eigen-decomposition $\mathbf{D}_{t_i} = \mathbf{V} \mathbf{\Lambda}_{t_i} \mathbf{V}^\top$, we can transform the original process \mathbf{X}_t^α , which is expressed in the canonical basis, into a new process $\hat{\mathbf{X}}_t^\alpha = \mathbf{V}^\top \mathbf{X}_t^\alpha$ that resides in the space spanned by the eigenbasis \mathbf{V} . With this transformation, the dynamics in (10) can be rewritten, for any interval $[t_i, t_{i+1})$, as:

$$d\hat{\mathbf{X}}_t^\alpha = \left[-\mathbf{\Lambda}_{t_i} \hat{\mathbf{X}}_t^\alpha + \alpha_{t_i} \right] dt + d\hat{\mathbf{W}}_t, \quad (21)$$

where $\hat{\mathbf{X}}_t^\alpha = \mathbf{V}^\top \mathbf{X}_t^\alpha$, $\hat{\alpha}_{t_i} = \mathbf{V}^\top \alpha_{t_i}$, $\hat{\mathbf{W}}_t = \mathbf{V}^\top \mathbf{W}_t$ and initial condition $\hat{\mathbf{X}}_0^\alpha \sim \mathcal{N}(\hat{\mu}_0, \hat{\Sigma}_0)$ with $\hat{\mu}_0 = \mathbf{V}^\top \mu_0$ and $\hat{\Sigma}_0 = \mathbf{V}^\top \Sigma_0 \mathbf{V}$. Since \mathbf{V} is orthonormal, $\hat{\mathbf{W}}_t$ retains the distribution $\hat{\mathbf{W}}_t \stackrel{d}{=} \mathbf{W}_t$ for all $t \in [0, T]$, allowing $\hat{\mathbf{W}}_t$ to be treated as a standard Wiener process. Now, given that $\mathbf{\Lambda}_{t_i}$ is diagonal, the linear SDE in equation (21) admits a closed-form solution for any $t \in [t_i, t_{i+1})$:

$$\hat{\mathbf{X}}_t^\alpha = e^{-(t-t_i)\mathbf{\Lambda}_{t_i}} \left(\hat{\mathbf{X}}_{t_i}^\alpha + \int_{t_i}^t e^{(s-t_i)\mathbf{\Lambda}_{t_i}} \hat{\alpha}_{t_i} ds + \int_{t_i}^t e^{(s-t_i)\mathbf{\Lambda}_{t_i}} d\hat{\mathbf{W}}_s \right). \quad (22)$$

Since the initial condition $\hat{\mathbf{X}}_0^\alpha$ is Gaussian and the SDE is linear with Gaussian noise, the process $\hat{\mathbf{X}}_t^\alpha$ remains Gaussian. Therefore, its first two moments—the mean and covariance—can be derived from the solution above. To derive the moments, we firstly evaluate the deterministic integral involving $\hat{\alpha}_{t_i}$:

$$\int_{t_i}^t e^{(s-t_i)\mathbf{\Lambda}_{t_i}} \hat{\alpha}_{t_i} ds = -\mathbf{\Lambda}_{t_i}^{-1} \left(\mathbf{I} - e^{(t-t_i)\mathbf{\Lambda}_{t_i}} \right) \hat{\alpha}_{t_i}. \quad (23)$$

Taking the expectation of $\hat{\mathbf{X}}_t^\alpha$, and using the martingale property of the Wiener process $\hat{\mathbf{W}}_t$, we obtain:

$$\hat{\mu}_t = \mathbb{E}_{\hat{\mathbf{X}}^\alpha \sim (21)} \left[\hat{\mathbf{X}}_t^\alpha \right] = e^{-(t-t_i)\mathbf{\Lambda}_{t_i}} \hat{\mu}_{t_i} - e^{-(t-t_i)\mathbf{\Lambda}_{t_i}} \mathbf{\Lambda}_{t_i}^{-1} \left(\mathbf{I} - e^{(t-t_i)\mathbf{\Lambda}_{t_i}} \right) \hat{\alpha}_{t_i}. \quad (24)$$

Next, compute the covariance of $\hat{\mathbf{X}}_t^\alpha$:

$$\hat{\Sigma}_t = \mathbb{E}_{\hat{\mathbf{X}}^\alpha \sim (21)} \left[e^{-2(t-t_i)\mathbf{\Lambda}_{t_i}} \left(\mathbf{X}_{t_i} - \mu_{t_i} + \int_{t_i}^t e^{(s-t_i)\mathbf{\Lambda}_{t_i}} d\hat{\mathbf{W}}_s \right) \left(\mathbf{X}_{t_i} - \mu_{t_i} + \int_{t_i}^t e^{(s-t_i)\mathbf{\Lambda}_{t_i}} d\hat{\mathbf{W}}_s \right)^\top \right] \quad (25)$$

$$= e^{-2(t-t_i)\mathbf{\Lambda}_{t_i}} \mathbb{E}_{\hat{\mathbf{X}}^\alpha \sim (21)} \left[(\mathbf{X}_{t_i} - \mu_{t_i}) (\mathbf{X}_{t_i} - \mu_{t_i})^\top + \left\| \int_{t_i}^t e^{(s-t_i)\mathbf{\Lambda}_{t_i}} d\hat{\mathbf{W}}_s \right\|_2^2 \right] \quad (26)$$

$$\stackrel{(i)}{=} e^{-2(t-t_i)\mathbf{\Lambda}_{t_i}} \mathbb{E}_{\hat{\mathbf{X}}^\alpha \sim (21)} \left[(\mathbf{X}_{t_i} - \mu_{t_i}) (\mathbf{X}_{t_i} - \mu_{t_i})^\top + \int_{t_i}^t e^{2(s-t_i)\mathbf{\Lambda}_{t_i}} ds \right] \quad (27)$$

$$\stackrel{(ii)}{=} e^{-2(t-t_i)\mathbf{\Lambda}_{t_i}} \hat{\Sigma}_{t_i} - \frac{1}{2} e^{-2(t-t_i)\mathbf{\Lambda}_{t_i}} \mathbf{\Lambda}_{t_i}^{-1} \left(\mathbf{I} - e^{2(t-t_i)\mathbf{\Lambda}_{t_i}} \right), \quad (28)$$

where (i) follows from the martingale property of $\hat{\mathbf{W}}_t$ and (ii) follows from Itô isometry:

$$\mathbb{E}_{\hat{\mathbf{X}}^\alpha \sim (21)} \left[\left\| \int_{t_i}^t e^{(s-t_i)\mathbf{\Lambda}_{t_i}} d\hat{\mathbf{W}}_s \right\|_2^2 \right] = \mathbb{E}_{\hat{\mathbf{X}}^\alpha \sim (21)} \left[\int_{t_i}^t e^{2(s-t_i)\mathbf{\Lambda}_{t_i}} ds \right]. \quad (29)$$

Using the recursive forms for the mean and covariance, we can determine these moments at each discrete time step t_i . For the mean $\hat{\mu}_{t_i}$, the recurrence relation is:

$$\hat{\mu}_{t_1} = e^{-(t_1-t_0)\mathbf{\Lambda}_{t_0}} \hat{\mu}_{t_0} - e^{-(t_1-t_0)\mathbf{\Lambda}_{t_1}} \mathbf{\Lambda}_{t_0}^{-1} \left(\mathbf{I} - e^{(t_1-t_0)\mathbf{\Lambda}_{t_0}} \right) \hat{\alpha}_{t_0} \quad (30)$$

$$\begin{aligned} \hat{\mu}_{t_2} &= e^{-\sum_{j=0}^1 (t_{j+1}-t_j)\mathbf{\Lambda}_{t_j}} \hat{\mu}_{t_0} \\ &\quad - e^{-\sum_{j=0}^1 (t_{j+1}-t_j)\mathbf{\Lambda}_{t_j}} \mathbf{\Lambda}_{t_0}^{-1} \left(\mathbf{I} - e^{(t_1-t_0)\mathbf{\Lambda}_{t_0}} \right) \hat{\alpha}_{t_0} \\ &\quad - e^{-(t_2-t_1)\mathbf{\Lambda}_{t_1}} \mathbf{\Lambda}_{t_1}^{-1} \left(\mathbf{I} - e^{(t_2-t_1)\mathbf{\Lambda}_{t_1}} \right) \hat{\alpha}_{t_1} \end{aligned} \quad (31)$$

\vdots

$$\hat{\mu}_{t_i} = e^{-\sum_{j=0}^{i-1} (t_{j+1}-t_j)\mathbf{\Lambda}_{t_j}} \hat{\mu}_{t_0} - \sum_{l=0}^{i-1} e^{-\sum_{j=l}^{i-1} (t_{j+1}-t_j)\mathbf{\Lambda}_{t_j}} \mathbf{\Lambda}_{t_l}^{-1} \left(\mathbf{I} - e^{(t_{l+1}-t_l)\mathbf{\Lambda}_{t_l}} \right) \hat{\alpha}_{t_l} \quad (32)$$

Similarly, for the covariance $\hat{\Sigma}_{t_i}$, the recurrence relation is:

$$\hat{\Sigma}_{t_1} = e^{-2(t_1-t_0)\mathbf{\Lambda}_{t_0}} \hat{\Sigma}_{t_0} - \frac{1}{2} e^{-2(t_1-t_0)\mathbf{\Lambda}_{t_1}} \mathbf{\Lambda}_{t_0}^{-1} \left(\mathbf{I} - e^{2(t_1-t_0)\mathbf{\Lambda}_{t_0}} \right) \quad (33)$$

$$\begin{aligned} \hat{\Sigma}_{t_2} &= e^{-\sum_{j=0}^1 2(t_{j+1}-t_j)\mathbf{\Lambda}_{t_j}} \hat{\Sigma}_{t_0} \\ &\quad - \frac{1}{2} e^{-\sum_{j=0}^1 2(t_{j+1}-t_j)\mathbf{\Lambda}_{t_j}} \mathbf{\Lambda}_{t_0}^{-1} \left(\mathbf{I} - e^{2(t_1-t_0)\mathbf{\Lambda}_{t_0}} \right) - \frac{1}{2} e^{-2(t_2-t_1)\mathbf{\Lambda}_{t_1}} \mathbf{\Lambda}_{t_1}^{-1} \left(\mathbf{I} - e^{2(t_2-t_1)\mathbf{\Lambda}_{t_1}} \right) \end{aligned} \quad (34)$$

\vdots

$$\hat{\Sigma}_{t_i} = e^{-2\sum_{j=0}^{i-1} (t_{j+1}-t_j)\mathbf{\Lambda}_{t_j}} \hat{\Sigma}_{t_0} - \frac{1}{2} \sum_{l=0}^{i-1} e^{-2\sum_{j=l}^{i-1} (t_{j+1}-t_j)\mathbf{\Lambda}_{t_j}} \mathbf{\Lambda}_{t_l}^{-1} \left(\mathbf{I} - e^{2(t_{l+1}-t_l)\mathbf{\Lambda}_{t_l}} \right). \quad (35)$$

