UNIBRAIN: SYNERGISTIC EEG-FMRI MODELING VIA TEMPORAL-SPATIAL CONSTRAINED REPRESENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Electroencephalography (EEG) resolves millisecond-scale dynamics of neural activity, whereas functional magnetic resonance imaging (fMRI) localizes activity to distinct brain regions. A unified multimodal representation should draw its temporal structure from EEG and its spatial organization from fMRI. Existing fusion models map both modalities into a shared space and harmonize them with modality-agnostic objectives. Because EEG and fMRI lack a one-to-one token-level correspondence, such training blurs EEG-specific temporal fluctuations and fMRI-specific spatial organization. We propose UniBrain, a self-supervised foundation model in which each modality constrains the representation axis that it measures most reliably. Temporal Pattern Alignment (TPA) anchors latent trajectories to EEG fluctuation patterns through a differentiable monotonic alignment that tolerates unequal sequence lengths. Spatial Functional Connectivity Alignment (SFCA) anchors latent relational structure to functional connectivity derived from fMRI. Pretrained on simultaneously acquired EEG–fMRI recordings, UniBrain outperforms unimodal and multimodal foundation models on 4 paired and 6 unimodal benchmarks. Joint pretraining also improves decoding when only a single modality is available, showing that the two modalities refine a shared representation instead of coexisting as independent streams. Code will be released on GitHub.

1 INTRODUCTION

Electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) are widely used noninvasive techniques for investigating human neural activity (Michel & Murray, 2012; Ogawa et al., 1990). Although both modalities reflect neural processes, they operate through distinct biophysical mechanisms and, in effect, partition the observable dimensions of neural activity: EEG records scalp voltage fluctuations associated with synchronized electrophysiological activity, resolving millisecond-scale dynamics, whereas fMRI measures blood-oxygen-level-dependent (BOLD) signals that are coupled to neural activity (Logothetis et al., 2001), localizing activity to distinct brain regions. Consequently, neither modality alone can fully characterize the temporal evolution and spatial organization of brain activity (Cichy & Oliva, 2020; Jorge et al., 2014).

The complementary properties of EEG and fMRI motivate their integration for learning more comprehensive brain representations. Classical fusion methods establish cross-modal associations at selected feature levels, including EEG-informed fMRI (Goldman et al., 2002; Jorge et al., 2014) and fMRI-constrained EEG source imaging (Henson et al., 2010). However, these correspondence methods do not construct a unified cross-modal representation space for learning and decoding. More recently, self-supervised multimodal frameworks have enabled neural signal models to learn transferable representations from multimodal datasets and support a broader range of downstream decoding tasks. For example, NOBEL maps heterogeneous neural signals into a shared token space using a large language model backbone (Tang et al., 2026), whereas MCSP employs cross-domain and cross-modal self-supervision across spatial, temporal, and spectral domains (Wei et al., 2025).

Despite this progress, the spatiotemporal mismatch between EEG and fMRI remains insufficiently addressed. Recent models accommodate heterogeneous input resolutions through modality-specific

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

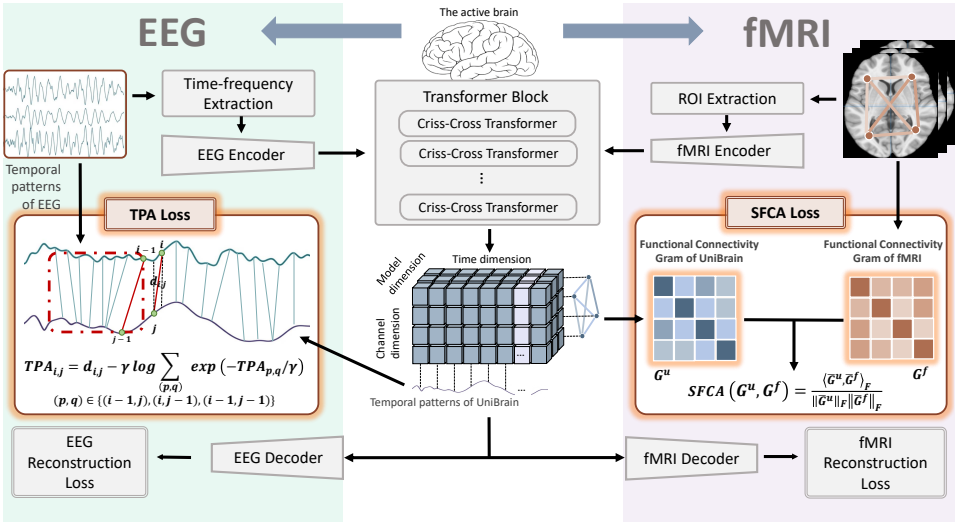

Figure 1: Overview of the UniBrain framework. EEG undergoes time–frequency extraction and fMRI undergoes ROI extraction. The resulting representations are fused into a shared uni-space and refined by stacked criss-cross Transformer blocks. TPA aligns uni-space temporal patterns with EEG features despite unequal sequence lengths and SFCA aligns the uni-space virtual-neuron relation matrix with the fMRI-derived brain-region functional-connectivity matrix. The refined uni-space is decoded to reconstruct both EEG and fMRI signals.

encoders or unified codebooks. But once the two streams enter a shared space, they are harmonized with modality-agnostic objectives. Such harmonization obscures the complementary information contained in each unique modalities, including the temporal fluctuation patterns of EEG and the functional connectivity structure of fMRI. We argue that a unified EEG–fMRI representation should inherit its temporal structure from EEG and its spatial organization from fMRI.

To operationalize this principle, we propose UniBrain, a self-supervised multimodal framework that encodes EEG and fMRI into a shared unified representation space and lets each modality supervises its informative dimension. Temporal Pattern Alignment (TPA) anchors the temporal fluctuations of the uni-space to EEG features through a differentiable monotonic alignment that tolerates unequal sequence lengths. Spatial Functional Connectivity Alignment (SFCA) anchors the relational structure among virtual neurons to functional connectivity derived from fMRI brain-region signals.

UniBrain is pretrained in a self-supervised manner on paired EEG–fMRI recordings from five publicly available datasets. Across 4 paired EEG–fMRI benchmarks and 6 unimodal benchmarks, UniBrain outperforms existing unimodal and multimodal foundation models, and ablation studies confirm that both constraints contribute consistently. Notably, joint pretraining improves decoding even when only a single modality is available, indicating that the two modalities refine a shared representation rather than merely coexisting within it.

In summary, our contributions are as follows:

- **Paired EEG–fMRI pretraining:** We pretrain UniBrain on simultaneously acquired EEG–fMRI recordings from five public datasets, spanning resting state, motor imagery, sleep, inner speech, and naturalistic viewing. The resulting model achieves the best balanced accuracy across the four paired EEG–fMRI decoding benchmarks considered in this work.
- **Temporal representation constraint:** TPA anchors temporal fluctuations in the uni-space to EEG features, preserving temporally informative structure despite unequal sequence lengths.
- **Spatial representation constraint:** SFCA anchors virtual-neuron relationships in the representation space to fMRI-derived functional connectivity, preserving spatial functional structure during multimodal fusion.

2 Related Work

EEG–fMRI Complementarity and Classical Fusion The integration of EEG and fMRI has long attracted attention in systems neuroscience. Classical fusion strategies operate on physically derived features: EEG-informed fMRI employs instantaneous EEG band-power regressors to identify BOLD correlates of fast neural rhythms (Goldman et al., 2002; Jorge et al., 2014), whereas fMRI-constrained EEG/MEG source imaging improves the spatial resolution of electrophysiological reconstructions (Henson et al., 2010). Concurrent recordings further indicate that EEG-derived and fMRI-derived functional connectivity networks exhibit coupled dynamics during cognitive tasks (Wirsich et al., 2020), motivating their joint modeling at the representation level. However, these classical methods establish cross-modal associations at predefined feature stages, rather than learning a unified representation space that can support diverse downstream decoding tasks.

