

000 001 002 003 004 005 MIND THE STATE: TOWARDS UNIFIED, CONTEXT- 006 AWARE EEG-TO-FMRI SYNTHESIS 007 008 009

010 **Anonymous authors**
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027

Paper under double-blind review
028

ABSTRACT

011 Functional magnetic resonance imaging (fMRI) provides dynamic measurements
012 of human brain activity at high spatial resolution and depth, but its use is con-
013 strained by high cost, limited accessibility, and strict acquisition requirements.
014 Synthesizing fMRI data from more accessible, non-invasive modalities such as
015 electroencephalography (EEG) offers a promising alternative, enabling inference of
016 deep brain activity from low-cost scalp recordings in naturalistic settings. Despite
017 recent progress, existing EEG-to-fMRI translation methods typically require train-
018 ing separate models for individual brain regions and offer limited consideration
019 of subject-level variability in brain dynamics. In this study, we propose **UniEFS**,
020 a **unified EEG-to-fMRI Synthesis** model that enables full-brain fMRI reconstruc-
021 tion while accommodating datasets with varying demographic and physiological
022 contexts within a single model. **UniEFS** leverages a pretrained fMRI decoder to
023 embed rich spatial priors, as well as condition-aware prompt tokens that encode
024 subject-level and experimental metadata to handle heterogeneous datasets. We
025 extensively evaluate our model performance on eyes-closed resting-state data and
026 demonstrate that it can reliably reconstruct temporally-resolved whole-brain fMRI
027 activity, with strong potential to generalize to task-based fMRI in a zero-shot
028 setting.

029 1 INTRODUCTION 030

031 The ability to non-invasively monitor brain activity is essential for advancing both neuroscience
032 research and clinical care. Electroencephalography (EEG) and functional magnetic resonance imaging
033 (fMRI) represent two ends of the neuroimaging spectrum. EEG captures fast, millisecond-scale
034 electrical signals from the scalp, offering a direct window into neural activity with excellent temporal
035 resolution and broad accessibility (Nicolas-Alonso & Gomez-Gil, 2012; Tong & Thankor, 2009).
036 However, it suffers from poor spatial resolution and limited sensitivity for mapping large-scale and
037 deep-brain circuits (Cohen, 2017; Chang & Chen, 2021). In contrast, fMRI provides rich spatial detail
038 by measuring blood oxygenation-level dependent (BOLD) signals driven by neurovascular coupling
039 across the entire 3D volume of the brain (Logothetis, 2008; Matthews et al., 2006). Yet fMRI is
040 expensive, infrastructure-intensive, and constrained by low temporal resolution. It is also largely
041 inaccessible in under-resourced communities and outpatient settings, and may be contraindicated for
042 patients with certain implants or conditions (Jalloul et al., 2023; van Beek et al., 2019; Geethanath &
043 Vaughan Jr, 2019). These complementary characteristics raise an intriguing question: *Can we equip*
044 *EEG with fMRI-like representational power?* If so, it would unlock a new paradigm for scalable,
045 high-resolution brain monitoring and decoding using only a lightweight, real-time, and non-invasive
046 sensor, transforming both clinical practice and cognitive neuroscience.

047 These factors motivate a growing interest in reconstructing fMRI signals from EEG, leveraging their
048 underlying physiological correlation and the representational power of deep learning to bridge the
049 spatial and temporal divide between these two modalities. A particularly underexplored area in this
050 field involves the **eyes-closed, resting-state condition**. This condition is of significant interest in both
051 research and clinical contexts: it offers a window into the brain’s intrinsic functional organization
052 and is widely used due to its simplicity and ease of implementation. It is especially valuable for its
053 applicability to diverse populations, including children and patients who may not tolerate or comply
with task-based paradigms. However, unlike task-based paradigms that provide clear temporal
anchors, resting-state brain activity is more spontaneous and variable, spanning a variety of internal

054
 055
 056
 057
 058
 brain states such as mind-wandering, vigilance fluctuations, or even light sleep, which makes it
 inherently more challenging to decode (Liu, 2016; Liu & Falahpour, 2020). For example, changes in
 vigilance introduce substantial non-stationarity: the decline from alertness into drowsiness and light
 sleep is accompanied by marked changes in the spectral content of EEG, and in the signal amplitude
 and network structure of fMRI (Liu & Falahpour, 2020; Martin et al., 2021).

059
 060
 061
 062
 063
 064
 065
 066
 067
 068
 069
 070
 071
 072
 073
 074
 075
 076
 077
 Within the body of existing work, NeuroBOLT (Li et al., 2024b) is - to our knowledge - the only study
 to date that has explored EEG-fMRI translation under the eyes-closed resting-state condition. While
 demonstrating promising results with multi-dimensional EEG representation learning, it requires
 training separate models for individual brain regions, limiting efficiency and scalability. Another very
 recent approach, CATD (Yao et al., 2025), demonstrated efficient cortical surface fMRI generation
 by conditioning a diffusion model on EEG. However, by design, its reliance on fMRI surface maps
 restricts reconstruction to the cortex, leaving subcortical regions outside the model's representational
 space. Subcortical brain areas are increasingly recognized as vital to healthy cognition as well as
 a wide range of disease processes (Favaretto et al., 2022; Koshiyama et al., 2018; Shepherd, 2013).
 This work also included resting-state data, albeit during eyes-open conditions, which may not be
 as conducive to more dramatic shifts in vigilance (e.g., falling asleep) compared to eyes-closed
 conditions. Moreover, these approaches share a common limitation: they treat the EEG-fMRI
 relationship as largely uniform across individuals. This overlooks inter-subject variability driven by
 demographic factors (e.g., age and sex) and from dynamic, time-varying physiological states (e.g.,
 drowsiness or vigilance), which are particularly pronounced during resting-state recordings. Such
 states have been shown to modulate both EEG and fMRI signals, as well as their correlations (Liu &
 Falahpour, 2020; Olbrich et al., 2009; Wong et al., 2013). Addressing this variability is therefore
 essential for developing scalable and generalizable EEG-to-fMRI translation frameworks that extend
 beyond specific conditions or cohorts. A more comprehensive review of related work is provided in
 Appendix A.

078
 079
 080
 081
 082
 083
 084
 085
 086
 087
 088
 089
 090
 091
 092
 093
 094
 095
 096
 097
 098
 099
 100
 101
 102
 103
 To address the issues discussed above, we propose a unified, context-aware framework for generalizable
 and efficient EEG-to-fMRI translation, operating in a frame-wise manner to reconstruct the fMRI
 frame at each time point. Rather than following prior work that relies solely on end-to-end training
 with scarce paired EEG-fMRI data, which may not adequately capture population-level variability,
 we adopt a two-stage strategy: **(1) uni-modal fMRI pretraining**, and **(2) cross-modal alignment**,
 where EEG signals are embedded to align with the learned fMRI latent space via a context-aware
 encoder. **Stage (1)** focuses on learning expressive and generalizable fMRI representations from
 larger-scale unpaired fMRI data with rich coverage of brain dynamics. A key motivation stems from
 the observation that fMRI activity exhibits structured spatial patterns even at the level of individual
 frames (Liu et al., 2018; 2013). In particular, co-activation pattern (CAP) analyses have revealed that
 groups of brain regions display recurring and instantaneous configurations of activation and deactivation
 (Liu et al., 2018). Building on this insight, and drawing inspiration from masked signal modeling
 (MSM) in vision, language, and neuroimaging (Chen et al., 2023; Xie et al., 2022; Radford et al.,
 2019; Yang et al., 2023; Jiang et al., 2024), we design a self-supervised masked modeling strategy
 that trains the model to recover masked brain regions from the visible context within each frame. This
 design encourages the model to capture transient spatial dependencies across regions and learn robust
 representations of instantaneous brain states. Given the domain shift between the pretraining corpus
 and the paired EEG-fMRI dataset, we then fine-tune the MAE on the fMRI portion of the EEG-fMRI
 dataset to obtain a domain-adapted encoder and decoder. In **Stage (2)**, we align EEG with this learned
 fMRI latent space and reconstruct fMRI with the pretrained decoder. To facilitate this, we introduce a
 context-aware EEG encoder that projects temporal and spectral features into the pretrained fMRI
 space, while explicitly incorporating auxiliary metadata. This contextual conditioning enables the
 model to account for individual variability in the EEG-fMRI relationship, thereby bridging the two
 modalities in a unified framework. We demonstrate that the resulting framework enables full-brain
 fMRI reconstruction from EEG within a unified model, without requiring region-specific supervision
 or subject-dependent customization. By leveraging uni-modal fMRI pretraining, domain adaptation,
 and latent alignment, UniEFS offers an effective and scalable solution for decoding intrinsic brain
 activity under eyes-closed resting-state conditions. The key contributions are summarized as follows:

104
 105
 106
 107
Context-aware EEG encoding. To better accommodate heterogeneity in EEG-fMRI data (e.g.,
 different acquisition sites, demographic attributes, and vigilance levels), we introduce prefix prompt
 tokens that encode dataset-specific and subject-level metadata, facilitating unified training across
 formats.

108 **Whole-brain EEG-to-fMRI synthesis.** We develop a unified framework that reconstructs fMRI
 109 activity, spanning hundreds of functional regions, from EEG using a single model. By first pretraining
 110 the fMRI decoder on unpaired fMRI data, we embed strong spatial priors to promote accurate and
 111 efficient reconstruction across all brain regions.

112 **Comprehensive evaluation of the predictive power.** We conduct extensive evaluations of fMRI
 113 time-series reconstruction performance across multiple brain areas, including cortical and subcor-
 114 tical regions, as well as whole-brain functional connectivity patterns. In addition, we evaluate the
 115 model’s ability to generalize across experimental conditions, demonstrating strong zero-shot transfer
 116 performance and the model’s broad predictive capacity.

118 2 METHODS

120 2.1 OVERVIEW

122 In this section, we describe the overall task setting and proposed framework of UniEFS. Our approach
 123 performs frame-by-frame fMRI prediction: given a sliding window of EEG signals preceding each
 124 fMRI time point, the model predicts the corresponding fMRI frame. This design enables flexible
 125 generation of fMRI sequences of arbitrary length. Our work mainly focuses on Regions-of-Interest-
 126 level (ROI-level) fMRI reconstruction, which offers a favorable trade-off between spatial resolution
 127 and efficiency. Compared to voxel-wise and surface-based methods, it reduces computational cost
 128 and improves signal-to-noise ratio (SNR), while also covering both cortical and subcortical regions
 129 for full-brain modeling. Moreover, as a representation adopted in recent fMRI foundation models
 130 (Dong et al., 2024; Caro et al., 2024; Thomas et al., 2022), ROI-level modeling provides a scalable
 131 and effective basis for future extensions.

132 However, achieving accurate frame-wise ROI-level reconstruction from EEG is non-trivial, due to the
 133 following key challenges. First, frame-wise reconstruction implicitly involves learning the projection
 134 from neuronal activity to its hemodynamic response, which is not uniform, varying across brain
 135 regions, individuals, and brain states. Second, paired EEG–fMRI datasets are scarce and moreover
 136 vary in their subject characteristics, potentially hindering generalization to broader populations and
 137 different conditions. To address these challenges, we propose a two-stage learning framework as
 138 illustrated in Figure 1: **(1) fMRI Pretraining and Adaptation via Masked Signal Modeling:** We first
 139 pretrain a powerful encoder-decoder model on unpaired fMRI datasets using a masked
 140 reconstruction objective. This stage enables the model to learn population-level representations
 141 of brain activity. To bridge the domain gap between pretraining and downstream application, we
 142 further fine-tune the pretrained model on the fMRI portion of the EEG-fMRI paired dataset, adapting
 143 the decoder to the target domain while preserving its generalization capacity. **(2) Context-aware**
 144 **EEG-to-fMRI Mapping:** In the second stage, we integrate a dedicated EEG encoder, conditioned
 145 on demographic and physiological priors, with the adapted fMRI decoder. The EEG encoder learns
 146 to map temporal and spectral features of EEG signals into the corresponding fMRI latent space,
 enabling full-brain fMRI reconstruction.

147 2.2 STAGE 1: FMRI PRETRAINING AND ADAPTATION VIA MASKED SIGNAL MODELING 148 (F-MSM)

150 **Pretraining.** Functional MRI measures brain activity via blood-oxygen-level-dependent (BOLD)
 151 signals represented as 3D volumes over time. To reduce dimensionality and improve signal-to-noise
 152 ratio (SNR), these signals are commonly summarized using brain parcellation techniques, which
 153 average voxel-wise signals within predefined regions of interest (ROIs), forming a 1D ROI vector per
 154 time point. This yields a parcellated fMRI matrix denoted as $Y \in \mathbb{R}^{P \times K}$, where P is the number of
 155 ROIs and K is the total number of time points. Here, we employ the Dictionaries of Functional Modes
 156 (DiFuMo) parcellation (Dadi et al., 2020) with $P=512$, which provides fine-grained, whole-brain
 157 coverage. During pretraining, each 1D ROI vector corresponding to a single time point is treated as an
 158 individual training sample, yielding K samples per fMRI scan. Although parcellation reduces voxel-
 159 level redundancy, functional dependencies and spatial correlations persist across brain regions due to
 160 the network-level organization of brain activity. To encourage the model to capture these intrinsic
 161 patterns, for each of the above ROI vectors, we adopt a high masking ratio (50%) during pretraining,
 forcing the network to infer random missing regional signals from the surrounding context. This

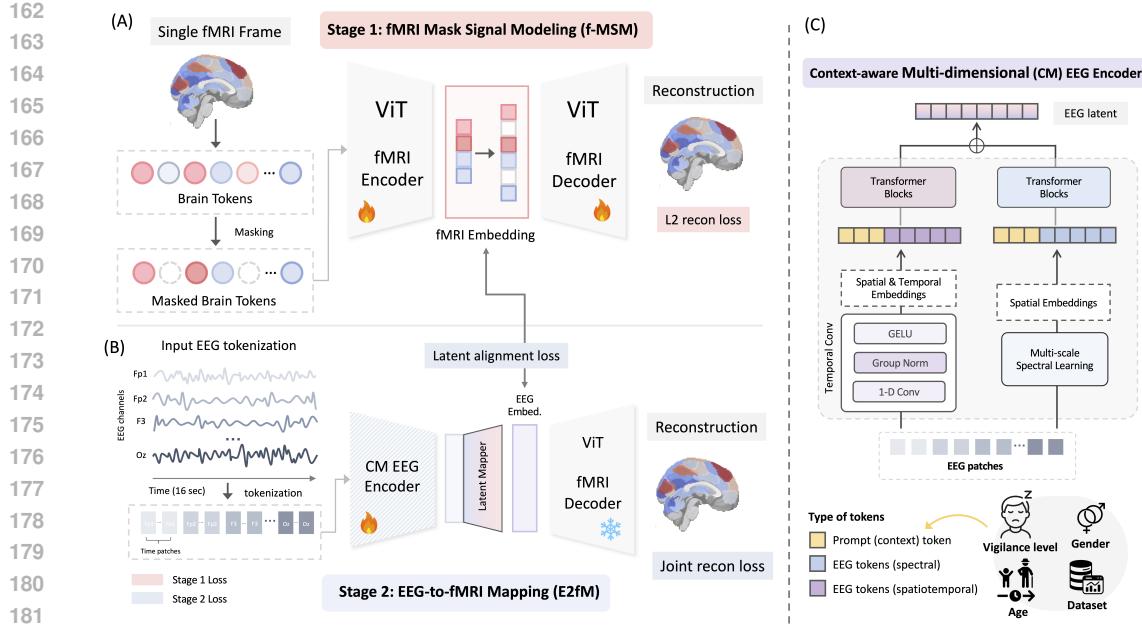


Figure 1: Overall framework. (A) Stage 1: Masked signal modeling on fMRI frames. (B) Stage 2: EEG-to-fMRI mapping. The pretrained fMRI encoder and decoder are frozen in this stage. (C) Context-aware EEG embedding.

design promotes the learning of expressive, population-level representations that generalize across individuals and tasks. For this reconstruction task, we employ a transformer-based architecture (Chen et al., 2023; 2024; Dosovitskiy et al., 2020), where each ROI is treated as an individual token. An encoder processes only the visible (unmasked) ROIs, and a lightweight decoder is trained to reconstruct the complete ROI vector based on the contextual information inferred from the unmasked regions.

