# Estimating Brain Activity with High Spatial and Temporal Resolution using a Naturalistic MEG-fMRI Encoding Model

**Beige Jerry Jin**
Carnegie Mellon University
jerryjin@andrew.cmu.edu

**Leila Wehbe**
Carnegie Mellon University
lwehbe@cmu.edu

## Abstract

Current non-invasive neuroimaging techniques trade off between spatial resolution and temporal resolution. While magnetoencephalography (MEG) can capture rapid neural dynamics and functional magnetic resonance imaging (fMRI) can spatially localize brain activity, a unified picture that preserves both high resolutions remains an unsolved challenge with existing source localization or MEG-fMRI fusion methods, especially for single-trial naturalistic data. We collected whole-head MEG when subjects listened passively to more than seven hours of narrative stories, using the same stimuli in an open fMRI dataset [LeBel et al., 2023]. We developed a transformer-based encoding model that combines the MEG and fMRI from these two naturalistic speech comprehension experiments to estimate latent cortical source responses with high spatiotemporal resolution. Our model is trained to predict MEG and fMRI from multiple subjects simultaneously, with a latent layer that represents our estimates of reconstructed cortical sources. Our model predicts MEG better than the common standard of single-modality encoding models, and it also yields source estimates with higher spatial and temporal fidelity than classic minimum-norm solutions in simulation experiments. We validated the estimated latent sources by showing their strong generalizability across unseen subjects and modalities. Estimated activity in our source space predicts electrocorticography (ECoG) better than an ECoG-trained encoding model in an entirely new dataset. By integrating the power of large naturalistic experiments, MEG, fMRI, and encoding models, we propose a practical route towards millisecond-and-millimeter brain mapping.

## 1 Introduction

Non-invasive neuroimaging is central to cognitive neuroscience, yet each modality remains constrained by a fundamental trade-off between spatial and temporal resolution. Magnetoencephalography (MEG), for instance, which is sensitive to the magnetic field induced by postsynaptic current in groups of spatially aligned neurons, offers millisecond-scale temporal precision but suffers from poor spatial detail. Conversely, the blood-oxygen-level dependent (BOLD) signal measured by functional magnetic resonance imaging (fMRI) provides millimeter-scale spatial maps but reflects a sluggish hemodynamic response that integrates neural activity over seconds [Hall et al., 2014]. Bridging their complementary strengths to obtain a unified, high spatiotemporal resolution view of neural source activity is critical for understanding complex processes such as speech comprehension, which recruits multiple subprocesses unfolding on the order of milliseconds across distributed cortical networks.

However, effectively integrating MEG and fMRI remains an unsolved challenge, particularly for single-trial naturalistic data. Conventional source localization methods often use fMRI data to

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Data on the Brain & Mind.

constrain the mathematically ill-posed MEG inverse problem. For example, fMRI activation maps can serve as spatial priors in minimum-norm estimation (MNE) to improve source localization [Liu et al., 1998, Dale et al., 2000, Suzuki and Yamashita, 2021, Moradi et al., 2024]. While commonly used for simple, event-related designs, these approaches ignore the stimuli and are often ineffective in recovering the continuously evolving neural dynamics in more naturalistic settings. A more promising approach, called *neurogenerative modeling*, aims to directly model cortical sources and then generate MEG and fMRI signals forward [Huster et al., 2012]. Yet, prominent frameworks like The Virtual Brain often rely on complex biophysical models and anatomical connectomes, which require much prior knowledge and make parameter fitting inefficient. Furthermore, their focus has largely been on resting-state dynamics rather than encoding responses to external stimuli, limiting their applicability to task-based naturalistic experiments [Ritter et al., 2013, Patow et al., 2024, Hashemi et al., 2025].

In this work, we introduce a novel framework that combines the strength of the neurogenerative approach with deep learning-based encoding model paradigms. We propose a novel encoding model that predicts MEG and fMRI signals for multiple subjects as a function of stimulus features, constrained by the requirement that both modalities originate from the same source estimates in a latent source space. This is achieved by incorporating anatomical information and biophysical forward models for MEG and fMRI. Thus, we effectively estimate the source activity that is high-resolution in both time and space. Crucially, the resulting source estimates are intended to be approximations that are physiologically faithful to the true brain signals, rather than a transformation of them. Note that the ground-truth neural activity at this resolution is inaccessible given current non-invasive techniques. Instead, we validate our estimates by showing that they generalize across experiments and subjects to predict invasive data recorded with electrocorticography (ECoG) from epileptic patients. Although ECoG provides a very partial coverage of the brain, each electrode provides signals that are highly resolved in time and space, and is thus an ideal test bed for our estimated brain space. We find that our model can produce powerful predictions of ECoG signal, outperforming models trained directly on ECoG data. This validates the promise of our approach in faithfully recovering the underlying brain source activity.

## 2 Methods

### 2.1 Source Space

As a standard practice in source estimation, we first define *source spaces* that specify the location of possible neural sources. For each subject, we construct a subject-specific source space according to their structural MRI scan with an octahedron-based subsampling method from `MNE-Python` [Gramfort et al., 2014]. This process yields a set of equally spaced sources on the cortical surface[1]. We also define a source space on the "fsaverage" brain, a standard "average" brain template in `FreeSurfer` [Fischl, 2012], using the same procedure. Furthermore, each source is modeled as an *equivalent current dipole*, a common approximation for the net postsynaptic currents generated by a small group of spatially aligned and synchronously firing neurons. Thus, *source estimates* refer to the estimated amplitudes of these dipoles over time. We additionally assume that all dipoles are oriented perpendicularly to the brain surface.

With these anatomically-derived source spaces, we compute two matrices. Using `MNE-Python`, we calculate the source morphing matrix $\boldsymbol{M}^{\mathrm{S}_i}$, which transforms the source estimates from the "fsaverage" source space to subject $\mathrm{S}_i$'s subject-specific source space. If $\mathrm{S}_i$ has MEG recordings, we also compute the lead-field matrix $\boldsymbol{L}^{\mathrm{S}_i}$, which maps $\mathrm{S}_i$'s source estimates to MEG sensor signals according to Maxwell's equations.

### 2.2 Model Architecture

We build a transformer-based encoding model to predict MEG and fMRI simultaneously for multiple subjects from stimulus input and source estimates, as schematized in Figure 1.

**Input Layer** We use three concatenated feature streams to represent the naturalistic stories and serve as model input. The first feature space consists of a 768-dimensional contextual word embedding,

[1]This discretization is a standard step in source localization as the exact location of brain sources cannot be precisely pinpointed and has to be hypothesized as a mesh.

