

000 BRAINFCIR: FUNCTIONAL CONTEXT INFORMED REP- 001 RESENTATION LEARNING FOR INTRACRANIAL NEURAL 002 SIGNALS 003

006 **Anonymous authors**

007 Paper under double-blind review

011 ABSTRACT

013 Intracranial neural recordings (e.g., stereo-ElectroEncephaloGraphy (sEEG)) have
014 offered a unique window to measure neural signals across multiple brain regions
015 simultaneously. Recent works have focused on developing neurofoundation models
016 that learn generalizable representations across both subjects and tasks from such
017 recordings. These models achieve exciting advances, yet overlook the modular
018 functional organization of the brain, where neurons from multiple adjacent anatomical
019 regions collectively support specific cognitive functions (e.g., the Wernicke area
020 for speech perception). A key challenge remains how to effectively incorporate
021 this functional contextual information into representation learning to improve both
022 interpretability and decoding performance. To tackle this challenge, we propose
023 a novel pre-training framework, BrainFCIR, that explicitly integrates functional
024 context into model design via spatial-context-guided representation learning. We
025 evaluate BrainFCIR on the publicly available sEEG speech-perception dataset.
026 Extensive experiments show that BrainFCIR, as a unified representation learning
027 framework for intracranial sEEG signals, significantly outperforms previous decoding
028 methods. Overall, our work underscores the significance of functional context
029 in developing more biologically plausible and high-performing neural decoding
030 models. Code and checkpoints will be publicly available.

031 1 INTRODUCTION

033 Intracranial stereo-ElectroEncephaloGraphy (sEEG) provides a unique window into human brain
034 function by recording neural activity directly from deep brain structures with high temporal precision.
035 This methodology enables simultaneous monitoring across distributed brain regions, capturing
036 dynamic interactions that underlie complex cognitive and behavioral processes. The capacity to
037 model these high-resolution signals offers significant potential for advancing our understanding
038 of large-scale functional networks (Wang et al., 2023; Zhang et al., 2023; Mentzelopoulos et al.,
039 2024; Zheng et al., 2025; Chau et al., 2024), with important implications for developing next-
040 generation neurotechnologies such as closed-loop brain-computer interfaces. In contrast to non-
041 invasive approaches like functional Magnetic Resonance Imaging (fMRI) (Caro et al., 2023; Dong
042 et al., 2024) or ElectroEncephaloGraphy (EEG) (Jiang et al., 2024b; Wang et al., 2024c; Jiang et al.,
043 2024a), sEEG bypasses the signal attenuation caused by the skull and scalp, providing more direct
044 measurements of neural dynamics with millisecond-scale resolution. Furthermore, while intracortical
045 Micro-Electrode Array (MEA) typically samples activity from highly localized neuronal populations,
046 sEEG offers broader coverage across multiple brain systems, albeit with sparser spatial sampling. This
047 combination of high temporal resolution and wide spatial coverage presents both unique opportunities
048 and distinctive computational challenges for modeling the brain’s spatiotemporal organization.

049 The development of foundation models for intracranial sEEG has become an active area of research,
050 driven by the goal of learning generalizable spatiotemporal representations of neural activity. This
051 effort mirrors similar advances in modeling other neural data modalities, including intracortical MEA
052 (Azabou et al., 2023; Ye et al., 2023; 2025), non-invasive EEG (Jiang et al., 2024b; Wang et al., 2024c;
053 Jiang et al., 2024a), and fMRI (Caro et al., 2023; Dong et al., 2024). Contemporary approaches
frequently employ large-scale transformer architectures, pre-trained in a self-supervised manner, to
learn powerful representations from sEEG that demonstrate strong performance on downstream tasks

054 and robust cross-subject generalization (Wang et al., 2023; Zhang et al., 2023; Chau et al., 2024).
 055 A key question, however, concerns the optimal method for training these models to better capture
 056 functional connectivity, which is crucial for both neural decoding and functional groups identification.
 057

058 While prior works (Wang et al., 2023; Zhang et al., 2023; Li et al., 2025) have largely used mask-
 059 based reconstruction tasks to understand the spatial-temporal organization of sEEG recordings, these
 060 models may over-rely on the intra-channel temporal patterns, leaving it unclear whether these models
 061 effectively capture functional context. PopT (Chau et al., 2024) takes the first step in effectively
 062 modeling inter-channel context during the pre-training stage. However, their approach decouples
 063 temporal and spatial modeling into two isolated stages (i.e., BrainBERT (Wang et al., 2023) for
 064 temporal modeling and PopT (Chau et al., 2024) for spatial modeling), preventing effective interaction
 065 between them – temporal-transformed embeddings cannot be refined leveraging spatial context, which
 066 ultimately limits the effectiveness of subsequent spatial modeling. As such, developing models that
 067 enable capturing precise inter-channel interaction (i.e., functional context) and studying the channel
 068 cluster estimated by such a neurofoundation model remains unexplored.

069 To address these two issues, we propose a functional-context-informed neurofoundation model for
 070 intracranial sEEG recordings, BrainFCIR, which models spatial-temporal relationships through
 071 functional context discrimination. To quantify the effect of functional context modeling, we estimate
 072 functional connectivity via the pre-trained model. And we perform channel clustering to identify
 073 functional groups, which enhances neural decoding via channel selection and further demonstrates its
 074 superiority in capturing inter-channel interactions.

075 To validate the effectiveness of our proposed framework, we evaluate BrainFCIR on the widely used
 076 Brain Treebank dataset (Appendix A). Empirically, BrainFCIR outperforms existing neurofoundation
 077 models (Zhang et al., 2023; Chau et al., 2024) for sEEG recordings and identifies channel cluster that
 078 faithfully aligns with those containing target neural activity. Besides, we further evaluate the capability
 079 of cross-subject functional group identification, demonstrating the great potential of BrainFCIR to
 080 support functionally grouping sEEG channels from unseen subjects.

081 To sum up, the main contributions of our work comprise:

- 082 **1. Explicit functional-context modeling:** We develop a spatiotemporal transformer model,
 083 BrainFCIR, for intracranial sEEG recordings and an associated functional context discrimina-
 084 tion task. During the pre-training stage, we explicitly model functional context to encourage
 085 learning functional-context-informed representations.
- 086 **2. State-of-the-art (SOTA) performance:** Our model achieves SOTA performance in de-
 087 coding speech perception from intracranial sEEG signals on the Brain Treebank dataset,
 088 demonstrating robust effectiveness across diverse decoding tasks.
- 089 **3. Cross-subject functional group identification:** Our model shows the potential to offer an
 090 off-the-shelf functional group identification toolkit for unseen subjects. When pre-trained
 091 within the target subject, functional groups identified by our model further enhance decoding.

093 2 RELATED WORKS

095 2.1 SPATIOTEMPORAL MODELS OF INTRACRANIAL SEEG RECORDINGS

097 Several prior studies have proposed spatiotemporal models for sEEG modeling, employing various
 098 strategies to incorporate spatial information. While the initial version of Brant introduced by Zhang
 099 et al. (2023) did not explicitly encode spatial relationships, later iterations (Li et al., 2025) incorporated
 100 learnable positional embeddings along the spatial axis, albeit without integrating neuroanatomical
 101 priors. Zheng et al. (2025) adopted a region-level approach, in which all channels within the pre-
 102 defined brain region were pooled, effectively collapsing spatial variability and eliminating the need
 103 for fine-grained spatial encoding. In comparison, Mentzelopoulos et al. (2024) and Chau et al.
 104 (2024) modeled space at the single-channel level by deriving token-wise spatial encodings from
 105 neuroanatomical coordinates of each channel, thereby incorporating anatomical context.

106 Despite these advances, to the best of our knowledge, most previous sEEG modeling studies (Zhang
 107 et al., 2023; Li et al., 2025; Zheng et al., 2025) rely on the mask-based reconstruction task to learn
 108 spatial-temporal relationships of intracranial sEEG recordings, which may over-rely on the intra-

108 channel temporal patterns. PopT (Chau et al., 2024) takes the first step in effectively capturing
 109 inter-channel context, despite their isolated spatial-temporal modeling strategies preventing effective
 110 interaction between spatial and temporal modeling.
 111

112 2.2 SELF-SUPERVISED LEARNING IN BCI

113 Recently, the pre-trained temporal-spatial models (i.e., foundation models) have drawn significant
 114 attention across diverse neural modalities, including EEG (Jiang et al., 2024b; Wang et al., 2024b;c),
 115 fMRI (Caro et al., 2023; Dong et al., 2024), neural spike (Ye et al., 2023; Zhang et al., 2024), etc.
 116 Since these neural modalities either have lower spatial resolution compared to sEEG recordings (e.g.,
 117 EEG, fMRI) or are typically implanted within a specific brain region (e.g., neural spike), they mainly
 118 leverage masked self-supervised pre-training for spatiotemporal models. However, intracranial sEEG
 119 recordings are inherently different from those recordings, providing a unique window into human
 120 brain function by recording neural activity directly from deep brain structures with high temporal &
 121 spatial resolution. This feature requires sEEG modeling methods to accurately identify channels
 122 from different functional modules, thereby enhancing our understanding of channel interactions
 123 within each functional group and improving neural decoding performance. As such, we develop
 124 a spatiotemporal transformer model paired with a functional context discrimination task, which
 125 enables the effective capture of inter-channel relationships for downstream decoding. Finally, to
 126 further quantify the effect of functional context modeling, we identify functional groups based on the
 127 estimated functional connectivity, further enhancing downstream decoding.
 128

