Unlocking Volition: Proactive Intention Decoding via Interpretable Graph Learning of Multi-Region ECoG

Kaizhong Zheng[1] **Jiajun Ma**[1] **Hongru Liu**[2] **Xinjian Li**[3]
Xi Wei[1] **Liangjun Chen**[1*] **Badong Chen**[1]

[1]Xi'an Jiaotong University [2]Tsinghua University [3]Zhejiang University
{kzzheng, xiwei, liangjunchen}@xjtu.edu.cn majiajun@stu.xjtu.edu.cn
chenbd@mail.xjtu.edu.cn hr-liu23@mails.tsinghua.edu.cn lxjbio@zju.edu.cn

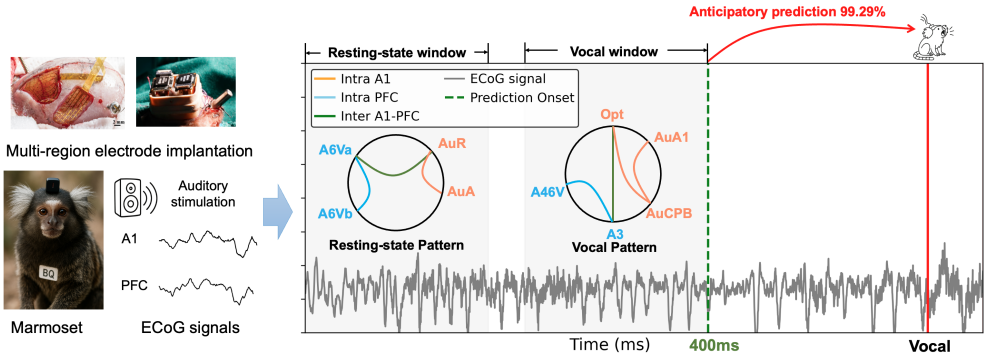


Figure 1: Overview of anticipatory marmoset vocalization prediction from multi-region ECoG recordings through ECoG-IBGT. Using common marmosets implanted with high-density electrode arrays across multiple cortical regions, we recorded high-fidelity ECoG signals from A1 and PFC under an auditory stimulation paradigm; our approach achieved 99.29% accuracy in predicting vocalization 400 ms before onset and revealed biologically meaningful patterns distinguishing vocalization from rest.

Abstract

Current brain–machine interfaces (BMI) face fundamental limitations due to inherent latency from reliance on delayed motor cortical signals, and computational overhead, restricting their effectiveness in real-time applications such as rehabilitation therapy. Recent neuroscience indicates that prefrontal and sensory cortical activities precede motor execution, offering an opportunity for proactive intent prediction. However, challenges remain in acquiring multi-region neural data, efficiently decoding high-dimensional signals, and ensuring model interpretability. To address these, we developed a high-density electrocorticography (ECoG) paradigm in marmosets and introduced an information-bottleneck-driven graph transformer (ECoG-IBGT), reframing neural decoding as graph classification. ECoG-IBGT achieves 99.29% accuracy 400 ms before vocal onset under within-dataset evaluation, and attains 95.98%, 88.03%, 72.13%, and 70.7% under cross-dataset, leave-session-out, cross-subject, and external evaluation, respectively, laying the foundation for reliable, low-latency BMIs. The code is available at https://github.com/ZKZ-Brain/ECoG-IBGT.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Human–computer interaction (HCI) is undergoing a profound transformation, with brain–machine interfaces (BMIs) increasingly envisioned as a pathway toward seamless and adaptive communication between humans and machines. Traditionally, most BMIs rely on decoding motor cortex activity to control external devices (Volkova et al., 2019). While motor–cortex–based decoding has achieved notable progress in both timeliness and accuracy, such approaches are fundamentally constrained: they capture the consequences of intention—motor execution—rather than intention itself (Tauste Campo et al., 2015; Sun et al., 2015; Ariani et al., 2022). As a result, current BMI systems remain largely reactive, limiting their capacity for anticipatory, low-latency interaction (LaRocco & Paeng, 2020; Skomrock et al., 2018). To enable truly natural HCI, BMIs must move beyond "recognizing actions before they unfold" toward "identifying intentions at the moment they arise", thereby aligning interaction more closely with human cognitive mechanisms.

Recent neuroscientific evidence strongly motivates this paradigm shift. Studies in common marmosets have shown that upstream cortical regions, including the prefrontal cortex (PFC) and auditory cortex (AC), exhibit activation 200–300 ms prior to spontaneous vocalization (Mitelut et al., 2022; Tsunada & Eliades, 2025). These pre-vocal signals not only precede motor execution but also predict subsequent acoustic features, revealing a temporal window for proactive intention prediction. This suggests that intention-related information emerges in higher-order cortices before motor output, offering a compelling foundation for intention decoding. Thus we operationalize this evidence as a machine-learning task—Proactive Intention Decoding (PID)—that, given multi-regional neural windows at time $\boldsymbol{t}$, predicts whether (and how) a behavior will occur at $\boldsymbol{t} + \boldsymbol{\Delta}$.

By reframing intention decoding as a machine learning problem, we aim to move BMIs from "post hoc recognition" to "proactive prediction", opening the door to the next generation of intention-aware, real-time human–machine interaction systems. Despite this insight, computational frameworks that can operationalize proactive intention decoding remain scarce. Current BMI technologies rarely exploit these upstream signals, and the gap between neuroscientific findings and machine learning practice is striking (Rouzitalab et al., 2023). Closing this gap introduces a new challenge for the ML community: how to design models that can capture weak, distributed, and dynamic neural patterns to predict internal intentions in real time:

(1) **High-quality, multi-regional neural acquisition**: Intention-related activity in PFC and AC is far weaker and more entangled than motor cortex signals (Sun et al., 2015; Ariani et al., 2022). Accurate decoding requires high-throughput recordings with broad coverage, yet EEG lacks resolution and implantable electrodes face biocompatibility limits. ECoG provides a promising compromise—good fidelity with moderate invasiveness—but current systems still suffer from limited coverage and durability, leaving a shortage of large-scale, high-quality datasets (Wang et al., 2023).