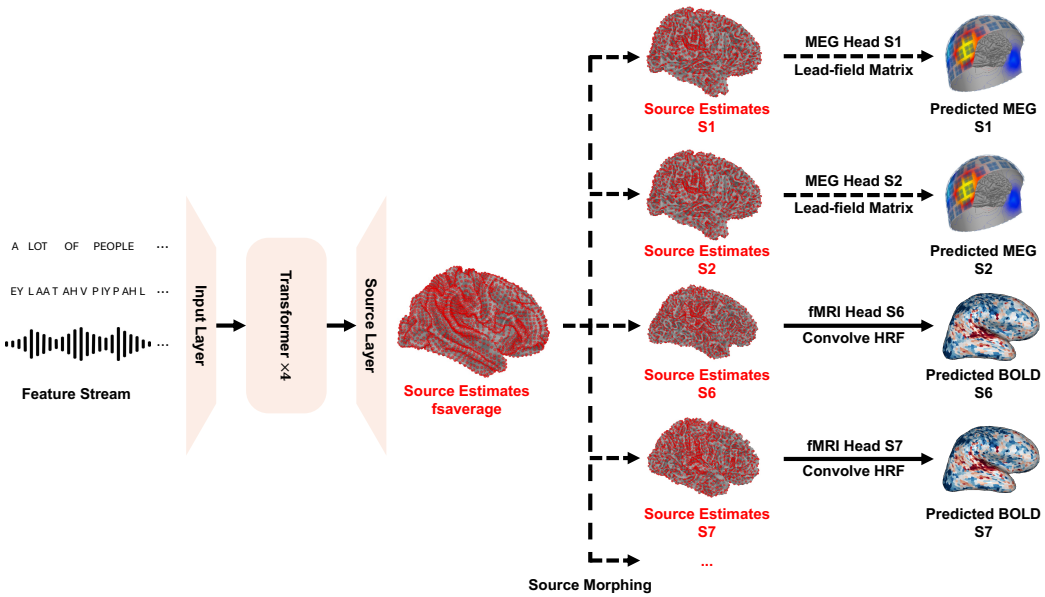


Figure 1: **Architecture of the MEG-fMRI encoding model.** Feature streams enter the network through the input layer and traverse four transformer layers before being projected into the "fsaverage" source space by the source layer. The source estimates in the "fsaverage" source space are then transformed into subject-specific source estimates by the source morphing matrix. The MEG head predicts sensor signals by multiplying the source estimates by the lead-field matrix. The fMRI head predicts BOLD responses by convolving the downsampled envelope of the source estimates with a learnable hemodynamic response function (HRF) kernel. The MEG and fMRI of multiple subjects (e.g., S1, S2, ...) are predicted simultaneously. Under the joint constraints of MEG and fMRI from multiple subjects, our model recovers the source estimates with high spatiotemporal resolution. Dashed arrows indicate steps that are pre-computed and not learnable.

obtained from the seventh hidden layer of GPT-2 [Radford et al., 2019] with a context window of 20 tokens. The second feature space is a phoneme feature space consisting of 44-dimensional one-hot vectors, where each dimension represents a phoneme in the CMU Pronouncing Dictionary or a non-speech sound [LeBel et al., 2023]. The third feature space is a 40-dimensional space of mel-spectrograms spanning 0-10 kHz, representing the perceived audio sound. These yield feature vectors $\boldsymbol{x}_t \in \mathbb{R}^{d_{\text{stim}}}$ with $d_{\text{stim}} = 852$, sampled at 50 Hz. The contextual word embeddings are repeated for the entirety of the word duration, and the phoneme embedding is repeated for the entirety of the phoneme duration. Then, the feature vectors go through the linear input layer before entering the transformer: $\boldsymbol{z}_t^{\text{in}} = \boldsymbol{W}^{\text{in}}\boldsymbol{x}_t + \text{bias}$, where $\boldsymbol{z}_t^{\text{in}} \in \mathbb{R}^{d_{\text{model}}}$, $\boldsymbol{W}^{\text{in}} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{stim}}}$ and $d_{\text{model}} = 256$.

**Transformer Encoder** To capture the dependency between features and feature-dependent latency in neural responses, we use four standard transformer encoder layers with two heads, feed-forward size = 512, and dropout = 0.2. Each attention block uses a causal sliding window of 500 tokens, so that the transformer could make use of the preceding 10 s of the stimulus features. We add learnable positional embeddings to the keys within each attention block, so that the transformer can learn feature-dependent neural latency.

**Source Layer and Source Morphing** The output of the transformer $\boldsymbol{z}_t^{\text{out}} \in \mathbb{R}^{d_{\text{model}}}$ is then projected to the source space through the linear source layer: $\boldsymbol{s}_t = \boldsymbol{W}^{\text{src}}\boldsymbol{z}_t^{\text{out}} + \text{bias}$, where $\boldsymbol{s}_t \in \mathbb{R}^{d_{\text{src}}}$, $\boldsymbol{W}^{\text{src}} \in \mathbb{R}^{d_{\text{src}} \times d_{\text{model}}}$ and $d_{\text{src}} = 8196$. The resulting $\boldsymbol{s}_t$ would represent the *source estimates* at time $t$ in the "fsaverage" source space. To obtain subject-specific predictions, we then use the pre-computed source morphing matrix $\boldsymbol{M}^{\text{S}_i} \in \mathbb{R}^{d_{\text{src}} \times d_{\text{src}}}$ to transform $\boldsymbol{s}_t$ into $\text{S}_i$'s source estimates $\boldsymbol{s}_t^{\text{S}_i}$: $\boldsymbol{s}_t^{\text{S}_i} = \boldsymbol{M}^{\text{S}_i}\boldsymbol{s}_t$.

**MEG Head** We get MEG sensor-wise prediction for $\mathrm{S}_i$ by $\hat{\boldsymbol{m}}_t^{\mathrm{S}_i} = \boldsymbol{L}^{\mathrm{S}_i}\, \boldsymbol{s}_t^{\mathrm{S}_i}$, where $\hat{\boldsymbol{m}}_t^{\mathrm{S}_i} \in \mathbb{R}^{d_{\text{MEG}}}$ and $\boldsymbol{L} \in \mathbb{R}^{d_{\text{MEG}} \times d_{\text{src}}}$ is the pre-computed lead-field matrix for $\mathrm{S}_i$. Importantly, $\boldsymbol{L}^{\mathrm{S}_i}$ contains subject-specific anatomical information and guarantees that $\boldsymbol{s}_t^{\mathrm{S}_i}$ represents the source estimates in our defined source space. This, together with $\boldsymbol{M}^{\mathrm{S}_i}$, in turn guarantees that $\boldsymbol{s}_t$ represents the source estimates in the "fsaverage" source space.

**fMRI Head** We calculate the envelope of source estimates as $\boldsymbol{p}_t^{\mathrm{S}_i} = |\boldsymbol{s}_t^{\mathrm{S}_i} + j\,\mathcal{H}(\boldsymbol{s}_t^{\mathrm{S}_i})|$ where $\mathcal{H}$ is the Hilbert transform and $j$ is the imaginary unit. Then, we downsample $\boldsymbol{p}_t^{\mathrm{S}_i}$ and convolve it with a double-gamma hemodynamic response function (HRF; whose parameters are learnable) to yield source-level BOLD predictions $\hat{\boldsymbol{y}}_\tau^{\mathrm{S}_i} \in \mathbb{R}^{d_{\text{src}}}$ for $\mathrm{S}_i$. Note that $\tau$ corresponds to the slower sampling rate of fMRI.

### 2.3 Stimuli

27 stories of "Moth Radio Hour" stories serve as stimuli, as in LeBel et al. [2023]. Each story has a length of around 10-15 minutes. These stories are partitioned into 21 training stories, one validation story, four test stories, and one anchor story.