129

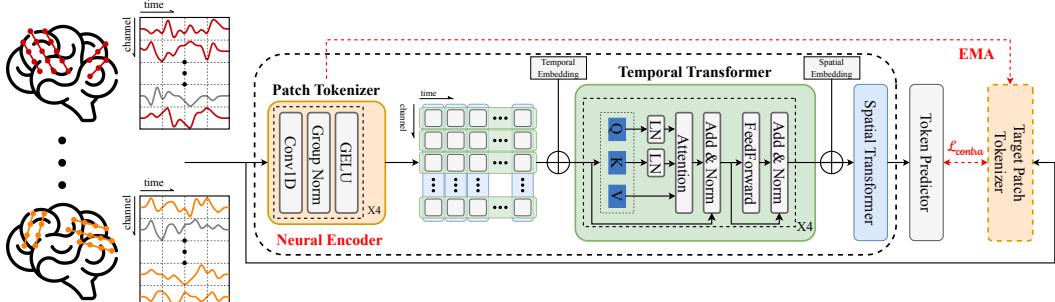
130 3 METHOD

131

132 To effectively capture functional context in representation learning, we develop a spatiotemporal
 133 transformer model and a novel pre-training framework that effectively guides the model to capture
 134 the functional context, thus enhancing downstream decoding. We first describe our spatial-temporal
 135 transformer model architecture. We then detail how our self-supervised pre-training procedure guides
 136 the model to learn functional-context-informed representations. Finally, we discuss our evaluation
 137 schemes.
 138

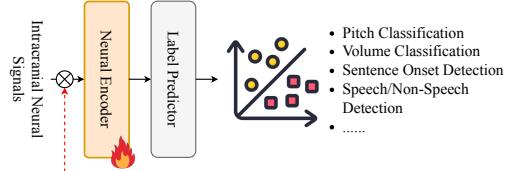
139

(a) Functional Context Pre-training

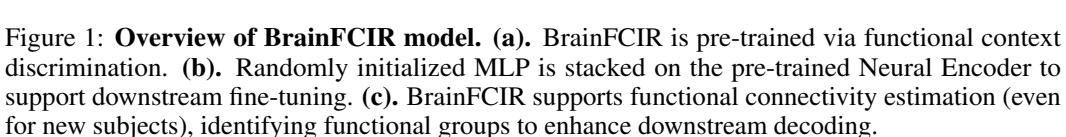


140

(b) Downstream Fine-tuning



141



142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

159 **Figure 1: Overview of BrainFCIR model.** (a). BrainFCIR is pre-trained via functional context
 160 discrimination. (b). Randomly initialized MLP is stacked on the pre-trained Neural Encoder to
 161 support downstream fine-tuning. (c). BrainFCIR supports functional connectivity estimation (even
 162 for new subjects), identifying functional groups to enhance downstream decoding.

162 3.1 MODEL ARCHITECTURE
163

164 Our model architecture and tokenization scheme are shown in Figure 1 (a). Given a multivariate
165 time series of intracranial sEEG activity $X \in \mathbb{R}^{C \times T}$, where C denotes the number of recording
166 channels and T denotes the total timestamps, we first tokenize channels as univariate signals (i.e.,
167 agnostic to space), following common practice (Zhang et al., 2023; Chau et al., 2024; Jiang et al.,
168 2024b). We create temporal patches of each channel that are of length L (e.g., 250 milliseconds),
169 yielding $\mathcal{X}_p = \{\mathbf{x}_{i,j}^p \in \mathbb{R}^L | i = 1, \dots, C; j = 1, \dots, N\}$, where $N = \lfloor \frac{T}{L} \rfloor$, the number of patches is
170 $|\mathcal{X}_p| = C \times N$, and $\mathbf{x}_{i,j}^p \in \mathbb{R}^L$ indicates the i -th patch of length L for the j -th channel.

171 The model architecture of BrainFCIR comprises two parts: (1) Patch Tokenizer; (2) Temporal &
172 Spatial Transformer. The Patch Tokenizer consists of a stack of convolution blocks. In the first step
173 of tokenization, each patch $\mathcal{X}_{i,j}^p$ is passed through the Patch Tokenizer (shared across patches). In
174 practice this tokenizer can take any form; here we choose a temporal convolution neural network
175 (CNN) both to account for the input signal’s continuous nature and because of prior domain knowledge
176 about the importance of oscillatory features in neural activity (Jacobs & Kahana, 2010; Buzsaki &
177 Draguhn, 2004). In each convolution block, the temporal convolution layer is stacked with group
178 normalization (Wu & He, 2018), and Gaussian Error Linear Unit (GELU) activation (Hendrycks &
179 Gimpel, 2016). We denote the patch embeddings from the Patch Tokenizer as

$$180 \mathcal{E}_p = \{\mathbf{e}_{i,j}^p \in \mathbb{R}^d | i = 1, \dots, C; j = 1, \dots, N\}, \quad (1)$$

181 where d is the dimension of embeddings.
182

183 In order to enable the model to be aware of the temporal information of patch embeddings, we
184 utilize the parameter-free temporal embeddings introduced by (Vaswani et al., 2017), i.e., $\mathcal{E}_t =$
185 $\{\mathbf{e}_1^t, \dots, \mathbf{e}_{t_{max}}^t\}$. Note that t_{max} is the hyperparameter determining the maximum number of time
186 patches and $t_{max} \geq N$. Given one arbitrary patch embedding $\mathbf{e}_{i,j}^p$ in Equation 1, we add the
187 corresponding temporal embedding to it:

$$188 \mathcal{E} = \{\mathbf{e}_{i,j}^p + \mathbf{e}_j^t | i = 1, \dots, C; j = 1, \dots, N\}, \quad (2)$$

189 which forms the input embeddings \mathcal{E} for the Temporal Transformer. Then, the embeddings will be
190 directly fed into the Transformer encoder (Vaswani et al., 2017) to get the temporal-transformed
191 embeddings $\mathcal{E} = \{\mathbf{e}_{i,j} | i = 1, \dots, C; j = 1, \dots, N\}$.
192

193 The functions supported by the same anatomical brain regions in different subjects are roughly
194 similar (e.g., the superior temporal gyrus (STG) consistently participates in auditory perception),
195 although fine-grained functional sub-organization may vary between subjects (Buzsáki, 2006). To
196 incorporate such anatomical priors into our model, we encode standardized anatomical coordinates
197 (e.g., LPI coordinates, MNI coordinates) using sinusoidal position encoding, generating a set of
198 spatial embeddings $\mathcal{E}_s = \{\mathbf{e}_i^s \in \mathbb{R}^{d_s} | i = 1, \dots, C\}$. These spatial embeddings are then added to the
199 corresponding temporal-transformed embeddings to form the combined input $\mathcal{E} = \{\mathbf{e}_{i,j} + \mathbf{e}_i^s | i =$
200 $1, \dots, C; j = 1, \dots, N\}$, which is subsequently processed by the Spatial Transformer to produce the
201 spatial-transformed embeddings \mathcal{E} .

202 To improve the stability and efficiency of transformer training, we adopt several optimizations
203 proposed by Dehghani et al. (2023). These include applying layer normalization to queries and
204 keys before computing dot-product attention, which prevents extreme values in attention logits and
205 promotes more stable gradient dynamics during learning:

$$206 \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{\text{LN}(Q)\text{LN}(K)^T}{\sqrt{d_{head}}}\right)V, \quad (3)$$

207 where d_{head} is the dimension of attention head and LN denotes layer normalization (Ba et al., 2016).
208

210 3.2 REPLACED FUNCTIONAL CONTEXT DISCRIMINATION
211

212 To encourage the model to effectively capture inter-channel functional context, we train BrainFCIR
213 using a replaced functional context discrimination task (Figure 1 (a)), which differs from prior work in
214 two ways. First, we follow JEPA (Assran et al., 2023) to encourage the model to discriminate replaced
215 functional context in the latent token space instead of the observation space (Zhang et al., 2023;
Li et al., 2025), which enhances the quality of learned representations. Second, unlike some prior

intracranial sEEG models (Wang et al., 2023; Zheng et al., 2025; Chau et al., 2024), we simultaneously train both the tokenizer and spatial-temporal Transformer to perform replaced functional context discrimination.

During pre-training, for each sEEG sample $\mathcal{X} \in \mathbb{R}^{C \times T}$, 10% of channels are randomly selected to have their activity replaced by activity from unrelated time points. To ensure balanced label distribution, we designate only 10% of unreplaced channels as positive samples during pre-training. The modified sample is directly fed into the Patch Tokenizer to get the patch embeddings \mathcal{E}_p . Then, the patch embedding \mathcal{E}_p is directly fed into the spatial-temporal Transformer to get the transformed embeddings \mathcal{E} . While the patch embeddings \mathcal{E}_p are obtained using our original Patch Tokenizer (left part in Figure 1 (a)), we use a separate Target Patch Tokenizer (right part in Figure 1 (a)) for the target embeddings $\hat{\mathcal{E}}_p$ to provide self-supervision signals. The Target Patch Tokenizer is updated with an exponential moving average (EMA) of the original Patch Tokenizer weights. To encourage the model to learn functional context, the model is trained to discriminate replaced channels based on the spatial-temporal transformed embeddings $\mathcal{E} \in \{e_{i,j} | i = 1, \dots, C; j = 1, \dots, N\}$, the target patch embeddings $\hat{\mathcal{E}}_p \in \{\hat{e}_{i,j}^p | i = 1, \dots, C; j = 1, \dots, N\}$ and the replaced label $y \in \{-1, 1\}$, where -1 indicates unreplaced, 1 indicates replaced. The contrastive loss is defined as follows:

$$\mathcal{L}_{contra} = \sum_{i,j} [1 + y \cdot \langle \ell_2(\text{Linear}(e_{i,j})), \ell_2(\hat{e}_{i,j}^p) \rangle], \quad (4)$$

where ℓ_2 represents ℓ_2 normalization and $\langle \cdot, \cdot \rangle$ the inner product. Combined with ℓ_2 , $\langle \cdot, \cdot \rangle$ calculates the cosine similarity between $e_{i,j}$ and $\hat{e}_{i,j}^p$, which takes value within $[-1, 1]$ -range. y is used to adjust whether to minimize or maximize such similarity, and the shift item 1 is added to ensure $\mathcal{L}_{contra} \geq 0$.