Self-Supervised Brain Representation Models Self-supervised foundation models for EEG and fMRI have achieved substantial progress in unimodal representation learning. However, they remain limited to a single modality and therefore cannot exploit complementary information from the other modality (Jiang et al., 2024; Wang et al., 2025; Ortega Caro et al., 2024; Dong et al., 2024). The emergence of multimodal brain foundation models has begun to address this limitation. MCSP (Wei et al., 2025) introduces cross-domain and cross-modal self-supervised losses to bridge the spatial, temporal, and spectral domains. Its cross-domain self-supervised loss employs domain-specific augmentations and contrastive objectives to enhance intra-modal representations, whereas its cross-modal self-supervised loss distills complementary information between fMRI and EEG through feature-convergence objectives. NOBEL (Tang et al., 2026) combines dedicated fMRI and EEG encoders, tokenizes their outputs, and feeds them into a large language model backbone to support instruction-conditioned neural decoding. Brain-OF (Guo et al., 2026) employs an any-resolution neural signal sampler to project signals with heterogeneous sampling rates and channel counts into a shared semantic space. Although these methods explore brain representation learning, their training objectives do not fully exploit the respective temporal and spatial advantages of each modality. This limitation restricts their ability to model the intrinsic temporal dynamics and spatial functional organization of neural signals.

To ensure that the representation space captures cross-modally consistent temporal fluctuations and spatial relationships, optimizing encoders and constraining signal reconstruction alone is insufficient. There is an urgent need for a mechanism that explicitly constrains the representation space along both temporal and spatial dimensions, preserving the spatiotemporal details of each modality while enabling effective multimodal fusion.

3 Method

3.1 Definitions and overview

As illustrated in Figure 1, UniBrain learns a shared representation that retains complementary temporal and spatial structure from concurrent EEG and fMRI. Given $\mathcal{X} = (\mathbf{X}^e, \mathbf{X}^f)$, where $\mathbf{X}^e \in \mathbb{R}^{B\times C_e\times T_e}$ is an EEG segment and $\mathbf{X}^f \in \mathbb{R}^{B\times R\times n_{\mathrm{TR}}}$ contains n fMRI repetition time (TR) volumes over R fMRI regions of interest (ROIs), UniBrain performs modality-specific encoding, uni-space fusion, spatiotemporal refinement, and reconstruction. Here, B is the batch size, C_e and T_e are the number of EEG channels and temporal samples. Temporal Pattern Alignment (TPA) preserves temporal fluctuations in EEG features, while Spatial Functional Connectivity Alignment (SFCA) aligns uni-space virtual-neuron relations with relations derived from fMRI brain-region functional connectivity. For an unavailable modality, a learned mask token replaces its latent representation, enabling EEG-only, fMRI-only, and paired inputs within one architecture.

3.2 Modality-specific Encoding and Uni-space Construction

EEG encoding. We partition each EEG channel into non-overlapping windows of length L and independently standardize every channel-window. A hierarchical convolutional encoder processes each window and cross-attention injects high-, middle-, and low-frequency features at corresponding temporal resolutions. With default downsampling ratios $(8, 4, 2)$ and $L = 128$, each window is

compressed into two D-dimensional tokens. We combine the resulting channel-wise features with learned electrode embeddings derived from sensor position and type, and aggregate them using learned virtual-neuron queries:

$$\mathbf{Z}^e \in \mathbb{R}^{B\times N_v\times W_e\times D} \tag{1}$$

where N_v denotes the number of virtual neurons and W_e denotes the EEG latent length.

fMRI encoding. For each fMRI volume, we standardize activations across ROIs and combine them with learned embeddings of normalized MNI coordinates and ROI type. We compress the resulting sequence with a hierarchical convolutional encoder and use a self-attention sequence mixer to model interactions among the compressed ROI features:

$$\mathbf{Z}^f \in \mathbb{R}^{B\times N_v\times W_f\times D}, \qquad W_f = n_{\mathrm{TR}} \tag{2}$$

The two modalities therefore share virtual-neuron and feature dimensions while retaining their modality-specific latent lengths.

Fusion and refinement. We concatenate $\mathbf{Z}^e$ and $\mathbf{Z}^f$ along their position dimension and project the result to a fixed-length uni-space by single linear layer:

$$\mathbf{Z}^u = \mathrm{Proj}_W\Big(\big[\mathbf{Z}^e;\mathbf{Z}^f\big]_{\mathrm{pos}}\Big) \in \mathbb{R}^{B\times N_v\times W_u\times D} \tag{3}$$

Here, Proj_W is a learned projection along the position dimension. We refine this representation with a six-layer criss-cross Transformer. Each layer splits the feature dimension into two subspaces: one attends across virtual neurons at each latent position and the other attends across latent positions for each virtual neuron according to Wang et al. (2025). A position-wise feed-forward network and residual connections follow the attention operations. This factorization captures nonlinear spatial and temporal interactions while preserving the uni-space shape. The default hyperparameters are as follows: $N_v = 32$, $D = 256$, and $W_u = 10$.

3.3 Temporal Pattern Alignment

Direct token-wise temporal matching is unsuitable because EEG and fMRI induce latent sequences of different lengths. TPA instead aligns their temporal fluctuation patterns through a differentiable monotonic sequence-alignment objective. For each virtual neuron n, we treat $\mathbf{Z}^u_{b,n,:,:} \in \mathbb{R}^{W_u\times D}$ and $\mathbf{Z}^e_{b,n,:,:} \in \mathbb{R}^{W_e\times D}$ as multivariate trajectories. After standardizing each feature dimension along its sequence axis, we define the local cost

$$d_{i,j}^{(b,n)} = \frac{1}{D}\left\|\tilde{\mathbf{Z}}^u_{b,n,i,:} - \tilde{\mathbf{Z}}^e_{b,n,j,:}\right\|_2^2 \tag{4}$$

Since the local cost computation is discrete and incompatible with gradient backpropagation, an approximation scheme is introduced to transform the state transition process into a continuous formulation: TPA aggregates local costs over monotonic paths with a temperature-controlled soft minimum. With cumulative cost $A_{i,j}^{(b,n)}$, the recurrence is presented as

$$A_{i,j}^{(b,n)} = d_{i,j}^{(b,n)} - \gamma\log\sum_{(p,q)\in\{(i-1,j),(i,j-1),(i-1,j-1)\}}\exp\left(-\frac{A_{p,q}^{(b,n)}}{\gamma}\right) \tag{5}$$

where b, n, i, and j respectively index the batch, virtual neuron, uni-space position, and EEG latent position. (p, q) indexes one of the three valid predecessor states in the alignment lattice. The temperature $\gamma > 0$ controls the smoothness of the minimum. We set $A_{0,0} = 0$ and assign infinite cost to infeasible boundary states. Inspired by Cuturi & Blondel (2017), we use the normalized alignment divergence $\mathcal{A}_\gamma$, which subtracts self-alignment terms to remove the soft-minimum bias. The resulting objective is

$$\mathcal{L}_{\mathrm{TPA}} = \frac{1}{BN_v}\sum_{b=1}^{B}\sum_{n=1}^{N_v}\mathcal{A}_\gamma\left(\tilde{\mathbf{Z}}^u_{b,n,:,:},\tilde{\mathbf{Z}}^e_{b,n,:,:}\right) \tag{6}$$

TPA encourages uni-space trajectories to preserve temporal variation in EEG features while allowing smooth shifts and unequal sequence lengths.

