NEST: A LARGE-SCALE BENCHMARK FOR PREDICTING FMRI IN NATIVE ANATOMY FROM EEG AND STRUCTURAL MRI

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting functional MRI (fMRI) from non-invasive electrophysiology (EEG) could enable cost-effective, high-resolution brain imaging, yet existing benchmarks omit subject-specific anatomies and rely on small cohorts that include fewer than 40 subjects. We propose NEST (**N**ative-anatomy **E**EG–**s**MRI–fMRI **T**ri-modal benchmark), a large-scale benchmark for predicting 4D fMRI in native anatomy from EEG and structural MRI during naturalistic viewing. By aligning separately recorded EEG and fMRI through a shared film timeline, rather than requiring simultaneous recording, NEST scales to 904 and 963 subjects watching two films, over $40\times$ more subjects than the largest public EEG–fMRI dataset. Models are evaluated on unseen subjects by voxel-level reconstruction, regional temporal correlation, and functional connectivity. We also propose NESTOR, which uses modality-specific conditioning, in which the subject's sMRI serves as a static anatomical condition and the EEG as a time-varying condition. NESTOR outperforms existing EEG-to-fMRI models by a clear margin, highlighting the importance of the subject's native anatomy. We demonstrate that NESTOR transfers to unseen movies in a zero-shot setting, indicating that it truly predicts fMRI from EEG signals rather than merely memorizing stimulus-specific fMRI patterns[1].

1 INTRODUCTION

Functional MRI (fMRI) is the dominant tool for whole-brain functional imaging, but it is expensive, motion-sensitive, and confined to the scanner (Glover, 2011; Poldrack et al., 2017). Electroencephalography (EEG) is portable and inexpensive but under-samples the brain, with only 64–256 scalp channels and limited spatial resolution (Michel & Brunet, 2019). The two modalities are complementary, since EEG offers portability and millisecond timing while fMRI offers whole-brain coverage and spatial localization. Cross-modal translation from EEG to fMRI (Calhas & Henriques, 2023; Li et al., 2024b; 2026a;b; Ma et al., 2025; Yao et al., 2025) aims to overcome this divide by synthesizing whole-brain functional maps directly from portable recordings. Existing methods, however, take no structural MRI as input, so none of them sees the anatomy of the subject whose brain it predicts. Meanwhile, stimulus-conditioned fMRI encoders (Cichy et al., 2019; Gifford et al., 2023; d'Ascoli et al., 2026) address an orthogonal task, predicting fMRI from the external stimulus itself rather than from the subject's actual neural response.

Existing EEG-to-fMRI works predict BOLD signals in predefined regions of interest (ROIs) of a group atlas (Li et al., 2024b; 2026a;b) or on a template cortical surface (Yao et al., 2025). Both formulations discard individual anatomy by construction. Consequently, they are poorly suited for analyses conducted in native space, such as presurgical (Bookheimer, 2007) or precision functional mapping (Gordon et al., 2017). This limitation is especially acute in lesioned or atrophic brains, where template normalization is least reliable (Brett et al., 2001). The other line predicts whole-brain volumes in each scan's native space from EEG alone (Calhas & Henriques, 2023; Lanzino et al., 2024; He et al., 2025; Roos et al., 2025), so a model must infer the subject's anatomy from a signal that carries little of it and can at best reproduce an average brain. Because static anatomy dominates voxel intensities, such a prediction can score well on voxel metrics without tracking the BOLD time

[1] `https://anonymous.4open.science/r/NESTOR`

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

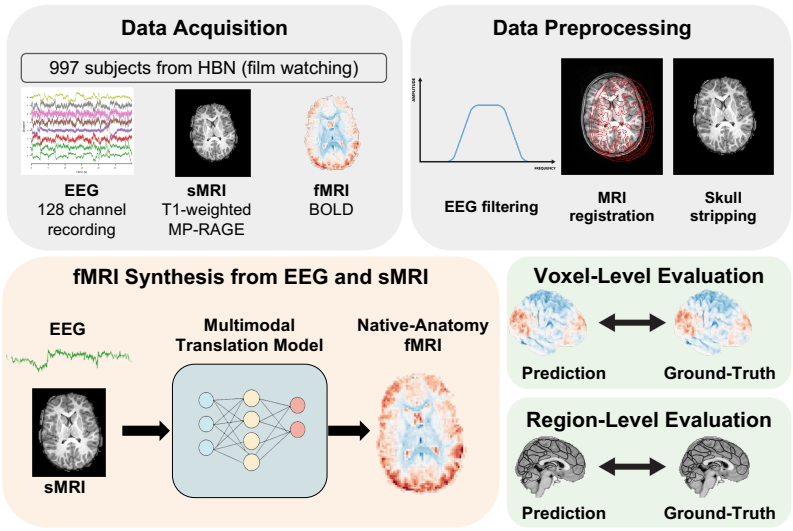

Figure 1: **Overview of NEST.** NEST provides matched (EEG, sMRI, fMRI) triplets from 904 and 963 subjects for two films (*Despicable Me* and *The Present*) in the Healthy Brain Network (Alexander et al., 2017), supporting the prediction of naturalistic-viewing brain dynamics in native anatomy. **(a) Data Acquisition.** Each subject contributes 129-channel EEG, a T1-weighted MP-RAGE structural MRI, and fMRI, with EEG and fMRI recorded in separate sessions while the subject watched the same films. **(b) Data Preprocessing.** A standardized pipeline filters the EEG and aligns the fMRI to the subject's own T1-weighted space with fMRIPrep (Esteban et al., 2019), with no warping to a group template. **(c) fMRI Synthesis.** A multimodal model translates EEG and sMRI into fMRI. **(d) Evaluation.** Predictions are scored against the measured fMRI at the voxel level (masked MSE, PSNR) and at the ROI level by temporal correlation and functional connectivity.

course. Both lines also depend on simultaneous EEG–fMRI recordings, which are difficult to acquire at scale (Li et al., 2026b), so their cohorts span only tens of subjects, limiting both the training data and the number of unseen subjects available for testing.

To ground prediction in each subject's own anatomy and to train on hundreds of subjects while testing on dozens of unseen ones, we introduce **NEST** (**N**ative-anatomy **E**EG–**s**MRI–fMRI **T**ri-modal benchmark), a benchmark for predicting 4D fMRI in native anatomy from EEG and structural MRI (sMRI) (see Figure 1). NEST comprises two per-film cohorts of 904 (*Despicable Me*) and 963 (*The Present*) subjects from the Healthy Brain Network (Alexander et al., 2017), each with 129-channel EEG, T1-weighted sMRI, and film-watching fMRI, partitioned into subject-disjoint train/val/test splits and released entirely in each subject's native anatomical space. Because EEG and fMRI are collected in separate visits, the two modalities are aligned based on the shared movie timeline rather than simultaneous recording. Finally, our evaluation protocol scores models on held-out subjects by voxel-level reconstruction, and region-level temporal correlation and functional connectivity.

We propose **NESTOR** that maps a subject's EEG and structural MRI to fMRI. To our knowledge, NESTOR represents the first framework to achieve this cross-modal synthesis. Given a subject's structural MRI and EEG, NESTOR predicts fMRI by combining the anatomical features supplied by MRI, and the activity fluctuation and localization provided by EEG using spatial and temporal bases. NESTOR achieves the best score on all six metrics for *Despicable Me* and on five of six for *The Present*. Trained on one film and tested zero-shot on the other, NESTOR performs comparably to training on the test film and still exceeds the baselines trained on that film on five of six metrics. The contribution of this paper is summarized as follows:

- We introduce the task of predicting 4D fMRI from EEG and the subject's structural MRI. This formulation keeps predictions in the subject's native anatomy rather than on a group template and avoids inferring that anatomy from EEG alone.
- We build NEST from the public Healthy Brain Network release by aligning separately recorded EEG and fMRI through a shared film, which yields over $40\times$ more subjects than any dataset used in prior EEG-to-fMRI benchmarks.

Table 1: Comparison with existing EEG-to-fMRI benchmarks. Prior benchmarks predict either regions of a group atlas or volumes on each scan's raw grid, from EEG alone and tens of simultaneously recorded subjects. NEST is, to our knowledge, the first to give the model each subject's structural MRI and to predict volumes aligned to it, with over 900 subjects per film.

Benchmark	Dataset	Task	# recordings	EEG (# ch.)	Full Brain	Native Space	sMRI input
Calhas & Henriques (2023)	Deligianni et al. (2014)	resting (eyes open)	10	64	✓	✓	×
	Walz et al. (2015)	object detection	10	43	✓	✓	×
	Pereira et al. (2020)	decision making	20	64	✓	✓	×
Li et al. (2024b)	Li et al. (2024b)	resting state	22	32 (26 used)	×	×	×
		auditory task	10	32 (31 used)	×	×	×
Li et al. (2026b)	Li et al. (2024b)	resting state	22	32 (23 used)	×	×	×
	Li et al. (2026b)	resting state	11	32 (23 used)	×	×	×
NEST (ours)	Alexander et al. (2017)	*Despicable Me*	904	129	✓	✓	✓
	Alexander et al. (2017)	*The Present*	963	129	✓	✓	✓

- We propose NESTOR, which uses modality-specific conditioning, with the subject's sMRI supplying the anatomy of every predicted volume and the EEG driving how the volumes change over time. NESTOR outperforms existing EEG-to-fMRI models on nearly all metrics and transfers zero-shot to the other film.