Now, since $\hat{\mathbf{X}}_t^\alpha = \mathbf{V}^\top \mathbf{X}_t^\alpha$, with $\hat{\mu}_0 = \mathbf{V}^\top \mu_0$ and $\hat{\Sigma}_0 = \mathbf{V}^\top \Sigma_0 \mathbf{V}$, we can express the mean and covariance in the original canonical basis as follows. For the mean $\hat{\mu}_{t \in \mathcal{T}}$, which is given by

$$\mathbf{V} \hat{\mu}_{t_i} = \mathbf{V} \left(e^{-\sum_{j=0}^{i-1} (t_{j+1}-t_j)\mathbf{\Lambda}_{t_j}} \hat{\mu}_{t_0} - \sum_{l=0}^{i-1} e^{-\sum_{j=l}^{i-1} (t_{j+1}-t_j)\mathbf{\Lambda}_{t_j}} \mathbf{\Lambda}_{t_l}^{-1} \left(\mathbf{I} - e^{(t_{l+1}-t_l)\mathbf{\Lambda}_{t_l}} \right) \hat{\alpha}_{t_l} \right) \quad (36)$$

$$= \mathbf{V} \left(e^{-\sum_{j=0}^{i-1} (t_{j+1}-t_j)\mathbf{\Lambda}_{t_j}} \mathbf{V}^\top \mu_0 - \sum_{l=0}^{i-1} e^{-\sum_{j=l}^{i-1} (t_{j+1}-t_j)\mathbf{\Lambda}_{t_j}} \mathbf{\Lambda}_{t_l}^{-1} \left(\mathbf{I} - e^{(t_{l+1}-t_l)\mathbf{\Lambda}_{t_l}} \right) \mathbf{V}^\top \alpha_{t_l} \right) \quad (37)$$

$$= e^{-\sum_{j=0}^{i-1} (t_{j+1}-t_j)\mathbf{D}_{t_j}} \mu_0 - \mathbf{V} \left(\sum_{l=0}^{i-1} e^{-\sum_{j=l}^{i-1} (t_{j+1}-t_j)\mathbf{\Lambda}_{t_j}} \mathbf{\Lambda}_{t_l}^{-1} \left(\mathbf{I} - e^{(t_{l+1}-t_l)\mathbf{\Lambda}_{t_l}} \right) \mathbf{V}^\top \alpha_{t_l} \right) \quad (38)$$

$$= e^{-\sum_{j=0}^{i-1} (t_{j+1}-t_j)\mathbf{D}_{t_j}} \mu_0 - \mathbf{V} \left(\sum_{l=0}^{i-1} e^{-\sum_{j=l}^{i-1} (t_{j+1}-t_j)\mathbf{\Lambda}_{t_j}} \mathbf{V}^\top \mathbf{D}_{t_l}^{-1} \mathbf{V} \left(\mathbf{I} - e^{(t_{l+1}-t_l)\mathbf{\Lambda}_{t_l}} \right) \mathbf{V}^\top \alpha_{t_l} \right) \quad (39)$$

$$= e^{-\sum_{j=0}^{i-1} (t_{j+1}-t_j)\mathbf{D}_{t_j}} \mu_0 - \sum_{l=0}^{i-1} e^{-\sum_{j=l}^{i-1} (t_{j+1}-t_j)\mathbf{D}_{t_j}} \mathbf{D}_{t_l}^{-1} \left(\mathbf{I} - e^{(t_{l+1}-t_l)\mathbf{D}_{t_l}} \right) \alpha_{t_l} \quad (40)$$

$$= \mu_{t_i} \quad (41)$$

where we used $\mathbf{D}_{t_j} = \mathbf{V} \mathbf{\Lambda}_{t_j} \mathbf{V}^\top$ and the orthonormality of \mathbf{V} . Similarly, for the covariance $\hat{\Sigma}_{t \in \mathcal{T}}$, we have

$$\mathbf{V}^{\hat{\Sigma}_{t_i}} \mathbf{V}^\top = \mathbf{V} \left(e^{-2 \sum_{j=0}^{i-1} (t_{j+1} - t_j) \mathbf{A}_{t_j}} \hat{\Sigma}_{t_0} - \frac{1}{2} \sum_{l=0}^{i-1} e^{-2 \sum_{j=l}^{i-1} (t_{j+1} - t_j) \mathbf{A}_{t_j}} \mathbf{A}_{t_l}^{-1} \left(\mathbf{I} - e^{2(t_{l+1} - t_l) \mathbf{A}_{t_l}} \right) \right) \mathbf{V}^\top \quad (42)$$

$$= \mathbf{V} \left(e^{-2 \sum_{j=0}^{i-1} (t_{j+1} - t_j) \mathbf{A}_{t_j}} \mathbf{V}^\top \Sigma_0 \mathbf{V} - \frac{1}{2} \sum_{l=0}^{i-1} e^{-2 \sum_{j=l}^{i-1} (t_{j+1} - t_j) \mathbf{A}_{t_j}} \mathbf{A}_{t_l}^{-1} \left(\mathbf{I} - e^{2(t_{l+1} - t_l) \mathbf{A}_{t_l}} \right) \right) \mathbf{V}^\top \quad (43)$$

$$= e^{-2 \sum_{j=0}^{i-1} (t_{j+1} - t_j) \mathbf{D}_{t_j}} \Sigma_0 - \mathbf{V} \left(\frac{1}{2} \sum_{l=0}^{i-1} e^{-2 \sum_{j=l}^{i-1} (t_{j+1} - t_j) \mathbf{A}_{t_j}} \mathbf{A}_{t_l}^{-1} \left(\mathbf{I} - e^{2(t_{l+1} - t_l) \mathbf{A}_{t_l}} \right) \right) \mathbf{V}^\top \quad (44)$$

$$= e^{-2 \sum_{j=0}^{i-1} (t_{j+1} - t_j) \mathbf{D}_{t_j}} \Sigma_0 - \mathbf{V} \left(\sum_{l=0}^{i-1} e^{-2 \sum_{j=l}^{i-1} (t_{j+1} - t_j) \mathbf{A}_{t_j}} \mathbf{V}^\top \mathbf{D}_{t_l}^{-1} \mathbf{V} \left(\mathbf{I} - e^{2(t_{l+1} - t_l) \mathbf{A}_{t_l}} \right) \right) \mathbf{V}^\top \quad (45)$$

$$= e^{-2 \sum_{j=0}^{i-1} (t_{j+1} - t_j) \mathbf{D}_{t_j}} \Sigma_0 - \frac{1}{2} \sum_{l=0}^{i-1} e^{-2 \sum_{j=l}^{i-1} (t_{j+1} - t_j) \mathbf{D}_{t_j}} \mathbf{D}_{t_l}^{-1} \left(\mathbf{I} - e^{2(t_{l+1} - t_l) \mathbf{D}_{t_l}} \right) \quad (46)$$

$$= \Sigma_{t_i} \quad (47)$$

Thus, both the mean μ_{t_i} and the covariance Σ_{t_i} of \mathbf{X}_t^α at each time step t_i are correctly recovered, completing the proof. \square

B.3 DERIVATION OF ELBO

We start the derivation by integrating the mixture distribution in (8) into the SOC problem (6) as follows:

$$\log p(\mathcal{Y}_{\text{tar}} | \mathbf{X}_{[0,T]}^\theta) = \log \int \zeta_\psi(\mathbf{y}_t | \tilde{\alpha}_t) \pi_{\tilde{\theta}}(\tilde{\alpha}_t | \mathbf{X}_t^\theta) d\tilde{\alpha}_t \quad (48)$$

$$= \log \int \zeta_\psi(\mathbf{y}_t | \tilde{\alpha}_t) \frac{1}{\mathbf{Z}(\mathbf{X}_t^\theta)} [p_\theta(\tilde{\alpha}_t | \mathbf{X}_t^\theta)^\lambda q_{\tilde{\theta}}(\tilde{\alpha}_t | \mathcal{Y}_{\text{tar}})^{1-\lambda}] d\tilde{\alpha}_t \quad (49)$$

$$= \log \int \zeta_\psi(\mathbf{y}_t | \tilde{\alpha}_t) \left[\frac{p_\theta(\tilde{\alpha}_t | \mathbf{X}_t^\theta)^\lambda q_{\tilde{\theta}}(\tilde{\alpha}_t | \mathcal{Y}_{\text{tar}})^{1-\lambda}}{\mathbf{Z}(\mathbf{X}_t^\theta) h(\tilde{\alpha}_t)} \right] h(\tilde{\alpha}_t) d\tilde{\alpha}_t - \log \mathbf{Z}(\mathbf{X}_t^\theta) \quad (50)$$

$$\stackrel{(i)}{\geq} \int [\log \zeta_\psi(\mathbf{y}_t | \tilde{\alpha}_t) + \lambda \log p_\theta(\tilde{\alpha}_t | \mathbf{X}_t^\theta) + (1 - \lambda) \log q_{\tilde{\theta}}(\tilde{\alpha}_t | \mathcal{Y}_{\text{tar}}) - \log h(\tilde{\alpha}_t)] h(\tilde{\alpha}_t) d\tilde{\alpha}_t - \log \mathbf{Z}(\mathbf{X}_t^\theta) \quad (51)$$

$$\stackrel{(ii)}{=} \int [\log \zeta_\psi(\mathbf{y}_t | \tilde{\alpha}_t) + (\lambda - 1) \log p_\theta(\tilde{\alpha}_t | \mathbf{X}_t^\theta) + (1 - \lambda) \log q_{\tilde{\theta}}(\tilde{\alpha}_t | \mathcal{Y}_{\text{tar}})] p(\tilde{\alpha}_t | \mathbf{X}_t^\theta) d\tilde{\alpha}_t - \log \mathbf{Z}(\mathbf{X}_t^\theta) \quad (52)$$

$$\stackrel{(iii)}{=} \int [\log \zeta_\psi(\mathbf{y}_t | \tilde{\alpha}_t) + (1 - \lambda) \log q_{\tilde{\theta}}(\tilde{\alpha}_t | \mathcal{Y}_{\text{tar}})] p_\theta(\tilde{\alpha}_t | \mathbf{X}_t^\theta) d\tilde{\alpha}_t + (1 - \lambda) C - \log \mathbf{Z}(\mathbf{X}_t^\theta) \quad (53)$$

$$\stackrel{(iv)}{\geq} \mathbb{E}_{\tilde{\alpha}_t \sim p(\tilde{\alpha}_t | \mathbf{X}_t^\theta)} \left[\underbrace{\log \zeta_\psi(\mathbf{y}_t | \tilde{\alpha}_t)}_{\text{MAE}} + (1 - \lambda) \underbrace{\log q_{\tilde{\theta}}(\tilde{\alpha}_t | \mathcal{Y}_{\text{tar}})}_{\text{JEPA}} \right] \quad (54)$$

$$= \mathbb{E}_{\tilde{\alpha}_t \sim p(\tilde{\alpha}_t | \mathbf{X}_t^\theta)} \left[\frac{1}{2\sigma_\zeta^2} \|\mathbf{y}_t - \mathbf{D}_\psi(\tilde{\alpha}_t)\|^2 + \frac{(1 - \lambda)}{2\sigma_q^2} \|\tilde{\alpha}_t - \alpha^{\tilde{\theta}}\|^2 \right], \quad (55)$$

where (i) follows from Jensen's inequality, and (ii) follows by setting proposal distribution $h = p_\theta$, (iii) follows from the definition of p_θ , since the entropy of Gaussian with constant covariance:

$$\int (\lambda - 1) \log p_\theta(\tilde{\alpha}_t | \mathbf{X}_t^\theta) p_\theta(\tilde{\alpha}_t | \mathbf{X}_t^\theta) d\tilde{\alpha}_t = (1 - \lambda) \int -\log p_\theta(\tilde{\alpha}_t | \mathbf{X}_t^\theta) p_\theta(\tilde{\alpha}_t | \mathbf{X}_t^\theta) d\tilde{\alpha}_t = (1 - \lambda) C \geq 0. \quad (56)$$