3.4 Spatial Functional Connectivity Alignment

SFCA therefore aligns the uni-space virtual-neuron relation matrix with a relation matrix derived from fMRI brain-region functional connectivity. We use a single-layer unidirectional Gated Recurrent Unit (GRU) to compress the temporal dimension, taking the hidden state at the last time step as the spatial representation $\mathbf{A}_b^u$ of the uni-space (Cho et al., 2014). Its virtual-neuron relation matrix is constructed as

$$\mathbf{G}_b^u = \frac{1}{D}\mathbf{A}_b^u(\mathbf{A}_b^u)^\top \in \mathbb{R}^{N_v \times N_v} \tag{7}$$

We derive the fMRI target $\mathbf{G}_b^f \in \mathbb{R}^{N_v \times N_v}$ by constructing a functional-connectivity matrix from fMRI across brain-region. After double centering $\bar{\mathbf{G}}_b^q = \mathbf{H}\mathbf{G}_b^q\mathbf{H}$ for $q \in \{u, f\}$, with $\mathbf{H} = \mathbf{I} - \frac{1}{N_v}\mathbf{1}\mathbf{1}^\top$, we measure normalized relation similarity by linear centered kernel alignment (Kornblith et al., 2019):

$$s(\mathbf{G}_b^u, \mathbf{G}_b^f) = \frac{\langle \bar{\mathbf{G}}_b^u, \bar{\mathbf{G}}_b^f \rangle_F}{\|\bar{\mathbf{G}}_b^u\|_F \|\bar{\mathbf{G}}_b^f\|_F + \epsilon} \tag{8}$$

Here, q selects the uni-space (u) or fMRI (f) relation matrix, $\mathbf{I}$ is the $N_v \times N_v$ identity matrix, $\mathbf{1}$ is an N_v-dimensional all-ones vector, $\langle \cdot, \cdot \rangle_F$ and $\| \cdot \|_F$ denote the Frobenius inner product and norm, and $\epsilon > 0$ prevents division by zero. SFCA minimizes the discrepancy between the two relation matrices:

$$\mathcal{L}_{\mathrm{SFCA}} = \frac{1}{B}\sum_{b=1}^{B}\Big[1 - s(\mathbf{G}_b^u, \mathbf{G}_b^f)\Big] \tag{9}$$

Matching relation matrices rather than individual latent coordinates allows the uni-space to retain fMRI functional organization without imposing a one-to-one correspondence between latent features and brain regions.

3.5 Reconstruction and Joint Optimization

The refined uni-space is decoded into both modalities. For EEG, a learned projection maps W_u positions to W_e positions before a sensor-conditioned decoder reconstructs channel-window signals. For fMRI, an ROI decoder maps uni-space to ROI-wise activity and restores the original ROI order. The EEG reconstruction loss $\mathcal{L}_e$ is the time-domain ℓ_1 error. The fMRI loss $\mathcal{L}_f$ is the ROI-wise MSE. The complete objective is

$$\mathcal{L}_{\mathrm{UniBrain}} = \mathcal{L}_e + \lambda_f\mathcal{L}_f + \lambda_{\mathrm{TPA}}\mathcal{L}_{\mathrm{TPA}} + \lambda_{\mathrm{SFCA}}\mathcal{L}_{\mathrm{SFCA}}. \tag{10}$$

Here, λ_f, λ_{TPA}, and λ_{SFCA} are scalar loss weights for fMRI reconstruction, TPA, and SFCA, respectively. All modules are optimized end-to-end. Reconstruction and alignment terms are evaluated only when their required modalities are present.

4 Experiment Setup

4.1 Pretraining

Pretrain datasets. The model is pretrained on five publicly available EEG–fMRI datasets, including 94c5t (Deligianni et al., 2014), ds002338 (Lioi et al., 2020), ds003768 (Gu et al., 2023), ds006033 (Simistira Liwicki et al., 2025), and CineBrain (Gao et al., 2025). They cover a wide spectrum of BCI paradigms, including resting state, motor imagery, sleep, inner speech, and naturalistic viewing. As shown in Table 1, the pretraining datasets comprise a total of 109.27 hours of data across diverse EEG and fMRI acquisition configurations. Among them, 80% of the data is used for training, 10% of the data is used for validation, and 10% of the data is used for testing. More details can be seen in Appendix A.1.

Preprocessing. The raw EEG data were resampled to a uniform sampling rate of 250 Hz. The 0.1–100 Hz bandpass filter, 50/60 notch filters, and z-score normalization were applied to reduce noise. For each dataset, the EEG signals were segmented according to the corresponding fMRI repetition time (TR) for cross-modal alignment. Then zero-padding was applied to ensure the unified length of EEG segments across all datasets. The fMRI BOLD signals were z-score normalized for each

TR and then processed using the Schaefer 2018 1,000-parcel, 17-network atlas (Schaefer et al., 2018). ROI-level signals were extracted using the Nilearn label masker (Abraham et al., 2014). The corresponding MNI centroid coordinates were min-max normalized to the range of [-1,1]. For the HCP-YA dataset (Van Essen et al., 2013), the CIFTI-format data could not be processed using volumetric masks. Therefore, the cortical vertices were averaged according to a predefined lookup table to map them onto the Schaefer-1000 parcellation, while subcortical regions were excluded. Concurrently acquired EEG and fMRI signals were strictly aligned based on their timestamps. When either EEG or fMRI recordings are unavailable, the UniBrain model automatically replaces the missing modality with mask tokens.

Table 1: Characteristics of the datasets used for UniBrain pretraining and downstream evaluation.

Usage	Modality	Dataset	BCI Paradigm	Subjects	Hours	EEG Channels	fMRI TR
Pretraining	*Simultaneous EEG–fMRI*	94c5t	Resting state	16	2.88	63	2.16 s
		ds002338	Motor imagery	17	7.60	63	1.0 s
		ds003768	Sleep state	33	60.60	32	2.1 s
		ds006033	Inner speech	3	2.19	63	2.0 s
		CineBrain	Video watching	6	36.00	64	0.8 s
Downstream evaluation	*Simultaneous EEG–fMRI*	ds002336	Motor imagery	10	5.37	63	2.0 s
		ds006040	Sustained attention	28	27.90	63	2.0 s
		ds007216	Sustained attention	24	32.00	31	2.0 s
		NATView	Naturalistic viewing	22	47.00	61	2.1 s
	EEG-only	FACED	Emotion detection	123	155.30	32	–
		ISRUC	Sleep staging	118	1,008.00	6	–
		PhysioNet-MI	Motor imagery	1,983	15,261.00	6	–
	fMRI-only	ADHD-200	Disease diagnosis	973	102.00	–	1.5–3.0 s
		HCP-YA-Emo	Emotion detection	176	12.43	–	0.72 s
		HCP-YA-Mot	Motor imagery	284	21.36	–	0.72 s

4.2 Evaluation

Downstream Datasets. The transferability of our pretrained model was evaluated across three categories of downstream datasets: simultaneous EEG–fMRI datasets, EEG-only datasets, and fMRI-only datasets. As shown in Table 1, all datasets excluded from pre-training are employed for downstream task testing to prevent information leakage. The paired EEG–fMRI benchmarks comprise motor imagery on ds002336 (Lioi et al., 2020), sustained attention on ds006040 (Cha et al., 2026) and ds007216 (Shareef-Trudeau et al., 2026), and naturalistic viewing on NatView (Telesford et al., 2023). The EEG-only benchmarks comprise emotion detection on FACED (Chen et al., 2023), sleep staging on ISRUC (Khalighi et al., 2016), and motor imagery on PhysioNet-MI (Schalk et al., 2004; Goldberger et al., 2000). The fMRI-only benchmarks comprise disease diagnosis on ADHD-200 (The ADHD-200 Consortium, 2012) and emotion and motor decoding on HCP-YA (Van Essen et al., 2013). Consistent with pretraining, bandpass filtering, notch filtering, resampling and normalization were applied for downstream dataset preprocessing.