2 Related Work

EEG-to-fMRI translation has moved from linear to deep models while remaining limited by the scarcity of simultaneous EEG–fMRI data (Li et al., 2026b). Early work relates EEG to fMRI in cohorts of at most 20 subjects (Deligianni et al., 2014; Meir-Hasson et al., 2014; Walz et al., 2015; Pereira et al., 2020). Deep sequence-to-sequence models predict a few deep-region time courses from a delay-shifted EEG window with one model per subject (Kovalev et al., 2022; Li et al., 2024a), transcoding models fuse the two modalities in a shared source space (Liu & Sajda, 2023), and volume-wise synthesis generates whole fMRI volumes from small corpora with attention, transformer, and state-space U-Net decoders (Calhas & Henriques, 2023; Lanzino et al., 2024; He et al., 2025). NeuroBOLT (Li et al., 2024b) and UnEBOLT (Li et al., 2026a) predict ROI values across subjects from the 16 s of EEG preceding each frame, on 22 resting-state and 10 auditory-task subjects. CATD (Yao et al., 2025) generates cortical-surface fMRI with an EEG-conditioned diffusion model, and UniEFS (Li et al., 2026b) unifies region-specific models into one full-brain ROI-level model with an HCP-pretrained fMRI decoder and context tokens for age, sex, and vigilance, trained on 29 simultaneously recorded subjects. E2fNet (Roos et al., 2025) predicts whole volumes rather than regions with an encoder–decoder network, evaluated on the three datasets assembled by Calhas & Henriques (2023), whose largest cohort has 20 subjects.

Three limits are common to these studies (see Table 1). Cohorts stay at tens of subjects, EEG montages have at most 64 channels, and most work is in MNI152 space. The first two follow from the cost of simultaneous acquisition (Li et al., 2026b). The third follows from standard preprocessing, which normalizes subjects to a template for group statistics (Esteban et al., 2019) and discards the morphology on which presurgical and precision functional mapping operate (Brett et al., 2001; Gordon et al., 2017). NEST instead pairs EEG and fMRI from separate sessions of the same subject through a shared film, which yields 904 subjects for *Despicable Me* and 963 for *The Present*, with 129-channel EEG and voxel-level fMRI targets in each subject's native anatomy, at the cost of EEG–fMRI simultaneity (see Section 3.1).

3 NEST: A Benchmark for Predicting fMRI in Native Anatomy

NEST is a large-scale benchmark for generating 4D fMRI during naturalistic viewing from subject-paired electroencephalography (EEG) and structural MRI (sMRI). We curate two per-film cohorts from the Child Mind Institute Healthy Brain Network (HBN) (Alexander et al., 2017): 904 subjects for *Despicable Me* and 963 subjects for *The Present* (with 870 subjects contributing both films and 997 subjects contributing at least one), each contributing 129-channel EEG, T1-weighted MP-RAGE sMRI, and film-watching fMRI, partitioned subject-disjointly into $787/40/77$ and $841/40/82$

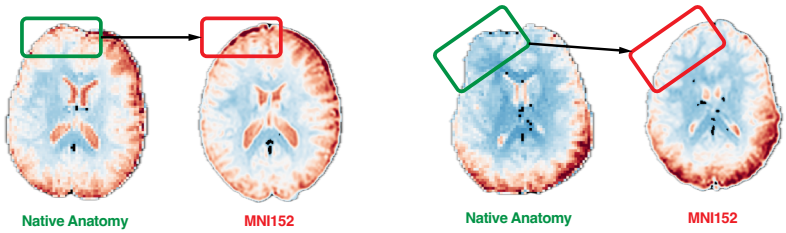

Figure 2: **Warping to a group template distorts each subject's anatomy.** Each pair shows one subject in its native space (left) and after MNI152 normalization (right), and the two subjects are shown at different axial levels. The green and red boxes mark the same anatomical region in the two coordinate systems, and the arrow shows how normalization displaces and reshapes it. Panels are cropped and scaled separately, so compare within a pair. NEST is released without this normalization.

train/val/test splits. Table 1 shows NEST's advantagshoe over prior EEG–fMRI corpora, being over $40\times$ larger by subject count, on a 129-channel montage rather than 64 or fewer, and including both RoI and voxel level metrics. Unlike existing benchmarks, NEST is not simultaneously acquired. Rather, EEG and fMRI come from separate sessions of the same subject and are aligned in movie time through the shared films.

3.1 Task Formulation

Let $\mathbf{x}^{\text{eeg}} \in \mathbb{R}^{C\times T_e}$ denote a segment of the subject's EEG with C electrodes sampled at f_s Hz over T_e samples, recorded while the subject watched a segment of a film, and let $\mathbf{x}^{\text{sMRI}} \in \mathbb{R}^{H\times W\times D}$ denote the subject's T1-weighted structural scan in its native space. The benchmark asks a model to predict the subject's fMRI over the same segment of the film,

$$\hat{\mathbf{y}} \;=\; g_\theta\big(\mathbf{x}^{\text{eeg}},\, \mathbf{x}^{\text{sMRI}}\big) \;\in\; \mathbb{R}^{H\times W\times D\times T_f}, \tag{1}$$

where g_θ is the translation model under evaluation, $T_f \ll T_e$ reflects the lower temporal resolution of fMRI, and the output grid matches the native spatial dimensions of $\mathbf{x}^{\text{sMRI}}$. The target $\mathbf{y}$ is the subject's measured fMRI over the corresponding segment, recorded in a separate session and aligned to the EEG via shared movie timestamps (see Section 3.3). Consequently, only stimulus-driven functional activity is predictable from the EEG. Session-specific spontaneous fluctuations are uncorrelated across visits and constitute irreducible target variance. Because evaluation is conducted on unseen test subjects, g_θ must generalize across individuals. Providing $\mathbf{x}^{\text{sMRI}}$ as structural input relieves the model from inferring cortical anatomy from scalp voltages, focusing the task on mapping functional dynamics onto a known structural substrate.

3.2 Subject-Native Anatomy

NEST provides all volumetric modalities—including structural scans, functional runs, and brain masks—entirely in each subject's native coordinate frame, omitting nonlinear warping to group templates such as MNI152. As illustrated across representative subjects in Figure 2, retaining native space preserves fine-grained cortical folding and individual morphological variation that are routinely smoothed out by template normalization. Consequently, our evaluation suite evaluates predictions directly against subject-native references, allowing trained models to generalize to unseen subjects without requiring template warping.

3.3 Data

We build NEST from the HBN (Healthy Brain Network) MRI/EEG release[2] (Alexander et al., 2017), a transdiagnostic pediatric cohort acquired at four sites with 129-channel EGI EEG at 500 Hz, T1-weighted MP-RAGE at 0.8 mm, and multiband fMRI at 2.4 mm with temporal resolution 0.8 s. We use the two naturalistic-viewing tasks available in both modalities, aligning fMRI and EEG to the movie timestamps for a clip of *Despicable Me* (10.0 min fMRI vs. 2.9 min EEG) and the short film *The Present* (3.3 min vs. 3.4 min). A subject enters a per-film cohort if the release contains a usable

[2] https://fcon_1000.projects.nitrc.org/indi/cmi_healthy_brain_network

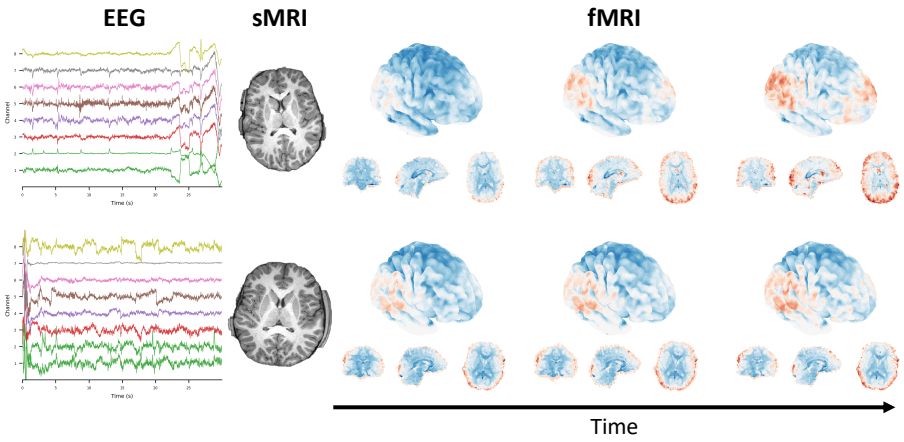

Figure 3: **Example of NEST. EEG:** a window of EEG recorded during the film (a representative subset of channels shown for visual clarity). **sMRI:** an axial slice of the subject-native, skull-stripped T1-weighted MP-RAGE used as structural input. **fMRI:** Activity at three time steps from the paired window. Top: 3D surface renderings project activity onto a shared template for visualization. Bottom: Orthogonal slices (sagittal, coronal, axial) display the same data in per-subject native space, highlighting differences in brain size, ventricle morphology, and cortical folding across rows. NEST releases all three modalities in subject-native anatomy. The template-surface 3D renderings here are used for visualization only.

T1 together with that film's EEG and fMRI (see Table 6) and every preprocessing stage completed for that film (fMRIPrep, the EEG segment, and the subject-native DiFuMo-64 atlas). This yields 904 subjects for *Despicable Me* and 963 for *The Present*, with 870 present in both and 997 in at least one. The pool is partitioned into subject-disjoint training, validation, and test splits (approximately 85%, 5%, and 10%), and each subject retains the same role in both films. Please refer to Tables 4 and 5.