Finally, (iv) follows from $(1 - \lambda)C \geq 0$ and since the normalization constant $\mathbf{Z}(\mathbf{X}_t^\theta)$ is calculated as:

$$\mathbf{Z}(\mathbf{X}_t^\theta) = \int \zeta_\psi(\tilde{\alpha}_t | \mathbf{X}_t^\theta)^\lambda q_{\tilde{\theta}}(\tilde{\alpha}_t | \mathcal{Y}_{\text{tar}})^{1-\lambda} d\tilde{\alpha}_t \quad (57)$$

$$= \int \mathbf{C}_1 \exp \left[-\frac{\lambda}{2\sigma_p^2} \|\tilde{\alpha}_t - \alpha_t^\theta\|^2 - \frac{(1-\lambda)}{2\sigma_q^2} \|\tilde{\alpha}_t - \alpha_{\tilde{\theta}}\|^2 \right] \quad (58)$$

$$= \int \mathbf{C}_1 \exp \left[-\frac{1}{2}(\tilde{\alpha}_t -)^\top \mathbf{S}^{-1}(\tilde{\alpha}_t -) + \frac{1}{2} \left({}^\top \mathbf{S}^{-1} - \frac{\lambda}{\sigma_p^2} \|\alpha_t^\theta\|^2 - \frac{1-\lambda}{\sigma_q^2} \|\alpha_{\tilde{\theta}}\|^2 \right) \right] \quad (59)$$

$$= \mathbf{C}_3 \exp \left[\frac{1}{2} \left({}^\top \mathbf{S}^{-1} - \frac{\lambda}{\sigma_p^2} \|\alpha_t^\theta\|^2 - \frac{1-\lambda}{\sigma_q^2} \|\alpha_{\tilde{\theta}}\|^2 \right) \right], \quad (60)$$

$$\text{where } \mathbf{C}_1 = \frac{1}{(2\pi)^{d/2} (\sigma_1^2)^{\frac{\lambda d}{2}} (\sigma_3^2)^{\frac{(1-\lambda)d}{2}}}, \mathbf{C}_3 = \frac{1}{\left(\frac{\lambda}{\sigma_1^2} + \frac{1-\lambda}{\sigma_3^2} \right)^{d/2} (\sigma_1^2)^{\frac{\lambda d}{2}} (\sigma_3^2)^{\frac{(1-\lambda)d}{2}}},$$

$$= \mathbf{S} \left(\frac{\lambda}{\sigma_p^2} \mathbf{X}_t^\theta + \frac{1-\lambda}{\sigma_q^2} \mathbf{T}_{\tilde{\theta}}(t, \mathcal{Y}_{\text{tar}}) \right), \text{ and } \mathbf{S} = \left(\frac{\lambda}{\sigma_p^2} + \frac{1-\lambda}{\sigma_q^2} \right)^{-1} \mathbf{I}. \quad (61)$$

Consequently, we get

$$\begin{aligned} \mathbf{Z}(\mathbf{X}_t^\theta) &= \mathbf{C}_3 \exp \left[\frac{1}{2} \left(\frac{\left(\frac{\lambda}{\sigma_p^2} \alpha_t^\theta + \frac{1-\lambda}{\sigma_q^2} \alpha_{\tilde{\theta}} \right)^\top \left(\frac{\lambda}{\sigma_p^2} \alpha_t^\theta + \frac{1-\lambda}{\sigma_q^2} \alpha_{\tilde{\theta}} \right)}{\left(\frac{\lambda}{\sigma_p^2} + \frac{1-\lambda}{\sigma_q^2} \right)} - \frac{\lambda}{\sigma_p^2} \|\alpha_t^\theta\|^2 - \frac{1-\lambda}{\sigma_q^2} \|\alpha_{\tilde{\theta}}\|^2 \right) \right] \\ &= \mathbf{C}_3 \exp \left[-\frac{\frac{\lambda(1-\lambda)}{\sigma_1^2 \sigma_3^2}}{2 \left(\frac{\lambda}{\sigma_1^2} + \frac{1-\lambda}{\sigma_3^2} \right)} \left\| \alpha_t^\theta - \alpha_{\tilde{\theta}} \right\|^2 \right]. \end{aligned} \quad (62)$$

It implies that $-\log \tilde{\alpha}(\mathbf{X}_t^\theta) \geq 0$. Hence we can derive the desired inequality in (9):

$$-\log p(\mathcal{Y}_{\text{tar}} | \mathcal{Y}_{\text{ctx}}) \leq \mathbb{E}_{\mathbf{X}^\theta \sim (10)} \left[\int_0^T \frac{1}{2} \|\alpha_t^\theta\|^2 dt - \sum_{t \in \mathcal{T}} \mathbb{E}_{p(\tilde{\alpha}_t | \mathbf{X}_t^\theta)} [\log \zeta_\psi(\mathbf{y}_t | \tilde{\alpha}_t) + (1-\lambda) \log q_{\tilde{\theta}}(\tilde{\alpha}_t | \mathcal{Y}_{\text{tar}})] \right] \quad (63)$$

$$= \mathbb{E}_{\mathbf{X}^\theta \sim (10)} \left[\int_0^T \frac{1}{2} \|\alpha_t^\theta\|^2 dt - \sum_{t \in \mathcal{T}} \mathbb{E}_{\tilde{\alpha}_t \sim p(\tilde{\alpha}_t | \mathbf{X}_t^\theta)} \left[\frac{1}{2\sigma_\zeta^2} \|\mathbf{y}_t - \mathbf{D}_\psi(\tilde{\alpha}_t)\|^2 + \frac{(1-\lambda)}{2\sigma_q^2} \|\tilde{\alpha}_t - \alpha_{\tilde{\theta}}\|^2 \right] \right] \quad (64)$$

$$= \mathcal{L}(\theta, \psi). \quad (65)$$

For stable learning, we train our model with rescaled training objective:

$$\hat{\mathcal{L}}(\theta, \psi) = \mathbb{E}_{\mathbf{X}^\theta \sim (10)} \left[\int_0^T \sigma_q^2 \|\alpha_t^\theta\|^2 dt - \sum_{t \in \mathcal{T}_{\text{obs}}} \mathbb{E}_{\tilde{\alpha}_t \sim p(\tilde{\alpha}_t | \mathbf{X}_t^\theta)} \left[\underbrace{\|\mathbf{y}_t - \mathbf{D}_\psi(\tilde{\alpha}_t)\|^2}_{\text{reconstruction}} + \tau \underbrace{\|\tilde{\alpha}_t - \alpha_{\tilde{\theta}}\|^2}_{\text{regularization}} \right] \right], \quad (66)$$

Here, $\tau = \frac{(1-\lambda)\sigma_\zeta^2}{\sigma_q^2}$ determines the balance between reconstruction and regularization. See Section 3.3 for details on how controlling the regularization influences the performance of BDO.

C PARALLEL SCAN ALGORITHM

The computation of the first two moments—the mean $\mu_{t \in \mathcal{T}}$ and covariance $\Sigma_{t \in \mathcal{T}}$ —of the controlled distributions can be efficiently parallelized using the scan (all-prefix-sums) algorithm (Blelloch, 1990). Leveraging the associativity of the underlying operations, we reduce the computational complexity

from $\mathcal{O}(k)$ to $\mathcal{O}(\log k)$ time with respect to the number of time steps k . We have established the linear recurrence in Theorem 2.2 for the mean and covariance at each time step t_i :

$$\mathbf{m}_{t_i} = \hat{\mathbf{A}}_i \mathbf{m}_{t_{i-1}} + \hat{\mathbf{B}}_i \alpha_{t_i}, \quad (67)$$

$$\Sigma_{t_i} = \bar{\mathbf{A}}_i \Sigma_{t_{i-1}} + \bar{\mathbf{B}}_i \mathbf{I}, \quad (68)$$

where we define $\Delta_i(t) = t - t_i$, $\hat{\mathbf{A}}_i = e^{-\Delta_{i-1}(t_i)\Lambda_{t_i}}$, $\hat{\mathbf{B}}_i = -e^{-\Delta_{i-1}(t_i)\Lambda_{t_i}} \Lambda_{t_i}^{-1} (\mathbf{I} - e^{\Delta_{i-1}(t_i)\Lambda_{t_i}})$, $\bar{\mathbf{A}}_i = e^{-2\Delta_{i-1}(t_i)\Lambda_{t_i}}$ and $\bar{\mathbf{B}}_i = -\frac{1}{2}e^{-2\Delta_{i-1}(t_i)\Lambda_{t_i}} \Lambda_{t_i}^{-1} (\mathbf{I} - e^{2\Delta_{i-1}(t_i)\Lambda_{t_i}})$. To apply the parallel scan algorithm to our recurrence, we define two separate sequences of tuples for the mean and covariance computations for all $i \in \{1, \dots, k\}$:

$$\mathbf{M}_i = (\hat{\mathbf{A}}_i, \hat{\mathbf{B}}_i \alpha_{t_i}), \quad \mathbf{S}_i = (\bar{\mathbf{A}}_i, \bar{\mathbf{B}}_i) \quad (69)$$