(2) **Scalable decoding of high-dimensional signals**: Most pipelines are tailored to low-dimensional EEG (Wu et al., 2016) and struggle with multi-regional ECoG. Hand-crafted features incur high costs and neglect dynamic inter-regional dependencies, limiting both accuracy and speed (Wang et al., 2016; Das et al., 2019). Early upstream signals are weak (Hockley et al., 2025b). Thus, new frameworks are needed that can extract discriminative patterns from high-dimensional ECoG, capture cross-region temporal–spectral dynamics, and enable real-time inference.

(3) **Limited interpretability**: While GNNs offer structural transparency, their explanations remain local or edge-level (e.g., GATs, GNNExplainer) (Veličković et al., 2018b; Ying et al., 2019). This hampers neuroscientific validity and reliability. Embedding interpretability during training—e.g., by isolating behaviorally relevant ensembles and motifs—could improve robustness, yield biologically meaningful insights into intent formation, and reduce the risk of failure in real-world BMIs.

To overcome these challenges, inspired by the vocalization prediction paradigm, we designed a vocalization-driven paradigm in common marmosets implanted with ultra-dense dual-region ECoG arrays (128 channels; 1.0–1.52 mm spacing) targeting primary auditory cortex (A1) and medial prefrontal cortex (mPFC). From this dataset, we propose ECoG-IBGT—an information-bottleneck-driven graph transformer that reformulates neural decoding as graph classification. By integrating dynamic graph construction with a learnable IB subgraph generator, our approach achieves dimensionality reduction while preserving behaviorally-relevant multi-region dynamics.

To summarize, our key contributions are:

- **High-quality, multi-subject, multi-context ECoG dataset.** We collected a large-scale dataset using ultra-dense dual-region ECoG arrays (128 electrodes, 1.0–1.52 mm spacing) implanted in marmosets, covering naturalistic conditions such as free movement and social interaction. The six curated datasets provide a valuable high-throughput resource for studying spontaneous behavior and benchmarking early intention decoding, helping bridge the translational gap between neuroscience and real-world BMI applications.
- **Accurate and proactive intention decoding architecture.** ECoG-IBGT formulates neural decoding as graph classification with dynamic brain graphs and spatiotemporal attention, achieving 99.29% accuracy 400 ms before vocal onset and extending the prediction horizon for low-latency BMI applications.
- **Built-in subgraph-level interpretability for neural decoding.** We propose an information-bottleneck subgraph generator that learns compact, interpretable subnetworks, revealing key cross-regional interactions and enabling scalable decoding for cognitive BMI applications.

2 Related Work

Decades of research have shown that multi-area recordings in primates reveal temporally precise information flow during flexible behaviors, supporting distributed models of decision-making and sensorimotor integration (Siegel et al., 2015). The common marmoset is particularly suited for studying vocal behavior, as distributed circuits across auditory, premotor, and prefrontal regions orchestrate perception, planning, and context-dependent modulation of vocal output Roy & Wang (2016); Jovanovic et al. (2022); Zanini et al. (2023). While non-invasive EEG captures low-frequency movement-related potentials with moderate success (Bulea et al., 2014; Liu et al., 2018; Shafiul Hasan et al., 2020), ECoG offers higher spatial-temporal resolution and signal-to-noise ratio, enabling continuous trajectory decoding (Chen et al., 2014), grasp-force prediction (Chen et al., 2014), and muscle activity estimation (Shin et al., 2012; Umeda et al., 2019). Chronic wireless ECoG in freely behaving marmosets has further revealed coordinated activity across multiple regions closely linked to natural vocalizations and motor behavior (Walker et al., 2021).

Recent neural decoding approaches using EEG increasingly rely on graph neural networks (GNNs) to decode brain activity Liu et al. (2025); Nejedly et al. (2025). However, existing GNN interpretability methods face two critical limitations: first, explanations often suffer from poor reproducibility across runs or datasets; second, treating node and edge importance independently can introduce logical inconsistencies, undermining the biological interpretability of inferred subnetworks. To address these challenges, we introduce a Connectivity Loss that enforces mutual alignment between node and edge masks, ensuring that edge-level and node-level importance are consistent. In addition, our Interpretability Encoder integrates the masks directly into the learning process, which promotes the extraction of robust, biologically meaningful, and highly stable biomarkers.

3 Methods

3.1 Problem Definition

The aim of this study is to predict the marmosets' intention before vocalization. To facilitate mathematical modeling, we reformulated the problem as a graph classification task. Specifically, each marmoset's ECoG recordings consist of 128 channels. For each vocalization event, we extract a single δ-second, 128-channel window ending at 400 ms before onset and computing pairwise functional connectivity across the 128 channels yields one graph. Rest samples are constructed analogously from non-overlapping baseline segments (Figure 2A). Each brain functional graph G contains n regions (channels), defined as $G = (X, A)$, where $X \in \mathbb{R}^{n \times n}$ are node features and $A \in \{0, 1\}^{n \times n}$ is an adjacency matrix, as shown in Figure 2A. Thus, our goal is to train a neural network to learn a graph representation h_G from the brain functional graph G_{ECoG} to predict the binary vocalization label y:

$$f : G_{ECoG} \rightarrow h_G \rightarrow y_{vocal}, \quad y \in \{0, 1\}. \tag{1}$$

The overall framework of ECoG-IBGT is depicted in Figure 2B. We employ the concept of information bottleneck to achieve the dual objective of removing redundant information and accurately predicting the vocalization state. ECoG-IBGT comprises three main modules: IB-subgraph generator,