To further demonstrate the effectiveness of our proposed contrastive loss to capture functional context, we also design two alternatives to train the spatial-temporal model, i.e., \mathcal{L}_{diff} and \mathcal{L}_{mse} . The difference loss \mathcal{L}_{diff} also encourages maximizing the similarity between embeddings of unreplaced channels:

$$\mathcal{L}_{diff} = \sum_{i,j} [\text{BCE}(\text{Linear}(\|e_{i,j} - \hat{e}_{i,j}^p\|_2^2))], \quad (5)$$

where $\|\cdot\|_2^2$ represents the squared ℓ_2 norm value and BCE represents the binary cross entropy loss.

Besides, following the commonly adopted JEPA pre-training framework (Assran et al., 2023; Dong et al., 2024), we randomly select patches to mask. Around 50% of patch embeddings \mathcal{E}_p are patch-wise chosen and masked. The masked position is termed as \mathcal{M} . Then, a shared learnable embedding $e_{[M]} \in \mathbb{R}^d$ is used to replace the original patch embeddings:

$$\mathcal{E}_m^f = \{e_i^m | i = 1, \dots, N\}, \quad e_i^m = m_i \odot e_{[M]} + (1 - m_i) \odot e_i^p, \quad (6)$$

where $\delta(\cdot)$ is the indicator function and $m_i = \delta(i \in \mathcal{M})$. After that, the masked embeddings \mathcal{E}_m will be fed into the spatial-temporal Transformer. The MSE loss \mathcal{L}_{mse} is the average mean-squared error between spatial-temporal transformed embeddings \mathcal{E} and target patch embeddings $\hat{\mathcal{E}}_p$:

$$\mathcal{L}_{mse} = \sum_{i,j \in \mathcal{M}} \|e_{i,j} - \hat{e}_{i,j}^p\|_2^2. \quad (7)$$

3.3 DOWNSTREAM EVALUATION

We evaluate the validity of our training procedure and the effectiveness of our learned model using several downstream decoding tasks in the Brain Treebank dataset (Wang et al., 2024a). We also quantify whether BrainFCIR effectively models the functional context via channel clustering. First, we validate our pre-trained model’s performance on four speech-perception related downstream tasks used by Wang et al. (2023); Chau et al. (2024): (1) Classification of low/high pitch; (2) Classification of low/high volume; (3) Identification of words that correspond to sentence onsets; (4) Classification of speech vs. non-speech audio. Classification performance is reported as an average across all hold-out test sessions (Appendix A), for 6 fine-tuning seeds each. As baselines, we compare our pre-trained model’s fine-tuned performance against a fine-tuned, randomly-initialized version of itself, as well as two advanced spatial-temporal sEEG models: Population Transformer (PopT) (Chau et al., 2024) and Brant (Zhang et al., 2023). Second, we ablate the loss item used in the replaced functional

270 context discrimination task to explore the effectiveness of our proposed contrastive loss. Third, we
 271 quantify the effectiveness of functional context modeling via channel clustering, and visualize the
 272 identified functional groups to strengthen the neuroscientific interpretation. When pre-trained with
 273 the target subject, the selected channels via channel clustering further enhance downstream decoding;
 274 when evaluated on new subjects, our model demonstrates the great potential to offer an off-the-shelf
 275 functional group identification toolkit for unseen subjects. Lastly, towards the goal of building
 276 intracranial sEEG neurofoundation models, we evaluate our modeling framework’s data scalability,
 277 overall model interpretability – the results of which are presented in Appendix E&I, respectively.
 278

279 4 EXPERIMENTS

281 4.1 DATASET

283 For our experiments, we used the publicly available Brain Treebank dataset (Wang et al., 2024a),
 284 which consists of intracranial recordings (2048 Hz) from 10 epilepsy patients collected over a
 285 total of 26 sessions as they watched Hollywood films. Film transcripts that are aligned to neural
 286 activity are also provided. The intracranial sEEG recordings cover multiple brain regions across both
 287 hemispheres, including the temporal and frontal lobes, which are known to support auditory and
 288 language processing. Neural data is provided at a sampling rate of 2048 Hz. We followed a similar
 289 preprocessing procedure as outlined by Wang et al. (2023; 2024a); Chau et al. (2024).

290 We evaluate our model against previous advanced baselines on four binary classification tasks (e.g.,
 291 sentence onset detection). Multi-channel sEEG signals are represented as $\mathcal{X} \in \mathbb{R}^{C \times T}$, and the paired
 292 label is $\mathcal{Y} \in \mathcal{Y}$, where \mathcal{Y} represents the label-set. We use ROC-AUC as the evaluation metric.
 293

294 4.2 IMPLEMENTATION DETAILS

296 **Preprocess.** The sEEG signals first undergo bandpass filtering between 0.5 Hz and 200 Hz to
 297 attenuate low-frequency drift and high-frequency noise. Following this, a 60 Hz notch filter is applied
 298 to suppress power-line interference. The signals are then resampled to 400 Hz and re-referenced
 299 (Li et al., 2018) according to the original setting to enhance the spatial resolution of the recordings.
 300 Finally, z-score normalization is applied independently to each channel to ensure consistent scaling
 301 across all channels, thereby facilitating stable model training.
 302

303 **Model Configurations.** Throughout both pre-training and fine-tuning, raw input patches are initially
 304 tokenized into the patch embedding space with a dimensionality of $d = 256$. These embeddings
 305 are then processed by a sequence of transformer modules: first by a Temporal Transformer and
 306 subsequently by a Spatial Transformer. Each of these transformer modules is implemented as a 4-layer
 307 encoder block. The architecture maintains a consistent model dimension of $d = 256$ across layers,
 308 while the inner feed-forward network (FFN) dimension is expanded to $d_{ff} = 1024$. Each multi-head
 309 self-attention layer utilizes 8 parallel attention heads to capture diverse contextual relationships. A
 310 comprehensive breakdown of the model’s hyperparameters and architectural specifics is provided in
 311 Appendix C.
 312

313 **Pre-training.** The pre-training model is trained using all recordings across all subjects, excluding
 314 those reserved for validation and testing in downstream tasks (Appendix A). When using recordings
 315 from all subjects, the model is trained on 8 GPUs (NVIDIA Tesla V100 32GB using Python 3.11.7
 316 and PyTorch 2.1.2 + CUDA 12.3) for ~ 12 hours in total.
 317

318 **Fine-tuning.** We split the task recordings into training, validation, and testing splits with a size
 319 roughly proportional to 80%, 10%, and 10%. All experiments are conducted on the same machine
 320 with the same set of random seeds. The train/validation/test splits are the same across different
 321 models. For each subject, models are trained for ~ 20 minutes. The best models are trained on the
 322 training set, selected from the validation set according to accuracy, and finally evaluated on the test
 323 set. For model comparison, we report the average and standard error values (of all subjects) on six
 324 random seeds to obtain comparable results.
 325

324 4.3 BRAINFCIR ENHANCES DECODING BY MODELING FUNCTIONAL CONTEXT
325

326 In Table 1, we report the average classification ROC-AUC over all test sessions and seeds. Our results
327 demonstrate that our model outperforms all alternative models by effectively capturing the functional
328 context. In comparison to randomly initialized versions of our model, our pre-training improves
329 downstream performance. To further quantify our model’s effectiveness in capturing functional
330 context, we forward sEEG samples from the target subject into the frozen model to estimate the
331 inter-channel functional connectivity (Appendix C). Based on the sparse inter-channel functional
332 connectivity, we perform hard clustering ($k = 10$) to extract the functional groups. When evaluating
333 the BrainFCIR model on the selected functional groups, our model further improves decoding
334 performance while greatly reducing the inference time, as only $\sim 20\%$ of channels are kept. Overall,
335 the results in Table 1 demonstrate that by effectively modeling the functional context, our model can
336 improve downstream task performance by learning functional-context-informed representations for
337 multi-regional neural activity.

338 Table 1: Results on the Brain Treebank dataset, with mean ROC-AUC and s.e.m. reported. Asterisks
339 indicate that the bolded model is significantly better than the second model ($p < 0.01$, paired T-test).

341 Method	342 Pitch	343 Volume	344 Sentence Onset	345 Speech/Non-Speech
346 Brant	347 0.61 ± 0.03	348 0.74 ± 0.03	349 0.80 ± 0.04	350 0.80 ± 0.03
351 LaBraM	352 0.69 ± 0.03	353 0.83 ± 0.02	354 0.87 ± 0.02	355 0.85 ± 0.02
356 CBraMod	357 0.71 ± 0.03	358 0.86 ± 0.03	359 0.88 ± 0.01	360 0.87 ± 0.02
361 PopT	362 0.74 ± 0.03	363 0.87 ± 0.03	364 0.90 ± 0.01	365 0.93 ± 0.02
366 BrainFCIR	367 0.77 ± 0.02	368 0.89 ± 0.02	369 0.94 ± 0.01	370 0.96 ± 0.01
371 w/o pre-training	372 0.59 ± 0.03	373 0.71 ± 0.05	374 0.80 ± 0.03	375 0.79 ± 0.05
376 w/ channel-selection	377 $0.78 \pm 0.02^*$	378 $0.91 \pm 0.02^*$	379 0.94 ± 0.01	380 $0.97 \pm 0.01^*$

390 4.4 CONTRASTIVE LOSS DURING PRE-TRAINING ENHANCES DOWNSTREAM PERFORMANCE
391

392 We investigated how our proposed contrastive loss effectively models the functional context. To
393 do so, we pre-trained our model using different loss items (including the original contrastive loss
394 \mathcal{L}_{contra} , the difference loss \mathcal{L}_{diff} , and the JEPA-style (Assran et al., 2023) MSE loss \mathcal{L}_{mse}), which
395 are detailed in Section 3.2. Then, we evaluated each pre-trained model’s performance on the same
396 speech perception tasks in Table 1.