Baselines. UniBrain is compared with three multimodal EEG–fMRI representation models: BrainOF (Guo et al., 2026), NOBEL (Tang et al., 2026), and MCSP (Wei et al., 2025). Two specialized EEG foundation models (LaBraM (Jiang et al., 2024) and CBraMod (Wang et al., 2025)) and two specialized fMRI foundation models (BrainLM (Ortega Caro et al., 2024) and Brain-JEPA (Dong et al., 2024)) are additionally included.

Downstream Settings. To ensure fair comparison, UniBrain is compared to all baselines using the same preprocessing procedures, training pipeline and downstream evaluation strategy. The only exception is that NOBEL employs the large language model instead of a discriminative classifier, following its original formulation (Tang et al., 2026). It maps brain tokens to a LoRA-tuned (Hu et al., 2022) Qwen2.5-1.5B-Instruct (Yang et al., 2024) and obtains predictions by greedy decoding and exact class-name string matching. To evaluate the model's generalization across subjects, a strict cross-subject split strategy was applied to all datasets where subjects in the training set do not appear in the validation or test sets. Models were trained for 30 epochs, and the model checkpoint that achieved the highest validation-set balanced accuracy was selected for testing. All results are reported as mean $\pm$ standard deviation across five runs with different random seeds, and significance

is assessed with two-sided paired t-tests against the strongest baseline in each column (against the unconstrained backbone in ablation studies), with * for $p \leq 0.05$ and ** for $p \leq 0.01$. More details can be seen in Appendix A.2.

Table 2: Paired EEG–fMRI downstream decoding performance on four datasets. Results are reported as mean ± standard deviation. Boldface denotes the best result in each column. Significance level: * for $p \leq 0.05$ and ** for $p \leq 0.01$.

Modality	Method	Parameters	NATView	ds006040	ds007216	ds002336
EEG	LaBraM	5.8M	0.372 ± 0.037	0.752 ± 0.039	0.383 ± 0.042	0.605 ± 0.042
	CBraMod	4.9M	0.402 ± 0.039	0.782 ± 0.040	0.413 ± 0.043	0.635 ± 0.043
fMRI	BrainLM	5.0M	0.386 ± 0.017	0.771 ± 0.018	0.293 ± 0.042	0.504 ± 0.028
	Brain-JEPA	86.0M	0.388 ± 0.018	0.772 ± 0.019	0.384 ± 0.041	0.575 ± 0.027
EEG–fMRI	BrainOF	47.5M	0.411 ± 0.025	0.762 ± 0.035	0.409 ± 0.036	0.642 ± 0.030
	NOBEL	1544M	0.410 ± 0.026	0.780 ± 0.030	0.425 ± 0.038	0.630 ± 0.032
	MCSP	3.4M	0.393 ± 0.021	0.743 ± 0.029	0.388 ± 0.030	0.623 ± 0.025
	UniBrain	12.56M	$\mathbf{0.440 \pm 0.027}^{*}$	$\mathbf{0.816 \pm 0.031}^{*}$	$\mathbf{0.445 \pm 0.040}$	$\mathbf{0.674 \pm 0.034}^{*}$

5 Results

5.1 Performance Comparison

The downstream decoding performance is first evaluated when both EEG and fMRI are available. We compare UniBrain with dedicated EEG and fMRI foundation models as well as multimodal representation models on four EEG–fMRI benchmarks covering naturalistic viewing (NATView), sustained attention (ds006040 and ds007216), and motor imagery (ds002336). Table 2 reports balanced accuracy (BACC) under the same downstream evaluation setting.

UniBrain achieves the highest balanced accuracy on all four benchmarks. Relative to the strongest baseline in each column, it improves BACC by 2.9, 3.4, 2.0, and 3.2 percentage points on NATView, ds006040, ds007216, and ds002336, respectively; the gain on ds007216 does not reach statistical significance. These gains are observed against both unimodal and multimodal baselines, including MCSP, indicating that UniBrain is effective across the diverse paired decoding settings considered here. These results are achieved with 12.56M parameters, substantially fewer than BrainOF and NOBEL, demonstrating the strength of the UniBrain architecture in simultaneous EEG–fMRI modeling.

Table 3: Effect of joint EEG–fMRI pretraining and input modality on downstream balanced accuracy. Results are reported as mean ± standard deviation.

	NATView			**ds002336**		
Modality	EEG	fMRI	EEG–fMRI	EEG	fMRI	EEG–fMRI
EEG-only	0.395 ± 0.027	–	–	0.615 ± 0.036	–	–
fMRI-only	–	0.398 ± 0.024	–	–	0.634 ± 0.027	–
UniBrain	$\mathbf{0.417 \pm 0.022}$	$\mathbf{0.400 \pm 0.027}$	$\mathbf{0.431 \pm 0.025}$	$\mathbf{0.628 \pm 0.028}$	$\mathbf{0.645 \pm 0.030}$	$\mathbf{0.661 \pm 0.034}$

5.2 Performance Analysis of Pretraining and Input Modalities

Having established the model performance under paired-input conditions, we subsequently disentangle the respective effects of multimodal pretraining and multimodal inference. We train EEG-only and fMRI-only UniBrain variants and compare them with the jointly pretrained model on NATView and ds002336. For the jointly pretrained model, we further evaluate EEG-only, fMRI-only, and combined EEG–fMRI inputs. This factorial design tests whether joint pretraining benefits each modality in isolation and whether using both modalities at inference provides an additional gain.

Table 3 shows that joint pretraining improves EEG-only and fMRI-only performance on NATView and ds002336. When both modalities are supplied at inference, the jointly pretrained model further improves over the stronger unimodal input on both datasets, reaching 0.431 on NATView and 0.661

on ds002336. These results separate the benefits of joint pretraining from those of combined inputs and show that both contribute under the evaluated settings.

Table 4: Ablation of UniBrain components on paired EEG–fMRI downstream decoding. Results are reported as balanced accuracy (mean ± standard deviation). Boldface denotes the best result in each column. Significance level: * for $p \leq 0.05$ and ** for $p \leq 0.01$.

Method	NATView	ds006040	ds007216	ds002336
UniBrain_base	0.398 ± 0.029	0.771 ± 0.033	0.407 ± 0.042	0.633 ± 0.036
w/o TPA	0.425 ± 0.026	0.798 ± 0.030	0.432 ± 0.041	0.657 ± 0.035
w/o SFCA	0.416 ± 0.028	0.789 ± 0.032	0.423 ± 0.039	0.648 ± 0.033
UniBrain	$\mathbf{0.440 \pm 0.027}^{*}$	$\mathbf{0.816 \pm 0.031}^{*}$	$\mathbf{0.445 \pm 0.040}^{*}$	$\mathbf{0.674 \pm 0.034}^{*}$

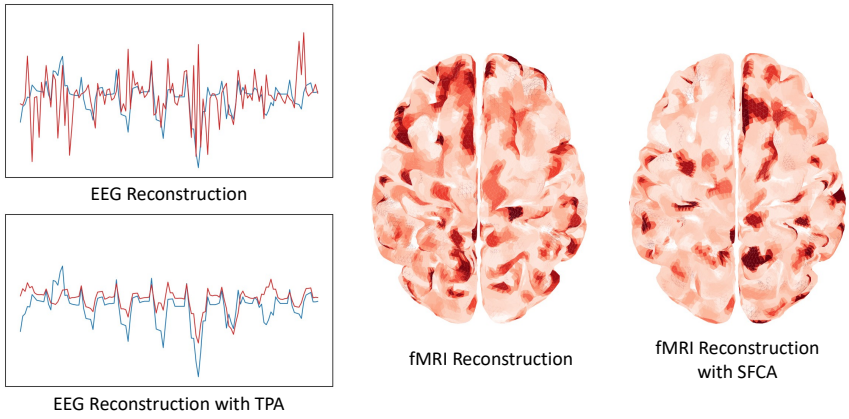

Figure 2: Qualitative comparison of EEG and fMRI reconstructions with and without constraints. For EEG, blue and red curves denote the original and reconstructed signals, respectively; for fMRI, more intense red indicates a larger deviation from the reference map.