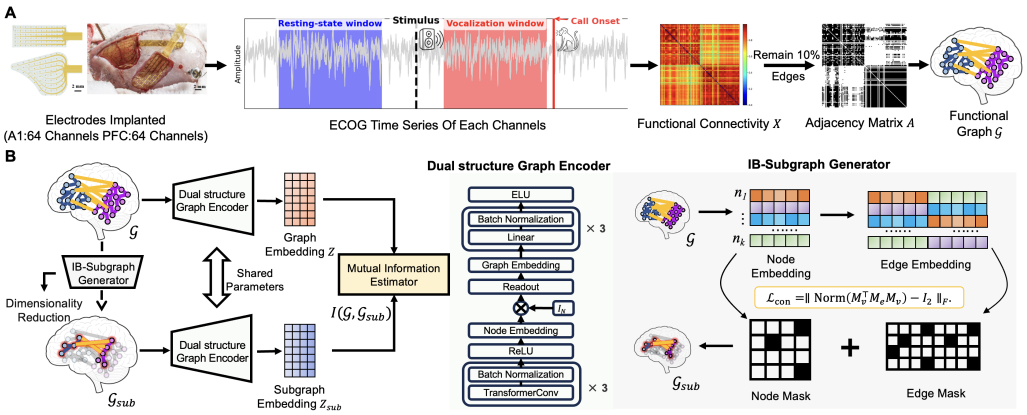


Figure 2: Overview of the graph-construction pipeline and proposed ECoG-IBGT framework. (A) ECoG signals are segmented into a resting-state window and a vocalization-state window. We compute pairwise Pearson's correlations across channels to obtain the functional connectivity matrix X. To construct the adjacency matrix A, we retain the top 10% of edges from X, yielding the functional graph $G = (A, X)$. (B) The original graph G is pruned by the IB-Subgraph Generator to produce the subgraph G_{sub}. Graph embeddings for both G and G_{sub} are generated via the Dual Structure Graph Encoder. Finally, we estimate the MI $I(G, G_{\text{sub}})$ using a mutual-information estimator.

graph encoder, and mutual information estimator. The IB-subgraph generator extracts a subgraph with soft edge masks M_e and node masks M_v from the original graph G. The graph encoder learns graph embeddings for both the G and the subgraph G_{sub}. Then the mutual information (MI) $I(G; G_{\text{sub}})$ between G and G_{sub} is calculated.

3.2 IB-Subgraph Generator

Current brain network subgraph methods typically do not jointly constrain node and edge selection, resulting in subgraphs that are neither interpretable nor compact. To address this limitation, we propose a novel subgraph generator that simultaneously prunes redundant nodes and sparsifies inter-node connections, yielding concise and functionally meaningful substructures.

The procedure of the IB-Subgraph Generator is shown in Figure 2B. Specifically, we compute node mask M_v and edge mask M_e:

$$M_v^{(i)} = \text{GumbelSoftmax}(\sigma(\text{MLP}(x_i))), \quad M_e^{(ij)} = \text{GumbelSoftmax}(\sigma(\text{MLP}(x_i; x_j))). \tag{2}$$

where σ represents the sigmoid function and x_i, x_j represent the node features.

To ensure the stability and robustness of the subgraph, we introduce the connectivity loss:

$$\mathcal{L}_{\text{conn}} = \left\| \text{Norm}(\hat{M}_v{}^{\top} M_e \hat{M}_v) - I_2 \right\|_F, \tag{3}$$

where $\hat{M}_v = [M_v, 1 - M_v]$.

We define $D = \hat{M}_v{}^{\top} M_e \hat{M}_v \in \mathbb{R}^{2\times 2}$, and it has two important entries:

$$D_{11} = \sum_{i,j} \hat{M}_v^i \, M_e^{ij} \, \hat{M}_v^j, \quad D_{22} = \sum_{i,j} (1 - \hat{M}_v^i) \, M_e^{ij} \, (1 - \hat{M}_v^j). \tag{4}$$

where the size of D_{11} quantifies connectivity within the subgraph, with a larger D_{11} indicating stronger connectivity in its internal structure. D_{22} represents connectivity outside the subgraph. Thus, minimizing connectivity loss preserves connectivity within the subgraph while promoting external sparsity, effectively removing redundant nodes and edges to yield a meaningful subgraph.

3.3 Dual Structure Graph Encoder

Dual structure graph encoder is used to obtain the graph embedding from the original graph and subgraph. We use the graph transformer as the backbone network. In addition, we employ an

interpretability-driven message passing mechanism to update the node embedding. The process of obtaining node embeddings in a traditional graph transformer can be expressed as follows:

$$h_v^{(\ell+1)} = \mathrm{ReLU}\Big(\mathrm{BN}\big(\mathrm{TFConv}(h_v^{(\ell)}, E)\big)\Big), \mathrm{TFConv}(h_v^{(\ell)}, E) = \sum_{k \in \mathcal{N}(v)} \alpha_{vk} \cdot W h_v. \tag{5}$$

where BN denotes batch normalization layer, TFConv represents graph transformer layer, α_{vk} is the attention weight, and W is the weight matrix used for value transformation.

To ensure interpretability and stability, we design an explainable message propagation architecture that seamlessly integrates explanation signals including node mask and edge mask during model training. Given node mask M_v and edge mask M_e, the message passing process is defined as:

$$h_v^{(\ell+1)} = \mathrm{ReLU}\Big(\mathrm{BN}\big(\mathrm{TFConv}(h_v^{(\ell)}, E, M_e)\big)\Big) \odot M_v. \tag{6}$$

3.4 Mutual Information Estimator

The graph IB objective includes two MI terms:

$$\min_{G_{\text{sub}}} I\left(G_{\text{sub}}, G\right) - \beta I\left(G_{\text{sub}}, y\right). \tag{7}$$

Minimizing $-I\left(\mathcal{G}_{\text{sub}}, y\right)$ (*a.k.a.*, maximizing $I\left(\mathcal{G}_{\text{sub}}, y\right)$) encourages G_{sub} to be maximally predictive of the graph label y. Mathematically, we have:

$$-I(G_{\text{sub}}, y) \leq \mathbb{E}_{y, G_{\text{sub}}}[-\log q_\theta(y|G_{\text{sub}})] := \mathcal{L}_{\text{CE}}(G_{\text{sub}}, y). \tag{8}$$

where $q_\theta\left(y|G_{\text{sub}}\right)$ is the variational approximation of the true mapping $p\left(y|G_{\text{sub}}\right)$ from G_{sub} to y. Eq. (8) indicates that minimizing $-I\left(G_{\text{sub}}, y\right)$ approximately equals minimizing the cross-entropy loss $\mathcal{L}_{\text{CE}}$.