Fine-tuning. After pretraining on public fMRI datasets, we adapt the model using the fMRI portion of our EEG–fMRI paired training set, with each scan preprocessed into ROI-level time series under the same parcellation scheme as used in pretraining.

Training Objective. Following He et al. (2022), the reconstruction loss (MSE) is computed solely on the masked tokens for both pretraining and finetuning.

2.3 STAGE 2: CONTEXT-AWARE EEG-TO-FMRI MAPPING

Following Li et al. (2024b), we extract the EEG window spanning a duration T , corresponding to the approximate latency of the hemodynamic response function (HRF), before each fMRI frame collection. This forms an EEG-fMRI paired input-output sample denoted as $\{X, Y^{(\text{paired})}\}$, where $X \in \mathbb{R}^{C \times T}$ represents the multichannel EEG input with C channels and T time points, and $Y^{(\text{paired})} \in \mathbb{R}^P$ denotes the corresponding parcellated fMRI ROI vector with P ROIs. The EEG input window X is first processed by the EEG encoder \mathcal{E}_{EEG} to generate a latent representation, which is then passed to the pretrained domain-adapted decoder $\mathcal{D}_{\text{fMRI}}$ obtained from f-MSM. Overall, given the full model $f_{\theta}(\cdot)$, the overall fMRI reconstruction task can be formulated as $\hat{Y}_t^{(\text{paired})} = f_{\theta}(X_{t-T:t-1})$, where $\hat{Y}_t^{(\text{paired})} \in \mathbb{R}^P$ is the reconstructed fMRI frame at time index t .

EEG Encoder. Our objective is to enable EEG-driven fMRI reconstruction by aligning EEG representations with the fMRI latent embedding space. To achieve this, we adapt the multi-dimensional encoder from NeuroBOLT (Li et al., 2024b) as the backbone encoder, a transformer-based architecture designed to capture rich and complementary spatial, temporal, and multi-scale spectral information from EEG signals. We first segment a EEG window X into non-overlapping patches using a window of length w , yielding a sequence of patches $x_{c,k} \in \mathbb{R}^w$ for each channel $c = 1, \dots, C$ and patch index

k = 1, …, $\lfloor T/w \rfloor$. These patches are then fed into (i) spatiotemporal module (a pretrained EEG encoder adapted from the EEG foundation model LaBraM (Jiang et al., 2024)) and (ii) multi-scale spectral transformer modules to generate two EEG latent embeddings $\mathbf{z}_{\text{EEG}_{st}}, \mathbf{z}_{\text{EEG}_{sp}} \in \mathbb{R}^{(C \times \frac{T}{w}) \times D}$ respectively, where D is the embedding dimension. Instead of applying global average pooling across the token dimension (Li et al., 2024b; Jiang et al., 2024; Yang et al., 2023), we retain the full sequence of token embeddings to preserve fine-grained spatial and temporal information. The embeddings from two modules are summed up and then passed through a latent mapping module, which consists of two linear projections to align the EEG embeddings with the dimensionality and structure of the fMRI latent space. This final latent representation is then passed to the fine-tuned decoder \mathcal{D} to reconstruct the full-brain fMRI signal.

Prefix Prompt Injection. To incorporate auxiliary information (Gao et al., 2024) and enhance the generalization of EEG representations, we introduce a set of learnable prefix prompts in the EEG encoder that are concatenated to the EEG patches prior to the transformer modules. These prompt tokens are designed to encode subject- and dataset-specific metadata and are optimized jointly with the rest of the model, enabling the network to adaptively condition its representation based on contextual information. Specifically, we include the following prompt tokens: **(1) Dataset tokens:** learnable embeddings of shape $\mathbb{R}^{J \times D}$, where J is a tunable hyperparameter, representing the number of dataset tokens ($J = 5$ in our experiments). Each dataset is assigned its own set of tokens, enabling the model to capture dataset-specific characteristics. **(2) Age token:** A single token generated by projecting the subject’s age through a linear embedding layer. **(3) Sex token:** A learnable token indicating the subject’s biological sex (e.g., male or female). **(4) Vigilance token:** A categorical learnable token encoding the vigilance level (drowsy, intermediate, alert) at each fMRI frame (TR), allowing the model to condition its reconstruction on frame-specific vigilance states. All prefix tokens share the same embedding dimension D as the EEG patch embeddings output by the EEG encoder. Let N_{prompt} be the total number of prompt tokens used (e.g., $N_{\text{prompt}} = 8$ when all components are included), such that $\mathbf{z}_{\text{prompt}} \in \mathbb{R}^{N_{\text{prompt}} \times D}$ represents all prompt tokens. These tokens are then concatenated to the EEG embedding \mathbf{z}_{EEG} along the token dimension to form the augmented token sequence $\mathbf{z} \in \mathbb{R}^{(N_{\text{prompt}} + N_{\text{EEG}}) \times D}$. This enriched EEG representation is then passed to the following transformers and to the alignment module for EEG-to-fMRI mapping.

EEG-fMRI Embedding Alignment. To align the EEG embeddings with the fMRI latent space, during training, we first obtain a reference embedding from fMRI by passing ground-truth fMRI vector $Y^{(\text{paired})}$ through the frozen fine-tuned fMRI encoder $\mathcal{E}_{\text{fMRI}}$. This yields the fMRI latent embedding:

$$\mathbf{E}_{\text{fMRI}} = \mathcal{E}_{\text{fMRI}}(Y^{(\text{paired})}) \in \mathbb{R}^{P \times D_{\text{fMRI}}}, \quad (1)$$

where D_{fMRI} is the dimension of the fMRI latent space.

In parallel, the EEG input X is first processed by the EEG encoder \mathcal{E}_{EEG} to obtain patch-wise latent embeddings \mathbf{z} . To align this representation with the fMRI latent space, we apply a linear projection module \mathcal{P} to map the enriched EEG embedding into the same shape as the fMRI embedding:

$$\mathbf{E}_{\text{EEG}}^{\text{proj}} = \mathcal{P}(\mathbf{z}) \in \mathbb{R}^{P \times D_{\text{fMRI}}}, \quad (2)$$

which is then used in the alignment and decoding modules.

To extract compact and semantically aligned representations between \mathbf{E}_{fMRI} and $\mathbf{E}_{\text{EEG}}^{\text{proj}}$, we adopt a LoRA-inspired low-rank projection mechanism (Hu et al., 2022). Specifically, we define two learnable matrices: $\mathbf{B} \in \mathbb{R}^{D_{\text{fMRI}} \times r}$ and $\mathbf{A} \in \mathbb{R}^{r \times D_{\text{comp}}}$, where r is the intermediate rank and D_{comp} is the final compressed dimension. The final low-rank embeddings are computed as:

$$\tilde{\mathbf{E}}_{\text{fMRI}} = \mathbf{E}_{\text{fMRI}} \cdot \mathbf{B} \cdot \mathbf{A} \in \mathbb{R}^{P \times D_{\text{comp}}}, \quad \tilde{\mathbf{E}}_{\text{EEG}} = \mathbf{E}_{\text{EEG}}^{\text{proj}} \cdot \mathbf{B} \cdot \mathbf{A} \in \mathbb{R}^{P \times D_{\text{comp}}}. \quad (3)$$

To encourage alignment between the low-rank EEG and fMRI embeddings, we optimize the mean squared error (MSE) between the two latent embeddings:

$$\mathcal{L}_{\text{align}} = \frac{1}{P} \sum_{p=1}^P \left\| \tilde{\mathbf{E}}_{\text{EEG}}^{(p)} - \tilde{\mathbf{E}}_{\text{fMRI}}^{(p)} \right\|_2^2, \quad (4)$$

270 where $\tilde{\mathbf{E}}^{(p)}$ is the embedding for the p -th ROI.
 271

272 **Training Objective.** Our framework is trained using a joint objective that combines **(1) a latent**

273 **alignment loss**, which uses MSE to align EEG and fMRI representations in latent space, and **(2) a**

274 **reconstruction loss**, which ensures accurate full-brain fMRI prediction by comparing the decoder’s

275 output to the ground truth fMRI vector. The **reconstruction loss** is formulated as a weighted

276 combination of MSE and spatial correlation (SCorr) to capture both absolute and relative accuracy:

$$\mathcal{L}_{\text{recon}} = \lambda_{\text{MSE}} \cdot \mathcal{L}_{\text{MSE}} + (1 - \lambda_{\text{MSE}}) \cdot \mathcal{L}_{\text{SCorr}}. \quad (5)$$

277 Here, the SCorr term is defined as:

$$\mathcal{L}_{\text{SCorr}} = 1 - \text{Corr}(\hat{Y}^{(\text{paired})}, Y^{(\text{paired})}), \quad (6)$$

278 where $\text{Corr}(\hat{Y}^{(\text{paired})}, Y^{(\text{paired})})$ denotes the Pearson correlation coefficient between the predicted
 279 fMRI ROI vector and the ground-truth vector. This term encourages the model to match not only
 280 the absolute magnitudes but also the *relative spatial pattern* of regional activations within each time
 281 point.

282 The overall training objective of Stage 2 is a weighted sum of the alignment and reconstruction losses:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{align}} \cdot \mathcal{L}_{\text{align}} + \lambda_{\text{recon}} \cdot \mathcal{L}_{\text{recon}}, \quad (7)$$

283 where λ_{align} and λ_{recon} are hyperparameters that balance the contribution of each loss term.

290 3 EXPERIMENTS

291 3.1 DATASETS AND PREPROCESSING

292 **Pretraining fMRI Dataset.** A subsample of resting-state fMRI (rs-fMRI) data from the HCP 1200-
 293 subject release (Van Essen et al., 2012) was used for pretraining. Subjects were scanned up to 4 times,
 294 twice on one day and twice on a second day. We included only those subjects who completed all
 295 four runs and were reported to have passed quality control in Xifra-Porxas et al. (2021); Power et al.
 296 (2017), resulting in 375 subjects ($n = 1500$ scans). The rs-fMRI scans in this dataset were acquired
 297 with a temporal resolution (TR) of 0.72 seconds, a duration of 1,200 frames per run (14.4 minutes),
 298 and a spatial resolution of 2 mm isotropic. Please refer to the Appendix B for preprocessing details.

299 **Resting-state Simultaneous EEG-fMRI Datasets.** **Dataset 1** is a shared dataset from Li et al.
 300 (2024b). This dataset comprises 29 simultaneous EEG-fMRI scans from 22 healthy subjects, with
 301 7 participants having two scans. Each scan lasts 20 minutes (575 TR, TR = 2.1 seconds). Scalp
 302 EEG was recorded using a 32-channel MR-compatible system (10–20 layout). Additional details
 303 on data acquisition and preprocessing can be found in Li et al. (2024b). **Dataset 2** is a different
 304 in-house rs-EEG-fMRI dataset. It comprises 10 scans from 7 healthy participants, with 3 individuals
 305 undergoing two scans each. During the scans, participants rested passively with their eyes closed.
 306 Written informed consent was obtained from all participants, and all procedures were approved by
 307 the Institutional Review Board. BOLD fMRI data were acquired on a 3T Siemens Prisma scanner
 308 using a multi-echo gradient-echo EPI sequence (TR = 2.1 seconds). Simultaneous scalp EEG was
 309 recorded using a 32-channel MR-compatible system (10–20 layout). To ensure consistency, this
 310 dataset was preprocessed using the same pipeline as Dataset 1. Further details on data collection and
 311 preprocessing are provided in Appendix B.

312 **Vigilance Score.** The vigilance state is a categorical score with three classes (drowsy, intermediate,
 313 and alert) assigned to each fMRI frame. This classification is derived from EEG data based on
 314 Vigilance Algorithm Leipzig (VIGALL 2.1 add-on for Brain Vision Analyzer¹) (Olbrich et al., 2015;
 315 Huang et al., 2015; Jawinski et al., 2019). VIGALL stages each 1-second segment of preprocessed
 316 EEG into one of five vigilance levels (A1, A2, A3, B1, B2/3), reflecting decreasing levels of arousal.
 317 These vigilance stage labels are then grouped into 63-second epochs (corresponding to 30 fMRI time
 318 points), and the distribution of stages within each epoch is used to assign a single vigilance class,
 319 i.e., alert, intermediate, or drowsy, to that epoch. This final label is then propagated to each of the
 320 30 fMRI frames within the corresponding epoch. Note that vigilance labels were shifted forward by
 321 5 seconds (~2 TRs) to account for the temporal delay between neural activity and the peak BOLD
 322 response. Please see the Appendix B for further details.

323 ¹<https://brainvision.com/products/analyzer-2/>

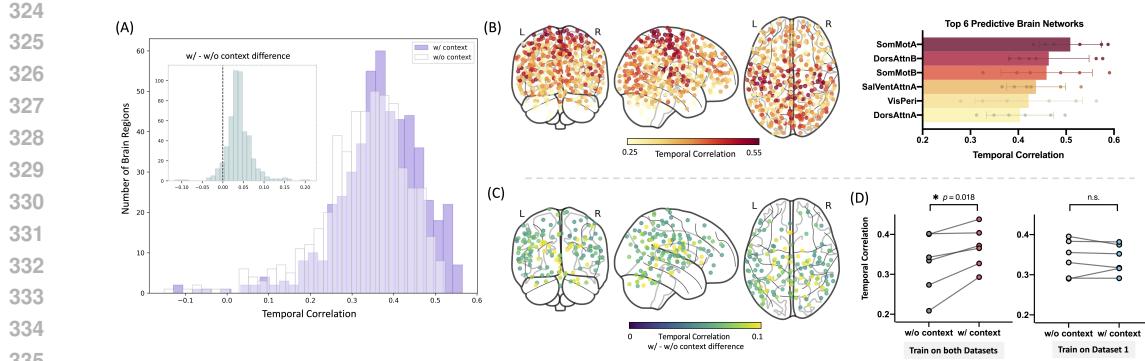


Figure 2: **Reconstruction performance.** (A) Distribution of average prediction performance for all brain regions with (w/) and without (w/o) context embedding. (B) Visualization of reconstruction performance and top predictive brain networks; dots represent brain regions. (C) Regional differences in performance. (D) Sample-wise differences in performance; points represent test scans.