5.3 Ablation Study

We next examine whether the two proposed constraints each contribute to the paired-decoding gains. Starting from the shared encoder–decoder backbone without TPA or SFCA, we add each constraint separately and then evaluate the complete model on the four EEG–fMRI benchmarks. This comparison isolates the contribution of each component while keeping the backbone and downstream protocol fixed. UniBrain_base denotes the backbone without temporal or spatial constraints.

As shown in Table 4, either constraint improves over the unconstrained backbone on every benchmark, and their combination is best in all four columns. Relative to the unconstrained backbone, the complete model improves BACC by 3.8–4.5 percentage points. Removing TPA from the complete model reduces BACC by 1.3–1.8 percentage points, whereas removing SFCA reduces it by 2.2–2.7 percentage points. Thus, both constraints contribute measurable gains in this ablation, and jointly applying them produces the strongest paired-decoding performance.

Figure 2 visualizes representative reconstructions of the EEG and the ROI of fMRI data. The trial was randomly selected from the downstream evaluation data to supplement and help interpret the quantitative results. In EEG reconstruction, the blue line denotes the original data and the red line denotes the reconstructed data. In fMRI reconstruction, red indicates the deviation from the original data. The more intense the red, the greater the deviation. For EEG, the reconstruction with TPA follows the major temporal fluctuations of the input signal while reducing high-frequency noise. For fMRI, the reconstruction with SFCA more closely preserves the spatial distribution of positive and negative activations in the reference map. These qualitative examples are consistent with the quantitative ablation results, illustrating that TPA and SFCA improve the temporal and spatial fidelity of the reconstructed multimodal signals, respectively. More visualization results are presented in Appendix A.8 to demonstrate the dataset independence of the qualitative conclusions.

5.4 Performance Analysis on Unimodal Datasets

To test whether representations learned from paired EEG–fMRI data transfer when only one modality is available downstream, we evaluate UniBrain separately on fMRI-only and EEG-only benchmarks. We retain the modality-appropriate input during fine-tuning and compare UniBrain with foundation models designed for that modality.

Table 5: Transfer performance on fMRI-only downstream datasets. Results are reported as mean $\pm$ standard deviation. Boldface denotes the best result in each column. Significance level: * for $p \leq 0.05$ and ** for $p \leq 0.01$.

Method	HCP-YA-Mot (4-class)		HCP-YA-Emo (2-class)		ADHD-200 (2-class)	
	BACC	F1-score	BACC	F1-score	BACC	F1-score
BrainLM	0.652 ± 0.023	0.641 ± 0.024	0.627 ± 0.026	0.624 ± 0.027	0.780 ± 0.019	0.792 ± 0.018
Brain-JEPA	0.681 ± 0.020	0.672 ± 0.021	0.634 ± 0.024	0.632 ± 0.025	0.788 ± 0.021	0.791 ± 0.020
UniBrain	$\mathbf{0.715 \pm 0.018}^{*}$	$\mathbf{0.718 \pm 0.019}^{**}$	$\mathbf{0.647 \pm 0.022}^{*}$	$\mathbf{0.641 \pm 0.023}$	$\mathbf{0.806 \pm 0.017}^{*}$	$\mathbf{0.809 \pm 0.016}^{*}$

fMRI-only transfer. We first evaluate transfer to three tasks spanning motor, emotion, and disease diagnosis. Table 5 compares UniBrain with BrainLM and Brain-JEPA using balanced accuracy and F1 score. UniBrain outperforms both fMRI baselines on all three tasks and on both reported metrics. Its largest improvement is on HCP-YA-Mot, where it exceeds Brain-JEPA by 3.4 and 4.6 percentage points on BACC and F1, respectively. The gains are smaller but consistent on HCP-YA-Emo and ADHD-200. These results show that the pretrained representation transfers to fMRI-only tasks across motor, emotion, and clinical diagnosis settings.

Table 6: Transfer performance on EEG-only downstream datasets. Results are reported as mean $\pm$ standard deviation. Boldface denotes the best result in each column. Significance level: * for $p \leq 0.05$ and ** for $p \leq 0.01$.

Method	ISRUC (5-class)		PhysioNet-MI (4-class)		FACED (9-class)	
	BACC	F1-score	BACC	F1-score	BACC	F1-score
LaBraM	0.763 ± 0.010	0.781 ± 0.013	0.617 ± 0.012	0.618 ± 0.014	0.527 ± 0.011	0.529 ± 0.009
CBraMod	0.772 ± 0.011	0.786 ± 0.009	0.629 ± 0.009	0.629 ± 0.010	0.537 ± 0.009	0.547 ± 0.010
UniBrain	$\mathbf{0.790 \pm 0.011}^{*}$	$\mathbf{0.805 \pm 0.010}^{**}$	$\mathbf{0.645 \pm 0.010}^{*}$	$\mathbf{0.646 \pm 0.010}^{*}$	$\mathbf{0.555 \pm 0.009}^{**}$	$\mathbf{0.566 \pm 0.009}^{**}$

EEG-only transfer. We then evaluate transfer to EEG-only sleep staging, motor-imagery, and emotion-classification benchmarks. Table 6 compares UniBrain with LaBraM and CBraMod using the same BACC and F1 metrics. UniBrain achieves the highest BACC and F1 score on all three EEG-only benchmarks. The consistent improvement across sleep, motor-imagery, and emotion tasks shows that the paired pretraining representation remains effective when fMRI is absent during downstream evaluation.

6 Conclusion

We introduced UniBrain, a self-supervised multimodal framework that learns unified brain representations by explicitly exploiting the complementary temporal and spatial information of EEG and fMRI. By using modality-specific temporal and spatial constraints during pretraining, UniBrain consistently improves downstream decoding performance across four paired EEG–fMRI benchmarks and supports effective transfer to unimodal settings. Ablation studies further confirm that both constraints contribute to the learned representations, with their combination providing the strongest overall performance. These results demonstrate the potential of modality-anchored representation learning for building generalizable multimodal brain foundation models, while larger-scale and broader external validation remains necessary.

USE OF GENERATIVE AI

In this work, we used generative AI tools for assisting with translation, suggesting experimental parameters, creating or editing software code, summarizing or analyzing existing literature, brainstorming, and editing a research paper to improve readability. We have not used generative AI tools for designing or providing feedback on research methodology or experiments, implementing methods, generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, cleaning or reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results. We have reviewed all AI-assisted work. We checked AI-assisted translations and wording edits by manually revising the manuscript. We checked AI-assisted literature summaries for accuracy and potential plagiarism through a manual literature survey. AI-assisted code edits were verified and tested for correctness by the authors and all reported results, numbers, and figures were confirmed against our experimental records. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8, 2014. doi: 10.3389/fninf.2014.00014.

Younghwa Cha, Yeji Lee, Eunhee Ji, SoHyun Han, Sunhyun Min, Hyoungkyu Kim, Minseo Cho, Hae Seong Lee, Youngjai Park, and Joon-Young Moon. Sustained attention task (gradCPT) dataset using simultaneous EEG-fMRI and DTI. *Scientific Data*, 13, 2026. doi: 10.1038/s41597-026-06616-6.

Jingjing Chen, Xiaobin Wang, Chen Huang, Xin Hu, Xinke Shen, and Dan Zhang. A large finer-grained affective computing EEG dataset. *Scientific Data*, 10, 2023. doi: 10.1038/s41597-023-02650-w.

Kyunghyun Cho, Bart Van Merriënboer, Çağlar Gulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1724–1734, 2014.

Radoslaw M. Cichy and Aude Oliva. A M/EEG-fMRI fusion primer: Resolving human brain responses in space and time. *Neuron*, 107(5):772–781, 2020.