Now, we define binary associative operators \otimes for the sequences $\{\mathbf{M}_i\}$ and $\{\mathbf{S}_i\}$:

$$\mathbf{M}_i \otimes \mathbf{M}_j = (\hat{\mathbf{A}}_i \circ \hat{\mathbf{A}}_j, \hat{\mathbf{A}}_i \circ \hat{\mathbf{B}}_j \alpha_{t_j} + \hat{\mathbf{B}}_i \alpha_{t_i}), \quad (70)$$

$$\mathbf{S}_i \otimes \mathbf{S}_j = (\bar{\mathbf{A}}_i \circ \bar{\mathbf{A}}_j, \bar{\mathbf{A}}_i \circ \bar{\mathbf{B}}_j + \bar{\mathbf{B}}_i), \quad (71)$$

where \circ denotes element-wise multiplication. We can verify that \otimes is an associative operator since it satisfies:

$$(\mathbf{M}_s \otimes \mathbf{M}_t) \otimes \mathbf{M}_u = (\hat{\mathbf{A}}_t \circ \hat{\mathbf{A}}_s, \hat{\mathbf{A}}_t \circ \hat{\mathbf{B}}_s \alpha_{t_s} + \hat{\mathbf{B}}_t \alpha_{t_t}) \otimes \mathbf{M}_u \quad (72)$$

$$= (\hat{\mathbf{A}}_u \circ (\hat{\mathbf{A}}_t \circ \hat{\mathbf{A}}_s), \hat{\mathbf{A}}_u \circ (\hat{\mathbf{A}}_t \circ \hat{\mathbf{B}}_s \alpha_{t_s} + \hat{\mathbf{B}}_t \alpha_{t_t}) + \hat{\mathbf{B}}_u \alpha_{t_u}) \quad (73)$$

$$= ((\hat{\mathbf{A}}_u \circ \hat{\mathbf{A}}_t) \circ \hat{\mathbf{A}}_s, (\hat{\mathbf{A}}_u \circ \hat{\mathbf{A}}_t) \circ \hat{\mathbf{B}}_s \alpha_{t_s} + \hat{\mathbf{A}}_u \circ \hat{\mathbf{B}}_t \alpha_{t_t} + \hat{\mathbf{B}}_u \alpha_{t_u}) \quad (74)$$

$$= \mathbf{M}_s \otimes (\mathbf{M}_t \otimes \mathbf{M}_u). \quad (75)$$

Thus, we get $(\mathbf{M}_s \otimes \mathbf{M}_t) \otimes \mathbf{M}_u = \mathbf{M}_s \otimes (\mathbf{M}_t \otimes \mathbf{M}_u)$, confirming associativity for \mathbf{M}_i . Similarly,

$$(\mathbf{S}_s \otimes \mathbf{S}_t) \otimes \mathbf{S}_u = (\bar{\mathbf{A}}_t \circ \bar{\mathbf{A}}_s, \bar{\mathbf{A}}_t \circ \bar{\mathbf{B}}_s \mathbf{I} + \bar{\mathbf{B}}_t \mathbf{I}) \otimes \mathbf{S}_u \quad (76)$$

$$= (\bar{\mathbf{A}}_u \circ (\bar{\mathbf{A}}_t \circ \bar{\mathbf{A}}_s), \bar{\mathbf{A}}_u \circ (\bar{\mathbf{A}}_t \circ \bar{\mathbf{B}}_s \mathbf{I} + \bar{\mathbf{B}}_t \mathbf{I}) + \bar{\mathbf{B}}_u \mathbf{I}) \quad (77)$$

$$= ((\bar{\mathbf{A}}_u \circ \bar{\mathbf{A}}_t) \circ \bar{\mathbf{A}}_s, (\bar{\mathbf{A}}_u \circ \bar{\mathbf{A}}_t) \circ \bar{\mathbf{B}}_s \mathbf{I} + \bar{\mathbf{A}}_u \circ \bar{\mathbf{B}}_t \mathbf{I} + \bar{\mathbf{B}}_u \mathbf{I}) \quad (78)$$

$$= \mathbf{S}_s \otimes (\mathbf{S}_t \otimes \mathbf{S}_u). \quad (79)$$

Hence, $(\mathbf{S}_s \otimes \mathbf{S}_t) \otimes \mathbf{S}_u = \mathbf{S}_s \otimes (\mathbf{S}_t \otimes \mathbf{S}_u)$, confirming associativity for \mathbf{S}_i . Now, we can apply the parallel scan described in Algorithm 1 for both $\mu_{t \in \mathcal{T}}$ and covariance $\Sigma_{t \in \mathcal{T}}$ based on the recurrence in (24, 28) and the defined associative operators \otimes . Employing the parallel scan algorithm offers significant computational benefits, especially for large-scale problems with numerous time steps k . The logarithmic time complexity ensures scalability, making it feasible to perform real-time computations or handle high-dimensional data efficiently.

D EXPERIMENTAL DETAILS

D.1 DATA PREPROCESSING

Preprocessing Pipeline. The preprocessing pipeline involved several standard steps, including skull-stripping, slice-timing correction, motion correction, non-linear registration, and intensity normalization. All data were aligned to the Montreal Neurological Institute (MNI) standard space for consistency. A whole-brain mask was applied to exclude non-brain tissues, such as the skull, from further analysis. The fMRI data were parcellated into 450 regions of interest (ROIs), comprising 400 cortical parcels based on the Schaefer-400 atlas [Schaefer et al. \(2017\)](#) and 50 subcortical parcels defined by Tian’s Scale III atlas [Tian et al. \(2020\)](#). The mean fMRI time-series for each ROI was extracted across all time points.

Data Normalization. To ensure comparability across participants and reduce inter-subject variability, we applied a two-step normalization process to the fMRI data. First, participant-wise zero-mean centering was performed by subtracting the mean signal from each ROI within each subject. Second, a robust scaling procedure was applied, where the median signal was subtracted, and the resulting values were divided by the interquartile range (IQR), computed across all participants for each ROI. After normalization, each fMRI sample was represented as a matrix of size $T \times N$, where T corresponds to the number of timesteps and N corresponds to the number of ROIs ($N = 450$).

Algorithm 1 Parallel Scan for Mean and Covariance

```

1: Input. Given time stamps  $\mathcal{T} = \{t_1, t_2, \dots, t_K\}$ ,
   initial mean  $\mu_{t_0}$  and covariance  $\Sigma_{t_0}$ , control policies
    $\{\alpha_{t_1}, \alpha_{t_2}, \dots, \alpha_{t_K}\}$ , matrices  $\{\Lambda_{t_1}, \Lambda_{t_2}, \dots, \Lambda_{t_K}\}$ .
2: Initialize sequences  $\{\mathbf{M}_i\}_{i=1}^K$  and  $\{\mathbf{S}_i\}_{i=1}^K$ :
3: for  $i = 1$  to  $K$  do in parallel
4:   Compute  $\Delta_i(t_i) = t_i - t_{i-1}$ .
5:   Compute  $\hat{\mathbf{A}}_i = e^{-\Delta_i(t_i)\Lambda_{t_i}}$ .
6:   Compute  $\hat{\mathbf{B}}_i = -e^{-\Delta_i(t_i)\Lambda_{t_i}}\Lambda_{t_i}^{-1}(\mathbf{I} - e^{\Delta_i(t_i)\Lambda_{t_i}})$ .
7:   Compute  $\bar{\mathbf{A}}_i = e^{-2\Delta_i(t_i)\Lambda_{t_i}}$ .
8:   Compute  $\bar{\mathbf{B}}_i = -\frac{1}{2}e^{-2\Delta_i(t_i)\Lambda_{t_i}}\Lambda_{t_i}^{-1}(\mathbf{I} - e^{2\Delta_i(t_i)\Lambda_{t_i}})$ .
9:   Set  $\mathbf{M}_i = (\hat{\mathbf{A}}_i, \hat{\mathbf{B}}_i\alpha_{t_i})$ .
10:  Set  $\mathbf{S}_i = (\bar{\mathbf{A}}_i, \bar{\mathbf{B}}_i)$ .
11: end for
12: Parallel Scan  $\{\mathbf{M}'_i\}_{i=1}^K$ 
   ParallelScan( $\{\mathbf{M}_i\}_{i=1}^K, \otimes$ )
13: Parallel Scan  $\{\mathbf{S}'_i\}_{i=1}^K$ 
   ParallelScan( $\{\mathbf{S}_i\}_{i=1}^K, \otimes$ )
14: for  $i = 1$  to  $K$  do in parallel
15:    $\mu_{t_i} = \mathbf{M}'^{(1)}_i \mu_{t_0} + \mathbf{M}'^{(2)}_i$ 
16:    $\Sigma_{t_i} = \mathbf{S}'^{(1)}_i \Sigma_{t_0} + \mathbf{S}'^{(2)}_i$ 
17: end for
18: Return  $\mu_{t \in \mathcal{T}}, \Sigma_{t \in \mathcal{T}}$ 

```

Algorithm 2 ParallelScan

```

1: Input. Sequence of tuples
    $\{\mathbf{T}_1, \mathbf{T}_2, \dots, \mathbf{T}_K\}$ , associative opera-
   tor  $\otimes$ .
2: Stage 1: Up-Sweep (Reduce).
3: for  $d = 0$  to  $\lceil \log_2 K \rceil - 1$  do
4:   for each subtree of height  $d$  in parallel do
5:     Let  $i = 2^{d+1}k + 2^{d+1} - 1$  for  $k =$ 
        $0, 1, \dots$ 
6:     if  $i < K$  then
7:        $\mathbf{T}_i = \mathbf{T}_{i-2^d} \otimes \mathbf{T}_i$ 
8:     end if
9:   end for
10: end for
11: Stage 2: Down-Sweep.
12:  $\mathbf{T}_K = \mathbf{I}$ , where  $\mathbf{I}$  is the identity element for
    $\otimes$ .
13: for  $d = \lceil \log_2 K \rceil - 1$  downto  $0$  do
14:   for each subtree of height  $d$  in parallel do
15:     Let  $i = 2^{d+1}k + 2^{d+1} - 1$  for  $k =$ 
        $0, 1, \dots$ 
16:     if  $i < K$  then
17:        $\mathbf{T}_{i-2^d} = \mathbf{T}_{i-2^d} \otimes \mathbf{T}_i$ 
18:     end if
19:   end for
20: end for
21: Return Scanned sequence
    $\{\mathbf{T}'_1, \mathbf{T}'_2, \dots, \mathbf{T}'_K\}$  where  $\mathbf{T}'_i =$ 
    $\mathbf{T}_1 \otimes \mathbf{T}_2 \otimes \dots \otimes \mathbf{T}_i$ .

```

UK Biobank (UKB) The UKB is a population-based prospective study comprising 500,000 participants in the United Kingdom, designed to investigate the genetic and environmental determinants of disease [Sudlow et al. \(2015\)](#). This study utilized 41,072 rs-fMRI scans from the publicly available, preprocessed UKB dataset [Alfaro-Almagro et al. \(2018\)](#). The preprocessing pipeline included non-linear registration to MNI space using FSL’s `applywarp` function, thereby ensuring standardized spatial alignment across participants [Jenkinson et al. \(2012\)](#).

Human Connectome Project in Aging (HCP-A) The HCP-A is a large-scale neuroimaging initiative focused on characterizing structural and functional connectivity changes associated with aging across a wide age range [Bookheimer et al. \(2019\)](#). This study accessed 724 rs-fMRI samples from healthy individuals between 36 and 89 years of age. Preprocessed rs-fMRI volumes provided by the HCP-A dataset were utilized for subsequent analyses.

Autism Brain Imaging Data Exchange (ABIDE) The ABIDE consortium aims to elucidate the neural mechanisms underlying autism spectrum disorder [Di Martino et al. \(2014\)](#). In the present work, 1,102 rs-fMRI samples were obtained from the Neuro Bureau Preprocessing Initiative [Craddock et al. \(2013a\)](#), which employs the Configurable Pipeline for the Analysis of Connectomes (C-PAC) [Craddock et al. \(2013b\)](#). The preprocessing steps included slice-timing correction, motion realignment, intensity normalization (with a 4D global mean set to 1000), and nuisance signal removal. Nuisance regression involved a 24-parameter motion model, component-based noise correction (CompCor) [Behzadi et al. \(2007\)](#) with five principal components derived from white matter and cerebrospinal fluid signals, and linear/quadratic trend removal. Functional-to-anatomical registration was performed via a boundary-based rigid-body approach, while anatomical-to-standard registration utilized ANTs. Band-pass filtering and global signal regression were not applied.