As for the MI term $I\left(G_{\text{sub}}, G\right)$, we use Hilbert-Schmidt independence criterion (HSIC) to calculate the mutual information between G and G_{sub}. HSIC directly estimates the MI without the need for discrete probability density function (PDF) estimation or any auxiliary neural networks. Formally, $I\left(G_{\text{sub}}, G\right)$ can be defined as:

$$I\left(G_{\text{sub}}, G\right) = (N-1)^{-2}\mathrm{tr}(K_G H K_{G_{\text{sub}}} H), \tag{9}$$

where N is batch size, tr is the trace operator, K_G and $K_{G_{\text{sub}}}$ are kernel matrices with $K_G^{ij} = k_G(Z^i, Z^j)$ and $K_{G_{\text{sub}}}^{ij} = k_{G_{\text{sub}}}(Z_{\text{sub}}^i, Z_{\text{sub}}^j)$, respectively and $H \in \mathbb{R}^{N \times N}$ is the centering matrix $H = I_{ij} - \frac{1}{N}$, Z and Z_{sub} represent the node embeddings derived from the graph encoder.

In this work, we use the radial basis function (RBF) kernel κ to obtain k_G and $k_{G_{\text{sub}}}$:

$$\kappa\left(z^i, z^j\right) = \exp\left(-\frac{\left\|z^i - z^j\right\|^2}{2\sigma^2}\right). \tag{10}$$

For the kernel width σ, we estimate 10 nearest distances of each sample and obtain the mean. We set σ to the average of these per-sample means.

Thus, the final objective of ECoG-IBGT is:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha\, \mathcal{L}_{\text{conn}} + \beta\, I\left(G_{\text{sub}}, G\right). \tag{11}$$

where α and β balance the contributions of each component.

4 Experiments

This section evaluates the effectiveness and presents the interpretability of ECoG-IBGT with extensive experiments. We aim to address the following research questions:

RQ1. Can the high-SNR, high spatio-temporal resolution ECoG signals acquired by our multi-region paradigm enable earlier and more accurate intention prediction?

RQ2. How does ECoG-IBGT perform compared with state-of-the-art models of various types?

RQ3. What interpretable neural mechanisms does ECoG-IBGT provide to guide dimensionality reduction and enable anticipatory prediction?

4.1 Experimental Settings

Datasets. Leveraging two adult common marmosets (named BQ and AZ) implanted with ECoG electrode (128 electrodes, 1.0–1.52 mm spacing) covering auditory cortex (64 channels) and PFC (64 channels), we employed a series of paradigms to achieve dense ECoG recordings. Signals were sampled at 500 Hz and band-pass filtered (1–200 Hz) to remove power-line and high-frequency noise. Table 1 summarizes the paradigm settings. For each marmoset, sessions correspond to pure-tone stimulation (i.e., 2,000, 4,000, 8,000, and 16,000 Hz), conspecific-call stimulation (i.e., playback of recorded marmoset vocalizations), and antiphonal-calling interactions. Each session lasted approximately 32 minutes, yielding a total of 233 minutes of recordings across all sessions. To balance the requirements of system latency and behavioral responsiveness, a 400 ms margin before vocal onset provided lead time for downstream execution, while also ensuring that data acquisition, feature extraction, and algorithmic inference can be completed in time to support real-time, closed-loop brain–machine interaction. Our data acquisition has passed the ethical certification of Zhejiang University. To focus on stimulus-evoked activity, we analyzed only trials in which vocalizations followed the stimulus, thereby excluding non-evoked or spontaneous vocalizations and minimizing potential confounds. Within these trials, vocalization windows refer to the periods preceding stimulus-induced vocalizations, while rest windows were sampled prior to stimulus onset to ensure comparability.

Table 1: Sessions across paradigms. B1–B7 and A1–A3 are session identifiers for the recordings. A1, chair-fixed; A2, free-moving.

Stimulus	Subject BQ	Subject AZ
Pure-tone recordings	B1-B4	–
Phee recordings	–	A1-A2
Antiphonal calling	B5-B7	A3

Based on the experimental subject, stimulation paradigm, and the marmoset's behavioral state, we assembled 6 datasets:

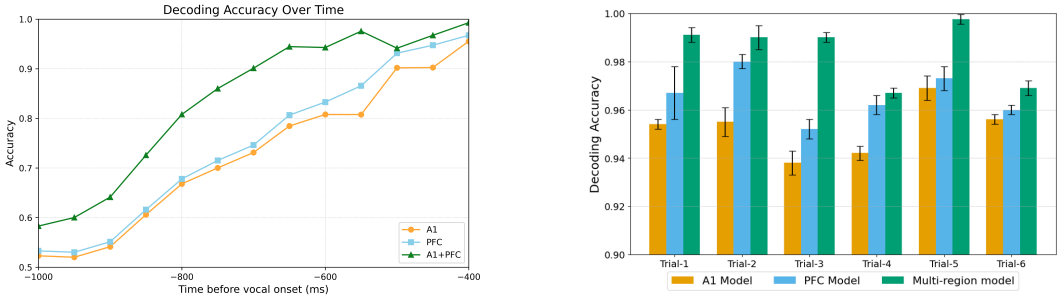
- **Dataset-1**: pooled 7 sessions (B1–B7) from BQ, yielding 770 samples (385 vocalization vs. 385 rest);
- **Dataset-2**: comprised 3 sessions (A1–A3) from AZ, also totaling 770 samples (385/385);
- **Dataset-3**: combined 4 sessions (B5–B7 and A3) in which antiphonal calling by paired marmosets elicited vocalizations, resulting in 1012 samples (506/506);
- **Dataset-4**: merged 6 sessions (B1–B4 with pure-tone and conspecific-call playback, plus A1–A2), likewise producing 1012 samples (506/506);
- **Dataset-5**: consisted of a single session (A2) under free-behavior conditions with playback-induced stimuli, containing 222 samples (111/111);
- **Dataset-6**: comprised 1 session (A1) under chair-restrained conditions with identical playback stimuli, also yielding 222 samples (111/111).