Marco Cuturi and Mathieu Blondel. Soft-DTW: A differentiable loss function for time-series. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 894–903, 2017.

Fani Deligianni, Maria Centeno, David W. Carmichael, and Jonathan D. Clayden. Relating resting-state fMRI and EEG whole-brain connectomes across frequency bands. *Frontiers in Neuroscience*, 8, 2014. doi: 10.3389/fnins.2014.00258.

Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Su Xian Chong, Fang Ji, Nathanael Ren Jie Tong, Christopher Li Hsian Chen, and Juan Helen Zhou. Brain-JEPA: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jianxiong Gao, Yichang Liu, Baofeng Yang, Jianfeng Feng, and Yanwei Fu. Cinebrain: A large-scale multi-modal brain dataset during naturalistic audiovisual narrative processing, 2025. URL https://arxiv.org/abs/2503.06940.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. doi: 10.1161/01.CIR.101.23.e215.

Robin I. Goldman, John M. Stern, Jerome Engel, and Mark S. Cohen. Simultaneous EEG and fMRI of the alpha rhythm. *NeuroReport*, 13(18):2487–2492, 2002. doi: 10.1097/00001756-200212200-00022.

Yameng Gu, Lucas E. Sainburg, Feng Han, and Xiao Liu. Simultaneous EEG and functional MRI data during rest and sleep from humans. *Data in Brief*, 48:109059, 2023. doi: 10.1016/j.dib.2023.109059.

Hanning Guo, Hanwen Bi, Farah Abdellatif, Andrei Galbenus, Jon Shah, Abigail Morrison, Jürgen Dammers, et al. Brain-of: An omnifunctional foundation model for fmri, eeg and meg. *arXiv preprint arXiv:2602.23410*, 2026.

Richard N. Henson, Guillaume Flandin, Karl J. Friston, and Jérémie Mattout. A parametric empirical Bayesian framework for fMRI-constrained MEG/EEG source reconstruction. *Human Brain Mapping*, 31(10):1512–1531, 2010. doi: 10.1002/hbm.20956.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Wei-Bang Jiang, Liming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. In *International Conference on Learning Representations*, volume 2024, pp. 16405–16426, 2024.

Joana Jorge, Wietske van der Zwaag, and Patrícia Figueiredo. EEG-fMRI integration for the study of human brain function. *NeuroImage*, 102:24–34, 2014.

Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. ISRUC-sleep: A comprehensive public dataset for sleep researchers. *Computer Methods and Programs in Biomedicine*, 124:180–192, 2016. doi: 10.1016/j.cmpb.2015.10.013.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Giulia Lioi, Claire Cury, Lorraine Perronnet, Marsel Mano, Elise Bannier, Anatole Lécuyer, and Christian Barillot. Simultaneous EEG-fMRI during a neurofeedback task, a brain imaging dataset for multimodal data integration. *Scientific Data*, 7, 2020. doi: 10.1038/s41597-020-0498-3.

Nikos K. Logothetis, Jon Pauls, Mark Augath, Torsten Trinath, and Axel Oeltermann. Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412(6843):150–157, 2001. doi: 10.1038/35084005.

Christoph M. Michel and Micah M. Murray. Towards the utilization of EEG as a brain imaging tool. *NeuroImage*, 61(2):371–385, 2012. doi: 10.1016/j.neuroimage.2011.12.039.

S Ogawa, T M Lee, A R Kay, and D W Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, 87(24):9868–9872, 1990. doi: 10.1073/pnas.87.24.9868.

Josue Ortega Caro, Antonio Henrique de Oliveira Fonseca, Syed Rizvi, Matteo Rosati, Christopher Averill, James Cross, Prateek Mittal, Emanuele Zappala, Rahul Dhodapkar, Chadi Abdallah, et al. Brainlm: A foundation model for brain activity recordings. In *International Conference on Learning Representations*, volume 2024, pp. 565–576, 2024.

Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and B T Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, 28(9):3095–3114, 2018. doi: 10.1093/cercor/bhx179.

G. Schalk, D.J. McFarland, T. Hinterberger, N. Birbaumer, and J.R. Wolpaw. BCI2000: A general-purpose brain-computer interface (BCI) system. *IEEE Transactions on Biomedical Engineering*, 51(6):1034–1043, 2004. doi: 10.1109/TBME.2004.827072.

Lotus Shareef-Trudeau, David Braun, Tiara Bounyarith, Janet Li, Huiling Peng, and Aaron Kucyi. A multi-session simultaneous EEG-fMRI dataset with repeated experience sampling. *bioRxiv*, 2026. doi: 10.64898/2026.02.04.703882.

Foteini Simistira Liwicki, Rajkumar Saini, Debashis Das Chakladar, Sumit Rakesh, Vibha Gupta, Marcus Liwicki, and Johan Eriksson. Simultaneous electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) data during an inner speech task. *Data in Brief*, 63: 112258, 2025. doi: 10.1016/j.dib.2025.112258.

Changli Tang, Shurui Li, Junliang Wang, Qinfan Xiao, Zhonghao Zhai, Lei Bai, Yu Qiao, Bowen Zhou, Wen Wu, Yuanning Li, et al. One brain, omni modalities: Towards unified non-invasive brain decoding with large language models. *arXiv preprint arXiv:2602.21522*, 2026.

Qawi K. Telesford, Eduardo Gonzalez-Moreira, Ting Xu, Yiwen Tian, Stanley J. Colcombe, Jessica Cloud, Brian E. Russ, Arnaud Falchier, Maximilian Nentwich, Jens Madsen, Lucas C. Parra, Charles E. Schroeder, Michael P. Milham, and Alexandre R. Franco. An open-access dataset of naturalistic viewing using simultaneous EEG-fMRI. *Scientific Data*, 10, 2023. doi: 10.1038/s41597-023-02458-8.

The ADHD-200 Consortium. The ADHD-200 consortium: A model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in Systems Neuroscience*, 6, 2012. doi: 10.3389/fnsys.2012.00062.

David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy E.J. Behrens, Essa Yacoub, and Kamil Ugurbil. The WU-minn human connectome project: An overview. *NeuroImage*, 80:62–79, 2013. doi: 10.1016/j.neuroimage.2013.05.041.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. In *International conference on learning representations*, volume 2025, pp. 75310–75346, 2025.

Xinxu Wei, Kanhao Zhao, Yong Jiao, Nancy B. Carlisle, Hua Xie, Gregory A. Fonzo, and Yu Zhang. Multi-modal cross-domain self-supervised pre-training for fMRI and EEG fusion. *Neural Networks*, 184:107066, 2025.

Jonathan Wirsich, Anne-Lise Giraud, and Sepideh Sadaghiani. Concurrent EEG- and fMRI-derived functional connectomes exhibit linked dynamics. *NeuroImage*, 219:116998, 2020.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

A Appendix

A.1 Dataset Details

Pretraining data. Each paired sample is indexed by one fMRI repetition time (TR). Its EEG component contains the samples acquired during that TR, whereas its fMRI component contains the corresponding ROI-wise BOLD measurement. The raw EEG segment length therefore varies with the acquisition TR; we zero-pad segments after resampling to form a common input length without changing the aligned acquisition interval. Each fMRI sample is represented as 1,000 Schaefer-atlas ROI signals and padded to 1,024 slots for batched processing. We retain the electrode coordinates and the atlas-centroid coordinates with every sample. Because EEG montages can vary even across sessions of the same dataset, batches group recordings with the same available-channel count rather than assuming a dataset-wide montage.

Data splits and task construction. We construct splits at the subject level, keeping every recording window from a participant in exactly one of the training, validation, or test sets. For the pretraining corpora, the 80/10/10 split reported in the main text is applied before training; the same subject-disjoint principle is used for downstream tasks. Supervised metadata are attached to individual TR-aligned samples. Samples without a valid task label are excluded from the corresponding supervised loss and evaluation, while they remain usable for self-supervised pretraining.