Attention Deficit Hyperactivity Disorder 200 (ADHD200) The ADHD200 dataset comprises 776 rs-fMRI and anatomical scans collected from individuals aged 7 to 21, including 491 typically developing individuals and 285 participants diagnosed with ADHD [Brown et al. \(2012\)](#). A total of 669 rs-fMRI datasets were selected for this study, specifically the preprocessed versions provided by the Neuro Bureau Preprocessing Initiative (Athena Pipeline) [Bellec et al. \(2017\)](#).

Human Connectome Project for Early Psychosis (HCP-EP) The HCP-EP is a neuroimaging initiative focused on understanding early psychosis, defined as the first five years following symptom onset, in individuals aged 16–35. The cohort includes participants with affective psychosis, non-affective psychosis, and healthy controls [Jacobs et al. \(2024\)](#); [Prunier & Shenton Martha; Breier \(2021\)](#). For this study, 176 rs-fMRI scans were analyzed. Preprocessing was conducted using fMRIPrep [Esteban et al. \(2019\)](#), followed by denoising with Nilearn [Nilearn contributors \(2025\)](#). The denoising process employed a 24-parameter motion model (including translations, rotations, their derivatives, and quadratic terms) and CompCor-derived components extracted from white matter and cerebrospinal fluid masks. Additionally, all confound variables were demeaned to ensure consistency.

D.2 PRE-TRAINING STAGE

Pre-training Data. For self-supervised pre-training, we utilized the large-scale UKB dataset, which comprises resting-state fMRI recordings and medical records from 41,072 participants ([Alfaro-Almagro et al., 2018](#)). We utilized 80% of the dataset for pre-training, while the remaining 20% held-out data was reserved for downstream evaluation. We used a fixed random seed (42) to ensure reproducibility when partitioning the UKB dataset into pre-training and held-out subsets. All experiments, including the reproduction of foundation model baselines, were conducted using the same dataset split to maintain consistency.

Irregular Multivariate Time-Series Sampling. We introduce irregularity in the time-series data by subsampling both the observation timestamps \mathcal{T}_{obs} and the corresponding fMRI signals \mathcal{Y}_{obs} . Unlike conventional approaches that assume uniformly spaced time points ([Caro et al., 2024](#); [Dong et al., 2024](#)), we select a uniformly sampled subset of timestamps from the full sequence, ensuring that only a fraction of the fMRI signal is observed. Specifically, from each full-length fMRI recording, we randomly sample 160 timesteps ($T = 160$), introducing variability in temporal resolution across different samples. This choice reflects the fundamental nature of brain dynamics, which evolve continuously rather than discretely, and encourages the model to infer missing states from incomplete sequences.

Temporal Masking. To encourage robust representation learning and improve generalization, we employ *temporal masking*, where a subset of the 160 sampled time points is randomly masked during training. We apply a masking ratio of $\gamma = 0.75$, meaning that 75% of the sampled timesteps are hidden while the model is trained to reconstruct them. In Figure 8, we vary γ across $[0.4, 0.5, 0.6, 0.7, 0.75, 0.8, 0.9]$ to examine the effect of masking ratio in learning robust representations. Actual reconstruction results are provided in the internal and external datasets as visualized in Figures 14 and 15.

Pre-training Algorithm. The pre-training of BDO follows the procedure outlined in Algorithm 3. Given an observed fMRI time-series \mathcal{Y}_{obs} , we employ a masked reconstruction strategy, where a random proportion γ of the temporal signals is masked to encourage the model to learn meaningful representations. The pre-training objective leverages amortized inference to approximate latent dynamics. At each iteration, a subset of observed time-series \mathcal{Y}_{ctx} is used as context, while the masked portion \mathcal{Y}_{tar} serves as the target for reconstruction. The encoder network \mathbf{T}_{θ} maps the context data to a sequence of latent states $\mathbf{z}_{t \in \mathcal{T}_{\text{ctx}}}$, which are then used to estimate drift terms and control policies, forming the basis for latent trajectory prediction. The decoder network \mathbf{D}_{ψ} reconstructs the missing target states, optimizing a training objective $\mathcal{L}(\theta, \psi)$ that aligns the predicted and true trajectories.

Pre-training Details. We trained BDO using a batch size of 128 and a total of 200 pre-training epochs. The learning rate was scheduled using a cosine decay scheduler ([Loshchilov & Hutter, 2016](#)) with a 10-epoch warm-up phase. During warm-up, the initial learning rate was set to 0.0001, which increased to a peak learning rate of 0.001 before gradually decaying to a minimum learning rate of 0.0001. For optimization, we employed the Adam optimizer ([Kingma & Ba, 2015](#)). Across all BDO configurations, we used a fixed number of basis $l = 100$ and consistently multiplied the observation times by a time-scale of 0.1 for all datasets. To update θ , Exponential Moving Average (EMA) momentum is used and linearly increased from 0.996 to 1.0. It is worth noting that our models required minimal hyperparameter tuning, which demonstrates that the proposed approximation scheme operates stably and robustly.

Algorithm 3 Pre-training BDO

```

1: Input. Time-series  $\mathcal{Y}_{\text{obs}} = \mathbf{y}_{t \in \mathcal{T}_{\text{obs}}}$ , masking ratio
    $\gamma$ , encoder network  $\mathbf{T}_{\theta}$ , decoder network  $\mathbf{D}_{\psi}$ 
2: for  $m = 1, \dots, M$  do
3:   Get  $\mathcal{Y}_{\text{ctx}}, \mathcal{Y}_{\text{tar}}$  by masking  $\gamma\%$  of temporal sig-
   nals.
4:   Sample  $\mathbf{z}_{t \in \mathcal{T}_{\text{ctx}}} \sim \prod_{t \in \mathcal{T}_{\text{ctx}}} q_{\theta}(\mathbf{z}_t | \mathcal{Y}_{\text{ctx}})$ .
5:   Compute  $\{\mathbf{D}_t, u_t, \alpha_t^{\theta}\}_{t \in \mathcal{T}_{\text{ctx}}}$ .
6:   Estimate  $\{\mu_t, \Sigma_t\}_{t \in \mathcal{T}_{\text{tar}}}$  with parallel scan algo-
   rithm.
7:   Sample  $\mathbf{X}_{t \in \mathcal{T}_{\text{tar}}}^{\theta} \sim \prod_{t \in \mathcal{T}_{\text{tar}}} \mathcal{N}(\mu_t, \Sigma_t)$ .
8:   Sample  $\hat{\mathbf{z}}_{t \in \mathcal{T}_{\text{tar}}} \sim \prod_{t \in \mathcal{T}_{\text{tar}}} p(\hat{\mathbf{z}}_t | \mathbf{X}_t^{\theta})$ .
9:   Compute  $\hat{\mathcal{L}}(\theta, \psi)$  using (66).
10:  Update  $(\theta, \psi)$  with  $\nabla_{\theta, \psi} \hat{\mathcal{L}}(\theta, \psi)$ .
11:  Apply  $\theta \leftarrow \text{EMA}(\theta)$ .
12: end for

```

Algorithm 4 Fine tuning BDO for downstream tasks

```

1: Input. Time-series and label  $(\mathcal{Y}_{\text{obs}}, \mathcal{O}_{\text{obs}})$ , pre-trained
   encoder network  $\mathbf{T}_{\theta^*}$ .
2: Sample  $\mathbf{z}_{t \in \mathcal{T}_{\text{obs}}} \sim \prod_{t \in \mathcal{T}_{\text{obs}}} q_{\theta^*}(\mathbf{z}_t | \mathcal{Y}_{\text{obs}})$ .
3: Compute optimal control policy  $\alpha_{t \in \mathcal{T}_{\text{obs}}} = \mathbf{B}_{\theta^*} \mathbf{z}_{t \in \mathcal{T}_{\text{obs}}}$ .
4: Compute the universal feature  $\mathbb{A} = \frac{1}{|\mathcal{T}_{\text{obs}}|} \sum_{t \in \mathcal{T}_{\text{obs}}} \alpha_t$ .
5: Predict  $\hat{\mathcal{O}}_{\text{obs}} = h_{\omega}(\mathbb{A})$ .
6: if Linear probing then
7:   Freeze the pre-trained encoder network  $\mathbf{T}_{\theta^*}$ .
8:   Compute  $\mathcal{L}(\theta^*, \omega) = \mathcal{L}_{\text{task}}(\mathcal{O}_{\text{obs}}, \hat{\mathcal{O}}_{\text{obs}})$  using (80).
9:   Update  $\omega$  with  $\nabla_{\omega} \mathcal{L}(\theta^*, \omega)$ .
10: else if Fine tuning then
11:   Unfreeze the pre-trained encoder network  $\mathbf{T}_{\theta^*}$ .
12:   Compute  $\mathcal{L}(\theta^*, \omega) = \mathcal{L}_{\text{task}}(\mathcal{O}_{\text{obs}}, \hat{\mathcal{O}}_{\text{obs}})$  using (80).
13:   Update  $(\theta^*, \omega)$  with  $\nabla_{\theta^*, \omega} \mathcal{L}(\theta^*, \omega)$ .
14: end if

```

D.3 MODEL ARCHITECTURE

To maintain the structural advantages of our formulation, we designed our encoder network architecture in a straightforward manner. In this regard, the networks used for pre-training BDO are listed below, where $N=450$ is the number of ROIs and d is the dimension of latent space \mathbb{R}^d as described in Table 5 for each model.

- **Encoder network q_{θ} :**
 Input (N) \rightarrow Linear(d) \rightarrow ReLU() \rightarrow LayerNorm(d) \rightarrow Linear(d)
 \rightarrow ReLU() \rightarrow LayerNorm(d) $\rightarrow 12 \times [\text{LayerNorm}(d) \rightarrow \text{Attn}(d) \rightarrow \text{FFN}(d)]$
- **FFN:**
 Input (d) \rightarrow LayerNorm(d) \rightarrow Linear($4 \times d$) \rightarrow GeLU() \rightarrow
 Linear(d) \rightarrow Residual(Input(d))
- **Attn:**
 Input (Q, K, V) \rightarrow Normalize(Q) \rightarrow Linear(Q) \rightarrow Linear(K) \rightarrow
 Linear(V) \rightarrow Attention(Q, K) \rightarrow Softmax(d) \rightarrow Dropout() \rightarrow
 Matmul(V) \rightarrow LayerNorm(d) \rightarrow Linear(d) \rightarrow Residual(Q)
- **Decoder network \mathbf{D}_{ψ} :**
 Input (d) \rightarrow Linear(N) \rightarrow ReLU() \rightarrow Dropout() \rightarrow Linear(d)

Table 5: Pre-training hyper-parameters

BDO Variants	Train EP	Warm-up EP	LR	Initial LR	Minimum LR	Batch Size	\mathbb{R}^d	# of base matrices (L)	EMA Momentum
BDO (5M)	200	10	0.001	0.0001	0.0001	128	192	100	[0.996, 1]
BDO (21M)	200	10	0.001	0.0001	0.0001	128	384	100	[0.996, 1]
BDO (86M)	200	10	0.001	0.0001	0.0001	128	768	100	[0.996, 1]

D.4 DOWNSTREAM EVALUATION STAGE

To assess the generalization and transferability of BDO, we conducted experiments across multiple datasets and tasks, encompassing both demographic and psychiatric prediction. Datasets used in this evaluation have distinct temporal resolutions and varying numbers of timesteps, reflecting the irregularity of real-world fMRI data acquisition. Additional details are described in Table 6. Note that in the downstream evaluation, irregular sampling and temporal masking were disabled. The full sequence of fMRI signals, timestamps, and corresponding labels was used, denoted as $(\mathcal{Y}_{\text{obs}}, \mathcal{T}_{\text{obs}}, \mathcal{O}_{\text{obs}})$.