Implementation Details. For experiments, we use a three-layer graph transformer module and set the number of heads M to 3 for each layer. We randomly split the data into 70% for training, 10% for validation, and the remaining 20% as the test set. In the training process of our proposed model, we use an Adam optimizer with an initial learning rate of 5×10^{-5} , a weight decay of 1×10^{-5} and set the loss-balancing hyperparameters α and β both to 1×10^{-5}. The batch size is set to 128. All models are trained for 150 epochs, and the epoch with the highest accuracy on the validation set is selected for performance comparison on the test set. The model is trained on an NVIDIA H20-NVLink. Please refer to the repository for the full implementation of ECoG-IBGT. All reported performances are the average of 5 random runs.

4.2 Multi-region Paradigm Decoding (RQ1)

Decoding performance of multi-region vs. single-region ECoG on different onset times. As shown in subfigure (a) of Figure 3, we compare the decoding accuracy of multi-region vs. single-region ECoG signals using a sliding 3 s window from -1 s to -400 ms relative to vocal onset (where -1 s and -400 ms denote the lead time at the window's end). The multi-region model outperforms single-region models at almost all anticipation time points, with the greatest gain observed between -800 ms and -550 ms.

Decoding performance of multi-region vs. single-region ECoG on different datasets . To validate reliability, we compare decoding performance across 6 datasets using multi-region vs single-region ECoG signals, as shown in subfigure (b) of Figure 3. The advantage of multi-region decoding is robust, consistently observed across datasets, with pronounced improvements in decoding accuracy over A1-only models.



(a) (b)

Figure 3: (a) Temporal anticipation: decoding performance as a function of lead time. (b) Session-wise comparison across 6 datasets of single- vs multi-region models.

4.3 Performance Evaluation (RQ2)

Comparative Study. We compare ECoG-IBGT with baselines of 5 types (see Table 2). ECoG-IBGT achieves the highest accuracy (99.29%), F1 (99.31%), AUC-ROC (99.99%) and AUC-PR (99.99%).**(a) ECoG-IBGT vs. traditional brain siginal methods.** ECoG-IBGT achieves more than 20% absolute improvements over SVM. **(b) ECoG-IBGT vs. GNN baselines.** ECoG-IBGT achieves more than 14% absolute improvements over GNN baselines such as GIN (Xu et al., 2019), GAT (Veličković et al., 2018a), GCN (Kipf & Welling, 2017) and GraphSAGE (Hamilton et al., 2017). **(c) ECoG-IBGT vs. temporal prediction model.** We compare two temporal models, EEGNet (Lawhern et al., 2018) and MedFormer (Wang et al., 2024). ECoG-IBGT achieves a 5.63% improvement over EEGNet (paired t-test, $p < 0.001$) and nearly 39% over MedFormer. **(d) ECoG-IBGT vs. graph-based information bottleneck methods** (Zheng et al., 2024; Wu et al., 2020; Seo et al., 2023). BrainIB also employs the information bottleneck method for brain network analysis. Our method achieves significantly higher accuracy than BrainIB, demonstrating the effectiveness of our unique design tailored to ECoG signals. **(e) ECoG-IBGT vs. graph-based dimensionality reduction methods.** We compare 5 graph-based dimensionality reduction methods including TopKPool (Cangea et al., 2018), DiffPool (Ying et al., 2018), SAGPool (Lee et al., 2019), Cluster-GT (Huang et al., 2024), and GraphPCA (Yang et al., 2024). The superiority of ECoG-IBGT against these methods suggests the effectiveness of dimensionality reduction. Moreover, ECoG-IBGT reduces inference time by at least 40% over temporal models.

Generalization Analysis. To further validate generalization, we conducted additional experiments on the BCI Competition III Dataset I benchmark. Our method achieved an accuracy of 70.7%, demonstrating its strong generalization capability. More details are shown in Appendix G.

In addition, we further investigate the generalization performance of cross-dataset and leave-session-out cross validation (see Appendix H for details). Results show that, despite the distributional shifts across recording conditions (antiphonal calling vs. pure-tone/conspecific playback; free-behavior vs. chair-restrained), our method maintains robust performance, achieving up to **95.98%** cross-dataset accuracy and **88.03%** leave-session-out cross validation accuracy.

To further assess whether ECoG-IBGT captures behavior-relevant motifs beyond a single individual, we evaluate cross-subject decoding where training and test sets are strictly separated by subject. Specifically, we train on all graphs from AZ and test on BQ (AZ→BQ). As shown in Table 3, ECoG-IBGT still achieves the best overall performance among all baselines, indicating that the learned subgraphs are not merely subject-specific artifacts but capture transferable, behavior-relevant multi-region motifs.

Table 2: Classification performance comparison. **Bold**: best; Underlined: second best.