3.2 EXPERIMENTAL SETUP

Data Preparation. From the preprocessed fMRI data in the paired and unpaired datasets, we extract time courses from regions of interest using the DiFuMo atlas (Dadi et al., 2020) with $P=512$ for full brain coverage. We regress out six motion-related confounds, apply a low-pass filter with a cutoff at 0.15 Hz (which captures the low-frequency band typically of interest in rs-fMRI studies), and z-normalize each ROI time series. For EEG preprocessing, we remove non-EEG channels, including ECG, EOG, and EMG, retaining 26 channels in Dataset 1 and 31 channels in Dataset 2. We select the 23 common EEG channels from these datasets for joint training. The EEG data are resampled to 200 Hz to improve computational efficiency while preserving relevant frequency information. For each fMRI time point, we extract a 16-second EEG window preceding the scan, strictly following the protocol described in Li et al. (2024b). EEG amplitude normalization is also performed, where the signal is divided by 100 to ensure that the majority of values fall within the range of -1 to 1.

Baselines. We compare our model with three open-source EEG-to-fMRI synthesis baselines (Kovalev et al., 2022; Li et al., 2024b;a) and state-of-the-art EEG encoders including recent foundation models (please refer to Appendix C.2 for details). These baseline models were originally designed or benchmarked in Li et al. (2024b) under a region-specific setting. To enable comparison on multi-region reconstruction, we adapt them by modifying the final projection layer to map the latent embeddings to the entire set of ROIs, thus extending them into multi-region baselines.

Implementation Details. All experiments are conducted on a single NVIDIA RTX A5000 GPU using Python 3.9.12, PyTorch 2.0.0, and CUDA 11.8. The training set for stage 1 consists of 1,200 scans, with 300 scans used for validation, resulting in approximately 720,000 training samples (one per time point). During stage 2, we train the model to predict fMRI signals across entire unseen scans using EEG, and use the same data partitioning strategy as in Li et al. (2024b) (an approximately 3:1:1 split for unseen-subject whole-scan reconstruction). We incorporate Dataset 2 as additional training data, resulting in a total of 28 training scans, 5 validation scans, and 6 test scans. Scans from the same individual are always assigned to the same split (training, validation, or test), **with train/validation/test subjects strictly disjoint**, since data from the same subject may have similar latent representations. For reproducibility, a fixed random seed is used across all experiments. Please refer to Appendix C for optimization and training details.

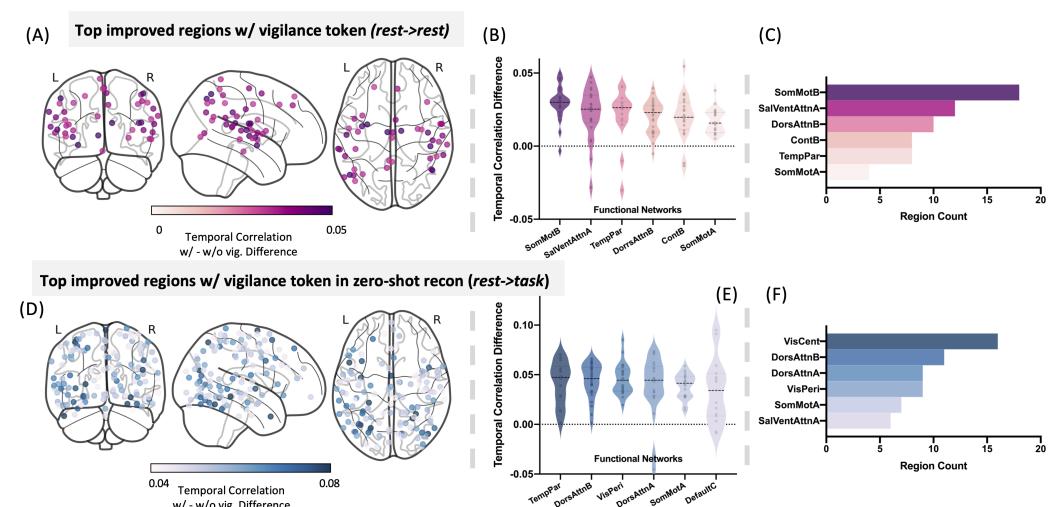
3.3 MAIN RESULTS

Our model was trained to predict held-out recordings across entire 20-minute scans using 23 EEG electrodes. We compare UniEFS with state-of-the-art EEG-to-fMRI translation baselines (Table 1) and EEG encoders (Table 5 in Appendix D.1), finding that UniEFS consistently outperforms the others in reconstructing regional time courses and has the second-best performance in recovering FC. We refer readers to Appendix E.1 for a detailed discussion on potential factors contributing to this observation and D.15 for visualization examples of reconstructed time series.

378

379 Table 1: Comparison of different models on full brain fMRI reconstruction. MM: Whether the
 380 model is originally designed for multi-region prediction; FB: Full brain; GM: Cortical gray matter;
 381 SC: Sub-cortical regions; CB: Cerebellum; Conn: Metric is applied on the full brain functional
 382 connectivity (FC) matrix; TCorr: Temporal correlation between prediction and ground truth; PCorr:
 383 Pixel-wise correlation between predicted and real FCs. **Bold**: the best; Underlined: the second best.
 384 Values are shown as mean \pm std. Paired t-test significance between our model and each baseline is
 385 shown using color codes: blue ($p < 0.05$), yellow ($p < 0.01$), red ($p < 0.001$), uncorrected.

Model Name	MM	FB TCorr \uparrow	GM TCorr \uparrow	SC TCorr \uparrow	CB TCorr \uparrow	Conn PCorr \uparrow	Conn MSE \downarrow
Ours	✓	0.367 \pm 0.052	0.394 \pm 0.060	0.276 \pm 0.082	0.247 \pm 0.060	0.527 \pm 0.084	0.233 \pm 0.072
NeuroBOLT (Li et al., 2024b)	✗	0.331 \pm 0.044	0.357 \pm 0.049	0.258 \pm 0.092	0.216 \pm 0.046	0.455 \pm 0.079	0.349 \pm 0.097
Li et al. (Li et al., 2024a)	✗	0.312 \pm 0.038	0.329 \pm 0.037	0.253 \pm 0.090	0.236 \pm 0.058	<u>0.535 \pm 0.077</u>	<u>0.217 \pm 0.065</u>
BEIRA (Kovalev et al., 2022)	✗	0.171 \pm 0.148	0.196 \pm 0.170	0.086 \pm 0.085	0.063 \pm 0.073	0.459 \pm 0.080	0.368 \pm 0.090



408 Figure 3: **Performance improvements with vigilance embedding.** (A-C) Improvements when train
 409 and test on resting-state condition. (D-F) Improvements when train on resting-state and test on task
 410 condition (zero-shot) (A, D) Brain regions showing the greatest performance improvement. (B, E)
 411 Distribution of region-wise improvement within brain networks. (C, F) The top 100 regions benefitting
 412 from vigilance embedding were selected and assigned to their corresponding brain networks. The 6
 413 networks with the highest counts among these top-ranked regions is shown.

414 Figure 2(A-C) presents a comprehensive evaluation of model performance, including region-wise
 415 distributions of predictive accuracy across brain areas with and without context embeddings. The
 416 results highlight the effectiveness of incorporating context embeddings, and indicate that activity
 417 in the somatomotor network is most reliably predicted from EEG signals, followed by the dorsal
 418 attention and salience/ventral attention networks. We further investigate the effectiveness of context
 419 embeddings under varying amounts of training data. As shown in Figure 2(D), when training on
 420 both datasets, the inclusion of context embeddings significantly improves prediction performance
 421 across individual scans (paired t-test, $p = 0.018$). However, when training is limited to Dataset 1 alone
 422 and evaluated on the same test scans, the performance difference between models with and without
 423 context embedding is negligible. This suggests that the benefit of incorporating context information
 424 emerges only when the training data exhibit sufficient variability, such as differences in subjects,
 425 vigilance levels, or population characteristics, allowing the model to meaningfully leverage auxiliary
 426 metadata.

427 To better understand the role of context information, we further analyze the effect of incorporating
 428 vigilance embeddings. Specifically, we compare model performance with and without vigilance
 429 conditioning to identify brain regions that benefit the most from this additional context. As shown in
 430 Figure 3, several regions, particularly within sensory-motor network, salience and attention-related
 431 networks, as well as one thalamus region, show marked improvement when vigilance information is
 included. These regions have been consistently reported in the fMRI literature as being associated

432 with vigilance, and the spatial distribution observed in Figure 3(A) closely overlaps with vigilance-
 433 related fMRI maps reported by previous studies (Liu & Falahpour, 2020; Schneider et al., 2016;
 434 Goodale et al., 2021). **Under zero-shot resting-state to auditory-task generation setting (Figure 3 (D-F)), we found Temporo-Parietal (TempPar) Network and Dorsal Attention Network (DAN) show the**
 435 **most pronounced improvements. Specifically, TempPar is known to support attentional reorienting,**
 436 **sensory-motor integration, and response preparation - functions that are strongly modulated by**
 437 **moment-to-moment vigilance and arousal. In the auditory task dataset, participants must detect**
 438 **auditory cues and make rapid button-press responses; thus, fluctuations in alertness directly impact**
 439 **both auditory processing efficiency and motor readiness, processes for which TempPar plays a central**
 440 **role. Likewise, DAN is among the networks most sensitive to arousal and sustained attention.** These
 441 **results support the biological plausibility of our approach and provide compelling evidence that**
 442 **incorporating vigilance context enables more accurate and interpretable EEG-to-fMRI translation,**
 443 **particularly in regions sensitive to fluctuations in arousal and attention, which is especially crucial to**
 444 **consider for resting-state data.**

446 3.4 GENERALIZATION PERFORMANCE AND ABLATION STUDIES

447 **Generalization.** We evaluate the zero-shot performance of our multi-region model on an unseen
 448 auditory task-based dataset from Li et al. (2024b) (Figure 4). Our model generalizes well to task-
 449 induced fMRI dynamics, capturing prominent brain activity features despite not being trained on
 450 task-based data. A comprehensive analysis of various rest-to-task transfer strategies, including fine-
 451 tuning, joint training and personalized-finetuning, is provided in Appendix D.2. To further validate
 452 the quality of the zero-shot generated fMRI and its ability to reflect true subject-specific patterns in
 453 FC, we additionally performed a connectome fingerprinting analysis, as described in Appendix D.3.
 454 Notably, our model’s zero-shot predicted fMRI also demonstrated high fingerprinting accuracy across
 455 full-brain, gray matter, and subcortical FC matrices (see Table 6 in Appendix). These findings suggest
 456 that the generated fMRI signals preserve individualized FC signatures, supporting their potential
 457 utility in downstream applications involving subject-specific brain representations, such as cognitive
 458 trait identification, behavioral decoding, and clinical profiling (Finn et al., 2015; Mantwill et al., 2022;
 459 Lu et al., 2024).

460
 461 Table 2: **Ablation study on model components.** Each row removes one input or loss function from the
 462 full model. **vig.:** vigilance token; **demo.:** demographic token; **d.t.:** dataset token; **w/o:** without. Paired
 463 t-test significance between our model and each ablation is shown using color codes: blue ($p < 0.05$),
 464 yellow ($p < 0.01$), red ($p < 0.001$), uncorrected.

465 Model Type	466 FB TCorr	467 GM TCorr	468 Conn PCorr	469 Conn MSE
470 Full (context + 5 d.t.)	0.367 ± 0.052	0.394 ± 0.060	0.527 ± 0.084	0.233 ± 0.072
471 w/o fine-tune	0.365 ± 0.056	0.391 ± 0.064	0.527 ± 0.066	0.267 ± 0.075
472 w/o pretrain	0.329 ± 0.033	0.361 ± 0.039	0.315 ± 0.074	0.439 ± 0.102
473 w/o vig.	0.357 ± 0.040	0.381 ± 0.044	0.532 ± 0.076	0.237 ± 0.070
474 w/o demo.	0.351 ± 0.060	0.378 ± 0.068	0.503 ± 0.081	0.271 ± 0.082
475 w/o demo. & vig.	0.327 ± 0.075	0.356 ± 0.089	0.491 ± 0.079	0.288 ± 0.089
476 w/o age	0.351 ± 0.047	0.376 ± 0.052	0.513 ± 0.088	0.261 ± 0.078
477 w/o sex	0.358 ± 0.041	0.382 ± 0.044	0.529 ± 0.079	0.235 ± 0.071
478 w/o d.t.	0.355 ± 0.049	0.380 ± 0.055	0.527 ± 0.081	0.234 ± 0.069
479 1 d.t.	0.349 ± 0.051	0.375 ± 0.059	0.494 ± 0.084	0.288 ± 0.085
480 10 d.t.	0.356 ± 0.048	0.381 ± 0.054	0.527 ± 0.081	0.236 ± 0.070
481 w/o $\mathcal{L}_{\text{align}}$	0.339 ± 0.052	0.367 ± 0.055	0.502 ± 0.082	0.280 ± 0.083
482 w/o $\mathcal{L}_{\text{recon}}$	0.147 ± 0.022	0.150 ± 0.025	0.172 ± 0.052	0.266 ± 0.042
483 w/o $\mathcal{L}_{\text{SCorr}}$	0.355 ± 0.049	0.384 ± 0.056	0.500 ± 0.083	0.293 ± 0.088
484 w/o \mathcal{L}_{MSE}	0.347 ± 0.050	0.376 ± 0.058	0.481 ± 0.085	0.317 ± 0.092

485 **Ablation Studies.** We evaluate the contribution of each component (Table 2). We first compared
 486 our full model against a variant in which the transformer-based fMRI decoder was trained from
 487 scratch, without pretraining. Results indicate that without pretraining, the model struggles to capture
 488 temporal correlations between brain regions, leading to significantly reduced performance. Next, we
 489 examined the impact of different components of the context prompt. We found that removing any of
 490 the context tokens resulted in a drop in performance, highlighting their importance. The number of
 491 dataset tokens serves as a tunable hyperparameter. We found that using five tokens was sufficient to
 492 effectively handle the two training datasets used in our experiments. Beyond these ablations, we also

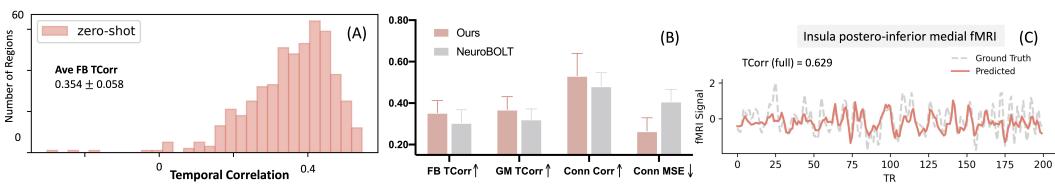


Figure 4: **Zero-shot reconstruction on task-condition data.** (A) Prediction performance distribution for all 512 regions. (B) Performance comparison with the baseline. (C) Example of reconstructed time series within the insula, shown for part of a scan.

analyzed the impact of data scale (Figure 10), mask ratio (Table 8), and patch size (Table 9) during fMRI pretraining, please see Appendix D.8, D.9, D.10 for details.

4 CONCLUSION

We introduce **UniEFS**, a unified framework for reconstructing full-brain fMRI activity from EEG. By leveraging large-scale fMRI pretraining, followed by domain adaptation and cross-modal alignment, our model effectively bridges the spatial-temporal gap between EEG and fMRI without relying on region-specific supervision or subject-dependent customization. We incorporate context-aware EEG encoding using metadata-informed prompt tokens, enabling the model to account for physiological and demographic variability that modulates EEG-fMRI correspondence. Our results demonstrate that UniEFS achieves state-of-the-art performance in time-resolved fMRI signal reconstruction across hundreds of brain regions, and its potential to recover functional connectivity. Our results highlight the effectiveness of combining self-supervised fMRI representation learning with context-conditioned EEG encoding for generalizable, scalable, context-aware, and interpretable cross-modality translation. UniEFS paves the way for real-world applications where high-resolution, fMRI-like insights could be derived from lightweight, portable EEG systems, enabling more accessible neuroimaging in clinical, cognitive, and mobile settings.