Table 6: Dataset Subject Demographics

Category	UKB	HCP-A	ABIDE	ADHD200	HCP-EP
# of subjects	41,072	724	1,102	669	176
Age, mean (SD)	54.98 (7.53)	60.35 (15.74)	17.05 (8.04)	11.61 (2.97)	23.39 (3.95)
Female, % (n)	52.30 (21,480)	56.08 (406)	14.79 (163)	36.17 (242)	38.07 (67)
Patient, % (n)	-	-	48.19 (531)	58.15 (389)	68.18 (120)
Target Population	Healthy Population	Healthy Population	ASD Healthy Population	ADHD Healthy Population	Psychotic Disorder Healthy Population

Internal Evaluation. For *internal evaluation*, we utilized a 20% held-out subset of the UKB dataset, which was excluded from pre-training. This evaluation focused on age regression and gender classification, leveraging both LP and FT to analyze how well the model retains and transfers knowledge acquired during pre-training.

External Evaluation. For *external evaluation*, we examined the ability of BDO to generalize to unseen datasets. Demographic and trait prediction was performed on the HCP-A dataset, where LP and FT were employed to assess model performance on age, gender, neuroticism, and flanker scores. Beyond demographic characteristics, we evaluated psychiatric diagnosis classification using 3 clinical fMRI datasets, including ABIDE, ADHD200, and HCP-EP. These evaluations relied on LP, as it provides a controlled assessment of the learned representations and their applicability to clinical classification tasks.

Random Splits. All the datasets are partitioned into training, validation, and test sets using a 6:2:2 ratio to ensure fair and reproducible evaluation. To maintain consistency, we perform partitioning with 3 consecutive random seeds, 0, 1, and 2.

- For classification tasks, such as gender classification, stratified sampling is applied to preserve class distributions across the training, validation, and test sets.
- For regression tasks, such as age regression, binning-based stratified sampling is employed. In this approach, the continuous target variable is first discretized into bins before applying stratified sampling, ensuring a balanced distribution of the target variable and mitigating potential biases from uneven data partitioning. Additionally, to improve numerical stability and facilitate optimization, the target variable is normalized using Z-score normalization, where the mean is subtracted, and the result is divided by the standard deviation.
- The distributions of the three random splits for age regression tasks with the UKB and HCP-A datasets, and six classification tasks with UKB gender, HCP-A gender, ABIDE diagnosis, ADHD200 diagnosis, and HCP-EP diagnosis are described in Figures 11–13.

Extracting the Universal Feature \mathbb{A} . To extract the *universal feature* \mathbb{A} , we define f as *mean-pooling* over the sequence of control signals $\alpha_{t \in \mathcal{T}}$, given by $\mathbb{A} := f(\alpha_{t \in \mathcal{T}}) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \alpha_t$. This formulation ensures that \mathbb{A} serves as a compact and transferable representation of the underlying spatio-temporal dynamics captured by the optimal control signals. To enhance biological interpretability, mean-pooling is chosen as it provides a *global summary* of the temporal evolution of the control sequence while suppressing high-frequency fluctuations that may arise due to local variations in α_t . Although we believe that mean-pooling provides a robust and scalable approach for summarizing temporal dynamics, we acknowledge that more sophisticated aggregation methods, such as weighted pooling or recurrent architectures, could further enhance downstream performance. These approaches may offer additional advantages for analyzing temporal dynamics, such as facilitating interpretability through attention weight analysis or capturing long-range dependencies. We leave the exploration of these advanced aggregation strategies for future work.

Downstream Evaluation Algorithm. To evaluate the effectiveness of BDO on downstream tasks, we follow the procedure outlined in Algorithm 4. Given an observed fMRI time-series \mathcal{Y}_{obs} and its corresponding labels \mathcal{O}_{obs} , we extract the universal feature representation \mathbb{A} using the pre-trained encoder \mathbf{T}_{θ^*} . This representation is subsequently used for classification or regression tasks through either LP or FT.

- In the LP setting, we freeze the pre-trained encoder \mathbf{T}_{θ^*} and train only the task-specific head $h_{\omega} : \mathbb{R}^d \rightarrow \mathbb{R}^N$ (single linear layer). The objective function $\mathcal{L}(\theta^*, \omega)$ measures the discrepancy between the predicted $\hat{\mathcal{O}}_{\text{obs}}$ and ground-truth \mathcal{O}_{obs} , and is optimized with respect to ω .

Table 7: Search space of end-to-end fine-tuning (FT) and linear probe (LP).

Configurations	FT	LP
Optimizer	AdamW (Loshchilov, 2017)	Adam (Kingma & Ba, 2015)
Training epochs	50	50
Batch size	[16, 32]	[16, 32, 64]
LR scheduler	cosine decay	cosine decay
LR	[0.001]	[0.01, 0.005]
Minimum LR	[0, 0.0001, 0.001]	[0.001, 0.005]
Weight decay	[0, 0.01]	[0]
Layer-wise LR decay	[0.85, 0.90, 0.95]	N.A.

- In the FT setting, the entire model, including \mathbf{T}_{θ^*} , is optimized. Both the encoder and task-specific head h_w are jointly updated to refine the feature extraction process for the target task.

Training Objective for Downstream tasks. The loss function for downstream tasks is defined based on the nature of the prediction problem: classification tasks use Binary Cross-Entropy (BCE) loss, while regression tasks employ Mean Squared Error (MSE) loss.

Model Selection. To determine the optimal model for each downstream task, we performed a grid search over key hyperparameters such as learning rate and batch size. For each task, we evaluated multiple configurations using the validation set and selected the model that achieved the best performance based on the predefined metric. The set of hyperparameters is provided in Table 7.

$$\mathcal{L}_{\text{task}}(\mathcal{O}_{\text{obs}}, \hat{\mathcal{O}}_{\text{obs}}) = \begin{cases} -\frac{1}{N} \sum_{i=1}^N [\mathcal{O}_{\text{obs},i} \log \hat{\mathcal{O}}_{\text{obs},i} + (1 - \mathcal{O}_{\text{obs},i}) \log(1 - \hat{\mathcal{O}}_{\text{obs},i})], & \text{if classification} \\ \frac{1}{N} \sum_{i=1}^N (\mathcal{O}_{\text{obs},i} - \hat{\mathcal{O}}_{\text{obs},i})^2, & \text{if regression} \end{cases} \quad (80)$$

D.5 GENERALIZATION FROM RESTING-STATE TO TASK-BASED fMRI

To conduct a challenging test of BDO’s generalization capabilities, we evaluated whether its representations, learned exclusively from unconstrained resting-state fMRI, could be effectively transferred to structured task-based fMRI, where brain dynamics are driven by explicit external stimuli.

For this experiment, we used the BDO-86M model, which was pre-trained solely on resting-state UKB data. We then evaluated its performance on three distinct and cognitively demanding task paradigms from the HCP-A dataset. The evaluation was performed under the LP setting.

The results are summarized in Table 8. While there is a moderate performance decrease compared to the in-domain resting-state baseline (HCP-A-Rest), the model still achieves strong predictive performance across all three task paradigms (HCP-A-VisMotor/FaceName/CARIT).

This successful transfer demonstrates that BDO learns fundamental, task-relevant neural dynamics that are not limited to the resting state. This underscores BDO’s broad applicability as a powerful feature extractor for diverse fMRI paradigms, even without any task-specific fine-tuning.

Table 8: Generalization from resting-state to task-based fMRI.

Dataset	Age (MSE) ↓	Age (Pearson) ↑	Gender (Acc.) ↑	Gender (F1) ↑
HCP-A-VisMotor	0.526 \pm 0.018	0.691 \pm 0.015	68.53 \pm 3.57	67.39 \pm 3.36
HCP-A-FaceName	0.459 \pm 0.012	0.732 \pm 0.009	66.20 \pm 3.44	65.29 \pm 3.72
HCP-A-CARIT	0.488 \pm 0.025	0.713 \pm 0.020	67.60 \pm 1.74	66.79 \pm 1.29
HCP-A-Rest	0.404 \pm 0.010	0.768 \pm 0.008	72.00 \pm 2.95	71.30 \pm 2.19

D.6 DETAILED SCALABILITY ANALYSIS

The numerical results used to generate the scalability analysis plot (Figure 7) are presented in Table 9. The table includes detailed linear probing performance for three BDO model variants (5M, 21M, and 86M), evaluated on HCP-A age regression and classification tasks across ABIDE, ADHD200, and HCP-EP datasets. Additionally, results from the data scalability experiment, conducted exclusively with the largest model (86M), are reported at varying proportions (25%, 50%, and 75%) of the total dataset. The entry labeled BDO (86M) corresponds to the model trained with the full dataset (100%), serving as the reference for both model and data scalability experiments.

Table 9: LP performance used for scalability analysis in Figure 7.

Variants	HCP-A	ABIDE	ADHD200	HCP-EP
	Age ($\rho \uparrow$)	ACC (%) \uparrow	ACC (%) \uparrow	ACC (%) \uparrow
BDO (5M)	0.635 \pm .031	62.42 \pm 2.68	59.65 \pm 2.30	73.33 \pm 7.50
BDO (21M)	0.729 \pm .011	63.79 \pm 1.83	61.15 \pm 1.97	71.43 \pm 4.04
BDO (25%)	0.686 \pm .010	61.06 \pm 1.05	57.39 \pm 3.90	72.38 \pm 5.95
BDO (50%)	0.702 \pm .014	63.03 \pm 1.63	56.89 \pm 3.38	74.29 \pm 9.90
BDO (75%)	0.734 \pm .011	65.45 \pm 2.70	58.15 \pm 1.78	74.29 \pm 7.56
BDO (86M)	0.768 \pm .008	66.67 \pm 1.13	61.40 \pm 1.97	76.19 \pm 4.86

D.7 COMPARISON OF SSL MODEL EFFICIENCY

Table 10: Comparison of pre-training efficiency and linear probing performance across SSL models.

Model (Parameters)	Age (Pearson) \uparrow	Gender (Acc.) \uparrow	GPU Hours (x 4 GPUs) \downarrow
MoCo (90M)	0.591	64.12	174 hrs
BYOL (90M)	0.619	64.81	165 hrs
BrainLM (85M)	0.636	65.28	496 hrs
BrainMass (90M)	0.630	66.20	244 hrs
BDO (86M)	0.768	72.00	15 hrs

This subsection presents the detailed experimental settings and the exact numerical results used to construct Figure 1, which shows that BDO surpasses other SSL models in both resource and parameter efficiency.

To evaluate the efficiency of SSL models, we measured the pre-training time using 4 NVIDIA RTX 3090 GPUs, calculated in GPU hours as the total CUDA time recorded with the PyTorch library and multiplied by the number of GPUs. Each model was trained for 200 epochs using the largest batch size that fully utilized available GPU memory.