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	AUC-PR	Time (ms)
LR	76.41 ±4.63	74.04 ± 6.45	80.05 ±3.51	76.84 ±4.64	86.57 ±3.66	85.86 ±3.63	**0.002**
SVM	77.49 ±3.84	72.79 ± 4.48	86.27 ±3.41	78.92 ±3.75	88.27 ±2.35	87.96 ±2.05	0.011
GIN	84.64 ±6.47	84.38 ±11.41	90.25 ±11.01	86.02 ±5.71	93.39 ±6.09	92.81 ±8.21	0.217
GAT	81.43 ±1.12	94.05 ±0.25	69.12 ±1.87	78.18 ±1.32	94.69 ±0.31	94.90 ±0.28	0.304
GCN	79.87 ±0.89	96.30 ±0.18	63.73 ±2.05	76.68 ±1.45	96.66 ±0.15	97.22 ±0.12	0.173
GraphSAGE	85.13 ±1.11	98.46 ±4.70	71.94 ±4.59	82.89 ±1.48	96.62 ±1.85	97.02 ±2.00	0.181
GIB	71.86 ±4.57	79.01 ±15.46	70.34 ± 24.89	68.56 ±11.57	80.41 ±3.57	81.11 ±3.13	0.085
PGIB	65.58 ±10.16	76.52 ±16.62	64.26 ±41.90	54.31 ± 31.61	80.76 ±3.41	81.39 ±3.12	0.119
BrainIB	78.57 ±0.42	73.21 ±1.98	91.28 ±0.52	81.20 ±0.87	90.22 ±0.67	91.21 ±0.55	3.308
EEGNet	93.66 ±0.31	93.76 ±0.26	93.65 ±0.32	93.65 ±0.32	98.04 ±0.12	98.47 ±0.09	0.619
MedFormer	60.48 ±1.98	59.90 ±2.15	59.55 ±2.22	59.53 ±2.17	61.84 ±1.89	61.20 ±1.95	0.604
TopKPool	89.29 ±0.67	93.92 ±0.28	83.95 ±0.92	88.02 ±0.65	97.10 ±0.13	96.98 ±0.15	0.089
DiffPool	79.22 ±1.33	78.63 ±1.47	82.52 ±1.25	80.49 ±1.38	87.62 ±0.82	87.93 ±0.76	0.096
SAGPool	90.68 ±0.91	94.18 ±0.26	87.31 ±0.85	90.48 ±0.72	95.20 ±0.34	96.75 ±0.22	9.827
Cluster-GT	81.07 ±1.29	80.34 ±2.85	80.54 ±3.12	80.35 ±1.22	89.37 ±1.22	87.94 ±1.51	0.524
GraphPCA	66.27 ±1.21	68.44 ±1.89	66.85 ±1.77	65.72 ±1.94	79.29 ±1.25	82.57 ±1.12	4.614
ECoG-IBGT	**99.29** ±0.30	**99.88** ±0.02	**98.75** ±0.07	**99.31** ±0.04	**99.99** ±0.01	**99.99** ±0.01	0.362

Table 3: Cross-subject AZ→BQ performance comparison. **Bold**: best; Underlined: second best.

Model	Acc	Prec	Rec	F1	AUC-ROC	AUC-PR
GIN	63.34 ±1.48	63.23 ±0.81	63.74 ±5.51	63.38 ±2.71	68.02 ±0.74	59.91 ±0.77
GAT	71.45 ±2.24	73.58 ±0.79	66.84 ±5.58	69.96 ±3.38	81.30 ±0.31	78.68 ±0.34
GCN	69.39 ±0.79	65.89 ±1.78	80.96 ±5.03	72.51 ±1.31	76.48 ±0.61	74.41 ±2.34
GraphSAGE	70.17 ±4.13	67.20 ±5.52	**81.56** ±4.91	**73.32** ±1.80	77.10 ±0.78	71.09 ±2.87
BrainIB	52.48 ±5.32	36.64 ±24.44	60.17 ±45.02	43.77 ±31.08	62.14 ±2.65	54.56 ±3.19
EEGNet	62.03 ±0.18	**77.13** ±1.34	62.03 ±0.18	55.91 ±0.39	**87.10** ±4.52	**88.60** ±3.82
MedFormer	54.70 ±10.26	56.93 ±12.01	48.35 ±20.99	49.78 ±13.95	57.37 ±15.34	58.27 ±11.66
TopKPool	68.71 ±6.64	67.98 ±4.73	70.29 ±19.41	67.75 ±11.67	71.64 ±9.07	63.99 ±5.71
DiffPool	62.56 ±7.74	67.84 ±10.98	51.47 ±22.75	55.08 ±16.11	70.81 ±10.10	68.82 ±9.90
SAGPool	58.90 ±7.91	57.74 ±5.92	58.51 ±18.26	57.50 ±11.89	63.93 ±0.19	54.13 ±0.07
Cluster-GT	68.12 ±0.49	65.94 ±0.50	75.00 ±2.47	70.15 ±0.94	72.66 ±0.45	62.68 ±0.62
GraphPCA	47.31 ±0.95	46.15 ±1.43	47.31 ±0.95	43.14 ±1.31	51.77 ±3.42	48.58 ±1.41
ECoG-IBGT	**72.13** ±1.50	72.31 ±1.38	71.82 ±4.30	71.99 ±2.14	77.53 ±1.57	72.21 ±4.88

To better analyze intention types, we conducted a balanced three-class classification task with 189 samples per class across resting state, stimulus-induced vocalizations, and responses to conspecific vocalizations, where ECoG-IBGT achieved the best performance across all metrics, demonstrating the strong applicability and effectiveness of our method (see Appendix H for details).

Ablation Study. We evaluated seven variants to quantify each component's contribution (see Table 4): the full model (ECoG-IBGT); a version without the mutual information estimation term $I(G_{\text{sub}}, G)$; one without the connectivity loss; and three ablations removing the node mask, edge mask, or subgraph generator. Each component substantially degrades performance when ablated, validating the rationale and effectiveness of the overall architecture.

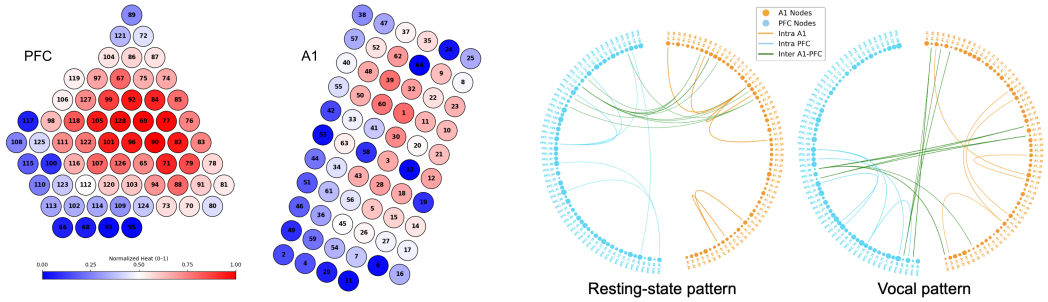
Table 4: Ablation study. **Bold**: best; Underlined: second best.