### 2.4 MEG Data

We recorded whole-head MEG from five subjects (S1–S5) as they passively listened to the stories. The validation and test stories were presented twice and the anchor story five times, and the repetitions were averaged. Data were acquired on a MEGIN TRIUX scanner ($d_{\text{MEG}} = 306$, 204 planar gradiometers, 102 magnetometers, 102 triple-sensor locations) at 1 kHz. The following preprocessing steps are performed with `MNE-Python` [Gramfort et al., 2013]: (1) temporal Signal Space Separation (tSSS) [Taulu and Simola, 2006]; (2) 1–150 Hz band-pass and 60/120 Hz notch filters; (3) independent component analysis (ICA) [Hyvarinen, 1999] removal of ocular, cardiac, and audio signal artifacts; (4) downsampling to 50 Hz. We also collected an anatomical T1 scan for each subject, which we use to reconstruct the subject's cortical surface using `FreeSurfer` [Fischl, 2012].

### 2.5 fMRI Data

We use an open fMRI dataset [LeBel et al., 2023] where eight subjects (S6–S13) underwent 3 T fMRI scans while passively listening to the stories. The reconstructed cortical surface of each subject based on the anatomical T1 scan and `FreeSurfer` [Fischl, 2012] is also provided. We project voxel-level data to the cortical surface using `pycortex` [Gao et al., 2015], and then average the nearest vertices of each source to get the source-level BOLD signals.

### 2.6 Model Training

For training efficiency and memory usage, we divide training stories into pieces of 40 seconds and train with a batch size of 8. We use the last 20 seconds to calculate loss to ensure the model's full exposure to past information. Training loss $L$ is defined as a weighted sum of MEG correlation loss (weighted by the repeatability of each sensor), fMRI correlation loss (weighted by the repeatability of each source), and smoothness loss $L_{\text{smooth}}$:

$$L = \alpha_1 \left(1 - \text{corr}(\boldsymbol{m}_t, \hat{\boldsymbol{m}}_t)\right) + \alpha_2 \left(1 - \text{corr}(\boldsymbol{y}_\tau, \hat{\boldsymbol{y}}_\tau)\right) + \alpha_3 L_{\text{smooth}}$$

where $\boldsymbol{m}_t$ and $\boldsymbol{y}_\tau$ denote the MEG and fMRI data[2] respectively, and $L_{\text{smooth}}$ is the mean squared difference between $s_t$ and $s_{t-1}$. We set $\alpha_1 = 1$ and $\alpha_3 = 0.0001$ throughout the training. For the first 20 epochs, we set $\alpha_2 = 0$ to let the model train only on MEG. For later epochs, we set $\alpha_2 = 1$ to let the model train on MEG and fMRI jointly with equal importance. This MEG-first curriculum enables our model to quickly capture the complex temporal dynamics under the guidance of MEG, and then refine the spatial pattern under the constraint of fMRI. We validate our model on the validation story and perform early stopping according to the validation loss, which has the same format as $L$.

[2] For more effective training, we denoise the fMRI data for training stories using the predicted BOLD signals from the fMRI linear encoding model in Section 3. Note that we still use the non-denoised fMRI data for validation and evaluation.

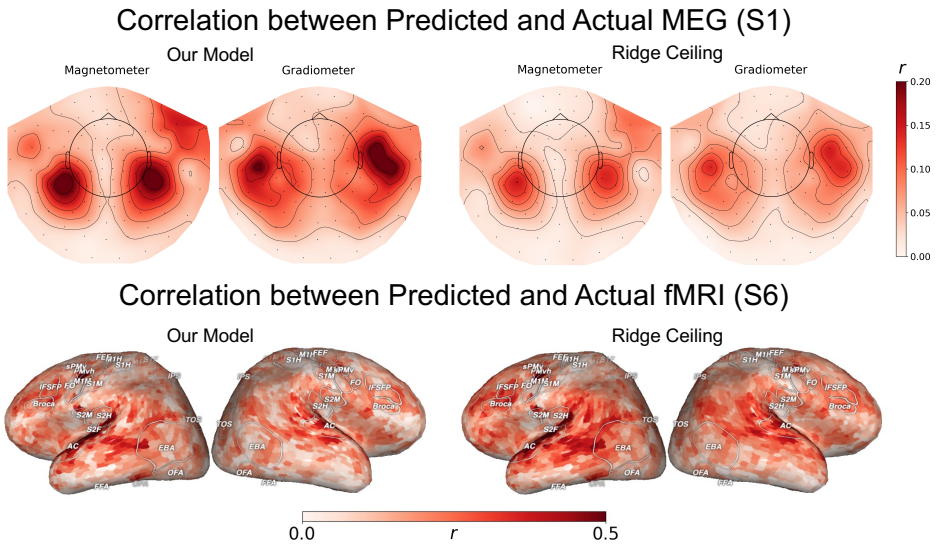


Figure 2: **Predictive performance on MEG and fMRI for two example subjects.** Our model is comparable to single-subject, single-modality ridge models, which serve as a ceiling. Top: performance on the MEG of S1. Both magnetometer and gradiometer sensors above the temporal lobe are predicted. Bottom: performance on the source-level BOLD signals of S6, shown on the inflated surface. Large parts of bilateral temporal and frontal areas are predicted.

## 3 Predictive Performance

We compare our model's predictive performance for MEG and fMRI against single-subject, single-modality linear encoding models. Since these models are not constrained by the other subject, the shared source space, or the other modality, they are not simple baselines to be surpassed, but rather a high-performance benchmark representing the upper limit ("ceiling") of what a linear model can achieve. **MEG Ridge Ceiling**: We shift the feature streams to represent the stimuli delayed by 0, 20, ..., 600 ms, and then use these 31 stimulus embedding matrices to train a stacked ridge regression model for each MEG sensor [Lin et al., 2024]. **fMRI Ridge Ceiling**: We concatenate features delayed by 2, 4, 6, and 8 s and fit a ridge regression for the source-level BOLD signals. This is the same model used for denoising training fMRI data (see Section 2.5).

We evaluate our model and the two Ridge Ceiling models by calculating the Pearson correlation $r$ between predicted and actual signals on held-out test stories for each subject. For MEG, we report $r$ per sensor; for fMRI, we report $r$ per source. The results show that our model is comparable to both Ridge Ceiling models (Figure 2 shows S1 and S6, for example). For S1's MEG, over temporal lobe sensors, the MEG Ridge Ceiling attains $r = 0.074 \pm 0.041$ whereas our model reaches $r = 0.109 \pm 0.064$. For S6's fMRI, the fMRI Ridge Ceiling yields $r = 0.267 \pm 0.074$ across the top quartile of sources, while our model achieves $r = 0.236 \pm 0.072$. The results for other subjects are reported in the appendix. We find that our model performs better than the MEG Ridge Ceiling, probably because the transformer encoder and the positional embeddings added to the keys enable our model to better capture nonlinearities and feature-dependent neural latency. However, our model performs slightly worse than the fMRI Ridge Ceiling, which might be due to the fact that our model forces all subjects to share the common "fsaverage" source estimates and does not have a mechanism to allow for spatial variance across subjects. Crucially, unavailable from either Ridge Ceiling models, our model delivers cortical source estimates with high spatial and temporal resolution.