Downstream task definitions. For simultaneous EEG–fMRI transfer, NATView evaluates visual-condition decoding, ds006040 and ds007216 evaluate sustained-attention decoding, and ds002336 evaluates motor-imagery decoding. The EEG-only benchmarks comprise five-class sleep staging on ISRUC, four-class motor-imagery decoding on PhysioNet-MI, and nine-class emotion recognition on FACED. The fMRI-only benchmarks comprise disease diagnosis on ADHD-200 and motor and emotion decoding on the HCP-YA subsets. ADHD-200 and HCP-YA-Emo are binary tasks; HCP-YA-Mot follows the four-class formulation reported in Table 5. These labels are derived from the task-event metadata and mapped to a common per-sample format; dataset-specific acquisition statistics are listed in Table 1.

A.2 Fine-tuning Details

Adaptation protocol. We fine-tune the complete UniBrain backbone together with a newly initialized task head. The head receives the complete uni-space token sequence: virtual-neuron and temporal-position axes are flattened into tokens, processed by two pre-normalized self-attention blocks, mean pooled, and passed to a two-layer MLP classifier. This design exposes the classifier to all latent spatiotemporal positions rather than pooling either modality before adaptation. For the unimodal experiments, the available modality is encoded normally and the missing stream is replaced with the learned modality-mask token. In the frozen-backbone analysis, only the task head is updated; the full-fine-tuning setting updates both the backbone and task head.

Optimization and model selection. For UniBrain, we use AdamW with learning rate 10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 0.01, and a 10% warm-up followed by cosine decay. The micro-batch size is 8 and gradient accumulation yields an effective batch size of 64. We train with bfloat16 and DeepSpeed ZeRO-2. All discriminative methods use the common 30-epoch schedule and subject-disjoint data splits described in the main text; each baseline retains its own architecture and task head. We select the checkpoint with the highest validation BACC and evaluate it once on the held-out test set. Reported values are the mean and standard deviation across five runs with different random seeds.

Metrics and generative baseline. BACC is the primary cross-method metric because it weights each class equally under class imbalance. We additionally report weighted F1 score for the unimodal classification benchmarks. NOBEL retains its generative formulation: brain tokens are mapped to a LoRA-tuned Qwen2.5-1.5B-Instruct model, and the predicted class is obtained by greedy decoding followed by exact matching to the class names. The resulting class predictions are evaluated with the same downstream metrics as the discriminative methods.

A.3 Pretraining Optimization

To document the optimization behavior underlying the preceding evaluations, we plot the pretraining loss over 100 epochs in Figure 3. This diagnostic assesses whether the training objective converges stably during pretraining.

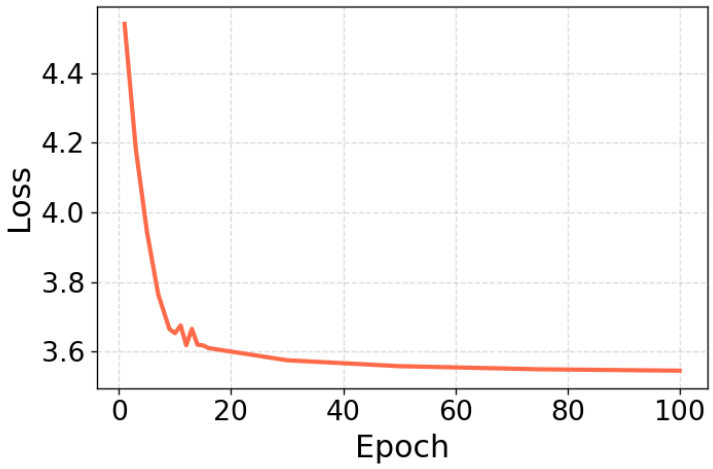

Figure 3: UniBrain pretraining loss over 100 epochs.

The loss decreases sharply during the first 10 epochs and continues to decline after the minor fluctuations observed between epochs 10 and 15. The curve remains stable through epoch 100, indicating that the pretraining objective is optimized stably under the training configuration used in this work.

A.4 Scaling Behavior of Pretraining Data

To characterize how pretraining-data scale affects downstream decoding, we pretrain UniBrain on progressively larger subsets of ds006040, ranging from 10^2 to 2×10^5 trials, and evaluate BACC on ds006040. All other model and downstream evaluation settings are kept fixed. This experiment measures the empirical scaling behavior of UniBrain over the available range of pretraining data.

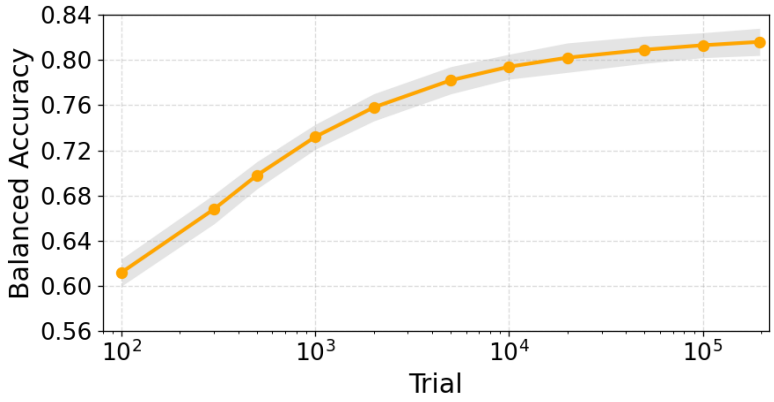

Figure 4: Scaling of ds006040 downstream balanced accuracy with the number of pretraining trials.

Figure 4 shows a monotonic increase in downstream BACC as the number of pretraining trials grows. Performance rises from 0.612 with 10^2 trials to 0.794 with 10^4 trials, and exceeds 0.815 at 2×10^5 trials. The improvement is largest at smaller scales and becomes more gradual beyond 10^4 trials. Over the evaluated range, these results show that increasing the pretraining set improves ds006040 decoding performance.

A.5 Effect of Temporal Representation Constraints

To select the temporal alignment objective used in UniBrain, we compare TPA with first-order and second-order difference losses and path-signature regularization. All variants use the same multimodal encoder–decoder backbone and differ only in the temporal constraint. We evaluate downstream decoding performance to determine whether a constraint is consistently effective across evaluation settings.

Table 7: Comparison of temporal representation constraints on paired EEG–fMRI downstream decoding. UniBrain_base denotes the backbone without temporal constraints. Results are reported as balanced accuracy (mean ± standard deviation); Boldface denotes the best result in each column.

Method	NATView	ds006040	ds007216	ds002336
UniBrain_base	0.398 ± 0.029	0.771 ± 0.033	0.407 ± 0.042	0.633 ± 0.036
UniBrain_1stDiff	0.403 ± 0.028	0.776 ± 0.032	0.411 ± 0.041	0.638 ± 0.035
UniBrain_2ndDiff	0.400 ± 0.029	0.773 ± 0.033	0.409 ± 0.041	0.636 ± 0.035
UniBrain_PathSig	0.407 ± 0.028	0.780 ± 0.032	0.414 ± 0.040	0.641 ± 0.034
UniBrain_TPA	**0.416** $\pm 0.028^{**}$	**0.789** $\pm 0.032^{**}$	**0.423** $\pm 0.039^{**}$	**0.648** $\pm 0.033^{**}$

Table 7 shows that TPA achieves the highest BACC on all four downstream benchmarks. Relative to the temporal-unconstrained backbone, it improves BACC by 1.8 points on NATView and ds006040, 1.6 points on ds007216, and 1.5 points on ds002336. TPA also consistently outperforms the first- and second-order difference objectives, as well as path-signature regularization. These results support the selection of TPA as the temporal constraint in UniBrain under the evaluated protocol.