397 First, we find that the choice of the pre-training objective has a substantial impact on downstream
398 performance (Figure 2). Specifically, models trained with \mathcal{L}_{contra} achieve the highest decoding
399 accuracy, demonstrating the importance of explicitly modeling similarity relationships across neural
400 states for capturing functional context. Second, \mathcal{L}_{diff} underperforms relative to \mathcal{L}_{contra} , likely due
401 to its sole reliance on differentiating samples without explicitly encouraging similarity among positive
402 pairs, which appears critical for learning functionally meaningful representations. Third, \mathcal{L}_{mse} , which
403 relies on a traditional reconstruction-based objective, yields the lowest performance. This suggests
404 that an over-reliance on intra-channel temporal dynamics—without explicit inter-channel relational
405 modeling—fails to capture the functional context necessary for robust speech perception decoding.

406 To statistically validate these observations, we performed paired T-tests for pairwise comparisons.
407 The analysis revealed a significant main effect of the loss type. Post-hoc tests confirmed that \mathcal{L}_{contra}
408 significantly outperformed both \mathcal{L}_{diff} ($p < 0.01$) and \mathcal{L}_{mse} ($p < 0.001$), while \mathcal{L}_{diff} also surpassed
409 \mathcal{L}_{mse} ($p < 0.001$). In summary, our results indicate that contrastive learning objectives – particularly
410 those that balance similarity encouragement and dissimilarity constraints—are most effective for
411 modeling functional context in neural signals, underscoring the importance of relational inductive
412 biases in self-supervised pre-training for neural decoding.

413 4.5 BRAINFCIR CAN IDENTIFY FUNCTIONAL MODULES THROUGH CHANNEL CLUSTER
414

415 We further validated the neurobiological plausibility and generalization capability of BrainFCIR by
416 visualizing its estimated functional connectivity (Figure 3) and examining how functional group
417 selection generalizes across varying pre-training cohort sizes (Figure 4). We projected the inferred

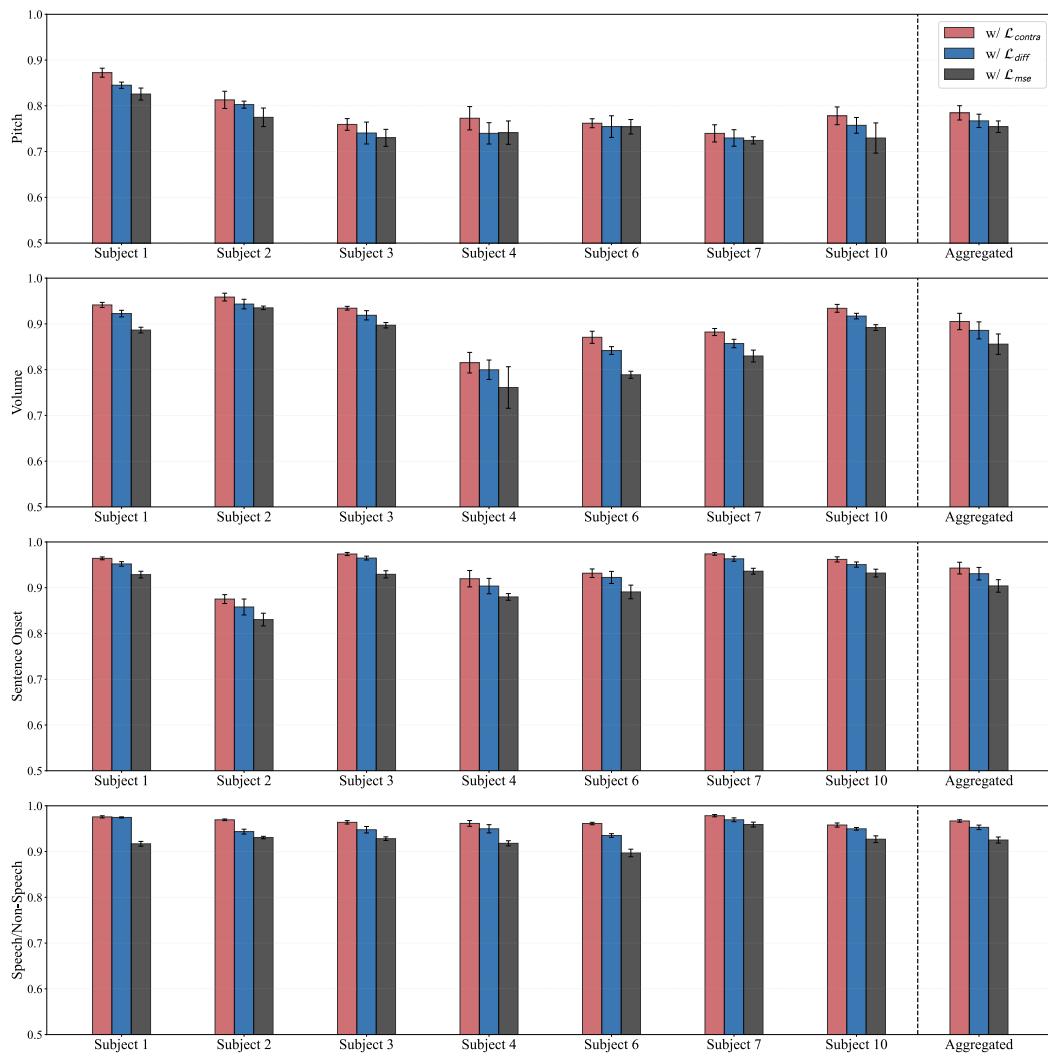


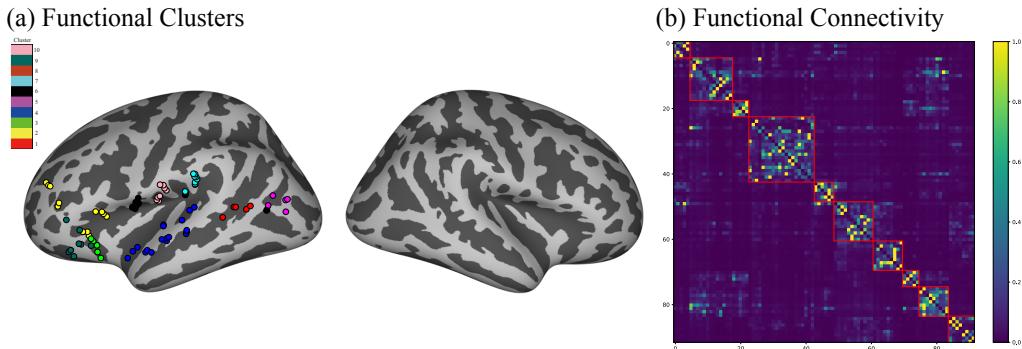
Figure 2: **Ablations on Functional Context Loss.** We pre-train BrainFCIR with different losses to encourage learning functional-context-informed representations. [We report ROC-AUC for each task across 6 random seeds.](#)

inter-channel functional connections onto cortical surface maps alongside the resulting functional clusters, enabling anatomical interpretation of the learned representations. Additionally, we systematically varied the number of subjects included in pre-training—with and without the target subject—to assess how cohort composition affects the quality of functionally-informed channel selection for downstream decoding.

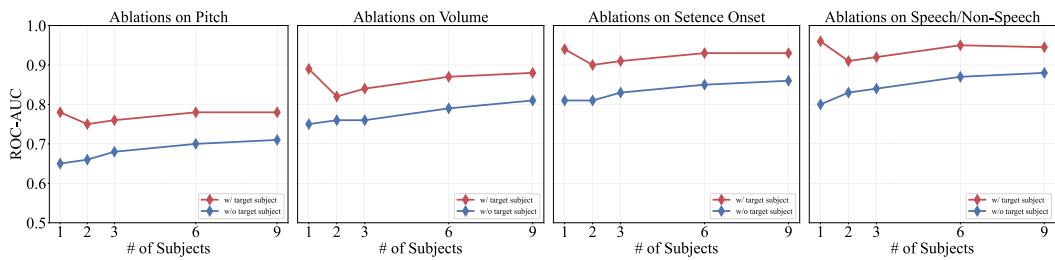
First, we observed that the functional connectivity patterns estimated by BrainFCIR yield clusters that align well with known neuroanatomical regions (Figure 3), suggesting that the model captures biologically meaningful neural groupings. Second, when the pre-training dataset included only the target subject, channel selection based on the resulting model achieved the highest decoding accuracy, reflecting optimal adaptation to subject-specific functional organization. Third, introducing additional subjects initially reduced performance, likely due to interference from inter-subject variability; however, as more subjects were added, decoding accuracy gradually recovered and approached the single-subject baseline, indicating that the model learns to distill shared functional principles across individuals. Fourth, in the more challenging zero-shot setting where the target subject was excluded from pre-training, decoding performance improved steadily with larger pre-training cohorts, yet

432 consistently lagged behind the subject-included condition—highlighting a persistent gap attributable
 433 to individual-specific functional specializations.

434 These results demonstrate that while BrainFCIR captures generalizable functional topology, fully
 435 leveraging subject-specific functional specializations still requires target-subject data. Nevertheless,
 436 the model’s ability to approach subject-specific performance with increasing cross-subject data
 437 underscores its potential for scalable neural decoding applications.