3.4 Preprocessing

EEG preprocessing. We preprocess EEG following standard protocols for brain foundation models (Jiang et al., 2024; El Ouahidi et al., 2025; Wang et al., 2025). The film segment of each 500 Hz recording is re-referenced to the common average, high-pass filtered at 0.3 Hz to remove slow drifts, notch filtered at 60 Hz to remove power-line interference, and resampled to 200 Hz (see Appendix C).

sMRI / fMRI preprocessing. Standard fMRI pipelines warp every subject onto a group template (*e.g.*, MNI152), which removes per-subject anatomy and is unsuitable for the clinical settings discussed in Section 3.2. NEST therefore preprocesses each subject entirely in their own space. We preprocess structural and functional MRI with fMRIPrep (Esteban et al., 2019), using BOLD derivatives aligned to each subject's T1-weighted anatomical space. Each BOLD run is linearly detrended and high-pass filtered at 0.008 Hz over the full run, with its temporal mean restored. No additional spatial smoothing or nuisance regression is applied. Following temporal preprocessing, BOLD volumes are resampled using trilinear interpolation from their approximately 2.4-mm T1w-space grid onto a subject-centered, 2.4-mm isotropic grid of ($72\times88\times72$) voxels for model input, and the sMRI is resampled onto the same grid. Per-subject MRIQC (Esteban et al., 2017) summaries are released so that users can set their own criteria. Mean framewise displacement has a median of 0.26 mm on *Despicable Me* and 0.24 mm on *The Present*, similar across splits (see Table 6), and every sMRI–BOLD registration was checked visually. Because targets retain unregressed motion and physiological noise uncorrelated with the EEG, we report benchmark results on both the full cohort and a low-motion subset (see Section 5).

3.5 Pairing EEG and fMRI in Movie Time

Each subject's EEG and fMRI recordings are aligned to movie time using onset markers, and each EEG window is paired with the fMRI frames that cover the same movie time. In our default setting, a window spans 96 s of movie time, corresponding to $T_e = 19{,}200$ EEG samples (at 200 Hz) and $T_f = 120$ fMRI frames (at TR $= 0.8$ s), and models may choose other windows or strides. The subject's preprocessed sMRI is paired across all segments as the static structural input, forming

training tuples $(\mathbf{x}^{\text{eeg}}, \mathbf{x}^{\text{sMRI}}, \mathbf{y})$ synchronized in movie time. We release the scripts and configuration files that build these tuples from the public HBN release (Appendix F). Figure 3 shows representative tuples from several subjects, illustrating both the per-subject anatomical variability NEST preserves and the heterogeneity of EEG and fMRI activity across the cohort.

3.6 Evaluation Protocol

We evaluate generated fMRI on held-out subjects. All movie segments from test subjects are reserved exclusively for scoring, requiring models to generalize across unseen anatomy and individual functional responses to identical stimuli. We report three families of metrics: voxel-level, functional structure, and temporal structure. Together, they capture differences in anatomy, connectivity, and dynamics. Each metric is computed once per subject over the full scored sequence of T frames, with one prediction per frame, and then averaged across subjects with equal weight.

Voxel-level metrics. We report mean squared error (MSE), peak signal-to-noise ratio (PSNR), and structural similarity index measurement (SSIM) (Wang et al., 2004). We warp the DiFuMo-64 atlas (Dadi et al., 2020) into each subject's native space and use its binarized support as the mask for all voxel metrics. This mask restricts scoring to the brain regions the atlas covers, mostly gray matter, and excludes non-brain background, where errors would reward reproducing static anatomy rather than predicting activity. These metrics quantify low-level intensity tracking but remain largely insensitive to higher-order temporal or network structure.

Region-level temporal correlation. We measure how closely the predicted regional time courses follow the ground truth by their temporal correlation (temporal r). Ground-truth and predicted fMRI volumes are reduced to $R{=}64$ regional time series $\mathbf{s}_r, \hat{\mathbf{s}}_r \in \mathbb{R}^T$ ($r \in \{1, \dots, R\}$) via DiFuMo-weighted spatial averaging. We compute the Pearson correlation between ground-truth and predicted time series for each region over the full scored sequence and average across regions, yielding Temporal $r = \frac{1}{R}\sum_{r=1}^{R} \rho(\mathbf{s}_r, \hat{\mathbf{s}}_r)$. Regions whose measured signal is constant are excluded, and a constant prediction scores zero. Temporal r captures whether the model recovers stimulus-evoked neural dynamics across localized functional areas.

Region-level functional connectivity. We evaluate macro-scale functional organization using functional connectivity (FC) matrices $\mathbf{C}, \hat{\mathbf{C}} \in \mathbb{R}^{R\times R}$, defined as pairwise Pearson correlations between the R regional time series over the full scored sequence. We compare the two matrices across unique off-diagonal entries using Pearson correlation (FC r) and mean squared error, with FC MSE $= \frac{2}{R(R-1)}\sum_{i<j}(C_{ij} - \hat{C}_{ij})^2$. FC r measures agreement in relative connection patterns, while FC MSE captures absolute discrepancy. Both metrics characterize co-fluctuation structure independently of frame-by-frame temporal alignment.

4 NESTOR: Anatomy-Conditioned EEG-to-fMRI Translation

We propose NESTOR, a multimodal translation model that predicts 4D fMRI from paired EEG and subject-native sMRI. The sMRI supplies the anatomy of every predicted frame, and the EEG drives how the frames change over time. Figure 4 shows the architecture of NESTOR.

4.1 Architecture

The EEG is encoded with the frozen EEG foundation model REVE (El Ouahidi et al., 2025), denoted $\mathcal{R}$, which processes the 96 s window as three consecutive 32 s chunks whose tokens are concatenated along time into the EEG latent $\mathbf{z}^{\text{eeg}} = \mathcal{R}(\mathbf{x}^{\text{eeg}})$, one sequence per electrode. A trainable EEG adapter $\mathcal{A}$ maps this latent to the EEG embedding $\mathbf{e} = \mathcal{A}(\mathbf{z}^{\text{eeg}})$, aligned with the T_f fMRI frames.

Both MRI modalities are encoded with the pretrained MAISI 3D autoencoder (Guo et al., 2025), whose frozen encoder $\mathcal{E}$ compresses the sMRI volume and each fMRI frame from $72{\times}88{\times}72$ voxels (2.4 mm) to a $4{\times}18{\times}22{\times}18$ latent, and whose frozen decoder $\mathcal{D}$ maps latents back to volumes. This gives the subject's sMRI latent $\mathbf{z}^{\text{sMRI}} = \mathcal{E}(\mathbf{x}^{\text{sMRI}})$ and the latent of the fMRI window $\mathbf{z} = \big(\mathcal{E}(\mathbf{y}_1), \dots, \mathcal{E}(\mathbf{y}_{T_f})\big) \in \mathbb{R}^{T_f\times 4\times 18\times 22\times 18}$. The sMRI latent is replicated across the T_f frames and refined by four spatiotemporal blocks $F_1, \dots, F_4$, each conditioned on $\mathbf{e}$,

$$\mathbf{h}^{(0)} = \mathbf{1}_{T_f} \otimes \mathbf{z}^{\text{sMRI}}, \qquad \mathbf{h}^{(\ell)} = F_\ell\big(\mathbf{h}^{(\ell-1)}, \mathbf{e}\big), \quad \ell = 1, \dots, 4, \qquad \mathbf{h} = \mathbf{h}^{(4)}. \tag{2}$$

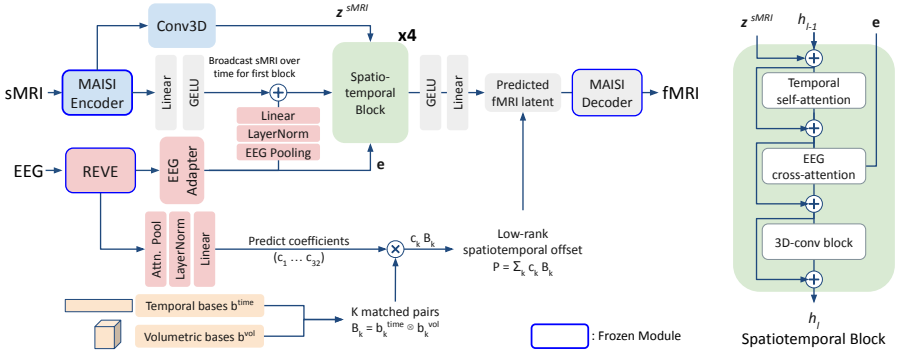

Figure 4: **Architecture of NESTOR.** The sMRI latent supplies the anatomy of every frame, and the EEG conditions four spatiotemporal blocks (right) through cross-attention and sets the coefficients of a low-rank offset $\mathbf{P}$ added to the predicted latent.

Each block contains self-attention across frames, a 3D convolution within each frame, and cross-attention to $\mathbf{e}$. All T_f frames are therefore predicted jointly from the whole 96 s EEG window rather than from separate 32 s chunks.

The EEG embedding predicts a single coefficient vector $\mathbf{c} \in \mathbb{R}^K$, with $K = 32$. The model learns a volumetric basis $\mathbf{B}^{\text{vol}} \in \mathbb{R}^{4 \cdot V \times K}$ and a temporal basis $\mathbf{B}^{\text{time}} \in \mathbb{R}^{T_f \times K}$, where $V = 18 \cdot 22 \cdot 18$ is the number of latent voxels, each with 4 channels. Together, they form a low-rank spatiotemporal offset

$$\mathbf{P} = \mathbf{B}^{\text{vol}} \operatorname{diag}(\mathbf{c}) \big(\mathbf{B}^{\text{time}}\big)^\top = \sum_{k=1}^{K} c_k \, \mathbf{b}_k^{\text{vol}} \big(\mathbf{b}_k^{\text{time}}\big)^\top \in \mathbb{R}^{4 \cdot V \times T_f}, \tag{3}$$

where $\mathbf{b}_k^{\text{vol}}$ and $\mathbf{b}_k^{\text{time}}$ are the k-th columns of $\mathbf{B}^{\text{vol}}$ and $\mathbf{B}^{\text{time}}$, respectively. Each of the K components pairs a learned volumetric pattern with a learned time course, and its strength is set by the corresponding EEG-predicted coefficient c_k. The offset thus gives the EEG a direct path to the prediction that bypasses the spatiotemporal blocks, with the basis components shared across subjects. The predicted fMRI latent $\hat{\mathbf{z}}$ and the predicted volumes $\hat{\mathbf{y}}$ of Eq.(1) are then

$$\hat{\mathbf{z}} = \text{MLP}(\mathbf{h}) + \mathbf{P}, \qquad \hat{\mathbf{y}}_t = \mathcal{D}(\hat{\mathbf{z}}_t), \quad t = 1, \ldots, T_f. \tag{4}$$

Appendix B details the EEG adapter and the spatiotemporal blocks.