## 4 Simulation Experiments

To test the fidelity of our source estimates under controlled conditions, we use the story feature streams to generate synthetic source activity from a linear model whose weights are randomly sampled according to the empirical feature covariance, with different lags applied to word embeddings, phonemes, and spectrograms to mimic hierarchical processing in the brain. From the synthetic

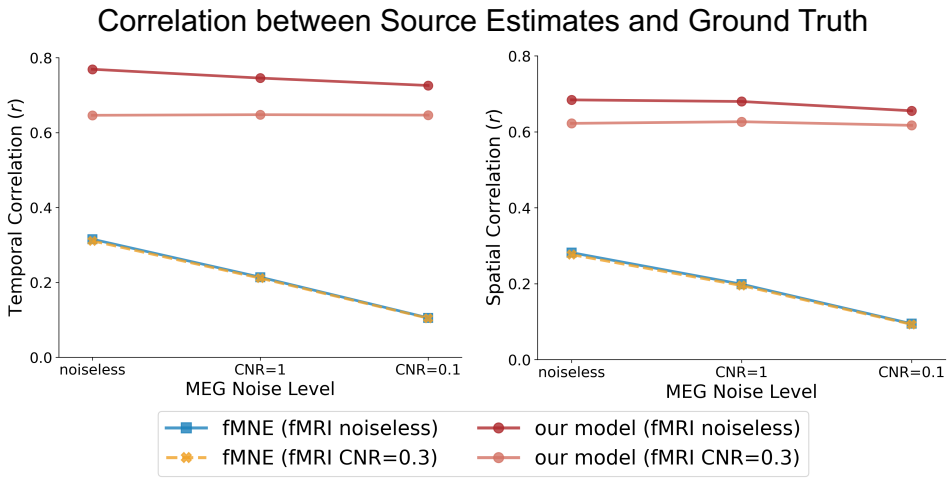


Figure 3: **Results of simulation experiments.** We report the mean Pearson $r$ computed over time within each source (left panel, temporal correlation) and over sources at each time point (right panel, spatial correlation) between the source estimates and the ground truth. Our model outperforms fMNE in both aspects under all noise levels.

source activity, we then generate MEG and fMRI signals at different noise levels measured by contrast-to-noise ratios (CNR): $\infty$ (noiseless) / 1 / 0.1 for MEG and $\infty$ (noiseless) / 0.3 for fMRI[3].

We train our model on the simulated MEG and fMRI of the training stories, and evaluate the source estimates of the test stories. We calculate the Pearson $r$ between the source estimates with the ground truth along the temporal dimension or the spatial dimension. We compare our model against fMRI-weighted MNE (fMNE) [Liu et al., 1998], which is built upon the classical MNE framework [Hämäläinen and Ilmoniemi, 1994] and incorporates fMRI activity to allow sources with higher BOLD responses to be more active. Figure 3 shows that our model recovers source activity more accurately than fMNE at all noise levels in both time and space.

## 5 ECoG Prediction

To validate the generalizability of our model, we test its predictive performance on a novel "Podcast" ECoG dataset [Zada et al., 2025], which features unseen subjects and a different neural recording modality. ECoG, which comes from a small contact in the brain, provides signals with high spatiotemporal resolution, serving as a valuable proxy for ground-truth source activity. In this dataset, neural activity was recorded from nine subjects via intracranial electrodes as they listened to a 30-minute audio podcast. Following the methodology described in Section 2.2, we extract semantic, phoneme, and spectrogram features from the audio stimulus. These features are then input into our trained model to generate source estimates in the "fsaverage" source space.

**Zero-Shot Prediction for ECoG** First, we test the model's ability to generate zero-shot predictions for ECoG signals. To achieve this, we map the electrode locations to the "fsaverage" surface and assign the time series of the nearest estimated source as the zero-shot prediction of that electrode. We then evaluate these predictions with two analyses. First, we calculate the Pearson $r$ between each predicted and actual electrode time series, with statistical significance assessed via a permutation test. The results reveal that 916 of 1268 electrodes showed a significant correlation, with the most accurately predicted electrodes located over the superior temporal sulcus (Figure 4, top left panel). Second, we design a binary classification task. For each electrode, we randomly select a one-minute prediction segment and calculate its absolute Pearson $r$ with the true corresponding ECoG segment as well as with a distractor (a non-corresponding segment from the same electrode). We then measure

[3]We define CNR as the ratio of the standard deviation of signal and noise: CNR $= \sigma_{\text{signal}}/\sigma_{\text{noise}}$ [Welvaert and Rosseel, 2013]. We choose these CNR levels because the mean CNR in our MEG data is around 0.1 and the mean CNR in our fMRI data is around 0.3.