451 **Figure 3: Results for Functional Module Identification.** (a). The visualization of functional clusters
 452 identified by our method ($k = 10$). (b). The functional connectivity estimated by our method.



463 **Figure 4: Ablations on Channel Selection.** We pre-train BrainFCIR while varying the number of
 464 subjects (either w/ or w/o the target subject). The averaged ROC-AUC across subjects is reported.

467 5 DISCUSSION

468 Our work demonstrates that explicitly modeling functional context through spatial-context-guided rep-
 469 resentation learning significantly advances intracranial sEEG decoding. BrainFCIR not only achieves
 470 state-of-the-art performance but also produces functionally coherent channel clusters that align with
 471 neuroanatomy. The contrastive objective proves essential for capturing inter-channel relationships,
 472 outperforming reconstruction-based losses. Furthermore, while subject-specific pre-training yields
 473 optimal decoding, our model effectively generalizes functional topology across subjects, with per-
 474 formance scaling steadily with cohort size. These findings highlight that incorporating functional
 475 context is critical for building biologically plausible and high-performing neural decoders. Future
 476 work will explore dynamic functional networks and extend the framework to other cognitive domains.

478 6 CONCLUSION

481 This paper proposes BrainFCIR, a novel neurofoundation model for intracranial sEEG that explicitly
 482 incorporates functional context into representation learning via spatial-context-guided pre-training.
 483 Comprehensive experiments demonstrate that BrainFCIR not only achieves state-of-the-art perfor-
 484 mance in speech perception decoding on the Brain Treebank dataset, but also identifies functionally
 485 coherent channel clusters that align with known neuroanatomy. In addition, the model shows promis-
 ing generalization in cross-subject functional group identification, with performance scaling robustly

486 as pre-training cohort size increases. Overall, our framework— informed by principles of brain
487 network organization—provides a biologically interpretable and high-performing approach for neural
488 decoding, moving toward more clinically applicable and transparent brain-computer interfaces.
489

490

491

492

493

494

495

496

497

498

499

500

501

502

503

504

505

506

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540 THE USAGE OF LLMs
541

542 Our writing process was assisted by DeepSeek-R1 (Guo et al., 2025), which was used to polish
543 textual clarity. Brief paragraphs were provided to the model, and its output was critically evaluated
544 before relevant revisions were adopted for the final version.

545
546 REPRODUCIBILITY STATEMENT
547

548 Code to train models and reproduce the results will be publicly available. To facilitate review, an
549 anonymous repository link will be provided during rebuttal stage, which is only visible to reviewers.

550
551 REFERENCES
552

553 Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat,
554 Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding
555 predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and*
556 *Pattern Recognition*, pp. 15619–15629, 2023.

557 Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael
558 Mendelson, Blake Richards, Matthew Perich, Guillaume Lajoie, and Eva Dyer. A unified, scalable
559 framework for neural population decoding. *Advances in Neural Information Processing Systems*,
560 36:44937–44956, 2023.

561 Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint*
562 *arXiv:1607.06450*, 2016.

563
564 György Buzsáki. *Rhythms of the Brain*. Oxford university press, 2006.

565
566 Gyorgy Buzsaki and Andreas Draguhn. Neuronal oscillations in cortical networks. *science*, 304
567 (5679):1926–1929, 2004.

568 Josue Ortega Caro, Antonio H de O Fonseca, Christopher Averill, Syed A Rizvi, Matteo Rosati,
569 James L Cross, Prateek Mittal, Emanuele Zappala, Daniel Levine, Rahul M Dhodapkar, et al.
570 Brainlm: A foundation model for brain activity recordings. *bioRxiv*, pp. 2023–09, 2023.

571
572 Geeling Chau, Christopher Wang, Sabera Talukder, Vighnesh Subramaniam, Saraswati Soedarmadji,
573 Yisong Yue, Boris Katz, and Andrei Barbu. Population transformer: Learning population-level
574 representations of neural activity. *ArXiv*, pp. arXiv–2406, 2024.

575 Xupeng Chen, Ran Wang, Amirhossein Khalilian-Gourtani, Leyao Yu, Patricia Dugan, Daniel
576 Friedman, Werner Doyle, Orrin Devinsky, Yao Wang, and Adeen Flinker. A neural speech
577 decoding framework leveraging deep learning and speech synthesis. *Nature Machine Intelligence*,
578 6(4):467–480, 2024.

579
580 Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer,
581 Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling
582 vision transformers to 22 billion parameters. In *International conference on machine learning*, pp.
583 7480–7512. PMLR, 2023.

584
585 Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah
586 Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated
587 labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of
588 interest. *Neuroimage*, 31(3):968–980, 2006.

589
590 Iain DeWitt and Josef P Rauschecker. Wernicke’s area revisited: parallel streams and word processing.
591 *Brain and language*, 127(2):181–191, 2013.

592
593 Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Chong, Fang Ji, Nathanael Tong,
594 Christopher Chen, and Juan Helen Zhou. Brain-jepa: Brain dynamics foundation model with
595 gradient positioning and spatiotemporal masking. *Advances in Neural Information Processing*
596 *Systems*, 37:86048–86073, 2024.

594 Angela D Friederici. The brain basis of language processing: from structure to function. *Physiological*
 595 *reviews*, 91(4):1357–1392, 2011.
 596

597 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu
 598 Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforce-
 599 ment learning. *Nature*, 645(8081):633–638, 2025.

600 Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint*
 601 *arXiv:1606.08415*, 2016.
 602

603 Joshua Jacobs and Michael J Kahana. Direct brain recordings fuel advances in cognitive electrophysi-
 604 ology. *Trends in cognitive sciences*, 14(4):162–171, 2010.
 605

606 Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. Neurolm: A universal multi-
 607 task foundation model for bridging the gap between language and eeg signals. *arXiv preprint*
 608 *arXiv:2409.00101*, 2024a.

609 Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic represen-
 610 tations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*, 2024b.
 611

612 Guangye Li, Shize Jiang, Sivylla E Paraskevopoulou, Meng Wang, Yang Xu, Zehan Wu, Liang Chen,
 613 Dingguo Zhang, and Gerwin Schalk. Optimal referencing for stereo-electroencephalographic
 614 (seeg) recordings. *NeuroImage*, 183:327–335, 2018.

615 Jiahe Li, Xin Chen, Fanqi Shen, Junru Chen, Yuxin Liu, Daoze Zhang, Zhizhang Yuan, Fang Zhao,
 616 Meng Li, and Yang Yang. Deep learning-powered electrical brain signals analysis: Advancing
 617 neurological diagnostics. *arXiv preprint arXiv:2502.17213*, 2025.
 618

619 Georgios Mentzelopoulos, Evangelos Chatzipantazis, Ashwin Ramayya, Michelle Hedlund, Vivek
 620 Buch, Kostas Daniilidis, Konrad Kording, and Flavia Vitale. Neural decoding from stereotactic eeg:
 621 accounting for electrode variability across subjects. *Advances in Neural Information Processing*
 622 *Systems*, 37:108600–108624, 2024.

623 Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm.
 624 *Advances in neural information processing systems*, 14, 2001.
 625

626 Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier
 627 Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn:
 628 Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
 629

630 Constantijn L Van der Burght, Tomás Goucha, Angela D Friederici, Jens Kreitewolf, and Gesa
 631 Hartwigsen. Intonation guides sentence processing in the left inferior frontal gyrus. *Cortex*, 117:
 632 122–134, 2019.

633 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz
 634 Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing*
 635 *systems*, 30, 2017.
 636

637 Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio
 638 Cases, and Andrei Barbu. Brainbert: Self-supervised representation learning for intracranial
 639 recordings. *arXiv preprint arXiv:2302.14367*, 2023.

640 Christopher Wang, Adam Uri Yaari, Aaditya K Singh, Vighnesh Subramaniam, Dana Rosenfarb,
 641 Jan DeWitt, Pranav Misra, Joseph R Madsen, Scellig Stone, Gabriel Kreiman, et al. Brain
 642 treebank: Large-scale intracranial recordings from naturalistic language stimuli. In *The Thirty-
 643 eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*,
 644 2024a.
 645

646 Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. Eegpt: Pretrained
 647 transformer for universal and reliable representation of eeg signals. *Advances in Neural Information*
 648 *Processing Systems*, 37:39249–39280, 2024b.

648 Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and
649 Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. *arXiv preprint*
650 *arXiv:2412.07236*, 2024c.

651

652 Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet:
653 Temporal 2d-variation modeling for general time series analysis. In *The eleventh international*
654 *conference on learning representations*, 2022.

655 Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on*
656 *computer vision (ECCV)*, pp. 3–19, 2018.

657

658 Joel Ye, Jennifer Collinger, Leila Wehbe, and Robert Gaunt. Neural data transformer 2: multi-context
659 pretraining for neural spiking activity. *Advances in Neural Information Processing Systems*, 36:
660 80352–80374, 2023.

661

662 Joel Ye, Fabio Rizzoglio, Adam Smoulder, Hongwei Mao, Xuan Ma, Patrick Marino, Raeed Chowd-
663 hury, Dalton Moore, Gary Blumenthal, William Hockeimer, et al. A generalist intracortical motor
664 decoder. *bioRxiv*, 2025.

665

666 Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant:
667 Foundation model for intracranial neural signal. *Advances in Neural Information Processing*
668 *Systems*, 36:26304–26321, 2023.

669

670 Yizi Zhang, Yanchen Wang, Donato Jiménez-Benetó, Zixuan Wang, Mehdi Azabou, Blake Richards,
671 Renee Tung, Olivier Winter, Eva Dyer, Liam Paninski, et al. Towards a “universal translator” for
672 neural dynamics at single-cell, single-spike resolution. *Advances in Neural Information Processing*
673 *Systems*, 37:80495–80521, 2024.