Variant	Accuracy	F1-score	AUC-ROC	Time (ms)
w/o MI estimation	93.15 ±1.43	92.99 ±1.62	98.86 ±0.27	0.356
w/o connectivity	92.69 ±1.31	92.50 ±1.50	98.76 ±0.30	0.356
w/o node masking	80.62 ±3.21	76.88 ±4.89	98.37 ±0.84	0.355
w/o edge masking	84.61 ±4.51	82.70 ±6.34	97.35 ±0.75	0.354
w/o subgraph gen.	94.87 ±1.12	94.11 ±1.36	99.15 ±0.36	0.342
w/o MI estimation + connectivity	91.77 ±3.98	91.04 ±4.30	98.07 ±0.37	0.331
w/o node masking + edge masking + connectivity	87.45 ±2.51	86.90 ±2.94	95.08 ±1.38	**0.325**
ECoG-IBGT	**99.29 ±0.30**	**99.31 ±0.04**	**99.99 ±0.01**	0.3623

Sensitivity Study. We evaluated the sensitivity of our graph construction strategy to different functional connectivity measures and edge-retention thresholds. Results indicate that Pearson correlation offers the best balance between predictive accuracy and computational efficiency, with stable and competitive performance under a 10% edge-retention threshold (see Appendix I for details).

4.4 Interpretability Analysis (RQ3)

Node Interpretability Analysis. To investigate the contribution of individual channels to vocalization prediction, we visualized the average node mask (Figure 4a). Node mask values were min–max normalized to the $[0, 1]$ range, and a color bar indicates the mapping. Nodes with mask scores above 0.5 were rendered in red hues (greater importance), whereas those below 0.5 were shown in blue (lower importance). Results show vocalization in marmosets is associated with widespread cortical engagement. Within the auditory system, the rostral, caudal, and primary auditory cortices exhibited strong involvement. In the PFC, the dorsal prefrontal region emerged as a key node, consistent with prior reports (Kato et al., 2018). Our masks effectively suppressed noisy electrodes by assigning them low weights, with details provided in the Appendix B.



(a) (b)

Figure 4: Node importance and connections motif analysis in PFC and A1 regions. (a) Spatial distribution of node-level importance scores. (b) Intra- and inter-regional connectivity patterns under resting-state versus vocal states.

Edge Interpretability Analysis. To distinguish the connectivity patterns between resting-state and vocalization, we compared the subgraph structures for each state. Specifically, we computed the mean edge importance across the 128-node functional graphs per state, retained the top 30 edges to form the dominant subgraphs, and visualized their differences (see Figure 4b). Meanwhile, Figure 4a presents normalized node-importance heatmaps for the PFC and A1 anatomical regions during each state. During vocalization, a circuit linking dorsolateral PFC (A46V) and primary somatosensory cortex (A3), with OPt as a multimodal hub (Andersen & Buneo, 2002; Colby & Goldberg, 1999), converges on auditory core and parabelt regions (AuA1, AuCPB) to form a predictive-control feedback loop. Frontal areas send corollary-discharge-like top-down signals to auditory cortex (Eliades & Wang, 2003, 2005; Tsunada et al., 2024), A3 provides somatosensory updates to motor plans (Tremblay et al., 2003), and auditory regions monitor vocal output for error correction (Tourville & Guenther, 2011). In rest, this loop is weakened and connectivity is instead dominated by a baseline fronto-auditory pathway between ventral PFC and rostral auditory fields, consistent with top-down modulation, predictive coding, and prior fronto-temporal findings (Friston, 2005; Asilador & Llano, 2021; Hockley et al., 2025a). High-density, multi-region ECoG revealed fine-scale vocal-production subcircuits, whose importance was further validated by edge perturbation experiments showing stronger accuracy drops after removing top-ranked edges than random edges. (see Appendix F).

5 Conclusion

In this work, we established a multi-dimensional experimental framework and collected high-quality dual-region ECoG recordings from the primary auditory cortex (A1) and prefrontal cortex (PFC) in common marmosets across multiple vocal paradigms. Based on these data, we developed an interpretable information-bottleneck graph transformer (ECoG-IBGT) for early multi-region neural

decoding. Our model achieved 99.29% accuracy 400 ms before vocal onset, significantly outperforming 16 state-of-the-art methods. Beyond predictive performance, ECoG-IBGT robustly identifies task-relevant cortical subregions and reveals distinct intra- and inter-regional connectivity patterns, strengthening the neurophysiological foundation of anticipatory prediction. Nonetheless, current findings are based on a limited number of subjects, which constrains the generalizability of the model across individuals and conditions.

6 Acknowledgments and Disclosures

This work was supported by the National Key Research and Development Program of China (Grant No.2023YFB4705502) and the National Natural Science Foundation of China (32541016, 62436005, U25A20540, 62473303).

Competing Interests. The authors declare no competing interests.

References

Andersen, R. A. and Buneo, C. A. Intentional maps in posterior parietal cortex. *Annual Review of Neuroscience*, 25:189–220, 2002. doi: 10.1146/annurev.neuro.25.112701.142922.

Ariani, G., Pruszynski, J. A., and Diedrichsen, J. Motor planning brings human primary somatosensory cortex into action-specific preparatory states. *eLife*, 11:e69517, 2022. doi: 10.7554/eLife.69517. Published 2022 Jan 12. PMID: 35018886; PMCID: PMC8786310.