4.2 Objective

NESTOR is trained with a latent reconstruction loss and two regional time-series losses, which supervise temporal dynamics without decoding latents into volumes. We derive regional signals $\mathbf{S}(\mathbf{z}) \in \mathbb{R}^{R \times T_f}$ by pooling latent channels under the subject-native DiFuMo-64 maps projected onto the latent grid, then combining the channels with fixed weights $\mathbf{W} \in \mathbb{R}^{R \times 4}$ fitted on training subjects (Appendix C). Let $\mathbf{s}_r(\mathbf{z}) \in \mathbb{R}^{T_f}$ denote region r's time series, and let $\mathbf{a}_t(\mathbf{z}) \in \mathbb{R}^R$ denote the centered, unit-L_2-normalized regional activity at time t. These losses avoid backpropagating through the frozen MAISI decoder for every frame, which would be prohibitively expensive.

With Pearson correlation ρ, the three losses are

$$\begin{aligned} \mathcal{L}_{\text{mse}} &= \|\hat{\mathbf{z}} - \mathbf{z}\|_2^2, \qquad \mathcal{L}_{\text{corr}} = \frac{1}{R} \sum_{r=1}^{R} \Big(1 - \rho\big(\mathbf{s}_r(\hat{\mathbf{z}}),\, \mathbf{s}_r(\mathbf{z})\big)\Big), \\ \mathcal{L}_{\text{gram}} &= 1 - \rho\big(\mathbf{g}(\hat{\mathbf{z}}),\, \mathbf{g}(\mathbf{z})\big), \qquad \mathbf{g}(\mathbf{z}) = \big(\mathbf{a}_t(\mathbf{z})^\top \mathbf{a}_{t'}(\mathbf{z})\big)_{t<t'}, \end{aligned} \tag{5}$$

where $\mathbf{g}(\mathbf{z}) \in \mathbb{R}^{T_f(T_f-1)/2}$ is the upper triangle of the temporal Gram matrix. $\mathcal{L}_{\text{mse}}$ constrains the absolute intensity and the static anatomy of the prediction, and $\mathcal{L}_{\text{corr}}$ matches the time course of each region up to a positive scale and an offset. $\mathcal{L}_{\text{gram}}$ matches the similarity between every pair of frames, which neither $\mathcal{L}_{\text{mse}}$, dominated by the static anatomy, nor the region-wise $\mathcal{L}_{\text{corr}}$ constrains directly. The total loss is

$$\mathcal{L} = \lambda_{\text{mse}} \mathcal{L}_{\text{mse}} + \lambda_{\text{corr}} \mathcal{L}_{\text{corr}} + \lambda_{\text{gram}} \mathcal{L}_{\text{gram}}, \tag{6}$$

with $\lambda_{\text{mse}} = 2$, $\lambda_{\text{corr}} = 1$, and $\lambda_{\text{gram}} = 0.5$, selected from $\{0.5, 1, 2\}$ on the validation set. Appendix E shows an ablation study on the objective function.

Table 2: Held-out test performance per movie condition, averaged across subjects. Cohort sizes are 77 (DM) and 82 (TP). DiFuMo-64 is used for region-level evaluation. Best results per film in bold. † UnEBOLT (Li et al., 2026a) directly predicts the brain signals for each RoI.

Movie	Model	Voxel-level			Region-level (DiFuMo-64)		
		MSE ($\times 10^{-3}$) ↓	PSNR ↑	SSIM ↑	Temporal r ↑	FC r ↑	FC MSE ↓
DM	UnEBOLT (Li et al., 2026a)	—†	—	—	0.0731	0.2547	0.1841
	E2fNet (Roos et al., 2025)	13.137	19.70	0.6560	−0.0212	0.1941	0.1716
	NESTOR (ours)	**8.102**	**22.12**	**0.7950**	**0.2896**	**0.3473**	**0.1643**
TP	UnEBOLT (Li et al., 2026a)	—	—	—	0.0854	0.2124	0.2121
	E2fNet (Roos et al., 2025)	16.866	18.81	0.6293	−0.0153	0.2335	**0.1697**
	NESTOR (ours)	**7.557**	**22.49**	**0.7824**	**0.3057**	**0.2459**	0.2039

5 EXPERIMENT

5.1 EXPERIMENTAL DETAILS

Baselines. We compare the translation baseline (see Section 4) with two recent EEG-to-fMRI models, UnEBOLT (Li et al., 2026a) and E2fNet (Roos et al., 2025), trained on NEST using their released code. Both predict one fMRI frame from the preceding 16 s of EEG and use neither MRI nor stimulus conditioning; we apply them at each frame, whereas NESTOR predicts all T_f frames in a window at once. UnEBOLT uses a LaBraM-initialized backbone (Jiang et al., 2024) to predict ROIs, mapping EEG projected to its 26-channel montage to 64 DiFuMo signals. Since it only predicts ROIs, we evaluate it only on ROI and connectivity metrics. E2fNet maps magnitude spectra from the EEG window to a whole-brain volume, with its decoder resized to our $72\times88\times72$ grid. Both retain their released objectives and use the baseline's subjects and fMRI targets (see Appendix C).

Training. Each model is trained and evaluated on a single film, *Despicable Me* (DM) or *The Present* (TP), under the same subject-disjoint splits. The translation baseline trains on 96 s windows with a 4 s stride, each pairing 120 fMRI frames with the EEG spanning the same movie time (see Section 3.1), for 20 epochs on one NVIDIA H100 GPU with an exponential moving average of the weights. The two framewise baselines train on every frame that has 16 s of measured EEG history. We select the best checkpoint for each model using validation performance, full-sequence Temp-Corr for the translation baseline, loss for UnEBOLT, and SSIM for E2fNet following the original implementation. Table 7 lists the training settings of all three models. The experiments use the full per-film cohorts, 904 (DM) and 963 (TP) subjects with $787/40/77$ and $841/40/82$ training, validation, and test subjects.

Evaluation. All models are scored on the available frames of each test run. For our translation baseline, we use frames 0 to 199 for DM and 0 to 239 for TP. Since the framewise baselines need 16 s of measured EEG before their first prediction, we do not evaluate on the first 20 frames for them. The translation baseline predicts 96 s windows sliding by 4 s, each window is decoded, and overlapping decoded volumes are averaged at every frame before scoring. Results are reported for the full test cohort, and Appendix Table 8 repeats them for runs whose fMRIPrep framewise displacement (FD), averaged over the EEG-paired frames (frames 0–212 for DM, the whole run for TP), is below 0.2 mm, without frame censoring.

5.2 PARTIAL RECOVERY OF MOVIE-RELATED BOLD STRUCTURE

Table 2 compares the three baselines on the same held-out test subjects. The translation baseline reaches a temporal correlation of 0.29 on *Despicable Me* and 0.31 on *The Present*, whereas UnEBOLT stays below 0.1. It also produces the most accurate volumes, with a PSNR 2 to 4 dB above that of E2fNet, the other model that predicts volumes. E2fNet reaches an SSIM of 0.63 to 0.66 while its temporal correlation stays near zero, so its volumes resemble the ground truth in structure without following its activity, as anticipated in Section 1. Connectivity separates the models less sharply, and E2fNet attains the lowest FC MSE on *The Present*. On low-motion runs, the temporal correlation rises to 0.34 and 0.43 for the translation baseline and to 0.10 and 0.11 for UnEBOLT (see Appendix D). The translation baseline also receives the subject's sMRI and a longer EEG context than the framewise baselines, so these gaps reflect differences in input as well as in architecture. Figure 5 shows qualitative results.

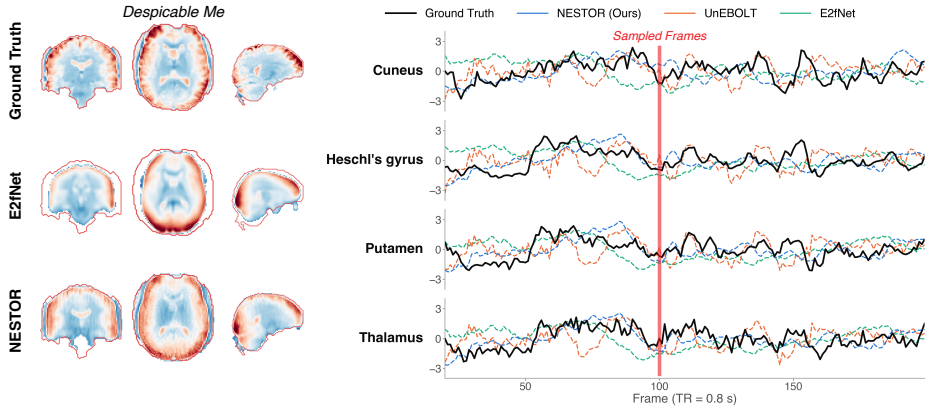

Figure 5: **Qualitative results on *Despicable Me*.** Left: Representative sagittal, coronal, and axial slices at $t = 100$ s for ground-truth fMRI and predictions from E2fNet and NESTOR. Right: Ground-truth and predicted DiFuMo-64 regional time courses across four representative regions (Cuneus, Heschl's gyrus, Putamen, and Thalamus). The vertical line indicates the time frame corresponding to the volumes displayed on the left.