674

675 Hui Zheng, Haiteng Wang, Weibang Jiang, Zhongtao Chen, Li He, Peiyang Lin, Penghu Wei,
676 Guoguang Zhao, and Yunzhe Liu. Du-in: Discrete units-guided mask modeling for decoding
677 speech from intracranial neural signals. *Advances in Neural Information Processing Systems*, 37:
678 79996–80033, 2025.

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702 A TASK DETAILS

704 Brain Treebank (Wang et al., 2024a) dataset is a publicly available dataset of 10 epilepsy patients
 705 while they were watching movies from a set of 21 animated/action Hollywood movies. Each subject
 706 watched one or more movies while their brain activity was recorded (measured by sEEG). There is a
 707 total of 26 sessions across all subjects, each being ~ 2 hours long on average.

709 A.1 PRE-TRAINING DETAILS

711 We detail the pre-training configurations of BrainFCIR (Figure 1 (a)). From the 26 available sessions,
 712 16 were used for training, 2 were held out as downstream validation, and the remaining 7 were held
 713 out for downstream testing, as specified in Table 2. We prepare the pre-training data by segmenting
 714 neural recordings for each session into non-overlapping intervals of 4 seconds, resulting in a total of
 715 27,698 pre-training segments – corresponding to ~ 30 hours.

716 Table 2: The session splits for pre-training in Brain Treebank.

718 Subject	719 Session	720 Duration (hours)	721 Split
722 Subject 1	Session 1	1.91	Train
	Session 2	2.90	Test
	Session 3	2.07	Train
724 Subject 2	Session 1	2.60	Train
	Session 2	2.42	Train
	Session 3	2.66	Train
	Session 4	3.00	Train
	Session 5	3.73	Train
	Session 6	1.85	Valid
	Session 7	3.52	Test
730 Subject 3	Session 1	1.90	Test
	Session 2	2.94	Train
	Session 3	4.06	Train
734 Subject 4	Session 1	1.87	Test
	Session 2	1.75	Train
	Session 3	1.31	Valid
736 Subject 5	Session 1	1.54	Train
738 Subject 6	Session 1	0.81	Train
	Session 2	1.32	Train
	Session 3	1.60	Test
741 Subject 7	Session 1	1.67	Test
	Session 2	1.77	Train
743 Subject 8	Session 1	1.41	Train
744 Subject 9	Session 1	1.00	Train
746 Subject 10	Session 1	1.57	Test
	Session 2	2.33	Train

749 A.2 FINE-TUNING DETAILS

751 We adopted the same task specification and analysis window in PopT (Chau et al., 2024), yielding
 752 5-second neural activity per trial. We report the number of training, validation, and test trials for each
 753 downstream task in Table 3. The number of positive and negative labels is balanced.

754 **Pitch.** The pitch of a given word is extracted using Librosa’s `pitch` function over a Mel-
 755 spectrogram (sampling rate 48,000 Hz, FFT window length of 2048, hop length of 512, and 128 mel

756 filters). For this task, for a given session, the positive examples consist of words in the top quartile of
 757 pitch, and the negative examples are the words in the bottom quartile.
 758

759 **Volume.** The volume of a given word is computed as the average intensity of root-mean-square
 760 (RMS) (`rms` function, frame and hop lengths 2048 and 512, respectively). For this task, for a given
 761 session, the positive examples are the words in the top quartile of volume, and the negative examples
 762 are the words in the bottom quartile.
 763

764 **Sent. Onset (Sentence Onset).** The negative examples are intervals of activity from 1s periods
 765 during which no speech occurs in the movie. The positive examples are intervals of brain activity that
 766 correspond with hearing the first word of a sentence.
 767

768 **Word Onset (Speech vs. Non-speech).** The negative examples are intervals of activity from 1s
 769 periods during which no speech occurs in the movie. The positive examples are intervals of brain
 770 activity that correspond with dialogue being spoken in the stimulus movie.
 771

772 For each task, we follow the evaluation protocol in PopT (Chau et al., 2024), using the specified
 773 movie for downstream classification. Since these tasks are binary classification (CLS) tasks, we
 774 flatten embeddings and add a linear head after either pre-trained or randomly initialized models.
 775 Training employs binary cross-entropy (BCE) loss, with results quantified using ROC-AUC scores.
 776

777 Table 3: The trial splits for fine-tuning in Brain Treebank.
 778

779 Subject	780 Pitch			781 Volume			782 Sent. Onset			783 Word Onset		
	784 Train	785 Valid	786 Test	787 Train	788 Valid	789 Test	790 Train	791 Valid	792 Test	793 Train	794 Valid	795 Test
796 Subject 1	797 4076	798 510	799 510	800 4076	801 510	802 510	803 2358	804 318	805 318	806 10130	807 1267	808 1267
809 Subject 2	810 2560	811 320	812 320	813 2560	814 320	815 320	816 1710	817 214	818 214	819 10236	820 1280	821 1280
822 Subject 3	823 4038	824 505	825 505	826 4038	827 505	828 505	829 3282	830 411	831 411	832 4128	833 517	834 517
836 Subject 4	837 996	838 125	839 125	840 996	841 125	842 125	843 866	844 109	845 109	846 3984	847 498	848 498
852 Subject 6	853 2536	854 317	855 317	856 2536	857 317	858 317	859 1694	860 212	861 212	862 5092	863 637	864 637
868 Subject 7	869 2932	870 367	871 367	872 2932	873 367	874 367	875 2068	876 259	877 259	878 4680	879 586	880 586
886 Subject 10	887 3328	888 417	889 417	890 3328	891 417	892 417	893 2664	894 333	895 333	896 3786	897 474	898 474

810 B BASELINE DETAILS 811

812 In experiments, we compare our model to the existing advanced neurofoundation models (Zhang
813 et al., 2023; Chau et al., 2024) on intracranial sEEG signals. The details of these baseline models are
814 given here:

815

- 816 • **Brant** (Zhang et al., 2023): A self-supervised model for sEEG recordings that can capture
817 both long-term temporal dependency and spatial correlation from neural signals. Brant is
818 primarily designed for medical use, serving as an sEEG foundation model. Although Brant
819 mainly evaluates its performance on the low-level modeling tasks (Wu et al., 2022) (e.g.,
820 neural signal forecasting, imputation, etc.), Brant achieves SOTA performance on some
821 high-level modeling tasks (e.g., seizure detection). As a foundation model in the sEEG
822 pre-training field, this model is suitable to serve as a baseline for comparison.
- 823 • **PopT** (Chau et al., 2024): A self-supervised model for sEEG that learns population-level
824 codes for arbitrary ensembles of neural recordings at scale. PopT stacks on top of pre-trained
825 temporal embeddings (Wang et al., 2023) and enhances downstream decoding by enabling
826 the learned aggregation of multiple spatially sparse channels. PopT serves as an sEEG
827 foundation model, achieving SOTA performance on Brain Treebank (Wang et al., 2024a).
828 As a foundation model in the sEEG pre-training field, this model is suitable to serve as a
829 baseline for comparison.

830 The detailed implementations of these baseline models are given here:

831

- 832 • For the Brant method (Zhang et al., 2023), the hyperparameters are optimized based on
833 the Brant-Tiny model for better performance. We changed the length of the patch segment
834 from 6 seconds to 1 second. Additionally, we replace the linear embedding layer with a
835 convolutional embedding layer, which is used in LaBraM (Jiang et al., 2024b). The numbers
836 of convolution filters are $\{96, 96, 96\}$; the sizes of convolution kernels are $\{9, 9, 3\}$; the
837 numbers of convolution strides are $\{5, 5, 1\}$.
- 838 • For the PopT method (Chau et al., 2024), the hyperparameters are the same as the original
839 implementation of the PopT model. The data samples are resampled to the specified
840 sampling rate (i.e., 2048 Hz).

841 When evaluating the decoding performance of these baseline models, we follow the same experiment
842 setup as our model; see Appendix C for more details.

843 For the self-supervised methods, the pre-training setup follows the original setup of each model:

844

- 845 • For the Brant model, we also use all sEEG recordings from all subjects within the Brain
846 Treebank dataset to pre-train it. While the total pre-training dataset is smaller than the
847 one used in the original paper, the number of subjects (i.e., the number of sEEG location
848 configurations) is greater than in the original paper. The data samples are 4 seconds.
- 849 • For the PopT model, we include neural recordings from all available subjects within the
850 Brain Treebank dataset for pre-training. The data samples are 4 seconds.