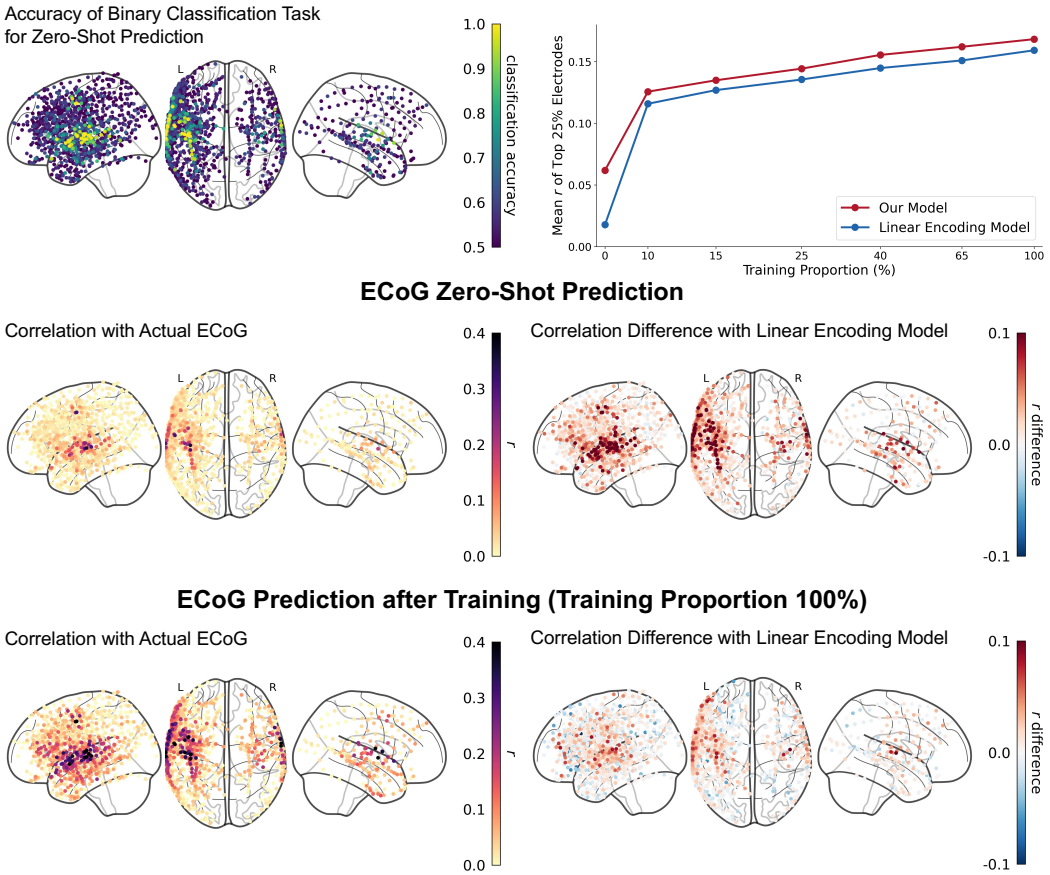


Figure 4: **Predictive performance on a new ECoG dataset.** Top left: Performance of our model's zero-shot prediction on a binary classification task. Top right: Mean Pearson $r$ of top 25% electrodes of our model and the linear encoding model under different amounts of training data. Notably, training proportion of 0% corresponds to zero-shot prediction. Middle: Correlation map of our model's zero-shot prediction and electrode-wise correlation difference with the linear encoding model. Bottom: Correlation map of our model's trained prediction with 100% training data and electrode-wise correlation difference with the linear encoding model.

how accurately the true segment could be identified by its higher correlation. Repeating this process 1000 times, we use the binomial test to determine significance. This evaluation finds that 683 of 1268 electrodes perform significantly above chance (Figure 4, middle left panel).

**Trained Prediction for ECoG** In addition to zero-shot predictions, we also train a linear mapping from the source estimates to the ECoG signals. Using a three-fold cross-validation scheme, we set aside 10 minutes of ECoG data as the test set in each fold. We then train a ridge regression model to predict each electrode from our source estimates using a varying proportion of the remaining data for training. We compare the performance of this approach against a linear encoding model for ECoG trained directly on stimulus features using stacked ridge regression [Lin et al., 2024]. As illustrated in Figure 4, when provided with an equal amount of training data, the predictions derived from our model's source estimates consistently outperform the linear encoding model. This suggests that our model provides a powerful inductive bias, generating representations that can be more readily mapped to neural activity. Notably, the performance at a training proportion of zero percent corresponds to the zero-shot prediction described in the last paragraph. In this scenario, our model's advantage is most pronounced, substantially outperforming the randomly initialized linear encoding model. Collectively, these findings demonstrate that our model's source estimates generalize effectively to new subjects and modalities, faithfully representing underlying cortical activity.

## 6 Related Work

**Source Localization** MNE [Hämäläinen and Ilmoniemi, 1994], beamformers [Van Veen et al., 1997], and Bayesian approaches [Wipf and Nagarajan, 2009] are classic source localization methods for MEG/EEG. To improve spatial precision, a common strategy is to incorporate fMRI data, using activation maps as spatial priors to bias source estimates toward hemodynamically active regions [Liu et al., 1998, Dale et al., 2000, Liu and He, 2008, Xu et al., 2018, Henson et al., 2010, Suzuki and Yamashita, 2021, Moradi et al., 2024]. However, these methods are typically applied to each time point separately, and they do not leverage information from the sensory input by design, making them less suitable for tracking the rich neural dynamics elicited by naturalistic stimuli.

**Multimodal Fusion** Instead of using fMRI to inform MEG/EEG inversion, multimodal fusion methods use symmetric models to jointly assess information from both modalities. For example, joint independent component analysis and joint tensor/matrix decomposition on MEG and fMRI could identify shared latent spatiotemporal components [Calhoun et al., 2006, Belyaeva et al., 2024]. Another popular approach is to align MEG and fMRI using representation similarity analysis, so that researchers can pinpoint MEG within a time window to particular cortical regions [Cichy et al., 2014, Cichy and Oliva, 2020, Leonardelli and Fairhall, 2022, Yeh et al., 2024]. However, these methods do not yield a direct, high-resolution estimate of the underlying neural source activity itself, nor do they explicitly model how that activity is driven by stimulus features.

**Neurogenerative Modeling** Conceptually similar to our approach, neurogenerative modeling builds models for latent neural sources and uses biophysical forward models to generate neuroimaging data [Huster et al., 2012, Castaldo et al., 2023, Kang and Park, 2024]. For example, The Virtual Brain specifies parameterized neural mass models for cortical regions, obtains connectivity between regions from anatomical scans, uses lead-field matrices and hemodynamic functions to generate MEG/EEG and fMRI data, and fits parameters to observed data [Ritter et al., 2013, Patow et al., 2024, Hashemi et al., 2025]. Key distinctions from our work are that these frameworks often rely heavily on prior knowledge about neural circuits, require complicated Bayesian inference for parameter estimation, and have primarily focused on resting-state dynamics rather than stimulus encoding.

**Encoding Models for Language** Voxel-wise or channel-wise linear encoding with single word or contextual embeddings has mapped semantic and syntactic processing in fMRI [Wehbe et al., 2014a, Huth et al., 2016, Toneva and Wehbe, 2019, Schrimpf et al., 2021, Reddy and Wehbe, 2021, Toneva et al., 2022, Caucheteux and King, 2022, Tang et al., 2023] and MEG [Wehbe et al., 2014b, Toneva and Wehbe, 2019, Toneva et al., 2022, Caucheteux and King, 2022]. These methods have mapped between embeddings and brain activity, pinpointing the brain regions that are predicted by the information in the embeddings, and the time points at which they are predicted. While MEG results have allowed researchers to paint somewhat of a spatiotemporal picture, it remains limited in the spatial resolution that it offers. The fMRI results, though with a high spatial resolution, still remain unconnected to the temporal course. Even those works that have used both fMRI and MEG have not combined them beyond comparing their results [e.g., Caucheteux and King, 2022].

## 7 Conclusions, Limitations, and Future Work

We have presented a transformer-based encoding model that successfully recovers source activity from naturalistic, multi-subject MEG and fMRI recordings. Our model demonstrates high predictive accuracy for held-out data, and its source estimates with high spatiotemporal resolution are validated through simulation experiments and ECoG prediction. By combining MEG and fMRI with a naturalistic encoding model, our work opens new avenues for non-invasively probing the dynamics of language and cognition without sacrificing spatial or temporal fidelity.

Our work has two limitations that point to future research. First, the current model relies on raw phonemes and spectrum features. While effective, incorporating contextualized representations from pretrained large audio-language models (e.g., Whisper [Radford et al., 2023]) could potentially improve the model's performance. Second, our source space is constrained to surface dipoles with fixed orientations. This simplification prevents the model from capturing activity in subcortical structures or from non-perpendicular cortical sources. Future work should incorporate volumetric or hybrid surface–volume source spaces with orientation freedom to improve neuroanatomical fidelity.

## References


Irina Belyaeva, Ben Gabrielson, Yu-Ping Wang, Tony W Wilson, Vince D Calhoun, Julia M Stephen, and Tülay Adali. Learning spatiotemporal brain dynamics in adolescents via multimodal meg and fmri data fusion using joint tensor/matrix decomposition. *IEEE Transactions on Biomedical Engineering*, 71(7):2189–2200, 2024.