Table 3: Zero-shot cross-stimulus transfer performance of NESTOR. Models trained on one movie condition are evaluated on the unseen counterpart test set (DM: *Despicable Me*; TP: *The Present*).

Movie (Train → Test)	Voxel-level			Region-level (DiFuMo-64)		
	MSE ($\times 10^{-3}$) ↓	PSNR ↑	SSIM ↑	Temporal r ↑	FC r ↑	FC MSE ↓
DM → DM	**8.102**	**22.13**	**0.7950**	0.2895	**0.3474**	**0.1643**
TP → DM	8.620	21.95	0.7945	**0.3036**	0.3070	0.1940
TP → TP	**7.633**	**22.48**	**0.7824**	**0.3057**	0.2459	0.2037
DM → TP	7.731	22.28	0.7802	0.2512	**0.2472**	**0.1712**

5.3 Zero-Shot Transfer

To test whether the model relies on memorizing the specific stimulus rather than learning a general mapping from EEG and structural MRI to fMRI, we evaluate zero-shot transfer across movies. Specifically, a translation baseline trained on one movie is evaluated on the test set of the alternate movie. As shown in Table 3, performance on the unseen movie remains comparable to within-stimulus evaluation, demonstrating that the translation generalizes across distinct naturalistic stimuli.

6 Discussion

We introduced NEST, a benchmark for predicting 4D fMRI in native anatomy from EEG and structural MRI, and proposed NESTOR, which takes the anatomy of each predicted volume from the sMRI and its changes over time from the EEG. Aligning separately recorded EEG and fMRI through a shared film scales NEST to over 900 subjects per film. NESTOR outperforms existing EEG-to-fMRI models on nearly all metrics and transfers to other movie-watching data without additional training (*i.e.*, zero-shot transfer).

However, NEST contains only film-viewing data, and whether these results hold for resting-state or task paradigms remains untested. Moreover, because EEG and fMRI were acquired in separate sessions, session-specific activity such as spontaneous fluctuations cannot be predicted from EEG. Extending NEST to additional paradigms and to simultaneously acquired EEG–fMRI data would relax both constraints. NESTOR further requires a structural MRI of each subject, although a single acquisition can be reused across multiple EEG sessions. In its absence, a standard anatomical template may be substituted, a setting we leave to future work. The principle underlying NESTOR, whereby structural MRI provides anatomy and EEG provides temporal dynamics, may extend to other combinations of neuroimaging modalities.

AI USE STATEMENT

In this work, we used generative AI tools solely for assistance in formatting, grammatical polishing, and editing of the manuscript. All AI-assisted text was thoroughly reviewed and proofread by the authors, who take full responsibility for the final content of this work.

REFERENCES

Lindsay M Alexander, Jasmine Escalera, Lei Ai, Charissa Andreotti, Karina Febre, Alexander Mangone, Natan Vega-Potler, Nicolas Langer, Alexis Alexander, Meagan Kovacs, et al. An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific Data*, 4(1):170181, 2017.

Susan Bookheimer. Pre-surgical language mapping with functional magnetic resonance imaging. *Neuropsychology Review*, 17(2):145–155, 2007.

Matthew Brett, Alex P Leff, Chris Rorden, and John Ashburner. Spatial normalization of brain images with focal lesions using cost function masking. *NeuroImage*, 14(2):486–500, 2001.

David Calhas and Rui Henriques. EEG to fMRI synthesis benefits from attentional graphs of electrode relationships. In *MLHC*, volume 219, pp. 76–93. PMLR, 2023.

Radoslaw Martin Cichy, Gemma Roig, Alex Andonian, Kshitij Dwivedi, Benjamin Lahner, Alex Lascelles, Yalda Mohsenzadeh, Kandan Ramakrishnan, and Aude Oliva. The Algonauts Project: A platform for communication between the sciences of biological and artificial intelligence. *arXiv preprint arXiv:1905.05675*, 2019.

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

Stéphane d'Ascoli, Jérémy Rapin, Yohann Benchetrit, Hubert Banville, and Jean-Remi King. TRIBE: TRImodal brain encoder for whole-brain fMRI response prediction. In *ICLR*, 2026.

Fani Deligianni, Maria Centeno, David W Carmichael, and Jonathan D Clayden. Relating resting-state fMRI and EEG whole-brain connectomes across frequency bands. *Frontiers in Neuroscience*, 8:258, 2014.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG – adapting to any setup with large-scale pretraining on 25,000 subjects. In *NeurIPS*, volume 38, 2025.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018.

Oscar Esteban, Daniel Birman, Marie Schaer, Oluwasanmi O. Koyejo, Russell A. Poldrack, and Krzysztof J. Gorgolewski. MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLoS ONE*, 12(9):e0184661, 2017.

Oscar Esteban, Christopher J Markiewicz, Ross W Blair, Craig A Moodie, A Ilkay Isik, Asier Erramuzpe, James D Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, et al. fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, 16(1):111–116, 2019.

Alessandro T Gifford, Benjamin Lahner, Sari Saba-Sadiya, Martina G Vilas, Alex Lascelles, Aude Oliva, Kendrick Kay, Gemma Roig, and Radoslaw M Cichy. The Algonauts Project 2023 challenge: How the human brain makes sense of natural scenes. *arXiv preprint arXiv:2301.03198*, 2023.

Gary H Glover. Overview of functional magnetic resonance imaging. *Neurosurgery Clinics of North America*, 22(2):133–139, 2011.

Evan M Gordon, Timothy O Laumann, Adrian W Gilmore, Dillan J Newbold, Deanna J Greene, Jeffrey J Berg, Mario Ortega, Catherine Hoyt-Drazen, Caterina Gratton, Haoxin Sun, et al. Precision functional mapping of individual human brains. *Neuron*, 95(4):791–807, 2017.

Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, and Daguang Xu. MAISI: Medical AI for synthetic imaging. In *WACV*, pp. 4430–4441, 2025.

Dongyi He, Shiyang Li, Bin Jiang, and He Yan. Spec2VolCAMU-Net: A spectrogram-to-volume model for EEG-to-fMRI reconstruction based on multi-directional time-frequency convolutional attention encoder and vision-Mamba U-Net. *Journal of Neural Engineering*, 22(5), 2025. doi: 10.1088/1741-2552/ae15bf.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *ICLR*, 2024.

Alexander Kovalev, Ilia Mikheev, and Alexei Ossadtchi. fMRI from EEG is only deep learning away: the use of interpretable DL to unravel EEG-fMRI relationships. *arXiv preprint arXiv:2211.02024*, 2022.

Romeo Lanzino, Federico Fontana, Luigi Cinque, Francesco Scarcello, and Atsuto Maki. NT-ViT: Neural transcoding vision transformers for EEG-to-fMRI synthesis. In *ECCVW*, 2024.

Yamin Li, Ange Lou, Ziyuan Xu, Shiyu Wang, and Catie Chang. Leveraging sinusoidal representation networks to predict fMRI signals from EEG. In *Medical Imaging 2024: Image Processing*, volume 12926, pp. 129263A. SPIE, 2024a.

Yamin Li, Ange Lou, Ziyuan Xu, Shengchao Zhang, Shiyu Wang, Dario J. Englot, Soheil Kolouri, Daniel Moyer, Roza G. Bayrak, and Catie Chang. NeuroBOLT: Resting-state EEG-to-fMRI synthesis with multi-dimensional feature mapping. In *NeurIPS*, volume 37, 2024b.

Yamin Li, Ange Lou, Chang Li, Shiyu Wang, Haatef Pourmotabbed, Ziyuan Xu, Shengchao Zhang, Dario J. Englot, Soheil Kolouri, Daniel Moyer, Roza G. Bayrak, and Catie Chang. UnEBOLT: A unified model for EEG-to-BOLD translation and functional connectivity reconstruction. In *MIDL*, 2026a.

Yamin Li, Shiyu Wang, Chang Li, Ange Lou, Haatef Pourmotabbed, Sarah Goodale, Dario J Englot, Daniel Moyer, Roza G. Bayrak, and Catie Chang. Mind the state: Towards unified, context-aware EEG-to-fMRI synthesis. In *ICML*, 2026b.

Xueqing Liu and Paul Sajda. Fusing simultaneously acquired EEG and fMRI via hierarchical deep transcoding. In *International Conference on Brain Informatics*, pp. 57–67, 2023.

Zhenren Ma, Jialin Li, Jiabao Wen, and Wenjing Yan. A deep generative framework based on a diffusion model integrated with Mamba block in modality transferring from resting-state EEG to fMRI. In *Proceedings of the 2025 14th International Conference on Bioinformatics and Biomedical Science*, pp. 130–134, 2025.

Yehudit Meir-Hasson, Sivan Kinreich, Ilana Podlipsky, Talma Hendler, and Nathan Intrator. An EEG finger-print of fMRI deep regional activation. *NeuroImage*, 102:128–141, 2014.

Christoph M. Michel and Denis Brunet. EEG source imaging: a practical review of the analysis steps. *Frontiers in Neurology*, 10:325, 2019.

Michael Pereira, Nathan Faivre, Iñaki Iturrate, Marco Wirthlin, Luana Serafini, Stéphanie Martin, Arnaud Desvachez, Olaf Blanke, Dimitri Van De Ville, and José del R Millán. Disentangling the origins of confidence in speeded perceptual judgments through multimodal imaging. *Proceedings of the National Academy of Sciences*, 117(15):8382–8390, 2020.