864 **C MODEL DETAILS**
865866 The BrainFCIR model (Table 4) is a novel neurofoundation model for intracranial sEEG recordings,
867 as shown in Figure 1 (a). The architecture of BrainFCIR contains two parts: (1) Patch Tokenizer, (2)
868 Temporal & Spatial Transformer, and (3) Channel Cluster Module. During the pre-training stage,
869 one additional "Token Predictor" (i.e., linear projection) is added after the "Spatial Transformer" for
870 functional context discrimination.
871872 **Functional Context Discrimination.** Since sEEG channels capture local and depth information
873 from different brain regions, their recordings inherently capture unique neural information with
874 minimal overlap. This makes the functional context discrimination task better suited for learning
875 inter-channel relationships compared to mask-based reconstruction approaches (Zhang et al., 2023;
876 Jiang et al., 2024b). To ensure balanced label distribution, we designate only 10% of unreplaced
877 channels as positive samples during pre-training.
878879 **Channel Cluster Module.** After pre-training with the spatial context task, we calculate the chan-
880 nel connectivity $\mathcal{P} \in \mathbb{R}^{C \times C}$ following Algorithm 1. Then, spectral cluster (Ng et al., 2001) is
881 applied to group channels into functional clusters, using scikit-learn’s (Pedregosa et al., 2011)
882 `cluster.SpectralClustering` with default function arguments.
883884 **Algorithm 1** The calculation of channel connectivity $\mathcal{P} \in \mathbb{R}^{C \times C}$.
885886 **Require:** $\{\mathcal{X}_i \in \mathbb{R}^{C \times T} | i = 1, \dots, N_{\text{samples}}\}$ $\triangleright N_{\text{samples}}$ is the number of samples.
887 $\mathcal{P} \leftarrow \mathbf{0}_{C \times C}$ $\triangleright \mathcal{P} \in \mathbb{R}^{C \times C}$ is initialized as 0s.
888 **while** $i \leq N_{\text{samples}}$ **do**
889 $\hat{\mathcal{P}} \leftarrow \text{model}(\mathcal{X}_i)$ $\triangleright \hat{\mathcal{P}} \in \mathbb{R}^{N_{\text{layer}} \times N_{\text{head}} \times C \times C}$ is spatial attention scores.
890 $\hat{\mathcal{P}} \leftarrow \text{mean}(\hat{\mathcal{P}}, \text{axes} = [0, 1])$ $\triangleright \hat{\mathcal{P}} \in \mathbb{R}^{C \times C}$ is averaged across [layer,head]-dimensions.
891 $\mathcal{P} \leftarrow \mathcal{P} + \hat{\mathcal{P}} / N_{\text{samples}}$
892 **end while**
893894 **D MODEL EFFICIENCY**
895896 Table 5 shows the FLOPs (with `thop` package) and per-trial inference time across all methods.
897 Our model achieves superior efficiency with the smallest parameter count (3.32M) and lower com-
898 putational footprint (19.13 GFLOPs), enabling faster inference (23.73 ms) compared to existing
899 approaches. With functional channel selection, computational cost drops significantly to 2.42 GFLOPs
900 while maintaining—or even improving—decoding performance, further reducing inference time to
901 16.37 ms. These results demonstrate that our method offers a highly efficient and practical solution
902 for decoding speech perception from intracranial sEEG recordings. All experiments were conducted
903 on a single NVIDIA V100 GPU.
904905 **E DATA SCALING**
906907 To evaluate data efficiency, we assessed downstream classification performance (Table 6) of BrainF-
908 CIR when pretrained on progressively larger fractions of the available data (5% to 75%). Performance
909 demonstrated a clear scaling trend with increased pretraining data. Data subsets were constructed
910 through incremental session-wise addition until the target percentage was met. This process was
911 repeated across 6 random seeds to ensure robustness. For smaller data fractions, we adjusted the
912 number of pretraining epochs to maintain a consistent total number of parameter updates.
913914 **F ABLATIONS ON CHANNEL SELECTION**
915916 The spatial-attention weights are not equivalent to traditional signal-level functional connectivity (FC)
917 measures (e.g., coherence). The averaged spatial-attention weights can be viewed as an alternative to

918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
Table 4: The hyperparameters for BrainFCIR training.

Module	Name	Value
Patch Tokenizer	# of Input Channels	{1,64,64}
	# of Output Channels	{64,64,64}
	Kernel Size	{9,9,3}
	Stride	{5,5,2}
	Padding	{4,4,1}
Temporal Transformer	Flatten Window	2
	# of Transformer Layers	4
	Hidden Size	128
	MLP Size	512
	MLP Dropout Ratio	{0.2,0.}
Spatial Transformer	# of Attention Heads	8
	Attention Head Size	64
	Attention Dropout Ratio	0.2
	# of Transformer Layers	4
	Hidden Size	128
Token Predictor	MLP Size	512
	MLP Dropout Ratio	{0.2,0.}
	# of Attention Heads	8
	Attention Head Size	64
	Attention Dropout Ratio	0.2
Linear Projection	128 → 128	
Optimizer	Batch Size	64
	Maximum Learning Rate	3e-4
	Minimum Learning Rate	5e-6
	Learning Rate Scheduler	Cosine
	Optimizer Type	AdamW
	Adam β	(0.9, 0.99)
	Weight Decay	0.05
	Total Epochs	100
	Warm-up Epochs	10
	EMA momentum schedule	linear
w/ channel-selection	EMA start momentum	0.996
	EMA final momentum	1

955
956
Table 5: Model Efficiency Analysis on Brain Treebank dataset.

Methods	Model Size	GFLOPs	Time (ms)
Brant	500M	116.2700±6.6485	54.11±5.21
PopT	20M	27.9417±1.7639	26.39±3.16
BrainFCIR	3.32M	19.1308±1.0945	23.73±2.92
w/ channel-selection	-	2.4153±0.1719	16.37±2.19

966
967
968
969
970
analyze connectivity. To further illustrate the advantages of our spatial-attention-based estimation of
functional connectivity, we additionally perform clustering on the coherence-based FC, ranking each
cluster using a downstream task. Since channel selection based on BrainFCIR estimation only uses
channels contained in the first cluster, we also report the decoding performance on the first cluster to
evaluate whether the coherence-based FC clustering can accurately identify functional boundaries.

971
Since coherence primarily estimates channel connectivity based on low-order correlations between
channels, it is more susceptible to the influence of channel anatomical proximity. Compared to our

972 Table 6: BrainFCIR’s downstream performance scales as a function of pre-training data size.
973

Data Percentage	Pitch	Volume	Sentence Onset	Speech/Non-Speech
100%	0.77±0.02	0.89±0.02	0.94±0.01	0.96±0.01
75%	0.76±0.02	0.87±0.02	0.93±0.01	0.95±0.01
50%	0.75±0.02	0.88±0.02	0.93±0.02	0.95±0.01
25%	0.71±0.03	0.83±0.04	0.88±0.04	0.90±0.03
10%	0.68±0.04	0.80±0.03	0.85±0.05	0.87±0.03
5%	0.64±0.04	0.75±0.04	0.83±0.04	0.83±0.05

982 Table 7: BrainFCIR’s downstream performance varies across different connectivity estimation for
983 channel selection.
984

Connectivity Type	Pitch	Volume	Sentence Onset	Speech/Non-Speech
-	0.77±0.02	0.89±0.02	0.94±0.01	0.96±0.01
BrainFCIR	0.78±0.02	0.91±0.02	0.94±0.01	0.97±0.01
Corherence	0.75±0.02	0.88±0.02	0.90±0.01	0.92±0.01

991
992
993 method (Table 7), it struggles to effectively estimate the precise boundaries of functional modules,
994 resulting in lower performance.995 Besides, we further execute ablations on the number of clusters (Table 8). Fewer clusters (e.g., 5)
996 reduce the spatial resolution of functional groups and may include irrelevant channels. More clusters
997 maintain performance but require more group-level evaluation and combination.998 Table 8: BrainFCIR’s downstream performance maintains when varying number of clusters for
999 channel selection.
1000

# of clusters	5	8	10	15	20
Pitch	0.77±0.02	0.78±0.02	0.78±0.02	0.78±0.02	0.78±0.02
Volume	0.90±0.02	0.90±0.02	0.91±0.02	0.91±0.02	0.91±0.02
Sentence Onset	0.94±0.01	0.94±0.01	0.94±0.01	0.94±0.01	0.94±0.01
Speech/Non-Speech	0.96±0.01	0.97±0.01	0.97±0.01	0.97±0.01	0.97±0.01

1008
1009
1010 G CROSS-SUBJECT TRANSFER
10111012 To evaluate the generalizability of our pretrained weights, we performed a leave-one-out (LOO)
1013 cross-validation. A model was pretrained on all but one subject, then fine-tuned and evaluated on
1014 the held-out subject. Results indicate that excluding a subject from pretraining does not significantly
1015 impact downstream performance (Table 9), demonstrating the robustness and potential utility of our
1016 approach for new, unseen data.1017
1018 H ADDITIONAL RESULTS
10191020 Since Brain Treebank dataset was collected while subjects watched movies, we extracted the audio
1021 portion of the movies to evaluate speech synthesis tasks (Chen et al., 2024). The evaluation results of
1022 different baselines are shown in Table 10. Our model still outperforms all baselines, demonstrating
1023 that the explicit modeling of functional connections helps in decoding cognitive tasks such as speech
1024 perception. We reported the Pearson Correlation Coefficient between the predicted mel-spectrogram
1025 and the ground truth.

1026

Table 9: BrainFCIR’s downstream performance in leave-one-out (LOO) setting.

1027

1028

1029

1030

1031

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

1042

1043

1044

1045

1046

1047

1048

1049

Setting	Pitch	Volume	Sentence Onset	Speech/Non-Speech
-	0.77±0.02	0.89±0.02	0.94±0.01	0.96±0.01
LOO	0.74±0.03	0.87±0.02	0.93±0.02	0.94±0.01

Table 10: BrainFCIR’s downstream performance on speech synthesis task.

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

1042

1043

1044

1045

1046

1047

1048

1049

I MODEL INTERPRETABILITY

To visualize electrode coverage across brain regions (Figure 5), we mapped intracranial electrode locations to anatomical regions using the Desikan-Killiany atlas (Desikan et al., 2006). For each subject, electrode coordinates were registered to the fsaverage surface template and assigned to corresponding cortical parcellations. ROI activation intensities were computed by normalizing electrode counts per region across subjects and experimental conditions. The resulting intensity maps were projected onto inflated cortical surfaces using Nilearn’s surface plotting functions. Brain visualizations display both hemispheres in lateral view with a color-coded intensity scale (red colormap) representing normalized electrode density, providing clear spatial representation of recording coverage across cortical areas for each experimental task.