ETHICS STATEMENT

All study procedures were conducted with approval from the Institutional Review Board (IRB), and the research is classified as posing minimal risk to participants since both EEG and fMRI are non-invasive neuroimaging modalities. Individuals with contraindications to MRI were precluded from study participation. During our data collection, participants rested passively with their eyes closed and received compensation for their participation. Written informed consent was obtained from all participants prior to enrollment. The protocols explicitly outlined the nature of the resting-state scanning, any possible discomfort, and the participants' right to withdraw at any time without penalty. We confirm that no vulnerable populations were targeted or exploited, and the study did not involve any experimental manipulations beyond routine neuroimaging procedures. Collected data were anonymized and handled in compliance with appropriate privacy and confidentiality safeguards to minimize the risk of re-identification or misuse.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, we have made our experimental setup as transparent and accessible as possible. Key model architectures, training protocols, and evaluation metrics are described in detail in Sections 2 and 3 of the main text. Additional implementation details and hyperparameter settings are provided in the Appendix B and C. Upon acceptance, we will release our complete source code, pretrained model weights, and the in-house EEG-fMRI dataset (i.e., Dataset 2) used in this study under an appropriate data-sharing agreement. This will allow the community to fully reproduce and build upon our results.

540 REFERENCES
541

542 Philip J Allen, Oliver Josephs, and Robert Turner. A method for removing imaging artifact from
543 continuous eeg recorded during functional mri. *Neuroimage*, 12(2):230–239, 2000.

544 Elena Bondi, Yidan Ding, Yisha Zhang, Eleonora Maggioni, and Bin He. Investigating the neu-
545 rovascular coupling across multiple motor execution and imagery conditions: a whole-brain
546 eeg-informed fmri analysis. *NeuroImage*, pp. 121311, 2025.

547 Susan Y Bookheimer, David H Salat, Melissa Terpstra, Beau M Ances, Deanna M Barch, Randy L
548 Buckner, Gregory C Burgess, Sandra W Curtiss, Mirella Diaz-Santos, Jennifer Stine Elam, et al.
549 The lifespan human connectome project in aging: an overview. *Neuroimage*, 185:335–348, 2019.

550 David Calhas and Rui Henriques. Eeg to fmri synthesis benefits from attentional graphs of electrode
551 relationships. *arXiv preprint arXiv:2203.03481*, 2022.

552 Josue Ortega Caro, Antonio Henrique de Oliveira Fonseca, Syed A Rizvi, Matteo Rosati, Christopher
553 Averill, James L Cross, Prateek Mittal, Emanuele Zappala, Rahul Madhav Dhadapkar, Chadi
554 Abdallah, and David van Dijk. BrainLM: A foundation model for brain activity recordings.
555 In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=RwI7ZEfr27>.

556 JF Cavanagh. Eeg: 3-stim auditory oddball and rest in parkinson’s. *OpenNeuro, OpenNeuro*, 2021.

557 Catie Chang and Jingyuan E Chen. Multimodal eeg-fmri: advancing insight into large-scale human
558 brain dynamics. *Current opinion in biomedical engineering*, 18:100279, 2021.

559 Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain:
560 Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of
561 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22710–22720, 2023.

562 Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic mindscapes: High-quality video recon-
563 struction from brain activity. *Advances in Neural Information Processing Systems*, 36, 2024.

564 Michael X Cohen. Where does eeg come from and what does it mean? *Trends in neurosciences*, 40
565 (4):208–218, 2017.

566 Kamalaker Dadi, Gaël Varoquaux, Antonia Machlouzarides-Shalit, Krzysztof J Gorgolewski, Demian
567 Wassermann, Bertrand Thirion, and Arthur Mensch. Fine-grain atlases of functional modes for
568 fmri analysis. *NeuroImage*, 221:117126, 2020.

569 Zijian Dong, Li Rulin, Yilei Wu, Thuan Tinh Nguyen, Joanna Su Xian Chong, Fang Ji, Nathanael
570 Ren Jie Tong, Christopher Li Hsian Chen, and Juan Helen Zhou. Brain-JEPA: Brain dynam-
571 ics foundation model with gradient positioning and spatiotemporal masking. In *The Thirty-
572 eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=gtU2eLSAm0>.

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

577 Chiara Favaretto, Michele Allegra, Gustavo Deco, Nicholas V Metcalf, Joseph C Griffis, Gordon L
578 Shulman, Andrea Brovelli, and Maurizio Corbetta. Subcortical-cortical dynamical states of the
579 human brain and their breakdown in stroke. *Nature communications*, 13(1):5069, 2022.

580 Emily S Finn, Xilin Shen, Dustin Scheinost, Monica D Rosenberg, Jessica Huang, Marvin M Chun,
581 Xenophon Papademetris, and R Todd Constable. Functional connectome fingerprinting: identifying
582 individuals using patterns of brain connectivity. *Nature neuroscience*, 18(11):1664–1671, 2015.

583 Shanghua Gao, Teddy Koker, Owen Queen, Tom Hartvigsen, Theodoros Tsilgkaridis, and Marinka
584 Zitnik. Units: A unified multi-task time series model. *Advances in Neural Information Processing
585 Systems*, 37:140589–140631, 2024.

594 Sairam Geethanath and John Thomas Vaughan Jr. Accessible magnetic resonance imaging: a review.
 595 *Journal of Magnetic Resonance Imaging*, 49(7):e65–e77, 2019.
 596

597 Matthew F Glasser, Stamatios N Sotiroopoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl,
 598 Jesper L Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R Polimeni, et al. The
 599 minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124,
 600 2013.

601 Sarah E Goodale, Nafis Ahmed, Chong Zhao, Jacco A de Zwart, Pinar S Özbay, Dante Picchioni,
 602 Jeff Duyn, Dario J Englot, Victoria L Morgan, and Catie Chang. fmri-based detection of alertness
 603 predicts behavioral response variability. *elife*, 10:e62376, 2021.

604

605 Michael P Harms, Leah H Somerville, Beau M Ances, Jesper Andersson, Deanna M Barch, Mat-
 606 teo Bastiani, Susan Y Bookheimer, Timothy B Brown, Randy L Buckner, Gregory C Burgess,
 607 et al. Extending the human connectome project across ages: Imaging protocols for the lifespan
 608 development and aging projects. *Neuroimage*, 183:972–984, 2018.

609

610 Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked
 611 autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer
 612 vision and pattern recognition*, pp. 16000–16009, 2022.

613

614 Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang,
 615 Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

616

617 Jue Huang, Christian Sander, Philippe Jawinski, Christine Ulke, Janek Spada, Ulrich Hegerl, and
 618 Tilman Hensch. Test-retest reliability of brain arousal regulation as assessed with vigall 2.0.
*619 *Neuropsychiatric Electrophysiology**, 1:1–13, 2015.

620

621 Mohammad Jalloul, Monica Miranda-Schaeubinger, Abass M Noor, Joel M Stein, Raisa Amiruddin,
 622 Hermon Miliard Derbew, Victoria L Mango, Adeyanju Akinola, Kelly Hart, Joseph Weygand, et al.
 623 Mri scarcity in low-and middle-income countries. *NMR in Biomedicine*, 36(12):e5022, 2023.

624

625 Philippe Jawinski, Holger Kirsten, Christian Sander, Janek Spada, Christine Ulke, Jue Huang, Ralph
 626 Burkhardt, Markus Scholz, Tilman Hensch, and Ulrich Hegerl. Human brain arousal in the resting
 627 state: a genome-wide association study. *Molecular psychiatry*, 24(11):1599–1609, 2019.

628

629 Weibang Jiang, Liming Zhao, and Bao liang Lu. Large brain model for learning generic representa-
 630 tions with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning
 631 Representations*, 2024. URL <https://openreview.net/forum?id=QzTpTRVtrP>.

632

633 Daisuke Koshiyama, Masaki Fukunaga, Naohiro Okada, Fumio Yamashita, Hidenaga Yamamori,
 634 Yuka Yasuda, Michiko Fujimoto, Kazutaka Ohi, Haruo Fujino, Yoshiyuki Watanabe, et al. Role of
 635 subcortical structures on cognitive and social function in schizophrenia. *Scientific reports*, 8(1):
 636 1183, 2018.

637

638 Alexander Kovalev, Ilia Mikheev, and Alexei Ossadtchi. fmri from eeg is only deep learning away:
 639 the use of interpretable dl to unravel eeg-fmri relationships. *arXiv preprint arXiv:2211.02024*,
 640 2022.

641

642 Romeo Lanzino, Federico Fontana, Luigi Cinque, Francesco Scarcello, and Atsuto Maki. Nt-vit:
 643 Neural transcoding vision transformers for eeg-to-fmri synthesis. *arXiv preprint arXiv:2409.11836*,
 644 2024.

645

646 Hongli Li, Man Ding, Ronghua Zhang, and Chunbo Xiu. Motor imagery eeg classification algorithm
 647 based on cnn-lstm feature fusion network. *Biomedical signal processing and control*, 72:103342,
 648 2022.

649

650 Yamin Li, Ange Lou, Ziyuan Xu, Shiyu Wang, and Catie Chang. Leveraging sinusoidal representation
 651 networks to predict fmri signals from eeg. In *Medical Imaging 2024: Image Processing*, volume
 652 12926, pp. 795–800. SPIE, 2024a.

648 Yamin Li, Ange Lou, Ziyuan Xu, Shengchao Zhang, Shiyu Wang, Dario Englot, Soheil Kolouri,
 649 Daniel Moyer, Roza Bayrak, and Catie Chang. Neurobolt: Resting-state eeg-to-fmri synthesis
 650 with multi-dimensional feature mapping. *Advances in Neural Information Processing Systems*, 37:
 651 23378–23405, 2024b.

652 Thomas T Liu. Noise contributions to the fmri signal: An overview. *NeuroImage*, 143:141–151,
 653 2016.

654 Thomas T Liu and Maryam Falahpour. Vigilance effects in resting-state fmri. *Frontiers in neuro-
 655 science*, 14:511047, 2020.

656 Xiao Liu, Catie Chang, and Jeff H Duyn. Decomposition of spontaneous brain activity into distinct
 657 fmri co-activation patterns. *Frontiers in systems neuroscience*, 7:62295, 2013.

658 Xiao Liu, Nanyin Zhang, Catie Chang, and Jeff H Duyn. Co-activation patterns in resting-state fmri
 659 signals. *Neuroimage*, 180:485–494, 2018.

660 Xueqing Liu and Paul Sajda. Fusing simultaneously acquired eeg and fmri via hierarchical deep
 661 transcoding. In *International Conference on Brain Informatics*, pp. 57–67. Springer, 2023a.

662 Xueqing Liu and Paul Sajda. Latent neural source recovery via transcoding of simultaneous eeg-fmri.
 663 In *International Conference on Brain Informatics*, pp. 318–330. Springer, 2023b.

664 Nikos K Logothetis. What we can do and what we cannot do with fmri. *Nature*, 453(7197):869–878,
 665 2008.

666 Jiayu Lu, Tianyi Yan, Lan Yang, Xi Zhang, Jiaxin Li, Dandan Li, Jie Xiang, and Bin Wang. Brain
 667 fingerprinting and cognitive behavior predicting using functional connectome of high inter-subject
 668 variability. *NeuroImage*, 295:120651, 2024.

669 Maron Mantwill, Martin Gell, Stephan Krohn, and Carsten Finke. Brain connectivity fingerprinting
 670 and behavioural prediction rest on distinct functional systems of the human connectome.
 671 *Communications biology*, 5(1):261, 2022.

672 Caroline G Martin, Biyu J He, and Catie Chang. State-related neural influences on fmri connectivity
 673 estimation. *NeuroImage*, 244:118590, 2021.

674 Paul M Matthews, Garry D Honey, and Edward T Bullmore. Applications of fmri in translational
 675 medicine and clinical practice. *Nature Reviews Neuroscience*, 7(9):732–744, 2006.

676 Yehudit Meir-Hasson, Sivan Kinreich, Ilana Podlipsky, Talma Hendler, and Nathan Intrator. An eeg
 677 finger-print of fmri deep regional activation. *Neuroimage*, 102:128–141, 2014.

678 Thomas M Moehlman, Jacco A de Zwart, Miranda G Chappel-Farley, Xiao Liu, Irene B McClain,
 679 Catie Chang, Hendrik Mandelkow, Pinar S Özbay, Nicholas L Johnson, Rebecca E Bieber, et al.
 680 All-night functional magnetic resonance imaging sleep studies. *Journal of neuroscience methods*,
 681 316:83–98, 2019.

682 Luis Fernando Nicolas-Alonso and Jaime Gomez-Gil. Brain computer interfaces, a review. *sensors*,
 683 12(2):1211–1279, 2012.

684 Sebastian Olbrich, Christoph Mulert, Susanne Karch, Maja Trenner, Gregor Leicht, Oliver Pogarell,
 685 and Ulrich Hegerl. Eeg-vigilance and bold effect during simultaneous eeg/fmri measurement.
 686 *neuroimage*, 45(2):319–332, 2009.

687 Sebastian Olbrich, Marie M Fischer, Christian Sander, Ulrich Hegerl, Hubert Wirtz, and Andrea
 688 Bosse-Henck. Objective markers for sleep propensity: comparison between the multiple sleep
 689 latency test and the vigilance algorithm leipzig. *Journal of sleep research*, 24(4):450–457, 2015.

690 Ayelet Or-Borichev, Guy Gurevitch, Ilana Klovatch, Ayam Greental, Yulia Lerner, Dino J Levy, and
 691 Talma Hendler. Neural and functional validation of fmri-informed eeg model of right inferior
 692 frontal gyrus activity. *NeuroImage*, 266:119822, 2023.

702 Wei Yan Peh, Yuanyuan Yao, and Justin Dauwels. Transformer convolutional neural networks for
 703 automated artifact detection in scalp eeg. In *2022 44th Annual International Conference of the*
 704 *IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3599–3602. IEEE, 2022.

705 Jonathan D Power, Mark Plitt, Timothy O Laumann, and Alex Martin. Sources and implications of
 706 whole-brain fmri signals in humans. *Neuroimage*, 146:609–625, 2017.

708 Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language
 709 models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

710 Max Schneider, Pamela Hathway, Laura Leuchs, Philipp G Sämann, Michael Czisch, and Victor I
 711 Spoormaker. Spontaneous pupil dilations during the resting state are associated with activation of
 712 the salience network. *NeuroImage*, 139:189–201, 2016.

714 Gordon MG Shepherd. Corticostriatal connectivity and its role in disease. *Nature Reviews Neuro-
 715 science*, 14(4):278–291, 2013.

717 Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie. Transformer-based spatial-temporal feature
 718 learning for eeg decoding. *arXiv preprint arXiv:2106.11170*, 2021.

719 Armin Thomas, Christopher Ré, and Russell Poldrack. Self-supervised learning of brain dynamics
 720 from broad neuroimaging data. *Advances in neural information processing systems*, 35:21255–
 721 21269, 2022.

722 Shobao Tong and Nitish V Thankor. *Quantitative EEG analysis methods and clinical applications*.
 723 Artech House, 2009.