Vince D Calhoun, Tülay Adali, Godfrey D Pearlson, and Kent A Kiehl. Neuronal chronometry of target detection: fusion of hemodynamic and event-related potential data. *Neuroimage*, 30(2): 544–553, 2006.

Francesca Castaldo, Francisco Páscoa Dos Santos, Ryan C Timms, Joana Cabral, Jakub Vohryzek, Gustavo Deco, Mark Woolrich, Karl Friston, Paul Verschure, and Vladimir Litvak. Multi-modal and multi-model interrogation of large-scale functional brain networks. *NeuroImage*, 277:120236, 2023.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):134, 2022.

Radoslaw M Cichy and Aude Oliva. Am/eeg-fmri fusion primer: resolving human brain responses in space and time. *Neuron*, 107(5):772–781, 2020.

Radoslaw M Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature neuroscience*, 17(3):455–462, 2014.

Anders M Dale, Arthur K Liu, Bruce R Fischl, Randy L Buckner, John W Belliveau, Jeffrey D Lewine, and Eric Halgren. Dynamic statistical parametric mapping: combining fmri and meg for high-resolution imaging of cortical activity. *neuron*, 26(1):55–67, 2000.

Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

James S Gao, Alexander G Huth, Mark D Lescroart, and Jack L Gallant. Pycortex: an interactive surface visualizer for fmri. *Frontiers in neuroinformatics*, 9:23, 2015.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. Meg and eeg data analysis with mne-python. *Frontiers in Neuroinformatics*, 7:267, 2013.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Lauri Parkkonen, and Matti S Hämäläinen. Mne software for processing meg and eeg data. *neuroimage*, 86:446–460, 2014.

Emma L. Hall, Siân E. Robson, Peter G. Morris, and Matthew J. Brookes. The relationship between MEG and fMRI. *NeuroImage*, 102:80–91, 2014.

Matti S Hämäläinen and Risto J Ilmoniemi. Interpreting magnetic fields of the brain: minimum norm estimates. *Medical & biological engineering & computing*, 32:35–42, 1994.

Meysam Hashemi, Damien Depannemaecker, Marisa Saggio, Paul Triebkorn, Giovanni Rabuffo, Jan Fousek, Abolfazl Ziaeemehr, Viktor Sip, Anastasios Athanasiadis, Martin Breyton, et al. Principles and operation of virtual brain twins. *IEEE Reviews in Biomedical Engineering*, pages 1–25, 2025.

Richard N Henson, Guillaume Flandin, Karl J Friston, and Jeremie Mattout. A parametric empirical bayesian framework for fmri-constrained meg/eeg source reconstruction. *Human brain mapping*, 31(10):1512–1531, 2010.

René J Huster, Stefan Debener, Tom Eichele, and Christoph S Herrmann. Methods for simultaneous eeg-fmri: an introductory review. *Journal of Neuroscience*, 32(18):6053–6060, 2012.

Alexander G. Huth, Wendy A. De Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532 (7600):453–458, 2016.

A. Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

Jiyoung Kang and Hae-Jeong Park. Integration of partially observed multimodal and multiscale neural signals for estimating a neural circuit using dynamic causal modeling. *PLOS Computational Biology*, 20(12):e1012655, 2024.

Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G. Huth. A natural language fMRI dataset for voxelwise encoding models. *Scientific Data*, 10(1):555, 2023.

Elisa Leonardelli and Scott L Fairhall. Similarity-based fmri-meg fusion reveals hierarchical organisation within the brain's semantic system. *NeuroImage*, 259:119405, 2022.

Ruogu Lin, Thomas Naselaris, Kendrick Kay, and Leila Wehbe. Stacked regressions and structured variance partitioning for interpretable brain maps. *NeuroImage*, 298:120772, 2024.

Arthur K Liu, John W Belliveau, and Anders M Dale. Spatiotemporal imaging of human brain activity using functional mri constrained magnetoencephalography data: Monte carlo simulations. *Proceedings of the National Academy of Sciences*, 95(15):8945–8950, 1998.

Zhongming Liu and Bin He. fmri–eeg integrated cortical source imaging by use of time-variant spatial constraints. *NeuroImage*, 39(3):1198–1214, 2008.

Narges Moradi, Bradley G Goodyear, and Roberto C Sotero. Deep eeg source localization via emd-based fmri high spatial frequency. *Plos one*, 19(3):e0299284, 2024.

Gustavo Patow, Ignacio Martin, Yonatan Sanz Perl, Morten L Kringelbach, and Gustavo Deco. Whole-brain modelling: an essential tool for understanding brain dynamics. *Nature Reviews Methods Primers*, 4(1):53, 2024.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

Aniketh Janardhan Reddy and Leila Wehbe. Can fMRI reveal the representation of syntactic structure in the brain? *Advances in Neural Information Processing Systems*, 34, 2021.

Petra Ritter, Michael Schirner, Anthony R McIntosh, and Viktor K Jirsa. The virtual brain integrates computational modeling and multimodal neuroimaging. *Brain connectivity*, 3(2):121–145, 2013.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45), 2021.

Keita Suzuki and Okito Yamashita. Meg current source reconstruction using a meta-analysis fmri prior. *Neuroimage*, 236:118034, 2021.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023.

Samu Taulu and Juha Simola. Spatiotemporal signal space separation method for rejecting nearby interference in meg measurements. *Physics in Medicine & Biology*, 51(7):1759, 2006.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in neural information processing systems*, 32, 2019.

Mariya Toneva, Tom M Mitchell, and Leila Wehbe. Combining computational controls with natural text reveals aspects of meaning composition. *Nature computational science*, 2(11):745–757, 2022.

Barry D Van Veen, Wim Van Drongelen, Moshe Yuchtman, and Akifumi Suzuki. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Transactions on biomedical engineering*, 44(9):867–880, 1997.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575, 2014a.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 233–243, 2014b.

Marijke Welvaert and Yves Rosseel. On the Definition of Signal-To-Noise Ratio and Contrast-To-Noise Ratio for fMRI Data. *PLoS ONE*, 8(11):e77089, 2013.

David Wipf and Srikantan Nagarajan. A unified bayesian framework for meg/eeg source imaging. *NeuroImage*, 44(3):947–966, 2009.

Jing Xu, Jingwei Sheng, Tianyi Qian, Yue-Jia Luo, and Jia-Hong Gao. EEG/MEG source imaging using fMRI informed time-variant constraints. *Human Brain Mapping*, 39(4):1700–1711, 2018.

Lu-Chun Yeh, Sushrut Thorat, and Marius V Peelen. Predicting cued and oddball visual search performance from fmri, meg, and dnn neural representational similarity. *Journal of Neuroscience*, 44(12), 2024.

Zaid Zada, Samuel A Nastase, Bobbi Aubrey, Itamar Jalon, Sebastian Michelmann, Haocheng Wang, Liat Hasenfratz, Werner Doyle, Daniel Friedman, Patricia Dugan, et al. The "podcast" ecog dataset for modeling neural activity during natural language comprehension. *Scientific Data*, 12(1):1135, 2025.