Table 10 presents the linear probing performance of each pre-trained model on age and gender prediction tasks on HCP-A dataset, alongside their respective pre-training times. Our results demonstrate that BDO achieves superior efficiency in pre-training, requiring significantly fewer GPU hours compared to other SSL methods while maintaining competitive or superior performance. This efficiency highlights the scalability of BDO, making it a practical choice for large-scale applications.

D.8 DETAILED MASK RATIO ANALYSIS

In our framework, the MAE objective plays a critical role by explicitly requiring the reconstruction of masked temporal segments. This encourages the model to capture detailed temporal dependencies and fine-grained dynamics inherent in fMRI signals. To directly validate the importance of the MAE objective, we extended our ablation study to include a zero mask ratio ($\gamma = 0$), which effectively removes the MAE reconstruction task. Starting from the no-masking condition, the HCP-A age regression performance in the LP setting for BDO-86M results are summarized below.

Table 11: Extended ablation study on the mask ratio γ .

Mask Ratio (γ)	Age (MSE) \downarrow	Age (Pearson) \uparrow
0.0 (No Masking)	0.793 ± 0.014	0.445 ± 0.020
0.2	0.487 ± 0.036	0.711 ± 0.027
0.4	0.513 ± 0.016	0.695 ± 0.015
0.6	0.476 ± 0.019	0.727 ± 0.011
0.75 (Optimal)	0.466 ± 0.025	0.738 ± 0.014
0.8	0.526 ± 0.014	0.686 ± 0.006

In the no-masking condition, we observed severely degraded downstream task performance, indicating that the model failed to learn meaningful, transferable representations. This result implies that without the reconstruction challenge introduced by masking, SSL pre-training becomes ineffective in capturing the complex temporal structures necessary for high-quality representation learning.

D.9 DISSECTING THE CONTRIBUTIONS OF MAE AND JEPA OBJECTIVES

To dissect the individual contributions of the MAE and JEPA objectives, we conducted an ablation study by controlling their relative influence with the balancing factor τ . A setting of $\tau = 0$ corresponds to an MAE-only model, and we also evaluated a JEPA-only model without the reconstruction term.

The results in Table 12 reveal a clear synergy. The MAE-only model ($\tau = 0$) establishes a strong performance baseline, while the JEPA-only model performs poorly, indicating the latent prediction task alone is insufficient. Crucially, the unified model with an optimal balance ($\tau = 0.03$) surpasses the MAE-only baseline. This demonstrates that the JEPA objective acts as a beneficial regularizer that refines the learned features via MAE, highlighting the complementary nature of the two objectives.

Table 12: Ablation study on the MAE and JEPA components.

	Age (MSE) \downarrow	Age (Pearson) \uparrow
JEPA-only	0.719 ± 0.040	0.521 ± 0.036
$\tau = 0$	0.480 ± 0.010	0.717 ± 0.006
$\tau = 0.03$	0.466 ± 0.025	0.738 ± 0.014
$\tau = 0.1$	0.663 ± 0.027	0.572 ± 0.028

D.10 NEUROBIOLOGICAL INTERPRETATION VIA INTEGRATED GRADIENTS

Integrated Gradients (IG) is an attribution analysis method from the field of explainable AI (XAI) that quantifies the contribution of each input feature—in our case, each brain ROI—to a model prediction (Sundararajan et al., 2017). The resulting IG score reflects how much a given ROI positively or negatively contributes to the model output, relative to a reference baseline input. Scores near zero indicate minimal influence. To highlight the most decisive features, we computed the absolute IG scores for each subject and normalized them across ROIs to enable comparison of relative importance. IG scores were computed from the trained models for both the HCP-A age regression and HCP-EP schizophrenia diagnosis tasks using the Captum (Kokhlikyan et al., 2020) library. The top 10 ROIs with the highest absolute IG scores are summarized in Table 13 and Table 14 for each task, alongside their corresponding Yeo-7 network and AAL atlas labels.

The spatial distribution of these attribution scores is visualized in Figure 10. For the age regression task in HCP-A, the analysis highlighted regions integral to motor, cognitive, and sensory functions known to undergo aging-related alterations, specifically the left precentral gyrus (Zhou et al., 2020), left medial superior frontal gyrus (Lamballais et al., 2020), and bilateral angular and occipital gyri (Fjell et al., 2009). In the HCP-EP diagnosis task, the analysis emphasized areas crucial to sensory perception, executive control, and self-awareness, all domains notably impaired in psychotic disorders. Prominent regions included the right postcentral gyrus (Ferro et al., 2015), bilateral superior occipital gyri (Tohid et al., 2015), right middle frontal gyrus (Stoyanov et al., 2021), and left superior parietal gyrus (Guo et al., 2014).

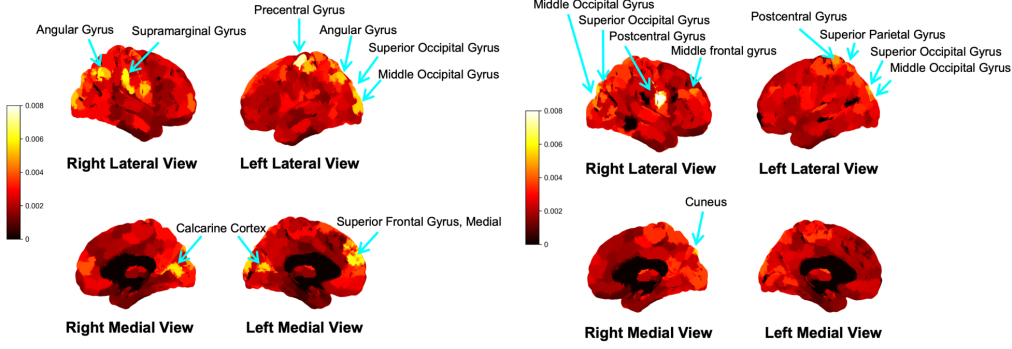


Figure 10: Brain surface visualization of IG scores. **(Left)** Age regression on the HCP-A dataset. **(Right)** Psychotic disorder diagnosis on the HCP-EP dataset.

Table 13: Top 10 ROIs with the highest IG scores in the HCP-A age prediction task.

Rank	Yeo-7 Network Label	IG Score	AAL Atlas Label
1	7Networks.LH.SomMot_26	0.0074	Precentral_L
2	7Networks.LH.Default_PFC_13	0.0061	Frontal_Sup_Medial_L
3	7Networks.RH.SalVentAttn_TempOccPar_6	0.0060	SupraMarginal_R
4	7Networks.RH.Default_Par_5	0.0059	Angular_R
5	7Networks.RH.Vis_19	0.0057	Calcarine_R
6	7Networks.RH.SalVentAttn_TempOccPar_5	0.0056	SupraMarginal_R
7	7Networks.LH.Vis_24	0.0056	Occipital_Sup_L
8	7Networks.LH.Vis_21	0.0055	Calcarine_L
9	7Networks.LH.Default_Par_6	0.0054	Angular_L
10	7Networks.LH.Vis_19	0.0054	Occipital_Mid_L

Table 14: Top 10 ROIs with the highest IG scores in the HCP-EP diagnosis prediction task.

Rank	Yeo-7 Network Label	IG Score	AAL Atlas Label
1	7Networks.RH.SomMot_16	0.0082	Postcentral_R
2	7Networks.RH.Vis_29	0.0056	Cuneus_R
3	7Networks.LH.Vis_29	0.0049	Occipital_Sup_L
4	7Networks.LH.Vis_27	0.0047	Occipital_Mid_L
5	7Networks.LH.SomMot_36	0.0047	Postcentral_L
6	7Networks.RH.SalVentAttn_PFC1_1	0.0046	Frontal_Mid_2_R
7	7Networks.RH.Vis_26	0.0045	Occipital_Sup_R
8	7Networks.LH.DorsAttn_Post_14	0.0042	Parietal_Sup_L
9	7Networks.LH.SomMot_14	0.0042	Postcentral_L
10	7Networks.RH.DorsAttn_Post_4	0.0041	Occipital_Mid_R

D.11 ADDITIONAL BASELINE

We attempted to include Brain-JEPA, one of the most prominent SSL models for fMRI analysis, as an additional baseline in our study. However, when Brain-JEPA was retrained under the same preprocessing pipeline used for all other models in this study as described in Section D.1, its reported performance could not be reproduced. For this reason, we excluded Brain-JEPA from the main text.

To isolate the effect of preprocessing, we conducted an auxiliary experiment in which both our model and Brain-JEPA were trained with the original Brain-JEPA preprocessing pipeline and evaluated via linear probing on the tasks from HCP-A. The comparison with Brain-JEPA is presented in Table 15. Note that BDO and Brain-JEPA share exactly the same Transformer backbone in this experiment. Under identical conditions such as preprocessing, model architecture, and model size, BDO consistently outperforms Brain-JEPA across all the tasks on HCP-A.

The only difference in preprocessing is the application of per-sample zero-mean normalization. The global mean may carry information relevant to demographic variables like age and gender, whereas task performance measures like Flanker are potentially more dependent on localized neural activity.

Table 15: Linear probing performance on HCP-A.

Methods	Age		Gender		Flanker	
	MSE ↓	ρ ↑	ACC (%) ↑	F1 (%) ↑	MSE ↓	ρ ↑
Brain-JEPA (86M)	0.408 ± .023	0.780 ± .004	68.92 ± 0.80	66.98 ± 3.72	0.994 ± .321	0.338 ± .029
BDO (86M)	0.298 ± .022	0.839 ± .010	74.48 ± 1.82	74.52 ± 3.81	0.966 ± .073	0.343 ± .059

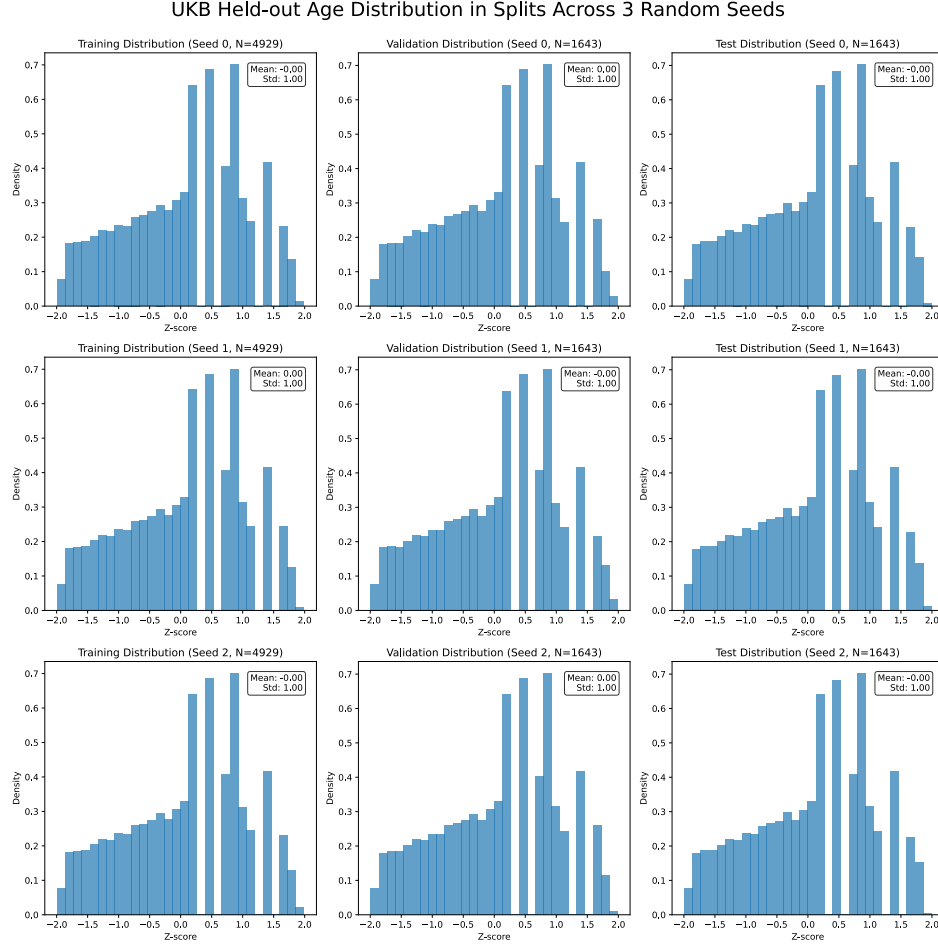


Figure 11: Age distribution across training, validation, and test splits for the UKB held-out age regression task under three different random seeds (0, 1, and 2). The dataset is partitioned using a 6:2:2 ratio, with binning-based stratified sampling applied to maintain a balanced target variable distribution. To enhance numerical stability, Z-score normalization is applied to the age variable. Each row represents a different random seed, illustrating the consistency of the sampling procedure across splits.