725 Edwin JR van Beek, Christiane Kuhl, Yoshimi Anzai, Patricia Desmond, Richard L Ehman, Qiyong
 726 Gong, Garry Gold, Vikas Gulani, Margaret Hall-Craggs, Tim Leiner, et al. Value of mri in medicine:
 727 More than just another test? *Journal of Magnetic Resonance Imaging*, 49(7):e14–e25, 2019.

728 David C Van Essen, Kamil Ugurbil, Edward Auerbach, Deanna Barch, Timothy EJ Behrens, Richard
 729 Bucholz, Acer Chang, Liyong Chen, Maurizio Corbetta, Sandra W Curtiss, et al. The human
 730 connectome project: a data acquisition perspective. *Neuroimage*, 62(4):2222–2231, 2012.

732 Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang
 733 Pan. CBramod: A criss-cross brain foundation model for EEG decoding. In *The Thirteenth
 734 International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=NPNUHgHF2w>.

736 Huilin Wei, Amirhossein Jafarian, Peter Zeidman, Vladimir Litvak, Adeel Razi, Dewen Hu, and
 737 Karl J Friston. Bayesian fusion and multimodal dcm for eeg and fmri. *Neuroimage*, 211:116595,
 738 2020.

740 Chi Wah Wong, Valur Olafsson, Omer Tal, and Thomas T Liu. The amplitude of the resting-state
 741 fmri global signal is related to eeg vigilance measures. *Neuroimage*, 83:983–990, 2013.

742 Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu.
 743 Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF
 744 conference on computer vision and pattern recognition*, pp. 9653–9663, 2022.

746 Alba Xifra-Porras, Michalis Kassinosopoulos, and Georgios D Mitsis. Physiological and motion
 747 signatures in static and time-varying functional connectivity and their subject identifiability. *Elife*,
 748 10:e62324, 2021.

749 Chaoqi Yang, M Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data
 750 learning in the wild. In *Thirty-seventh Conference on Neural Information Processing Systems*,
 751 2023. URL <https://openreview.net/forum?id=c2LZyTyddi>.

752 Weiheng Yao, Zhihan Lyu, Mufti Mahmud, Ning Zhong, Baiying Lei, and Shuqiang Wang. Catd:
 753 Unified representation learning for eeg-to-fmri cross-modal generation. *IEEE Transactions on
 754 Medical Imaging*, 2025.

756 A RELATED WORK
757

758 EEG-to-fMRI reconstruction, while currently an underexplored research area, has received growing
759 attention in recent years owing to advances in deep learning and cross-modality synthesis. Several
760 studies (Liu & Sajda, 2023a;b; Calhas & Henriques, 2022; Lanzino et al., 2024) have proposed
761 methods for reconstructing volume-wise fMRI spatial patterns from EEG signals. However, these
762 approaches generally lack quantitative evaluation of temporal dynamics, that is, how brain activity
763 evolves over time, or of temporal correlation across brain regions, which are critical readouts, under-
764 pinning analyses of functional connectivity, network dynamics, and brain state transitions. Without
765 assessing these aspects, it remains unclear whether such models can support broader neuroscientific
766 or clinical applications that rely on accurate reconstruction of brain-wide temporal structure. A
767 complementary line of work has investigated fMRI time series reconstruction in specific brain regions,
768 particularly in subcortical regions, such as early work using ridge regression (Meir-Hasson et al.,
769 2014; Or-Borichev et al., 2023), along with more recent deep-learning studies using seq-to-seq models
770 (Kovalev et al., 2022; Li et al., 2024a). Yet, these models are typically trained on a within-subject
771 basis, which hinders generalizability to new individuals. Furthermore, the majority of EEG-fMRI
772 synthesis efforts have been limited to task-based paradigms, where external cues help structure the
773 neural responses (Kovalev et al., 2022; Li et al., 2024a; Liu & Sajda, 2023a;b; Wei et al., 2020). As a
774 result, spontaneous resting-state activity, particularly in the natural eyes-closed condition, remains
775 largely unexplored. To bridge this gap, recent work Li et al. (2024b) introduced a transformer-based
776 framework for reconstructing fMRI time series in a few selected brain regions during eye-closed
777 resting-state. While this method shows promising generalization, it still requires training separate
778 models for each target region, limiting scalability and efficiency. A more recent work by Yao et al.
779 (2025) demonstrated efficient cortical surface fMRI generation by conditioning a diffusion model
780 on EEG. However, its reliance on fMRI surface maps restricts reconstruction to the cortex, leaving
781 subcortical regions outside the model’s representational space.

782 B DATASET AND PREPROCESSING
783

784 **Pretraining fMRI Dataset: Preprocessing** The HCP dataset (here, used for pretraining) had been
785 processed using the HCP Minimal Preprocessing Pipeline (Glasser et al., 2013). In addition to this
786 standard preprocessing, we removed low-order trends (polynomials up to the 4th order) to mitigate
787 scanner drift artifacts, and temporally downsampled the data by a factor of 2, resulting in a final
788 temporal resolution of 1.44 seconds and 600 frames per scan. This step was performed to render the
789 temporal resolution more comparable to conventional fMRI scans, including those of the EEG-fMRI
790 datasets used in our study. After extracting time courses from regions of interest using DiFuMo atlas
791 (Dadi et al., 2020), we additionally regress out six rigid-body head-motion parameters (translation
792 and rotation), apply a low-pass filter with a cutoff at 0.15 Hz (which captures the low-frequency band
793 typically of interest in rs-fMRI studies), and z-normalize each ROI time series.

794 **Details about paired EEG-fMRI Dataset 2** Dataset2 is an in-house resting-state EEG-fMRI
795 dataset. It comprises 10 scans from 7 healthy participants, with 3 individuals undergoing two scans
796 each. During the scans, participants rested passively with their eyes closed. Written informed consent
797 was obtained from all participants, and all procedures were approved by the Institutional Review
798 Board. MRI data were acquired on a 3T Siemens Prisma scanner. The T1-weighted structural images
799 were collected with the following parameters: TR = 2200 ms, TE = 4.25 ms, flip angle = 9 deg, 1
800 mm isotropic. BOLD fMRI images were collected using multi-echo gradient-echo EPI sequence
801 with TR = 2100 ms, echo times = 13.0, 29.4, and 45.7 ms, voxel size = $3 \times 3 \times 3 \text{ mm}^3$, slice gap
802 = 1 mm, matrix size = 82×50 , 30 axial slices. MRI scanner triggers were recorded together with
803 the EEG signals for data synchronization. The first seven volumes in fMRI data were dropped to
804 allow magnetization to reach steady state. The fMRI preprocessing steps are kept consistent with Li
805 et al. (2024b). Specifically, the steps included slice-timing and motion coregistration, noise reduction
806 using multi-echo ICA which is implemented in tedana 0.0.9a², alignment to an MNI152 standard
807 template (matrix shape: $91 \times 109 \times 91$), removal of low-order trends (up to 4th-order polynomials),
808 and spatial smoothing (to 3mm FWHM) using AFNI³. Simultaneous scalp EEG was acquired using

809 ²<https://tedana.readthedocs.io/en/stable/>

³<https://afni.nimh.nih.gov/afni>

a 32-channel MR-compatible system (10–20 layout, FCz reference; BrainAmps MR, Brain Products GmbH) at a sampling rate of 5 kHz. The EEG system was synchronized to the scanner’s 10 MHz clock to support gradient artifact correction. Preprocessing included removal of MR-related artifacts using BrainVision Analyzer 2 (Brain Products, Munich, Germany) (Moehlman et al., 2019; Allen et al., 2000), followed by downsampling to 250 Hz. No additional filtering was applied. The full set of 32 EEG channel labels is: ['FP1', 'FP2', 'F3', 'F4', 'C3', 'C4', 'P3', 'P4', 'O1', 'O2', 'F7', 'F8', 'T7', 'T8', 'P7', 'P8', 'FZ', 'CZ', 'PZ', 'OZ', 'FC1', 'FC2', 'CP1', 'CP2', 'FC5', 'FC6', 'CP5', 'CP6', 'TP9', 'TP10', 'POZ']. For joint training across Dataset 1 (Li et al., 2024b) and Dataset 2, we used the intersection of their channel sets, resulting in 23 overlapping channels: ['FP1', 'FP2', 'F3', 'F4', 'C3', 'C4', 'P3', 'P4', 'O1', 'O2', 'F7', 'F8', 'T7', 'T8', 'P7', 'P8', 'FZ', 'CZ', 'PZ', 'OZ', 'TP9', 'TP10', 'POZ'].

Auditory-task Dataset In Section 3.4, we evaluate whether our model, trained on resting-state data, can generalize to a different domain without additional training. To this end, we use only the test set from the auditory-task EEG-fMRI dataset from Li et al. (2024b). During the scans, binaural tones were presented with randomized inter-stimulus intervals (ISI), and the task included two versions differing only in tone timing: (1) a fast-ISI version (500 TRs per scan) and (2) a sparse-ISI version (693 TRs per scan). The test set comprises four scans. Two scans correspond to the fast-ISI version, and the other two to the sparse-ISI version. For additional details, please refer to Li et al. (2024b).

Vigilance States The vigilance state is represented as a categorical label with three classes—drowsy, intermediate, and alert—assigned to each fMRI frame. To derive these vigilance classes, we employed the automated VIGALL method (Huang et al., 2015; Jawinski et al., 2019; Olbrich et al., 2015), which classifies scalp EEG segments into five vigilance stages based on spatial power distributions. Specifically, we used the VIGALL 2.1 add-on in BrainVision Analyzer 2 to segment the preprocessed EEG into non-overlapping 1-second intervals and label each interval as one of five stages: A1, A2, A3, B1, or B2/3, corresponding to decreasing levels of alertness. Prior to staging, EEG signals were re-referenced to the average, and spherical spline interpolation was applied to reconstruct any missing channels required by the VIGALL standard. These vigilance stage labels are then grouped into 63-second epochs (corresponding to 30 fMRI time points) and the distribution of stages within each epoch is used to assign a single vigilance class, i.e., alert, intermediate, or drowsy, to that epoch, according to the following rules: (1) First, the five VIGALL stages were converted to integer values from 1 (most drowsy) to 5 (most alert); (2) The Wilcoxon signed-rank test was then applied to the integer values of each epoch to test for a significant difference of the median away from a (weighted) center value of 2.75; (3) Based on the test statistic, we assigned each epoch to one of the three vigilance classes using a z-threshold of ± 1.5 : epochs with significantly high or low median vigilance were labeled as alert or drowsy, respectively, while others were classified as intermediate; (4) Finally, consecutive epochs with the same vigilance label were merged to form continuous segments.

C MORE IMPLEMENTATION DETAILS

C.1 HYPERPARAMETERS

The default hyperparameters for the full pretraining model architecture are summarized in Table 3. The fMRI masked signal modeling (f-MSM) model is pretrained for 225 K iterations. The checkpoint achieving the highest spatial correlation between predicted and ground-truth signals on the validation set is selected as the final pretrained model, which is further fine-tuned for 20 epochs. The training hyperparameters of EEG-to-fMRI mapping are shown in Table 4.

Other Implementation Details We initialize EEG encoder’s spatiotemporal module using pre-trained weights from LaBraM-base (Jiang et al., 2024), with a token length of 200 (i.e., 1 second) and no overlap. For the multi-scale spectral module, we set the smallest scale size to $l_0 = 200$ (1 seconds without overlap), and use a multiscale level of 3. The functional connectivity metrics, i.e., Conn PCorr and Conn MSE, are calculated using the upper triangle of the correlation matrices, as they are symmetric.

864

865

Table 3: Hyperparameters used for stage 1: f-MSM pretraining and finetuning

866

867

Hyperparameters	Values
batch size	pretrain: 96, fine-tune: 16
learning rate	pretrain: 3e-4, fine-tune: 5.3e-5
weight decay	0.05
Optimizer	AdamW
patch size	1
encoder embedded dim	512
mask ratio	0.5
mlp ratio	2.0
decoder embedded dim	256
encoder depth	12
encoder heads	8
decoder depth	8
decoder heads	8

879

880

881

Table 4: Hyperparameters for stage 2: EEG-to-fMRI mapping

882

883

Hyperparameters	Values
Batch size	64
Peak learning rate	3e-4
Minimal learning rate	1e-6
Learning rate scheduler	Cosine
Optimizer	AdamW
Adam β	(0.9,0.99)
Weight decay	0.05
Drop path	0.1
Layer-wise learning rate decay	0.65
λ_{MSE}	0.5
λ_{align}	0.8
λ_{recon}	0.2

895

896

897

C.2 BASELINES

898

899

900

901

We compare our model against three publicly available and adaptable EEG-to-fMRI translation frameworks, all of which have been benchmarked in Li et al. (2024b). These are the only open-source methods compatible with the datasets and experimental setup in this study and Li et al. (2024b).

902

903

904

905

- **BEIRA** (Kovalev et al., 2022): BEIRA introduces a convolutional neural network (CNN)-based encoder-decoder architecture that translates EEG sequences into corresponding fMRI sequences in a sequence-to-sequence manner.
- **Li et al.** (Li et al., 2024a): This method extends BEIRA by incorporating an additional light-weight spectral representation learning module that leveraging sinusoidal activation function to better capture the frequency characteristics of EEG signals. It uses CNN-based downsampling and upsampling encoder-decoder blocks to perform the translation from EEG to fMRI during an eyes-open-eyes-closed task.
- **NeuroBOLT** (Li et al., 2024b): NeuroBOLT proposes a transformer-based multi-dimensional encoder for EEG-to-fMRI mapping in a seq-to-one format. It is a region-specific model, which means that models are trained separately for each region. It achieved state-of-the-art prediction performance in their resting-state dataset.

914

915

916

917

Among these baselines, the models by BEIRA and Li et al. were originally designed in a sequence-to-sequence format, where both the input and output are time series. To account for the hemodynamic delay of fMRI relative to EEG, the EEG sequence was temporally shifted by 6 seconds, i.e., the input EEG was delayed by 6 seconds to align with the corresponding fMRI response. For now

918 we do not include CATD (Yao et al., 2025) as a baseline, since CATD operates at the surface-map
 919 level restricted to the cortex, whereas our study focuses on ROI-based reconstruction covering the
 920 whole brain, including both cortical and subcortical regions, making the two settings not directly
 921 comparable. Moreover, since the implementation of CATD has not been publicly released, accurate
 922 reproduction is not currently feasible, which would preclude a fair comparison.

923 We also compare our model performance with state-of-the-art EEG encoders, and results are shown
 924 in D.1.

- 926 • **CBraMod** (Wang et al., 2025): CBraMod is a recent foundation model for EEG that
 927 follows the design of prior EEG foundation models by segmenting EEG signals into patches
 928 and pre-training via masked patch reconstruction. Building on this framework, CBraMod
 929 introduces two key innovations: (1) a criss-cross transformer backbone with parallel spatial
 930 and temporal attention mechanisms to separately capture heterogeneous dependencies in
 931 EEG, and (2) an asymmetric conditional positional encoding scheme that enables flexible
 932 adaptation to diverse EEG formats. Pre-trained on a large-scale EEG corpus, CBraMod
 933 outperforms state-of-the-art methods and demonstrates strong generalizability across up to
 934 10 downstream BCI tasks (12 public datasets). In our experiments, we initialize the model
 935 with these pre-trained weights to provide a warm start for the EEG-to-fMRI translation task.
- 936 • **LaBraM** (Jiang et al., 2024): LaBraM (Large Brain Model) is a unified foundation model for
 937 EEG that enables cross-dataset learning by segmenting EEG signals into channel patches and
 938 using vector-quantized neural spectrum prediction to encode them into compact neural codes.
 939 Pre-trained on 2,500 hours of EEG data from 20 datasets, LaBraM achieves state-of-the-art
 940 performance in various downstream tasks such as abnormal detection, event classification,
 941 emotion recognition, and gait prediction. In our experiment, we load the pre-trained weights
 942 as initialization (version: LaBraM-base).
- 943 • **BIOT** (Yang et al., 2023): BIOT is a transformer-based foundational architecture for
 944 biomedical signal encoding. It segments EEG signals into patches and learns spatiotemporal
 945 and spectral representations from EEG, which can be applied to various downstream tasks.
- 946 • **CNNTransformer** (Peh et al., 2022): CNNTransformer is a transformer convolutional
 947 neural network originally designed for automated artifact detection in EEG.
- 948 • **STTransformer** (Song et al., 2021): STTransformer is a transformer-based spatial-temporal
 949 feature learning neural network originally designed for EEG decoding.
- 950 • **FFCL** (Li et al., 2022): FFCL is a model combining learned latent features from CNN and
 951 LSTM models for the purpose of motor imagery EEG classification.