## A Predictive Performance for All Subjects

Here we report our model's predictive performance for MEG and fMRI for all subjects, against single-subject, single-modality Ridge Ceiling models (see Table 1, 2, Figure 2, 5, 6, 7).

Table 1: **MEG predictive performance** (Pearson $r$ over temporal lobe sensors)

| Subject | Our Model | MEG Ridge Ceiling |
|---|---|---|
| S1 | $0.109 \pm 0.064$ | $0.074 \pm 0.041$ |
| S2 | $0.073 \pm 0.043$ | $0.050 \pm 0.029$ |
| S3 | $0.089 \pm 0.037$ | $0.067 \pm 0.027$ |
| S4 | $0.054 \pm 0.024$ | $0.033 \pm 0.017$ |
| S5 | $0.088 \pm 0.033$ | $0.059 \pm 0.021$ |

Table 2: **fMRI predictive performance** (Pearson $r$ of top quartile of sources)

| Subject | Our Model | fMRI Ridge Ceiling |
|---|---|---|
| S6 | $0.236 \pm 0.072$ | $0.267 \pm 0.074$ |
| S7 | $0.234 \pm 0.056$ | $0.268 \pm 0.058$ |
| S8 | $0.224 \pm 0.067$ | $0.262 \pm 0.064$ |
| S9 | $0.208 \pm 0.063$ | $0.235 \pm 0.070$ |
| S10 | $0.162 \pm 0.063$ | $0.180 \pm 0.067$ |
| S11 | $0.180 \pm 0.062$ | $0.199 \pm 0.066$ |
| S12 | $0.181 \pm 0.055$ | $0.210 \pm 0.059$ |
| S13 | $0.130 \pm 0.037$ | $0.139 \pm 0.036$ |

## B Hyperparameter Tuning

We perform hyperparameter tuning and compare with our original design, which uses four transformer layers and two attention heads for each layer. As shown in the table, in the validation set, our four-layer model outperforms the three-layer model and the five-layer model. Similarly, two attention heads yield the lowest validation loss, compared with one head or four heads.

Although these results support our chosen architecture, we wish to note that the optimal hyperparameters will likely vary depending on factors such as dataset size and stimulus complexity, and we encourage researchers to perform their own tuning when adapting our framework for specific applications.

Table 3: **Results of hyperparameter tuning**

| Model | Total Validation Loss | MEG Validation Loss | fMRI Validation Loss |
|---|---|---|---|
| **4 layers, 2 heads (current)** | **1.807** | **0.918** | **0.889** |
| 3 layers, 2 heads | 1.819 | 0.929 | 0.890 |
| 5 layers, 2 heads | 1.813 | 0.921 | 0.893 |
| 4 layers, 1 head | 1.820 | 0.928 | 0.892 |
| 4 layers, 4 heads | 1.812 | 0.921 | 0.891 |

## C fMNE Implementation

We implement fMNE using `MNE-Python 1.8`. We make the inverse operator using the default settings for fixed orientation, depth weighted sources (`loose` = 0, `depth` = 0.8). We use the ground truth noise to compute the noise covariance for each MEG sensor. We set the regularization parameter `lambda2` to the inverse of the squared MEG CNR. We set the source covariance matrix as