HCP-A Age Distribution in Splits Across 3 Random Seeds

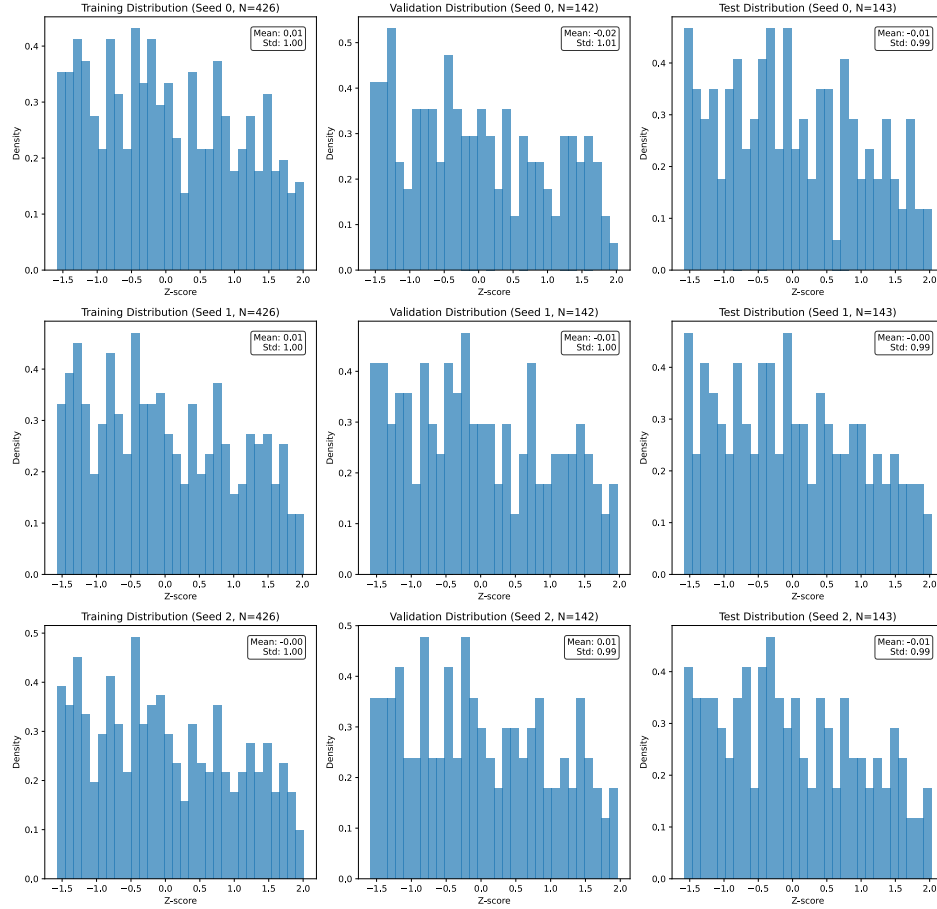


Figure 12: Age distribution across training, validation, and test splits for the HCP-A age regression task under three different random seeds (0, 1, and 2). The dataset is partitioned using a 6:2:2 ratio, with binning-based stratified sampling applied to maintain a balanced target variable distribution. To enhance numerical stability, Z-score normalization is applied to the age variable. Each row represents a different random seed, illustrating the consistency of the sampling procedure across splits.



Figure 13: Label distributions across six classification tasks (UKB held-out gender, HCP-A gender, ABIDE autism, ADHD200 ADHD, and HCP-EP psychotic disorder) for training, validation, and test splits. Each row corresponds to a different task, with columns representing the proportion of samples per class across data splits. Stratified sampling ensures that label distributions remain consistent across splits, despite variations in sample composition. To illustrate this, we visualize the distributions using a single random seed (0). Gender classification tasks are divided into Female/Male categories, while disease classification tasks distinguish between Control and Patient groups (ASD vs. Control for ABIDE, ADHD vs. Control for ADHD200, and Psychotic disorder vs. Control for HCP-EP).

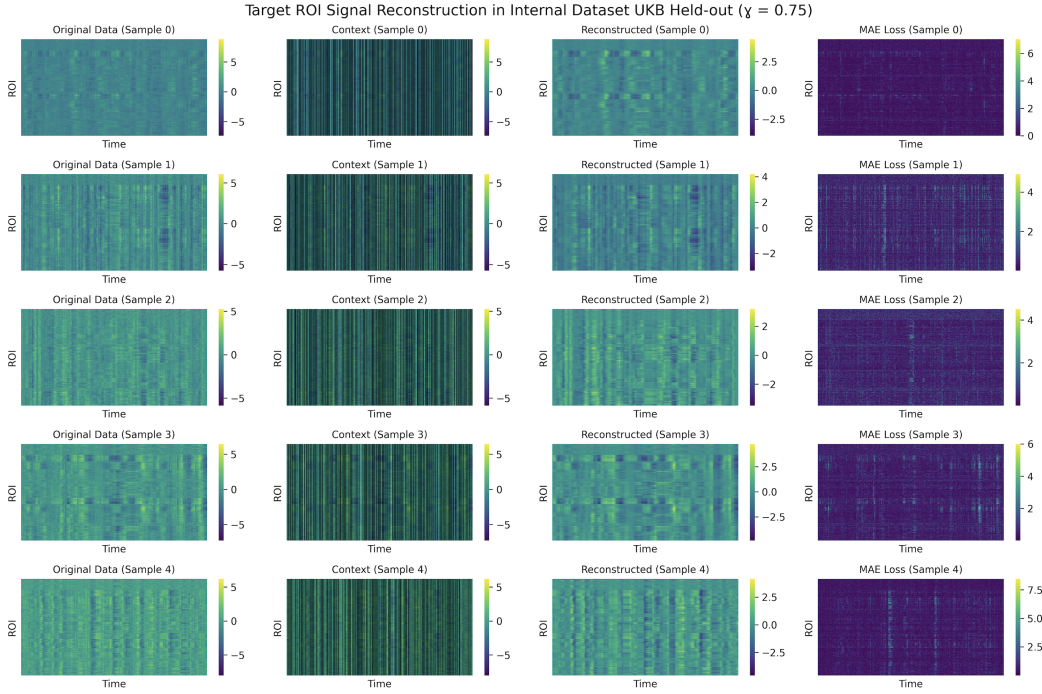


Figure 14: Reconstruction quality of BDO in the UKB held-out subset (internal dataset). Five samples are randomly drawn for visualization, with a mask ratio of $\gamma = 0.75$. Each column represents the original fMRI sample, context with masking patterns, reconstructed sample, and MAE (Mean Absolute Error) heatmaps. Although we set the mask ratio as high as 75%, the reconstruction quality remains robust, demonstrating that BDO efficiently captures the underlying brain dynamics and successfully reconstructs missing regions with high fidelity.

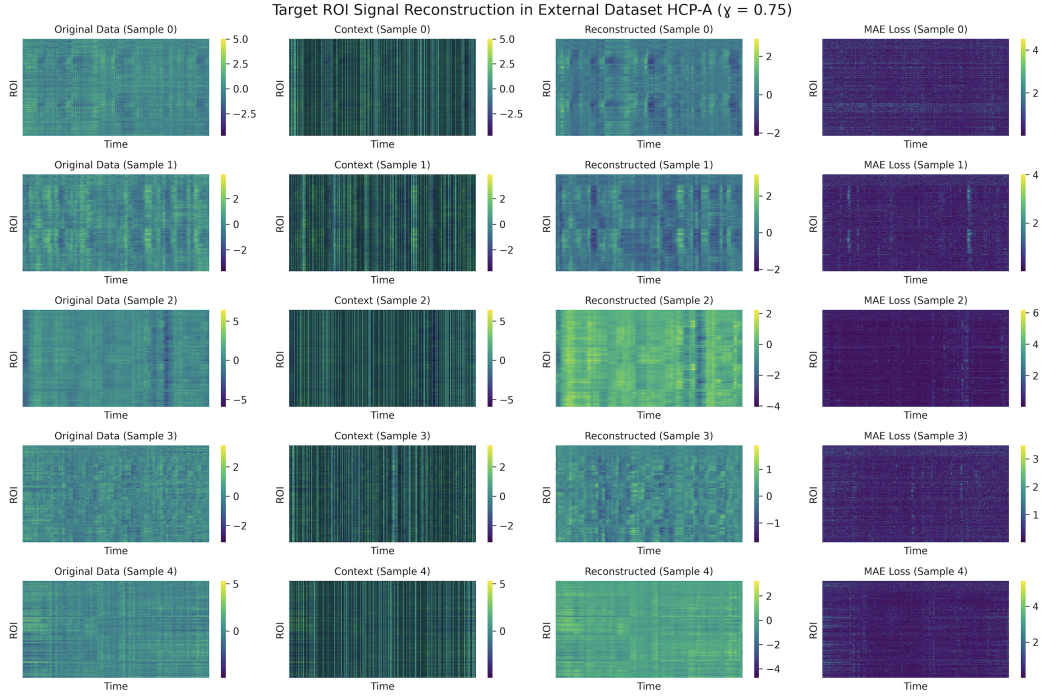


Figure 15: Reconstruction quality of BDO in HCP-A (external dataset). Five samples are randomly drawn for visualization, with a mask ratio of $\gamma = 0.75$. Each column represents the original fMRI sample, context with masking patterns, reconstructed sample, and MAE (Mean Absolute Error) heatmaps. Although we set the mask ratio as high as 75%, the reconstruction quality remains robust, demonstrating that BDO efficiently captures the underlying brain dynamics and successfully reconstructs missing regions with high fidelity.