954 In the NeuroBOLT experiments Li et al. (2024b), the authors adapted all baselines to a sequence-to-
 955 one format for evaluation. Following this approach, we apply the same adaptation and further attach
 956 a shared multi-ROI MLP decoder to each EEG encoder, enabling a single model to predict the full
 957 set of ROI signals for fair comparison. **Despite using pretrained EEG foundation model encoders,**
 958 **all baseline models are retrained from scratch to convergence under the same training protocol as**
 959 **ours (including optimizer, batch size, number of epochs, learning-rate schedule, and early stopping**
 960 **criteria). All baselines reached stable convergence under this standardized setup.**

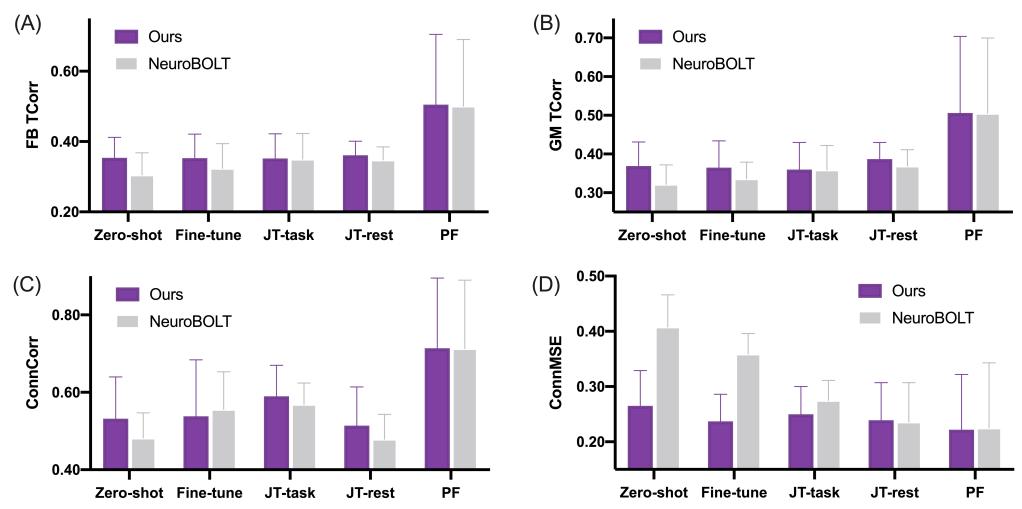
963 D ADDITIONAL RESULTS

965 D.1 PREDICTION PERFORMANCE OF OTHER EEG ENCODING BASELINES

967 In this section, we compare the performance of our model with that of other state-of-the-art EEG
 968 encoders, which were originally developed either as general-purpose EEG foundation models (Yang
 969 et al., 2023; Jiang et al., 2024; Wang et al., 2025) or were specifically designed for EEG decoding tasks
 970 (Song et al., 2021; Peh et al., 2022; Li et al., 2022). For a fair comparison, we attach a multi-region
 971 prediction head to each encoder to decode the full vector of fMRI ROIs. The results are reported in
 Table 5, where our model achieves superior performance on the majority of evaluation metrics.

972
973 Table 5: Full brain fMRI reconstruction: comparison with EEG encoding baselines. FB: Full brain;
974 GM: Cortical gray matter; SC: Sub-cortical regions; CB: Cerebellum; Conn: Metric is applied on
975 the upper triangle of the full-brain functional connectivity (FC) matrix; TCorr: Temporal correlation
976 between predicted and ground truth fMRI signals; PCorr: Pixel-wise correlation between predicted
977 and measured FC. **Bold**: the best; Underlined: the second best. Values are shown as mean \pm std.

Model Name	FB TCorr \uparrow	GM TCorr \uparrow	SC TCorr \uparrow	CB TCorr \uparrow	Conn PCorr \uparrow	Conn MSE \downarrow
Ours	0.367 ± 0.052	0.394 ± 0.060	0.276 ± 0.082	0.247 ± 0.060	0.527 ± 0.084	0.233 ± 0.072
CBraMod (Wang et al., 2025)	0.349 ± 0.033	0.369 ± 0.035	0.276 ± 0.081	0.257 ± 0.059	0.520 ± 0.062	0.234 ± 0.072
LaBraM-base (Jiang et al., 2024)	0.288 ± 0.041	0.320 ± 0.049	0.207 ± 0.082	0.161 ± 0.037	0.432 ± 0.069	0.401 ± 0.090
BIOT (Yang et al., 2023)	0.318 ± 0.068	0.353 ± 0.075	0.211 ± 0.101	0.175 ± 0.053	0.489 ± 0.077	0.299 ± 0.089
CNNTransformer (Peh et al., 2022)	0.319 ± 0.085	0.346 ± 0.092	0.242 ± 0.083	0.197 ± 0.081	0.518 ± 0.078	0.281 ± 0.070
STTransformer (Song et al., 2021)	0.091 ± 0.062	0.106 ± 0.074	0.052 ± 0.036	0.048 ± 0.015	0.436 ± 0.075	0.326 ± 0.033
FFCL (Li et al., 2022)	0.298 ± 0.034	0.321 ± 0.030	0.220 ± 0.074	0.194 ± 0.048	0.471 ± 0.075	0.319 ± 0.085



1002 Figure 5: **Comparison across rest-to-task transfer and training strategies.** Performance com-
1003 parison between our model and the state-of-the-art EEG-to-fMRI translation baseline, NeuroBOLT,
1004 under various evaluation setups: zero-shot transfer, fine-tuning, joint training on both rest and task
1005 scans with testing on task scans (JT-task), joint training and testing on resting-state scans (JT-rest),
1006 and personalized fine-tuning on individual task scans using a model pretrained on resting-state data
1007 (PF). (A) Full-brain temporal correlation (FB TCorr); (B) Gray matter temporal correlation (GM
1008 TCorr); (C) Spatial correlation of predicted and ground-truth functional connectivity (ConnCorr); (D)
1009 MSE of connectivity strength between real and reconstructed FC matrices (ConnMSE).

D.2 RESTING-STATE TO TASK-CONDITION GENERALIZATION

1013 In this section, we include a more detailed evaluation of generalization from resting-state EEG-fMRI
1014 to task-based EEG-fMRI using the auditory task dataset in Li et al. (2024b). We followed the same
1015 train-test split, resulting in 9 scans for training, 3 for validation, and 4 for testing. The results
1016 are summarized below in Figure 5 and benchmarked against the current state-of-the-art method,
1017 NeuroBOLT.

1018 Specifically, we considered four experimental settings: **(1) Zero-shot generalization:** The model is
1019 trained only on resting-state data and directly evaluated on task fMRI without any further training.
1020 **(2) Fine-tuning:** The model is pretrained on resting-state data and then fine-tuned on task data. **(3)**
1021 **Joint training:** The model is trained on a mixture of resting-state and task data. **(4) Personalized**
1022 **fine-tuning:** Starting from the resting-state pretrained model, we fine-tune individually for each test
1023 scan in the task dataset using 80% of the scan for fine-tuning, 10% for validation, and 10% for testing.
1024 As shown in Figure 5, our model shows strong generalization from resting-state to task-based fMRI
1025 in the zero-shot setting. Fine-tuning improves performance slightly, mainly in the FC reconstruction
part. Joint training helps task-fMRI FC reconstruction, but not necessarily for resting-state, which

1026 might be due to already richer variability of brain dynamics in the resting state. Although both models
 1027 perform similarly in personalized fine-tuning, overall our method still performs better in most metrics
 1028 especially in full-scan reconstruction scenario.

1030 D.3 CONNECTOME FINGERPRINTING VALIDATION OF ZERO-SHOT FC RECONSTRUCTION

1033 Table 6: Connectome fingerprinting accuracy across brain regions using ground-truth and zero-shot
 1034 predicted fMRI. Our model’s zero-shot outputs preserve subject-specific FC signatures. FB: full-
 1035 brain; GM: gray matter; SC: subcortical regions; FC: functional connectivity; Acc: fingerprinting
 1036 accuracy.

1037 Model	1038 FB-FC Acc	1039 GM-FC Acc	1040 SC-FC Acc
1038 Ground-truth fMRI	100% \pm 0%	100.00% \pm 0.00%	100% \pm 0%
1039 Zero-shot pred fMRI	90% \pm 10%	80.00% \pm 14.14%	90% \pm 10%

1041 To further validate the quality of the zero-shot generated fMRI under task conditions as described
 1042 in D.2, we performed a connectome fingerprinting analysis (Finn et al., 2015). This approach
 1043 assesses whether the predicted functional connectivity (FC) patterns retain one’s true brain portrait
 1044 and subject-specific patterns by attempting to identify individuals based on their FC profiles.

1045 Specifically, we selected 5 subjects from the auditory task dataset, each of whom had two scans
 1046 under **different conditions (fast and sparse auditory stimulus)**. In each trial, one scan per subject
 1047 was randomly assigned to a "database set", and the other scan formed the "target set". For each
 1048 target FC matrix, we computed Pearson correlations with all database matrices (using vectorized
 1049 upper-triangular edge values), and predicted subject identity by selecting the database matrix with the
 1050 highest similarity. This procedure was repeated across all 16 possible permutations (trials), and the
 1051 average identification accuracy was reported.

1052 As shown in Table 6, ground-truth fMRI achieved perfect identification accuracy. Notably, our model’s
 1053 zero-shot predicted fMRI also demonstrated high fingerprinting accuracy across full-brain, gray
 1054 matter, and subcortical FC matrices. These findings suggest that the generated fMRI signals preserve
 1055 individualized functional connectivity signatures, supporting their potential utility in downstream
 1056 applications involving subject-specific brain representations, such as cognitive trait identification,
 1057 behavioral decoding, and clinical profiling (Finn et al., 2015; Mantwill et al., 2022; Lu et al., 2024).

1058 D.4 DETAILED EVALUATION OF REST-TO-TASK ZERO-SHOT TRANSFER

1061 We provide a comprehensive analysis of the zero-shot fMRI reconstruction performance in Figure 6.
 1062 Specifically, panel (A) visualizes the region-wise temporal correlation across all 512 ROIs, and panel
 1063 (B) summarizes these results (TCorr values) by averaging ROIs within each corresponding brain
 1064 network to highlight network-level reconstruction quality.

1065 In addition to these region-aggregated TCorr scores, we further evaluate network-level time-series
 1066 reconstruction (see Figure 6(C-D)). Specifically, for each of the 17 networks defined in the Yeo atlas,
 1067 we directly average the reconstructed ROI time series within that network to obtain a single network-
 1068 level time series per subject. This provides a complementary assessment that focuses on larger-scale
 1069 temporal dynamics rather than ROI-level variations. Finally, we include example reconstructions
 1070 from the Salience/Ventral Attention Network, which contains several deep and non-surface regions
 1071 (e.g., anterior insula, dorsal ACC). These examples illustrate that our model is capable of recovering
 1072 coherent temporal structure even in networks that are typically challenging for EEG-based methods.

1073 D.5 ZERO-SHOT EVALUATION ON BROADER TASKS

1075 To more thoroughly evaluate the model’s zero-shot generalization ability beyond the auditory task
 1076 included in the paper, we conducted two additional zero-shot analyses on publicly available datasets,
 1077 targeting both paired EEG-fMRI settings and real-world EEG-only scenarios.

1078 **Parkinson’s disease EEG dataset** To evaluate real-world applicability where fMRI is unavailable,
 1079 along with generalizability to a clinical population, we tested our model on a Parkinson’s EEG

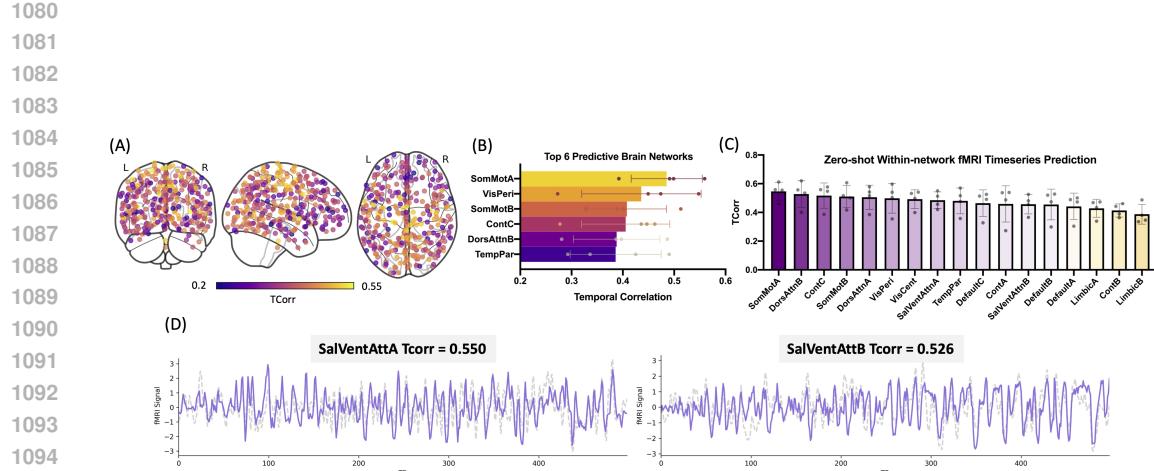


Figure 6: **Zero-shot fMRI reconstruction performance.** (A) Visualization of reconstruction performance across the whole brain; dots represent brain regions. (B) Top predictive brain networks; dots represent 4 evaluation participants in task dataset. (C) Network-wise fMRI timeseries reconstruction performance; dots represent 4 evaluation participants in task dataset. (D) Example visualization of zero-shot reconstructed network-wise fMRI time within Salience/Ventral Attention Network (gray: ground-truth, purple: prediction).