Russell A Poldrack, Chris I Baker, Joke Durnez, Krzysztof J Gorgolewski, Paul M Matthews, Marcus R Munafò, Thomas E Nichols, Jean-Baptiste Poline, Edward Vul, and Tal Yarkoni. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, 18(2):115–126, 2017.

Kristofer Grover Roos, Atsushi Fukuda, and Quan Huu Cap. From brainwaves to brain scans: A robust neural network for EEG-to-fMRI synthesis. *arXiv preprint arXiv:2502.08025*, 2025.

Jennifer M Walz, Robin I Goldman, Michael Carapezza, Jordan Muraskin, Truman R Brown, and Paul Sajda. Prestimulus EEG alpha oscillations modulate task-related fMRI BOLD responses to auditory stimuli. *NeuroImage*, 113:153–163, 2015.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *ICLR*, 2025.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, 2018.

Weiheng Yao, Zhihan Lyu, Mufti Mahmud, Ning Zhong, Baiying Lei, and Shuqiang Wang. CATD: Unified representation learning for EEG-to-fMRI cross-modal generation. *IEEE Transactions on Medical Imaging*, 44(7):2757–2767, 2025. doi: 10.1109/TMI.2025.3550206.

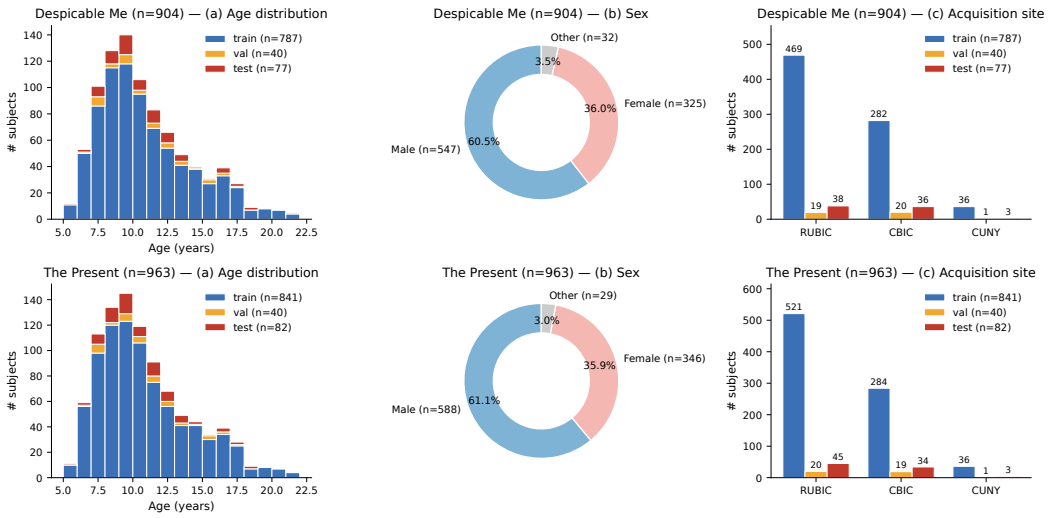

Figure 6: **Per-film demographic breakdown.** Top row: *Despicable Me* (n=904). Bottom row: *The Present* (n=963). Same columns in each row: (a) age distribution stacked by that film's train/val/test split, (b) sex composition, (c) per-site subject counts by split.

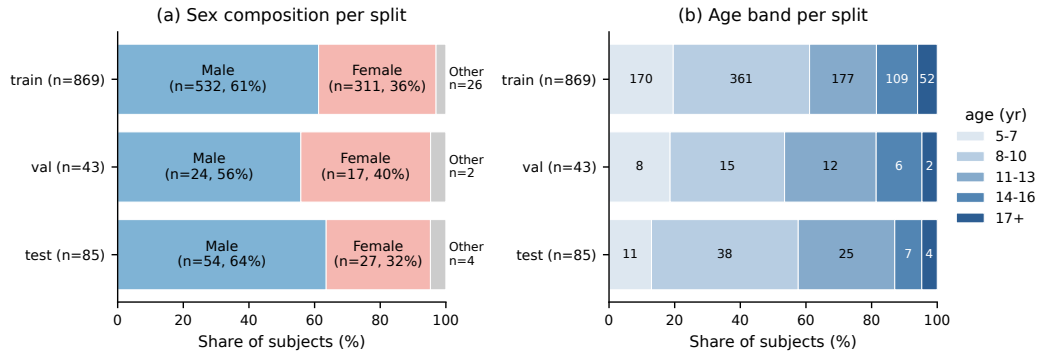

Figure 7: **Splits are demographically aligned.** (a) Sex composition is matched across train, val, and test (within ± 6 pp of the cohort mean of 61.2% Male); the small grey segments are the 32 self-reported *Other* subjects. (b) Age-band composition is likewise matched: each split is dominated by the 8–13 yr range that constitutes the bulk of HBN, with consistent tails at the 5–7 yr and $14+$ yr ends. These matched marginals support the design of subject-disjoint, demographically comparable evaluation splits.

A Cohort Details

The two per-film cohorts of NEST inherit HBN's transdiagnostic design. Beyond the age, sex, and site summaries reported in the main paper, the union of the two cohorts (n=997) is racially and ethnically heterogeneous (503 White, 119 Black, 26 Asian, 184 Other, 165 unreported; 249 Hispanic or Latino, 639 non-Hispanic, 109 unreported) and predominantly right-handed (832/997 Right, 123 Left, 10 Ambidextrous, 32 unreported) as shown in Figures 6 and 7. HBN additionally provides harmonized bifactor scores across four dimensions—a general p-factor along with internalizing, externalizing, and attention-specific factors—available for 943–947 of the 997 union subjects depending on the dimension, as the remainder are missing from the released participant table. Each score is a population z-score, so a value of 0 corresponds to the HBN mean, positive values to elevated symptomatology, and negative values to sub-population levels. Figure 8 shows the empirical distributions of the four scores, stacked by split. The cohort spans the full clinical-to-typical spectrum on every dimension, and the train/val/test partition covers the same range on each.

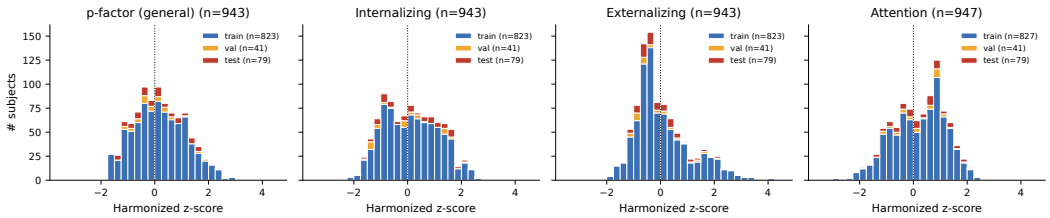

Figure 8: **Transdiagnostic spectrum of the NEST cohort.** Distributions of HBN harmonized bifactor z-scores stacked by split ($n = 943$–947 of the 997 subjects due to missing scores in the original HBN release). Train, validation, and test splits span identical ranges across both subclinical and elevated symptom profiles.

Table 4: Per-film cohorts and their subject-disjoint splits. A single subject-level partition of the 997-subject union pool (869/43/85) is intersected with each film's cohort to produce the per-film train/val/test splits, so a subject holds the same role in both films. The last row counts subjects present in both films. These are the splits used in every experiment of Section 5.

Cohort	Train	Val	Test	Total
Despicable Me	787	40	77	904
The Present	841	40	82	963
Both films (subset)	759	37	74	870

B Architecture Details of NESTOR

This section details the EEG adapter $\mathcal{A}$ and the spatiotemporal blocks $F_1, \ldots, F_4$ of Section 4, using its notation. All hidden features have width d=128.

B.1 EEG Adapter

REVE (El Ouahidi et al., 2025) splits each 32 s chunk of 200 Hz EEG into 35 overlapping 1 s patches and encodes every patch of every electrode into a 512-dimensional token. The 96 s window thus gives $\mathbf{z}^{\text{eeg}} \in \mathbb{R}^{C\times P\times 512}$ with C=129 electrodes and P=105 tokens per electrode, and $\mathbf{z}_j^{\text{eeg}} \in \mathbb{R}^{P\times 512}$ denotes the tokens of electrode j. For each electrode, the adapter computes

$$\mathbf{e}_j = \text{Interp}_{P\to T_f}\left(\text{FC}\left(\text{LN}(\mathbf{z}_j^{\text{eeg}})\right) + \boldsymbol{\epsilon}_j\right) \in \mathbb{R}^{T_f\times d}, \tag{7}$$

where LN is layer normalization, FC is a linear layer followed by GELU, $\boldsymbol{\epsilon}_j \in \mathbb{R}^d$ is a learned embedding of electrode j, and $\text{Interp}_{P\to T_f}$ interpolates linearly along time from the P token times to the T_f frame times. We write $\mathbf{e}_{t,j} \in \mathbb{R}^d$ for the t-th row of $\mathbf{e}_j$ and $\mathbf{e}_t = (\mathbf{e}_{t,1}, \ldots, \mathbf{e}_{t,C})$ for the C electrode tokens of frame t.