A.6 Effect of Spatial Representation Constraints

Having selected a temporal constraint, we next compare alternative objectives for spatial alignment. We evaluate SFCA against partial-correlation and graph-regularization objectives while fixing the multimodal encoder–decoder backbone and all other training components. This experiment tests whether the selected spatial objective yields consistent benefits across downstream tasks.

Table 8: Comparison of spatial representation constraints on paired EEG–fMRI downstream decoding. UniBrain_base denotes the backbone without spatial constraints. Results are reported as balanced accuracy (mean ± standard deviation); Boldface denotes the best result in each column.

Method	NATView	ds006040	ds007216	ds002336
UniBrain_base	0.398 ± 0.029	0.771 ± 0.033	0.407 ± 0.042	0.633 ± 0.036
UniBrain_partialCorr	0.400 ± 0.028	0.774 ± 0.032	0.410 ± 0.041	0.636 ± 0.035
UniBrain_GraphReg	0.412 ± 0.027	0.784 ± 0.031	0.416 ± 0.041	0.642 ± 0.035
UniBrain_SFCA	**0.425** $\pm 0.026^{*}$	**0.798** $\pm 0.030^{*}$	**0.432** $\pm 0.041^{**}$	**0.657** $\pm 0.035^{**}$

Table 8 shows that SFCA achieves the highest BACC on all four downstream benchmarks. Compared with the spatial-unconstrained backbone, SFCA improves BACC by 2.7 points on NATView and ds006040, 2.5 points on ds007216, and 2.4 points on ds002336. It also consistently exceeds partial-correlation and graph-regularization objectives, with a 1.3–1.6-point advantage over graph regularization. These results support the use of SFCA as the spatial alignment objective in UniBrain under the evaluated protocol.

A.7 Frozen-Backbone Probing versus Full Fine-tuning

To assess whether the pretrained representation is useful before task-specific adaptation, we compare frozen-backbone linear probing with full fine-tuning. The frozen setting updates only the downstream classifier, whereas the full setting updates the backbone and classifier jointly. We evaluate both settings for UniBrain, BrainOF, and NOBEL on the four paired benchmarks.

Under full fine-tuning, UniBrain achieves the best BACC on all four benchmarks. With the backbone frozen, UniBrain achieves the highest score on NATView (0.312), ds006040 (0.775), and ds002336 (0.613), while its ds007216 score (0.312) is close to NOBEL (0.315). Full fine-tuning improves

Table 9: Effect of backbone adaptation on paired EEG–fMRI downstream decoding. Results are reported as balanced accuracy (mean ± standard deviation); Boldface denotes the best result in each column.

Method	Finetune Setting	NATView	ds006040	ds007216	ds002336
BrainOF	Freeze	0.300 ± 0.023	0.739 ± 0.032	0.294 ± 0.033	0.597 ± 0.027
	Full Finetune	0.411 ± 0.025	0.762 ± 0.035	0.409 ± 0.036	0.642 ± 0.030
NOBEL	Freeze	0.308 ± 0.024	0.772 ± 0.027	0.315 ± 0.034	0.599 ± 0.029
	Full Finetune	0.410 ± 0.026	0.780 ± 0.030	0.425 ± 0.038	0.630 ± 0.032
UniBrain	Freeze	0.312 ± 0.024	0.775 ± 0.028	0.312 ± 0.036	0.613 ± 0.031
	Full Finetune	$\mathbf{0.440 \pm 0.027}$	$\mathbf{0.816 \pm 0.031}$	$\mathbf{0.445 \pm 0.040}$	$\mathbf{0.674 \pm 0.034}$

UniBrain by 4.1–13.3 BACC points relative to frozen probing across the four tasks. These results show that the pretrained features are usable with a frozen backbone and that task-specific adaptation provides additional gains.

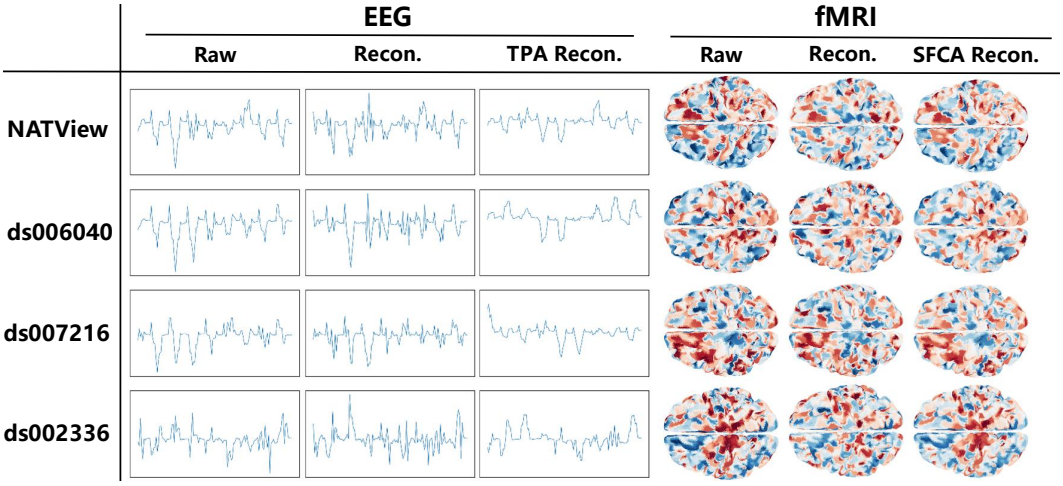

Figure 5: Qualitative reconstruction comparison on four downstream Simultaneous EEG–fMRI datasets.

A.8 Cross-Dataset Qualitative Reconstruction Analysis

Figure 5 presents qualitative reconstruction examples on all four simultaneous EEG–fMRI downstream datasets: NATView, ds006040, ds007216, and ds002336. For each dataset, we visualize a paired EEG–fMRI sample selected randomly from the held-out test set. Across the four datasets, TPA-constrained EEG reconstructions consistently preserve the dominant temporal fluctuations while suppressing spurious high-frequency variations, despite differences in task paradigm, EEG montage, and fMRI acquisition protocol. Likewise, SFCA-constrained fMRI reconstructions more closely retain the large-scale spatial distribution of positive and negative activations than unconstrained reconstructions. These examples indicate that the qualitative benefits of the proposed temporal and spatial constraints are consistently observed across the evaluated datasets.

A.9 Discussion

The present findings should be interpreted in light of several limitations. First, although pretraining and downstream evaluations use disjoint datasets, the training and test samples for each downstream task are drawn from the same dataset. This setting evaluates generalization to held-out subjects, but it does not establish robustness across datasets, acquisition sites, recording hardware, preprocessing

pipelines, or participant populations. Second, the baselines are restricted to pretrained foundation and representation models; comparisons with task-specific supervised methods trained from scratch on each benchmark would further characterize the practical benefit of pretraining. Third, SFCA currently uses a linear relation-alignment function. While this choice provides a stable and interpretable measure of representational similarity, it may not capture nonlinear functional dependencies in neural activity. Finally, the available simultaneous EEG–fMRI datasets remain limited in scale and diversity, and the current experiments do not yet characterize computational cost or robustness to incomplete or noisy recordings.

Future work will therefore focus on cross-dataset and cross-site pretraining and evaluation, with particular emphasis on separating the effects of data scale, modality composition, and model architecture. We will expand the baseline suite and assess statistical reliability across repeated runs and independent test sets. We will also investigate nonlinear relation kernels and alternative functional-connectivity formulations to model more complex spatial relationships. Additional analyses will evaluate scaling behavior, computational efficiency, signal-reconstruction fidelity, and robustness under incomplete or noisy EEG–fMRI observations. Extending the framework to additional modalities, such as MEG, may further test the generality of the proposed spatiotemporal representation-learning principle.