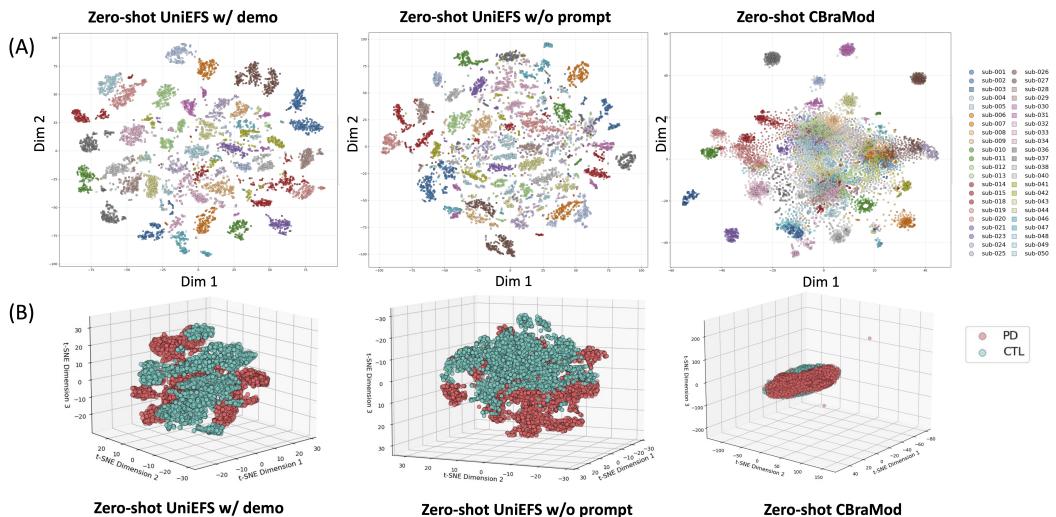


Figure 7: **Zero-shot EEG latent generated by UniEFS and CBraMod.** (A) 2-D t-SNE with each subject visualized with different colors. (B) 3-D t-SNE with colors distinguishing Parkinson patient (PD) and control (CTL) groups

1134 dataset (Cavanagh, 2021) with task conditions (oddball task) and demographic distribution (elderly
 1135 participants) entirely different from our training data. We refer the reader to the dataset paper for
 1136 details of the EEG collection, task condition, and demographic information.

1137 For this raw dataset, we performed standard EEG preprocessing, including band-pass filtering (0.1–75
 1138 Hz), 60 Hz notch filtering, and ICA-based artifact removal to eliminate eye-related components.
 1139 The cleaned EEG was then epoched into 16-seconds windows with a 2-second stride (i.e., 14-
 1140 second overlap), and fed each window into our trained UniEFS model to extract the fMRI-informed
 1141 EEG latent embedding. We then compared the t-SNE plots of our model’s embeddings versus the
 1142 embeddings from the state-of-the-art EEG foundation model CBraMod (Wang et al., 2025) as shown
 1143 in Figure 7. In summary, we have following key observations:

1144 (1) **Clear separation between Parkinson patients and controls:** In the zero-shot setting, our
 1145 latent representations cleanly separate the PD group from the control group, without any finetuning.
 1146 This indicates that the learned EEG→fMRI projection extracts neurophysiologically meaningful
 1147 structure that generalizes to entirely new populations. Such zero-shot group separation suggests
 1148 promising potential for downstream clinical applications (e.g., biomarker discovery, screening, or
 1149 disease monitoring) even when no fMRI is available.

1150 (2) **Strong individual-specific clustering:** Samples from the same subject cluster tightly together
 1151 in our latent space, indicating that the EEG→fMRI projection learns highly individualized EEG
 1152 representations even without explicit subject labels. Incorporating demographic embeddings (age,
 1153 sex) makes these subject-wise clusters even more compact, suggesting that the model can disentangle
 1154 individual-specific neural signatures from population-level factors. This pattern is consistent with our
 1155 fingerprinting results reported in the main paper, where subject identity can be reliably recovered
 1156 from the predicted fMRI signals. From the latent-space perspective, this further demonstrates that
 1157 the learned representation is not only discriminative but also structured in a way that preserves
 1158 stable, subject-specific traits. This also highlights the potential value of our representation for various
 1159 downstream applications.

1160 (3) **Better zero-shot structure than CBraMod:** Compared to CBraMod, our embeddings show
 1161 clearer intra-subject consistency and stronger inter-group separability under a completely zero-
 1162 shot setting, even though our model is trained on only 28 scans rather than a large-scale EEG
 1163 corpus. This highlights that the performance does not stem from data scale, but from the inductive
 1164 bias introduced by learning an fMRI-augmented EEG embedding, where projecting EEG into the
 1165 semantically structured fMRI latent space provides richer physiological constraints and induces
 1166 robust, discriminative, and subject-specific EEG representations, even in the absence of paired fMRI
 1167 during inference. This opens the possibility for using fMRI-informed EEG representations as a
 1168 scalable backbone for future EEG-only downstream tasks.

1169 **Paired EEG-fMRI dataset for motor conditions** We evaluated our model pretrained using resting-
 1170 state data on the Simultaneous EEG–fMRI Dataset for Multiple Motor Conditions (Bondi et al.,
 1171 2025), which contains task paradigms completely different from our auditory setting. As demo-
 1172 graphic information is not provided for this dataset, we evaluated the pretrained model without prompt
 1173 conditioning. The dataset paper contains full details on EEG–fMRI acquisition and preprocessing,
 1174 which we refer the reader to for additional information. Similarly, we directly fed the segmented EEG
 1175 window into our model pretrained on resting-state and assess the reconstructed fMRI signals. As
 1176 shown in Table 7, zero-shot predictions already produce meaningful temporal correlations and connec-
 1177 tivity structure, demonstrating stable cross-dataset generalization even when the task paradigm and
 1178 participant demographics differ substantially from training. Finetuning further improves subcortical
 1179 and cerebellar reconstruction (SC TCorr: 0.158→0.233; CB TCorr: 0.195→0.213) and substantially
 1180 strengthens connectivity estimation (ConnCorr: 0.363→0.509; ConnMSE: 0.257→0.139).

1181 In summary, these analyses together provide additional support that our model generalizes to novel
 1182 task structures, new subjects, and unseen EEG–fMRI dynamics.

1183 D.6 LEAVE-ONE-SITE-OUT TEST

1184 Figure 8 presents a detailed comparison between leave-one-site-out zero-shot generalization and
 1185 within-site training with three datasets (Rest Dataset 1, Rest Dataset 2, and Auditory-task Dataset),
 1186 along with a comprehensive ablation of the different prompt components. Panel (A) shows that

1188

1189

Table 7: Zero-shot vs. finetuned performance across metrics for motor-task fMRI reconstruction.

1190

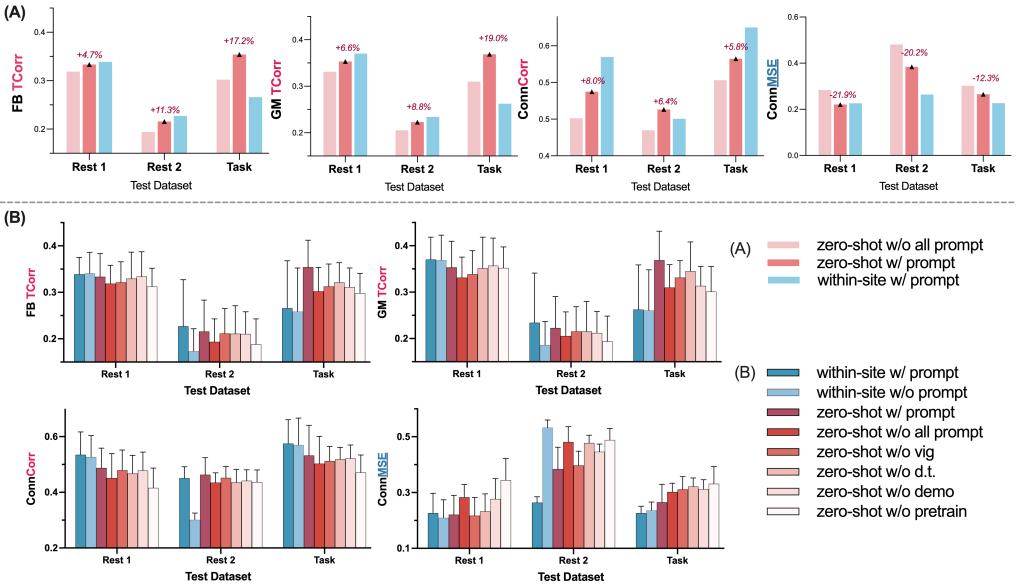
1191

1192

Condition	FB TCorr	GM TCorr	SC TCorr	CB TCorr	ConnCorr	ConnMSE
Zero-shot	0.237 \pm 0.083	0.266 \pm 0.087	0.158 \pm 0.083	0.195 \pm 0.091	0.363 \pm 0.114	0.257 \pm 0.024
Finetuned	0.254 \pm 0.052	0.272 \pm 0.057	0.233 \pm 0.064	0.213 \pm 0.059	0.509 \pm 0.060	0.139 \pm 0.013

1193

1194



1200

1201

1202

1203

1204

1205

1206

1207

1208

1209

1210

1211

1212

1213

1214

Figure 8: **Evaluation of leave-one-site-out vs. within-site performance under different prompt ablation settings.** (A) Zero-shot leave-one-site-out performance compared with within-site training for the full prompt model and the model without any prompts. Each bar shows the mean reconstruction metric across subjects, with percentage values indicating the improvement from adding prompts. (B) Comprehensive ablation analysis of each prompt component, including dataset tokens (d.t.), vigilance (vig), demographic prompts (demo), and pretraining, evaluated under both zero-shot and within-site settings. Error bars denote subject-wise standard deviation. Across test datasets (Rest 1, Rest 2, and Task), prompts consistently improve zero-shot generalization, and removing specific prompt types or pretraining reveals their individual contributions to mitigating cross-site domain shifts.

1215

1216

1217

1218

1219

1220

1221

1222

1223

1224

1225

1226

1227

1228

1229

1230

1231

1232

1233

1234

1235

1236

1237

1238

1239

1240

1241

adding prompt tokens consistently improves zero-shot performance across all metrics and datasets, reducing the cross-site performance gap and in several cases even matching or surpassing within-site results. This indicates that the prompt embeddings effectively capture site-, demographic-, and vigilance-related variability, enabling robust generalization when testing on unseen sites.

Panel (B) further decomposes the contribution of each prompt type. Removing individual prompts (dataset tokens, vigilance, demographic cues) leads to consistent drops in zero-shot performance, highlighting that each component contributes complementary contextual information. Notably, removing the pretraining stage causes the substantial degradation, confirming that fMRI self-supervised pretraining is essential for stabilizing cross-site representations.

D.7 PHYSIOLOGICAL PLAUSIBILITY CHECK

To further evaluate the physiological plausibility of the generated fMRI signal, we calculated and compared the power spectrum of real and generated fMRI signals across all regions. As shown in Figure 9, the predicted BOLD time series recovers the characteristic low-frequency PSD profile of real fMRI, including the dominant <0.1 Hz band and the expected log-linear decay. Moreover, the predicted and true PSDs exhibit very high correspondence across frequencies (Pearson $r = 0.93$, $p < 0.001$; Figure 9(B)), demonstrating that the model captures the correct temporal dynamics underlying

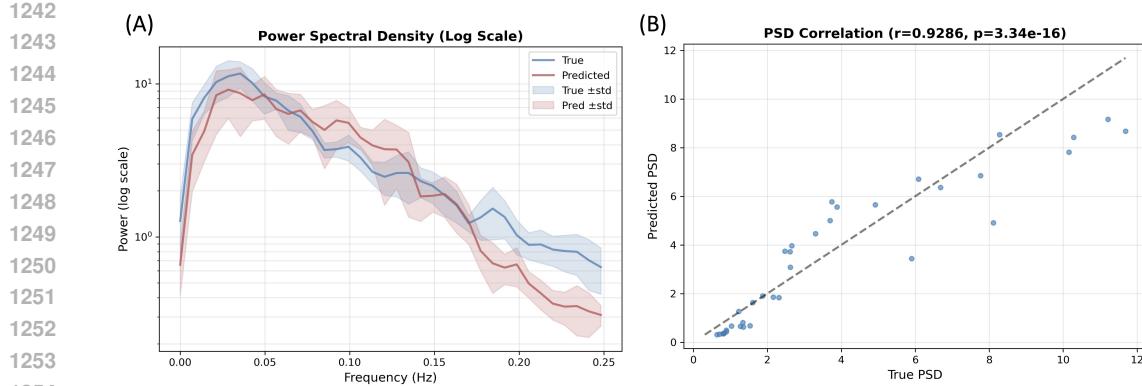


Figure 9: Real and generated fMRI spectrum comparison

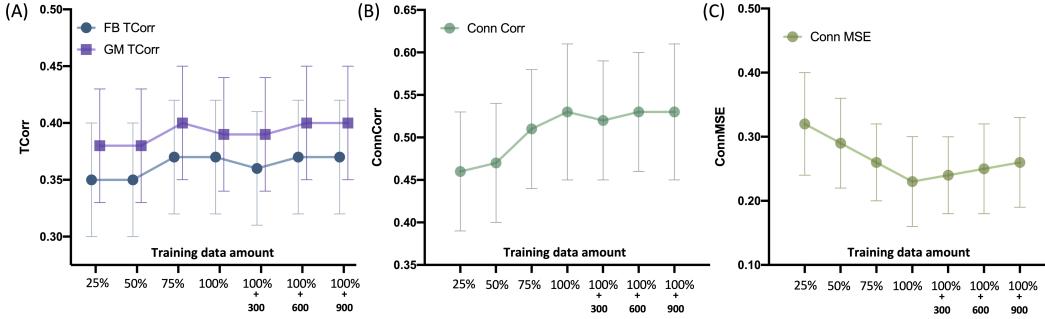


Figure 10: Impact of data amount during pretraining (under varying proportions of training scans: 25%, 50%, 75%, and 100%, where 100% corresponds to the dataset used in the main experiments; additional 300, 600, and 900 age-matched scans from HCP-A were also evaluated). (A) Impact on averaged temporal correlation within full-brain (FB TCorr) and gray matter (GM TCorr). (B) Spatial correlation (Conn Corr) between predicted and ground-truth functional connectivity (FC). (C) Mean squared error (Conn MSE) of connectivity strength in reconstructed FC matrices.

fMRI fluctuations. This analysis complements our TCorr and FC evaluations by confirming that the reconstructed signals preserve known neurophysiological properties of the BOLD response.

D.8 EFFECT OF DATA SCALING DURING PRETRAINING

To further assess the impact of sample size during uni-modal fMRI pretraining, we conducted additional experiments by varying the number of training samples. We also incorporated age-matched resting-state scans (ages 35–50) from the HCP-Aging dataset (Bookheimer et al., 2019; Harms et al., 2018) as supplementary pretraining data, allowing the model to learn from a broader range of fMRI variability and provide a more comprehensive analysis. As shown in Figure 10, we observed that using less than 75% of the training data leads to a noticeable performance drop across all metrics. However, once the training set exceeded 75% of the total data, the improvement in average temporal correlation became marginal. Interestingly, functional connectivity reconstruction, especially ConnMSE, continued to improve more consistently with additional data within the HCP-YA dataset. In contrast, supplementing with age-matched HCP-Aging scans did not lead to further improvements in EEG-to-fMRI translation, suggesting that the current data scale (1,200 training scans, 600 TRs each) may already be sufficient for effective pretraining. Therefore, we conclude that the current number of subjects in the HCP-YA dataset provides a substantial and efficient scale for fMRI representation learning and for capturing instantaneous spatial dependencies across brain regions.