B.2 Spatiotemporal Blocks

For $\ell \geq 1$, the hidden representation $\mathbf{h}^{(\ell)} \in \mathbb{R}^{T_f\times V\times d}$ holds one feature vector per frame and latent voxel. Before its first update, F_1 maps the replicated sMRI latent to width d and adds an EEG shift and positional embeddings,

$$\mathbf{h}_{t,v} \leftarrow \text{FC}\left(\mathbf{z}_v^{\text{sMRI}}\right) + \text{Lin}\left(\text{LN}(\bar{\mathbf{e}}_t)\right) + \mathbf{p}_v^{\text{space}} + \mathbf{p}_t^{\text{time}}, \qquad \bar{\mathbf{e}}_t = \frac{1}{C}\sum_{j=1}^{C}\mathbf{e}_{t,j}, \tag{8}$$

where $\mathbf{z}_v^{\text{sMRI}} \in \mathbb{R}^4$ is the sMRI latent at voxel v. The second term shifts all voxels of frame t by the same vector. For the positional embeddings, each voxel coordinate and the frame index are normalized to $[0, 1]$ and encoded by 256 sinusoidal features $\omega(\cdot)$ with frequencies $10000^{-k/128}$, $k = 0, \ldots, 127$. A separate two-layer perceptron f_i with SiLU (Elfwing et al., 2018) then maps these

Table 5: Demographics of the shared partition on the union pool (n=997). Sex includes HBN's self-reported *Other* category. Sites: RUBIC=Rutgers, CBIC=CitiGroup/Cornell, CUNY=City University of New York; Staten Island contributes no paired sessions and is omitted.

Split	# Subj.	# M / F / Oth.	Age median	Age range	RUBIC	CBIC	CUNY
Train	869	532/311/26	10.1	5.0–21.9	539	294	36
Val	43	24/17/2	10.3	6.9–21.2	22	20	1
Test	85	54/27/4	10.1	5.5–18.9	46	36	3
Total	997	610/355/32	10.1	5.0–21.9	607	350	40

Table 6: Quality-control summaries by film and split. fMRI rows use the MRIQC metrics (Esteban et al., 2017) distributed with HBN (Alexander et al., 2017) on the given film's run (FD denotes mean framewise displacement, TSNR temporal signal-to-noise ratio), sMRI rows MRIQC T1-weighted metrics restricted to the same per-film cohort (CNR contrast-to-noise, SNR total signal-to-noise ratio). Values are median (IQR) or percentage of subjects, and the first row of each block is the number of subjects with the metric available, since HBN's MRIQC tables are incomplete. Motion is substantial for this pediatric cohort. Users can threshold on the released summaries without changing the splits.

Film	Modality	Metric	Train	Validation	Test	Total
Despicable Me	# subjects in cohort		787	40	77	904
	fMRI	subjects with MRIQC	671	32	68	771
		FD mean (mm)	0.26 (0.17–0.46)	0.27 (0.16–0.48)	0.30 (0.18–0.47)	0.26 (0.17–0.46)
		FD mean > 0.5 mm	23%	25%	24%	23%
		TSNR	22.0 (14.5–29.4)	19.6 (14.8–29.2)	21.2 (13.6–29.7)	21.8 (14.4–29.4)
	sMRI	subjects with MRIQC	713	36	72	821
		T1w CNR	1.57 (1.26–1.80)	1.54 (1.22–1.73)	1.53 (1.26–1.78)	1.56 (1.26–1.79)
The Present	# subjects in cohort		841	40	82	963
	fMRI	subjects with MRIQC	720	34	73	827
		FD mean (mm)	0.23 (0.16–0.40)	0.26 (0.16–0.46)	0.24 (0.19–0.42)	0.24 (0.17–0.41)
		FD mean > 0.5 mm	19%	21%	19%	19%
		TSNR	26.8 (18.9–34.4)	26.6 (17.9–33.8)	25.7 (17.5–29.7)	26.8 (18.7–34.0)
	sMRI	subjects with MRIQC	762	36	78	876
		T1w CNR	1.57 (1.26–1.81)	1.54 (1.22–1.76)	1.56 (1.27–1.83)	1.57 (1.26–1.81)

features to width d for each axis i,

$$\mathbf{p}_v^{\text{space}} = \sum_{i=1}^{3} f_i\big(\omega(\xi_{v,i})\big), \qquad \mathbf{p}_t^{\text{time}} = f_4\Big(\omega\big(\tfrac{t-1}{T_f-1}\big)\Big), \tag{9}$$

where $(\xi_{v,1}, \xi_{v,2}, \xi_{v,3})$ are the normalized coordinates of voxel v. The temporal embedding thus encodes the position of a frame within the window, not within the film.

Each block F_ℓ then applies four updates in sequence,

$$\mathbf{h} \leftarrow \big(\mathbf{h} + \gamma(\mathbf{z}^{\text{sMRI}})\big) \odot (\mathbf{1} + \boldsymbol{\sigma}_\ell) + \boldsymbol{\beta}_\ell, \tag{10}$$

$$\mathbf{h}_{:,v} \leftarrow \mathbf{h}_{:,v} + \mathrm{MHA}\,\big(\mathrm{LN}(\mathbf{h}_{:,v}),\, \mathrm{LN}(\mathbf{h}_{:,v})\big), \qquad v = 1, \ldots, V, \tag{11}$$

$$\mathbf{h}_{t,:} \leftarrow \mathbf{h}_{t,:} + \mathrm{MHA}\,\big(\mathrm{LN}(\mathbf{h}_{t,:}),\, \mathrm{LN}(\mathbf{e}_t)\big), \qquad t = 1, \ldots, T_f, \tag{12}$$

$$\mathbf{h}_{t,:} \leftarrow \mathbf{h}_{t,:} + \mathrm{ConvBlock}(\mathbf{h}_{t,:}), \qquad t = 1, \ldots, T_f. \tag{13}$$

Here $\mathbf{h}_{:,v} \in \mathbb{R}^{T_f \times d}$ and $\mathbf{h}_{t,:} \in \mathbb{R}^{V \times d}$ are the features of voxel v across frames and of frame t across voxels. $\mathrm{MHA}(\mathbf{X}, \mathbf{Y})$ is multi-head attention with 8 heads, queries from $\mathbf{X}$, and keys and values from $\mathbf{Y}$, and each LN has its own parameters.

Let $\mathrm{Conv}^k_{a\to b}$ denote a $k\times k\times k$ convolution from a to b channels. The first update injects the anatomy into every block through the encoder $\gamma = \mathrm{Conv}^1_{32\to d} \circ \mathrm{GELU} \circ \mathrm{Conv}^3_{4\to 32}$, shared by all blocks, and then applies a learned channel-wise scale $\boldsymbol{\sigma}_\ell \in \mathbb{R}^d$ and shift $\boldsymbol{\beta}_\ell \in \mathbb{R}^d$ of block ℓ. The second update is the self-attention across frames, which treats the T_f frames of each latent voxel as a sequence without masking, so that a frame can depend on EEG from anywhere in the window. The third update is the cross-attention to the EEG embedding, in which the V voxels of frame t attend to the C electrode tokens $\mathbf{e}_t$ of the same frame. Its output projection is initialized to zero, so that it contributes nothing

at initialization. The fourth update is the 3D convolution within each frame on the $18\times22\times18$ latent grid,

$$\mathrm{ConvBlock} = \mathrm{Conv}^{1}_{d\to d} \circ \mathrm{GN} \circ \mathrm{GELU} \circ \mathrm{Conv}^{3}_{d\to d} \circ \mathrm{Conv}^{1}_{d\to d} \circ \mathrm{GN} \circ \mathrm{GELU}, \tag{14}$$

where GN (Wu & He, 2018) is group normalization with 32 groups and the $3\times3\times3$ convolution splits the channels into two groups. As the blocks have no separate feed-forward layer, this convolution also serves that role.

B.3 Low-Rank Spatiotemporal Offset

The coefficient vector $\mathbf{c}$ of the offset $\mathbf{P}$ in Eq. (3) is predicted from the EEG embedding by attention pooling over all frames and electrodes of the window,

$$\mathbf{c} = \mathrm{FC}\Big(\mathrm{LN}\Big(\sum_{t=1}^{T_f}\sum_{j=1}^{C}\alpha_{t,j}\,\mathbf{e}_{t,j}\Big)\Big), \qquad \alpha_{t,j} = \frac{\exp\left(\mathbf{q}^\top \mathbf{e}_{t,j}/\sqrt{d}\right)}{\sum_{t'=1}^{T_f}\sum_{j'=1}^{C}\exp\left(\mathbf{q}^\top \mathbf{e}_{t',j'}/\sqrt{d}\right)}, \tag{15}$$

where $\mathbf{q} \in \mathbb{R}^d$ is a learned query and FC maps from d to K channels. A single $\mathbf{c}$ is thus predicted for the whole window.

The offset is then formed by weighting each of the K learned components by its coefficient,

$$\mathbf{P} = \sum_{k=1}^{K} c_k\, \mathbf{b}^{\mathrm{vol}}_k \big(\mathbf{b}^{\mathrm{time}}_k\big)^\top, \qquad P_{v,t} = \sum_{k=1}^{K} c_k\, B^{\mathrm{vol}}_{v,k}\, B^{\mathrm{time}}_{t,k}, \tag{16}$$

so that entry (v, t) of $\mathbf{P}$ is the sum over components of the coefficient, the component's value at latent entry v, and its value at frame t. Each volumetric component $\mathbf{b}^{\mathrm{vol}}_k$ spans all 4 channels of the $18\times22\times18$ latent grid, and each temporal component $\mathbf{b}^{\mathrm{time}}_k$ is indexed by the position of a frame within the window, not within the film. $\mathbf{P}$ is reshaped to $T_f\times4\times18\times22\times18$ and added to the output of the MLP in the latent space, before decoding with $\mathcal{D}$.