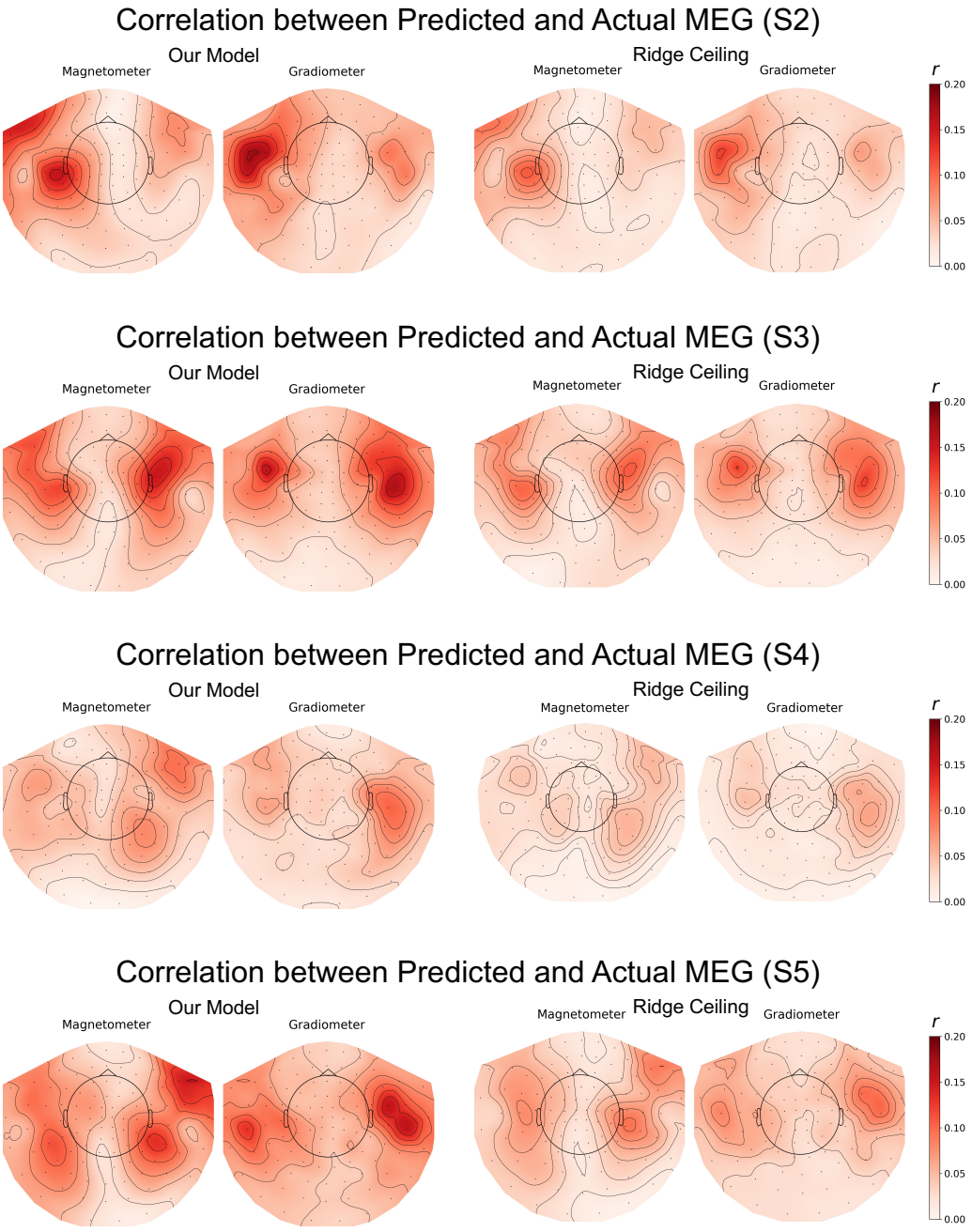


Figure 5: **Predictive performance on MEG for S2–S5.**

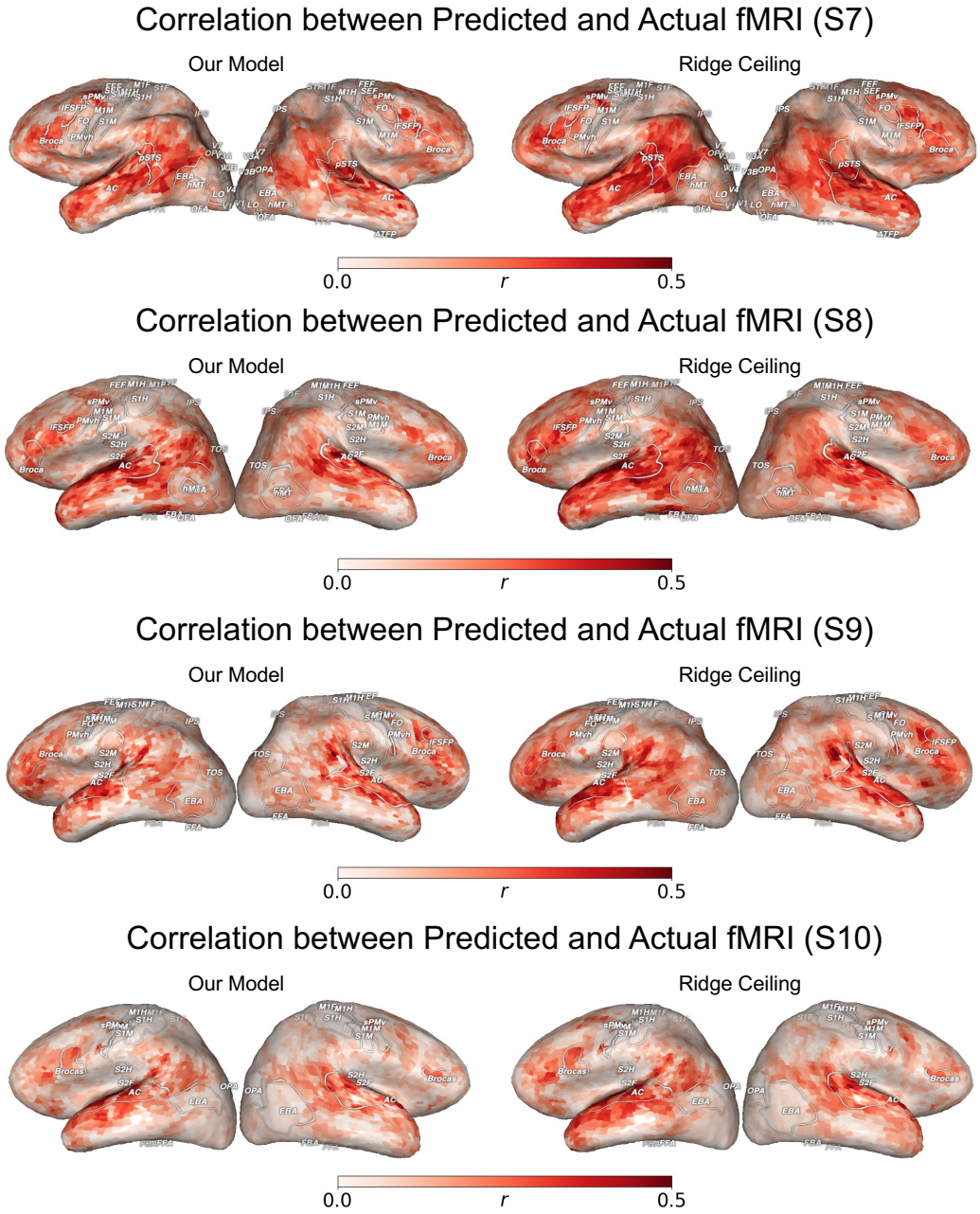


Figure 6: **Predictive performance on fMRI for S7–S10.**

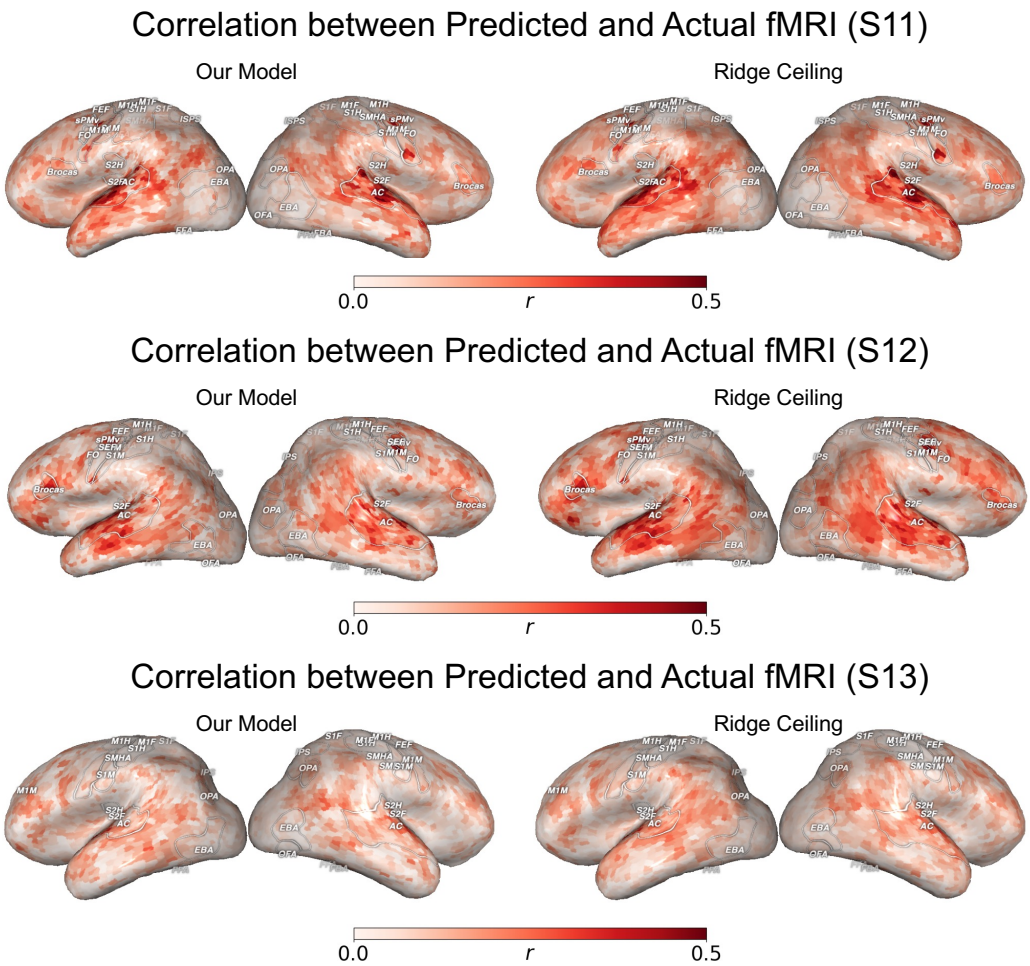


Figure 7: **Predictive performance on fMRI for S11–S13.**

a diagonal matrix whose elements are the BOLD signal of each source at the corresponding TR after normalization.