We performed channel clustering analysis across four distinct auditory-linguistic tasks: (1) Pitch; (2) Volume; (3) Sentence Onset; (4) Speech/Non-Speech. Based on final classification performance, we selected specific cluster groups and analyzed the spatial distribution of channels within these clusters. The clustering results revealed two distinct patterns. Pitch and volume classification tasks demonstrated consistent clustering patterns, while sentence onset detection and speech/non-speech classification tasks showed similar groupings to each other but differed from the pitch/volume conditions. For pitch and volume classification tasks, selected channels were predominantly distributed in bilateral auditory regions, specifically the superior temporal gyrus and transverse temporal gyrus. Secondary distributions were observed in Wernicke’s area and the middle frontal gyrus, suggesting engagement of both primary auditory processing and higher-order linguistic regions.

In contrast, sentence onset detection and speech/non-speech classification tasks showed channels primarily concentrated in the same bilateral auditory areas (superior temporal gyrus and transverse temporal gyrus), with comparable representation in Wernicke’s area. However, a striking difference emerged in the middle frontal gyrus, where virtually no channels were selected for these tasks, distinguishing them from the pitch/volume conditions. This differential pattern suggests distinct neural mechanisms underlying continuous acoustic feature processing versus discrete linguistic event detection, which is consistent with the results of previous neuroscience research (DeWitt & Rauschecker, 2013; Friederici, 2011; Van der Burght et al., 2019).

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

1080

1081

1082

1083

1084

1085

1086

1087

1088

1089

1090

1091

1092

1093

1094

(a) Pitch & Volume

1095

1096

1097

1098

1099

1100

1101

1102

1103

1104

1105

1106

(b) Sentence Onset & Speech/Non-Speech

1107

1108

1109

1110

1111

1112

1113

1114

1115

1116

1117

1118

1119

1120

1121

1122

1123

1124

1125

1126

1127

1128

1129

1130

1131

1132

1133

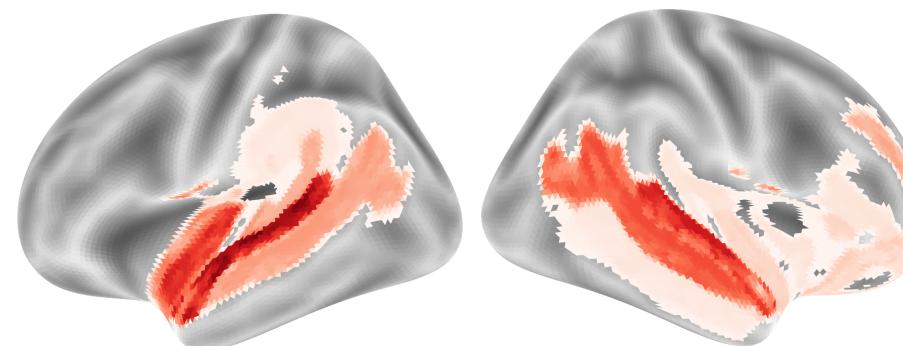
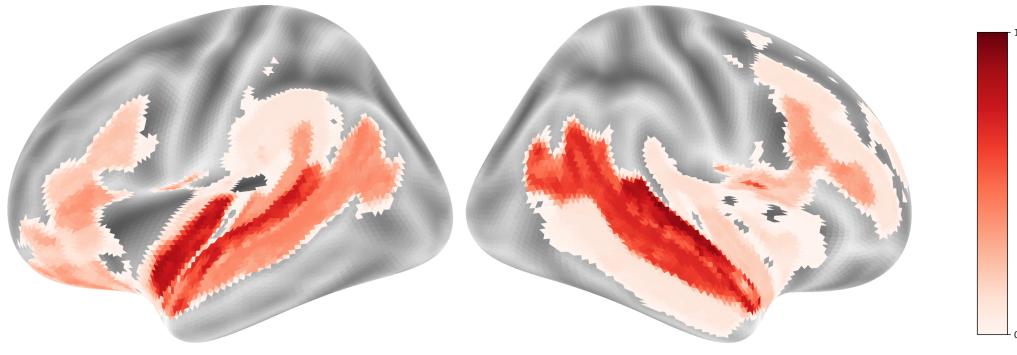


Figure 5: Visualization of the alignment between selected groups and language-related regions.

1134
1135

1136

1137

J SUBJECT-WISE CHANNEL CLUSTER

1138

1139

1140

1141

1142

1143

1144

1145

1146

1147

1148

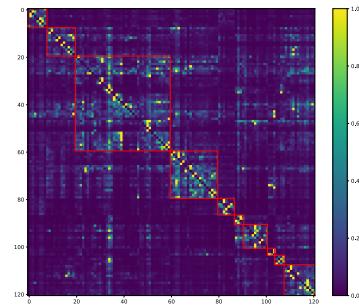
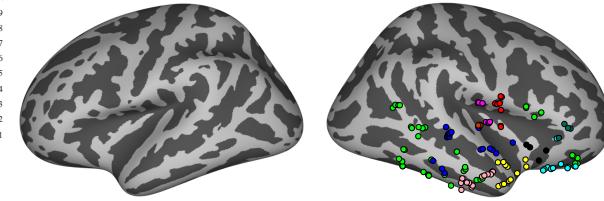
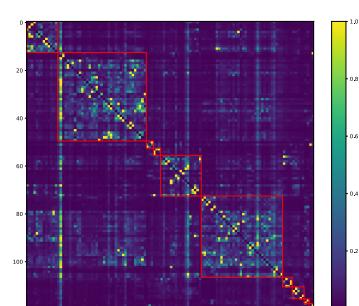
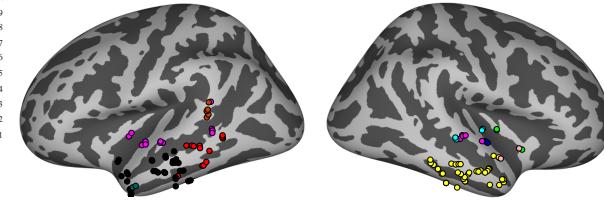
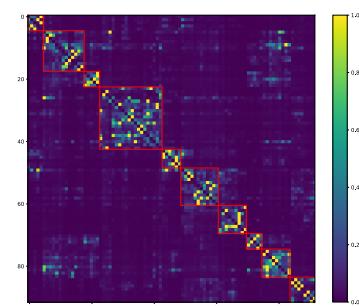
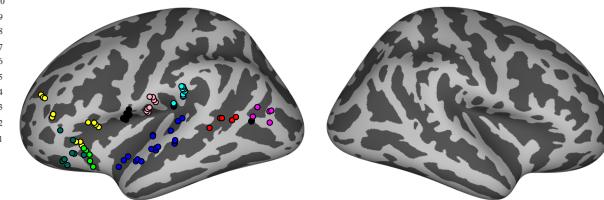
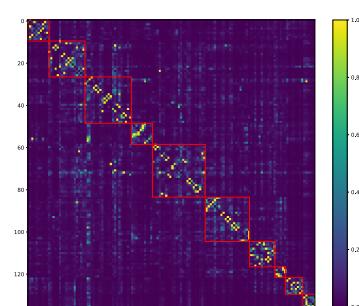
subj-01Cluster
10
9
8
7
6
5
4
3
2
1**subj-02**Cluster
10
9
8
7
6
5
4
3
2
1**subj-03**Cluster
10
9
8
7
6
5
4
3
2
1**subj-04**Cluster
10
9
8
7
6
5
4
3
2
1

Figure 6: Channel clusters from subjects (01-04).

1179

1180

1181

1182

1183

1184

1185

1186

1187

1188

1189

1190

1191

1192

1193

1194

1195

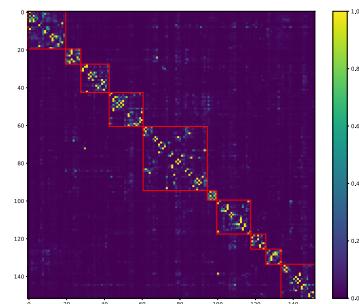
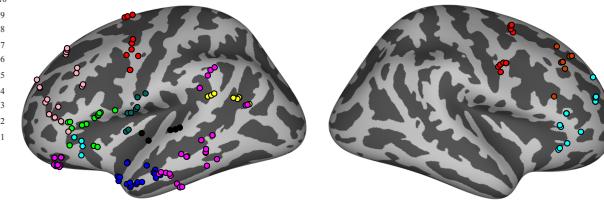
1196

1197

1198

1199

1200

subj-06

1201

1202

1203

1204

1205

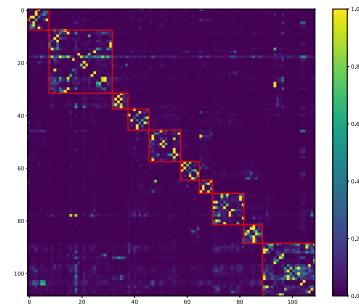
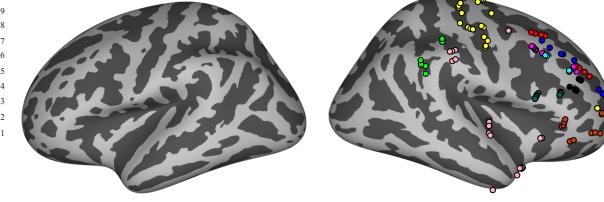
1206

1207

1208

1209

1210

subj-07

1211

1212

1213

1214

1215

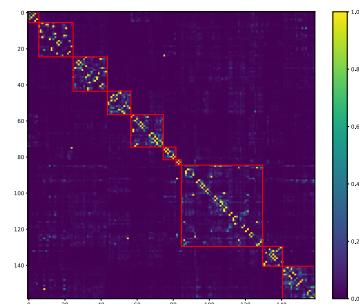
1216

1217

1218

1219

1220

subj-10

1221

1222

1223

1224

1225

1226

1227

1228

1229

1230

1231

1232

1233

1234

1235

1236

1237

1238

1239

1240

1241

Figure 7: Channel clusters from subjects (05,06,10).