C Implementation Details

C.1 EEG Preprocessing

We cut the EEG recordings to the stop and start markers denoted by subject metadata, giving the 128 recording channels and the Cz reference at 500 Hz. The segment is re-referenced to the common average, which makes Cz an ordinary data channel, so the model sees 129 channels. It is then high-pass filtered at 0.3 Hz to remove slow drifts, notch filtered at 60 Hz to remove power-line interference, and resampled to 200 Hz with a polyphase anti-aliasing filter. We clip amplitudes to $\pm500\,\mu$V when loaded, to suppress the electrode-pop artifacts common in pediatric recordings. The 200 Hz rate matches the frozen REVE encoder (El Ouahidi et al., 2025), which was pretrained on 200 Hz EEG. Because REVE takes arbitrary montages, all 129 channels are passed with their EGI electrode positions from REVE's released position bank, without channel selection or interpolation.

C.2 MRI Preprocessing

T1w and BOLD runs go through fMRIPrep 25.2.3 (Esteban et al., 2019). Only its T1w-space outputs enter NEST, so no sMRI or fMRI volume is warped to a template. fMRIPrep also estimates each subject's normalization to MNI152NLin2009cAsym, which we use only for evaluation. Initial non-steady-state volumes keep their indices but are excluded from every window. Each subject's grid is $72\times88\times72$ voxels at 2.4 mm, a $172.8\times211.2\times172.8$ mm field of view centred on the brain centroid. We resample volumes trilinearly onto it and normalize them to $[0, 1]$ using the per-run minimum and maximum, which we store and reuse at evaluation.

The DiFuMo-64 atlas (Dadi et al., 2020) is defined on the MNI152NLin6Asym template, so the atlas moves to the subject rather than the subject to the atlas. We compose a transform from MNI152NLin6Asym with the inverse of the fMRI mapping to MNI152NLin2009cAsym, applying it per-subject to the DiFuMo-64 soft masks with linear interpolation, and clip them at zero. The same maps pooled onto the latent grid define the ROI projection used in the training loss. At 9.6 mm the latent grid is coarse relative to acquisition, but DiFuMo modes are smooth over centimetres, so their weighted means survive the pooling, and no decode step is needed per training window.

Table 7: Training configuration of the three baselines. Architectures and inputs are described in Section 5. Unlisted UnEBOLT and E2fNet settings follow their released code, including UnEBOLT's per-run centering and scaling of each ROI target. r_t and r_s denote the Pearson correlation between predicted and ground-truth ROI signals along time and across ROIs.

	NESTOR	**UnEBOLT**	**E2fNet**
Backbone	width 128, depth 4, 8 heads	LaBraM-base, linear head	released encoder–decoder
Input per target	96 s EEG window, sMRI latent	trailing 16 s EEG, 26 channels	spectra of trailing 16 s EEG
Output	$4\times120\times18\times22\times18$ latent	64 ROI signals per frame	one $72\times88\times72$ volume
Training samples	120-frame windows, 4 s stride	every eligible frame	every eligible frame
Loss	Eq. (6)	$0.8\,\text{MSE} + 0.1(1-r_t) + 0.1(1-r_s)$	$0.5\,\text{MSE} + 0.5\,\text{SSIM}$ loss
Optimizer	AdamW	AdamW, layer-wise LR decay 0.65	AdamW
Learning rate	0.001	0.0003	0.001
Weight decay	0.01	0.05, none on bias/norm/tokens	0.01
Warm-up	none	5 epochs	50 updates
Schedule	cosine to 0	cosine to 0.000001	cosine, 5 restart cycles
Gradient clipping	norm 1.0	none	norm 1.0
Batch	2 windows	within-run blocks of ≤ 64 frames	32 frames
Epochs	20	40	50
Validation Criteria	temporal r	loss	SSIM

C.3 Latent and ROI Projection

We use the pre-trained MAISI autoencoder (Guo et al., 2025) without modifying its released weights. MAISI is a MONAI-based KL autoencoder with three resolution levels (64, 128, and 256 channels), two GroupNorm–SiLU residual blocks per level, no attention, and nearest-neighbor upsampling paired with $3\times3\times3$ convolutions between stages. Its two downsampling stages provide a $4\times$ spatial compression, mapping the 2.4 mm native grid ($72\times88\times72$) to a $4\times18\times22\times18$ latent representation.

When loading latents, we normalize to unit variance, and invert the normalization before decoding during prediction. The ROI projection matrix in the loss is fitted exclusively on training runs, separately for each movie condition. Specifically, for each of the 64 modes, a ridge regression ($\lambda = 0.0001 \times$ mean feature variance) maps the four latent channels pooled under that mode to the measured ROI signal, yielding a 64×4 weight matrix. Single-film models never access the alternate film's fMRI, even for computing these projection statistics.

C.4 Evaluation Protocol

The translation baseline trains on windows starting every 5 frames and keeps an exponential moving average (EMA) of its weights with decay 0.999. We report one of its 20 epoch checkpoints with the highest temporal r on the validation subjects. At test time, its windows start every 5 frames, each window is decoded onto the subject's 2.4 mm grid, and overlapping decoded volumes are averaged per frame. UnEBOLT and E2fNet predict each frame directly as they are originally trained. All models are scored on frames 20–199 (DM) and 20–239 (TP). Voxel metrics use the positive support of the subject's DiFuMo maps, and regional signals are DiFuMo-weighted means of the volumes. FC is left undefined for a subject if any region lacks a valid prediction. REVE's pretraining data include HBN EEG, and we did not audit participant overlap, so held-out refers to the fMRI targets and model selection.

D Results on Low-Motion Runs

Head motion is a well-documented confound in fMRI that can induce spurious temporal correlations and distort functional connectivity estimates. To verify that model predictions reflect true neural dynamics rather than motion artifacts, we evaluate performance on the low-motion subset of test runs (mean framewise displacement < 0.2 mm; Table 8). Across all models and both films, filtering out high-motion runs substantially reduces functional connectivity error (FC MSE), with Translation and UnEBOLT showing notable gains in temporal correlation.

Table 8: Test performance on low-motion runs (mean FD < 0.2 mm), averaged across subjects, with 33 (DM) and 30 (TP) subjects. † UnEBOLT directly predicts the brain signals for each RoI.

Movie	Model	Voxel-level			Region-level (DiFuMo-64)		
		MSE ($\times 10^{-3}$) ↓	PSNR ↑	SSIM ↑	Temporal r ↑	FC r ↑	FC MSE ↓
DM	Translation	**9.589**	**21.47**	**0.7841**	**0.3434**	**0.4037**	**0.1092**
	UnEBOLT	—†	—	—	0.0963	0.3528	0.1262
	E2fNet	14.550	19.15	0.6336	−0.0385	0.1756	0.1302
TP	Translation	**7.489**	**22.63**	**0.7724**	**0.4334**	0.3424	0.1487
	UnEBOLT	—	—	—	0.1131	**0.3570**	0.1540
	E2fNet	13.641	19.08	0.6130	−0.0131	0.2199	**0.1357**

Table 9: Ablation study on training loss terms in NESTOR, averaged across subjects. All variants include the base voxel reconstruction loss $\mathcal{L}_{\text{mse}}$. Omitting the regional correlation loss $\mathcal{L}_{\text{corr}}$ collapses temporal tracking ($r \approx 0$), while omitting the temporal Gram loss $\mathcal{L}_{\text{gram}}$ substantially degrades functional connectivity (FC) recovery. The full composite objective yields the strongest functional connectivity preservation while maintaining robust temporal tracking. Best results per movie and metric are shown in bold.

Movie	Loss Terms			Voxel-level			Region-level (DiFuMo-64)		
	$\mathcal{L}_{\text{mse}}$	$\mathcal{L}_{\text{corr}}$	$\mathcal{L}_{\text{gram}}$	MSE ($\times 10^{-3}$) ↓	PSNR ↑	SSIM ↑	Temporal r ↑	FC r ↑	FC MSE ↓
DM	✓	✓		**8.042**	**22.29**	0.8045	**0.3358**	0.3107	0.2277
	✓		✓	8.533	22.01	**0.8071**	−0.0015	0.1910	**0.1501**
	✓	✓	✓	8.102	22.12	0.7950	0.2896	**0.3473**	0.1643
TP	✓	✓		**7.548**	**22.50**	0.7870	**0.3136**	0.2156	0.2163
	✓		✓	8.533	22.36	**0.7924**	0.0607	0.1935	0.2309
	✓	✓	✓	7.557	22.49	0.7824	0.3057	**0.2459**	**0.2039**

E Ablation Study on Objective Function

We evaluate the individual contributions of each component in our composite training objective: voxel-level mean squared error ($\mathcal{L}_{\text{mse}}$), regional temporal correlation ($\mathcal{L}_{\text{corr}}$), and the temporal Gram matrix loss ($\mathcal{L}_{\text{gram}}$). As shown in Table 9, omitting $\mathcal{L}_{\text{corr}}$ collapses temporal tracking ($r \approx 0$), while omitting $\mathcal{L}_{\text{gram}}$ degrades functional connectivity recovery. The full composite objective yields the strongest functional connectivity preservation while maintaining robust temporal tracking.

F Access and License

Access. Code, configuration, frozen splits, quality-control tables, and baseline weights are available at `https://anonymous.4open.science/r/NESTOR`. The repository rebuilds NEST from the public HBN release with the inclusion filters, preprocessing, stimulus-locked pairing, and evaluation protocol described in this paper. Raw recordings are not re-hosted and remain governed by the HBN data-use agreement.

License and ethics. NEST inherits HBN's data-use terms[3] (Alexander et al., 2017): HBN releases data under Creative Commons BY 4.0 for participants who consented to commercial use and Creative Commons BY-NC-SA for those who did not. We adopt the stricter Creative Commons BY-NC-SA license to cover all subjects released in NEST. The underlying acquisitions were collected under IRB protocols approved by the Child Mind Institute.

[3] `https://fcon_1000.projects.nitrc.org/indi/cmi_healthy_brain_network/MRI_EEG.html`