1296 D.9 EFFECT OF MASKING RATIO DURING PRETRAINING
1297

1298 Table 8 shows the impact of different mask ratios on performance. We found that a mask ratio of 0.5
1299 yielded the best results. In typical Masked Autoencoder (MAE) training, a high mask ratio (e.g., 0.75)
1300 is often chosen to challenge the model to recover missing information and learn robust representations
1301 (He et al., 2022; Chen et al., 2023). However, in our case, a mask ratio of 0.75 did not yield the best
1302 performance for fMRI ROI data. The suboptimal performance of this high mask ratio may be due to
1303 the significant loss of inter-ROI correlation information, which is necessary for recovering full brain
1304 fMRI patterns. In other words, since we have already averaged the (voxel-wise) signals within regions
1305 to obtain the ROI data, much of the redundancy in voxel-wise signals has been reduced. Masking
1306 too much information hampers the model’s ability to capture the relationships between ROIs, which
1307 are essential for meaningful fMRI-to-EEG mapping. Preserving spatial continuity and functional
1308 connectivity is critical for the model to learn accurate representations. While when the mask ratio
1309 is small, it makes the task too easy for the model to learn complex patterns and may overfit to the
1310 existing information, leading to suboptimal generalization and performance on unseen data.

1312 Table 8: Influence of mask ratio in f-MSM

Mask ratio	FB TCorr \uparrow	GM TCorr \uparrow	SC TCorr \uparrow	CB TCorr \uparrow	Conn PCorr \uparrow	Conn MSE \downarrow
0.25	0.352 ± 0.044	0.378 ± 0.048	0.272 ± 0.080	0.234 ± 0.052	0.511 ± 0.078	0.274 ± 0.086
0.50	0.367 ± 0.052	0.394 ± 0.060	0.276 ± 0.082	0.247 ± 0.060	0.527 ± 0.084	0.233 ± 0.072
0.75	0.357 ± 0.058	0.385 ± 0.066	0.269 ± 0.090	0.241 ± 0.061	0.518 ± 0.086	0.263 ± 0.085

1312 D.10 IMPACT OF PATCH SIZE
1313

1323 Table 9: Influence of patch size in f-MSM

Patch size	FB TCorr \uparrow	GM TCorr \uparrow	SC TCorr \uparrow	CB TCorr \uparrow	Conn PCorr \uparrow	Conn MSE \downarrow
1	0.367 ± 0.052	0.394 ± 0.060	0.276 ± 0.082	0.247 ± 0.060	0.527 ± 0.084	0.233 ± 0.072
2	0.347 ± 0.069	0.375 ± 0.079	0.256 ± 0.088	0.222 ± 0.057	0.475 ± 0.083	0.322 ± 0.088
4	0.355 ± 0.040	0.378 ± 0.045	0.281 ± 0.078	0.245 ± 0.054	0.491 ± 0.093	0.282 ± 0.081
8	0.362 ± 0.062	0.387 ± 0.070	0.281 ± 0.087	0.245 ± 0.063	0.500 ± 0.082	0.257 ± 0.072

1324 In our default setting, the model uses a patch size of 1, where each token corresponds to a single
1325 brain region. This approach is grounded in the understanding that, unlike images - where adjacent
1326 pixels often share semantic content due to spatial continuity Dosovitskiy et al. (2020) - the ordering
1327 of regions of interest (ROIs) in a brain data vector does not inherently reflect anatomical proximity or
1328 functional similarity. Consequently, neighboring entries in the ROI vector may correspond to brain
1329 areas that are neither anatomically adjacent nor functionally related. By representing each ROI as a
1330 separate token, the model avoids imposing artificial spatial assumptions and allows for the learning of
1331 functional relationships based on actual connectivity patterns rather than from an arbitrary ordering.
1332 Here we compare the performance across different patch sizes in transformer of f-MSM. For patch
1333 sizes larger than 1, the patched data are transformed into embeddings using a 1D convolutional layer
1334 with the stride equal to the patch size. As shown in Table 9, a patch size of 1 achieves the best
1335 performance among most of the metrics compared with the larger patch sizes.

1343 D.11 COMPARISON BETWEEN DIFFERENT ALIGNMENT LOSS
1344

1345 Table 10 compares several alignment objectives used during Stage 2. Overall, MSE provides the
1346 most stable and accurate temporal reconstruction, achieving the highest or near-highest performance
1347 on FB TCorr, GM TCorr, and CB TCorr. While alternative objectives such as InfoNCE and cosine
1348 similarity show competitive performance on certain connectivity metrics (e.g., ConnCorr, ConnMSE),
1349 they exhibit substantially larger variance across subjects in subcortical regions, leading to potentially
less reliable overall behavior.

1350

1351

Table 10: Experiment on different alignment losses.

Loss Type	FB TCorr ↑	GM TCorr ↑	SC TCorr ↑	CB TCorr ↑	ConnCorr ↑	ConnMSE ↓
MSE (Ours)	0.367 ± 0.052	0.396 ± 0.058	0.251 ± 0.0371	0.247 ± 0.060	0.527 ± 0.084	0.233 ± 0.072
InfoNCE	0.346 ± 0.028	0.372 ± 0.017	0.270 ± 0.270	0.217 ± 0.074	0.537 ± 0.167	0.180 ± 0.001
Contrastive	0.329 ± 0.005	0.354 ± 0.007	0.256 ± 0.441	0.207 ± 0.018	0.516 ± 0.150	0.204 ± 0.002
Cosine	0.362 ± 0.054	0.391 ± 0.063	0.266 ± 0.074	0.232 ± 0.053	0.558 ± 0.091	0.212 ± 0.071
MSE+Cosine	0.352 ± 0.050	0.379 ± 0.076	0.256 ± 0.435	0.236 ± 0.039	0.511 ± 0.153	0.246 ± 0.070

1356

1357

1358

D.12 ABLATION STUDY ON EEG ENCODER MODULES

1360

Table 11 summarizes the ablation results on the EEG encoder. These results demonstrate that the encoder architecture is not arbitrary: the temporal–spatial module and the multi-scale spectral module capture distinct and complementary aspects of EEG signals, and both are empirically necessary for accurate EEG→fMRI reconstruction. This is consistent with the original NeuroBOLT findings and provides direct justification for our architectural design.

1365

1366

1367

Table 11: Ablation study on architectural components. Mean ± std. Colors denote significance vs. full model derived by paired t-test: red ($p < 0.001$), yellow ($p < 0.01$), blue ($p < 0.05$). Bold indicates best performance per column.

1370

Model Variant	FB Corr ↑	GM TCorr ↑	SC TCorr ↑	CB TCorr ↑	ConnCorr ↑	ConnMSE ↓
Ours (Full)	0.367 ± 0.052	0.396 ± 0.058	0.251 ± 0.0371	0.247 ± 0.060	0.527 ± 0.084	0.233 ± 0.072
Ours (no MSS)	0.189 ± 0.076	0.206 ± 0.092	0.126 ± 0.042	0.139 ± 0.074	0.122 ± 0.032	0.407 ± 0.098
Ours (no TS)	0.350 ± 0.065	0.376 ± 0.075	0.262 ± 0.083	0.226 ± 0.055	0.530 ± 0.074	0.230 ± 0.080

1374

1375

1376

D.13 IMPACT OF EEG INPUT LENGTH

1378

1379

1380

Table 12: Performance under different EEG input lengths.

1381

EEG Length	FB TCorr ↑	GM TCorr ↑	SC TCorr ↑	CB TCorr ↑	ConnCorr ↑	ConnMSE ↓
16 s	0.367 ± 0.052	0.394 ± 0.060	0.276 ± 0.082	0.247 ± 0.060	0.527 ± 0.084	0.233 ± 0.072
12 s	0.344 ± 0.049	0.368 ± 0.057	0.264 ± 0.069	0.235 ± 0.054	0.548 ± 0.065	0.187 ± 0.058
8 s	0.303 ± 0.059	0.330 ± 0.067	0.225 ± 0.055	0.211 ± 0.047	0.519 ± 0.074	0.246 ± 0.077
4 s	0.157 ± 0.030	0.157 ± 0.044	0.180 ± 0.039	0.166 ± 0.043	0.314 ± 0.040	0.249 ± 0.057

1386

1387

D.14 PERFORMANCE WITH DIFFERENT DiFuMo ATLAS GRANULARITIES

1389

Table 13 reports EEG-to-fMRI reconstruction performance across different DiFuMo atlas granularities (256, 512, and 1024 ROIs). Overall, the 256-ROI atlas yields the highest reconstruction accuracy across all metrics, followed by 512 ROIs, while 1024 ROIs shows the largest performance drop. This trend reflects the increasing difficulty of predicting finer-grained, higher-dimensional fMRI representations from limited EEG information: as ROI granularity increases, each region becomes smaller and noisier, making both temporal and connectivity reconstruction more challenging. These results suggest that moderate atlas resolutions (e.g., 256-512) strike a favorable balance between spatial detail and predictable signal quality.

1398

1399

D.15 UNSEEN RESTING-STATE SCAN RECONSTRUCTION EXAMPLES

1400

1401

1402

1403

In this section, we present several examples of held-out resting-state fMRI scan reconstructions, focusing on regions similar to those displayed in Li et al. (2024b) (Figure 11). These examples demonstrate the ability of our unified model to efficiently reconstruct entire resting-state fMRI scans across a wide time range using a single model.

1404

1405

Table 13: Performance with different DiFuMo atlas granularities.

1406

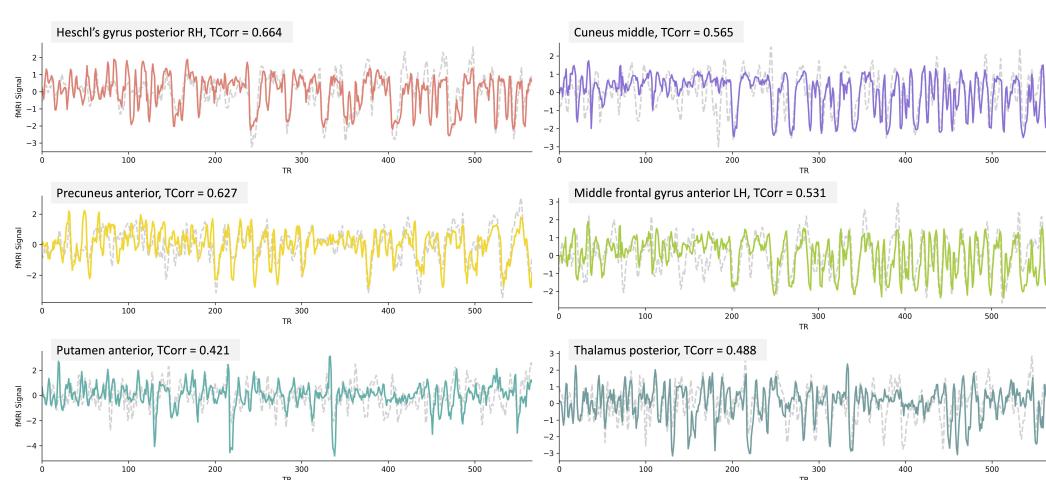
1407

1408

1409

1410

1411



1412

1413

1414

1415

1416

1417

1418

1419

1420

1421

1422

1423

1424

1425

1426

1427

1428

1429

1430

1431

1432

Figure 11: Held-out whole-scan reconstruction examples.

E ADDITIONAL DISCUSSION

E.1 DISCUSSION ON FC RECOVERY AND COMPARISON WITH BASELINES

While comparing our model with other EEG-to-fMRI translation baselines, we observe that UniEFS achieves the second-best performance in recovering functional connectivity (FC). One plausible reason our model does not outperform the CNN-based approach by Li et al. (2024a) in FC reconstruction is that FC is computed using Pearson correlation, which is highly sensitive to noise. Even minor prediction deviations can result in amplified discrepancies in pairwise correlations. The CNN baseline tends to produce smoother and more regularized outputs, which may suppress high-frequency fluctuations and thus yield more stable FC metrics—particularly in small-scale evaluation settings. In contrast, our model prioritizes frame-wise fidelity and regional dynamics, which may introduce local variability despite capturing more detailed temporal patterns. Notably, despite this, our model offers overall more consistent and strong performance across diverse evaluation settings.

E.2 DISCUSSION ON THE COMPLEMENTARY ROLES OF RECONSTRUCTION AND ALIGNMENT LOSS

From Table 2, we observe that during the alignment stage the reconstruction loss contributes the largest performance gain, while the improvement from the alignment loss is numerically smaller. Importantly, however, the gains from the alignment term are statistically significant and consistently positive across all subjects, indicating that its effect is systematic rather than noise-driven. This behavior aligns with the distinct roles of the two losses in our architecture. In stage-2, the fMRI decoder is frozen; thus, the reconstruction loss provides the primary source of task-specific semantic supervision, guiding the EEG encoder to produce latent representations that remain decodable through the fixed decoder. By contrast, the alignment loss encourages geometric proximity between EEG and fMRI latents but does not account for the decoder’s nonlinear inversion geometry. Because the pretrained fMRI latent space forms a curved and structured manifold, even small off-manifold deviations - while close in Euclidean distance - may decode into semantically incorrect fMRI patterns. Consequently, the reconstruction loss naturally leads to larger numerical improvements, whereas the alignment loss serves as a regularizer that improves latent-space geometry, stability, and cross-subject

1458 consistency. This explains why its contribution is more modest in magnitude yet remains statistically
 1459 reliable.
 1460

1461 E.3 CLARIFICATION ON RECONSTRUCTION VS. FUTURE-FRAME PREDICTION 1462

1463 The model reconstructs the fMRI signal associated with the neural activity expressed in the preceding
 1464 EEG window. This choice follows the physiology of the hemodynamic response, where the BOLD
 1465 signal at time t primarily reflects neural events that occurred several seconds earlier. Accordingly, our
 1466 formulation focuses on recovering the temporally aligned BOLD representation rather than predicting
 1467 future fMRI states beyond what is supported by the EEG window. This differs from multi-step
 1468 forecasting approaches, which aim to predict future frames and require additional considerations such
 1469 as modeling fMRI autocorrelation and controlling temporal information leakage. Such forecasting
 1470 extensions are conceptually distinct and would be an interesting future direction to explore.
 1471

1472 E.4 LIMITATIONS AND FUTURE WORK 1473

1474 For the resting condition, our model is trained on only two paired EEG-fMRI datasets with fewer than
 1475 32 EEG electrodes. The limited electrode coverage may impede the ability to accurately reconstruct
 1476 signals from subcortical regions. Since our context-aware embeddings are designed to accommodate
 1477 variability across datasets and populations, we plan to incorporate and collect additional resting-state
 1478 datasets, ideally with denser EEG electrode coverage, to further enhance the model’s capacity for
 1479 capturing fine-grained fMRI spatial dynamics, particularly in deep brain structures. Moreover, based
 1480 on our finding that the model exhibits strong zero-shot transfer ability, an exciting future direction
 1481 is to evaluate the pretrained resting-state model on diverse task-based datasets. Such downstream
 1482 evaluations would allow us to probe how well the learned EEG-to-fMRI mapping generalizes beyond
 1483 resting conditions, potentially enabling task-specific decoding and offering new insights into the
 1484 neural mechanisms that link spontaneous and task-evoked brain activity. In the longer term, this line
 1485 of work may also open avenues for clinical applications, such as noninvasive brain decoding and
 1486 monitoring of cognitive or pathological states in settings where fMRI is impractical.
 1487

1488 F THE USE OF LARGE LANGUAGE MODELS (LLMs) 1489

1490 We disclose that large language models (LLMs) were employed during manuscript preparation purely
 1491 for non-substantive editing tasks (e.g. grammar polishing and phrasing refinements). All scientific
 1492 content, analysis, experimental design, and writing decisions were conceived and carried out solely
 1493 by the authors.
 1494
 1495
 1496
 1497
 1498
 1499
 1500
 1501
 1502
 1503
 1504
 1505
 1506
 1507
 1508
 1509
 1510
 1511