Asilador, A. and Llano, D. A. Top-down inference in the auditory system: Potential roles for corticofugal projections. *Frontiers in Neural Circuits*, 14:615259, 2021. doi: 10.3389/fncir.2020.615259.

Bulea, T. C., Prasad, S., Kilicarslan, A., and Contreras-Vidal, J. L. Sitting and standing intention can be decoded from scalp eeg recorded prior to movement execution. *Frontiers in Neuroscience*, 8:376, 2014. doi: 10.3389/fnins.2014.00376.

Cangea, C., Veličković, P., Jovanović, N., Kipf, T., and Liò, P. Towards sparse hierarchical graph classifiers, 2018.

Chen, C., Shin, D., Watanabe, H., Nakanishi, Y., Kambara, H., Yoshimura, N., Nambu, A., Isa, T., Nishimura, Y., and Koike, Y. Decoding grasp force profile from electrocorticography signals in non-human primate sensorimotor cortex. *Neuroscience Research*, 83:1–7, 2014. doi: 10.1016/j.neures.2014.03.010.

Colby, C. L. and Goldberg, M. E. Space and attention in parietal cortex. *Annual Review of Neuroscience*, 22:319–349, 1999. doi: 10.1146/annurev.neuro.22.1.319.

Das, A., Sexton, D., Lainscsek, C., Cash, S. S., and Sejnowski, T. J. Characterizing brain connectivity from human electrocorticography recordings with unobserved inputs during epileptic seizures. *Neural Comput.*, 31(7):1271–1326, 2019. doi: 10.1162/neco_a_01205. Epub 2019 May 21.

Eliades, S. J. and Wang, X. Sensory-motor interaction in the primate auditory cortex during self-initiated vocalizations. *Journal of Neurophysiology*, 89(4):2194–2207, 2003. doi: 10.1152/jn.00627.2002.

Eliades, S. J. and Wang, X. Dynamics of auditory–vocal interaction in monkey auditory cortex. *Cerebral Cortex*, 15(10):1510–1523, 2005. doi: 10.1093/cercor/bhi030.

Friston, K. J. A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):815–836, 2005. doi: 10.1098/rstb.2005.1622.

Hamilton, W. L., Ying, R., and Leskovec, J. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, 2017. doi: 10.48550/arXiv.1706.02216.

Hockley, A., Bohórquez, L. H., and Malmierca, M. S. Top-down prediction signals from the medial prefrontal cortex govern auditory cortex prediction errors. *Cell Reports*, 44(4):115538, 2025a. doi: 10.1016/j.celrep.2025.115538.

Hockley, A., Bohórquez, L. H., and Malmierca, M. S. Top-down prediction signals from the medial prefrontal cortex govern auditory cortex prediction errors. *Cell Reports*, 44(4):115538, 2025b. ISSN 2211-1247. doi: https://doi.org/10.1016/j.celrep.2025.115538. URL `https://www.sciencedirect.com/science/article/pii/S2211124725003092`.

Huang, S., Song, Y., Zhou, J., and Lin, Z. Cluster-wise graph transformer with dual-granularity kernelized attention, 2024.

Jovanovic, V., Fishbein, A. R., de la Mothe, L., Lee, K. F., and Miller, C. T. Behavioral context affects social signal representations within single primate prefrontal cortex neurons. *Neuron*, 110 (8):1318–1326.e4, 2022. doi: 10.1016/j.neuron.2022.01.020.

Kato, M., Yokoyama, C., Kawasaki, A., Takeda, C., Koike, T., Onoe, H., and Iriki, A. Individual identity and affective valence in marmoset calls: in vivo brain imaging with vocal sound playback. *Animal Cognition*, 21(3):331–343, 2018. doi: 10.1007/s10071-018-1169-z.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Lal, T., Hinterberger, T., Widman, G., Schröder, M., Hill, J., Rosenstiel, W., Elger, C., Schölkopf, B., and Birbaumer, N. Methods towards invasive human brain computer interfaces. In *Advances in Neural Information Processing Systems*, 2004.

LaRocco, J. and Paeng, D.-G. Optimizing computer–brain interface parameters for non-invasive brain-to-brain interface. *Frontiers in Neuroinformatics*, 14:1, Feb 2020. doi: 10.3389/fninf.2020.00001. Epub 2020 Feb 7.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. Eegnet: a compact convolutional neural network for eeg-based brain-computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. doi: 10.1088/1741-2552/aace8c.

Lee, J., Lee, I., and Kang, J. Self-attention graph pooling. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3734–3743. PMLR, 09–15 Jun 2019.

Liu, D., Chen, W., Lee, K., Chavarriaga, R., Iwane, F., Bouri, M., Pei, Z., and Del R Millán, J. Eeg-based lower-limb movement onset decoding: Continuous classification and asynchronous detection. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(8):1626–1635, 2018. doi: 10.1109/TNSRE.2018.2855053.

Liu, G., Yan, Y., Cai, J., Wu, E. Q., Fang, S., Cheok, A. D., and Song, A. Gcd: Graph contrastive denoising module for gnns in eeg classification. *Expert Systems with Applications*, 265:126013, 2025.

Mitelut, C., Zhang, Y., Sekino, Y., Boyd, J. D., Bollanos, F., Swindale, N. V., Silasi, G., Saxena, S., and Murphy, T. H. Mesoscale cortex-wide neural dynamics predict self-initiated actions in mice several seconds prior to movement. *eLife*, 11:e76506, 2022. doi: 10.7554/eLife.76506.

Nejedly, P., Hrtonova, V., Pail, M., Cimbalnik, J., Daniel, P., Travnicek, V., Dolezalova, I., Mivalt, F., Kremen, V., Jurak, P., et al. Leveraging interictal multimodal features and graph neural networks for automated planning of epilepsy surgery. *Brain Communications*, 7(3):fcaf140, 2025.

Rouzitalab, A., Boulay, C. B., Park, J., and Sachs, A. J. Intracortical brain–computer interfaces in primates: A review and outlook. *Biomedical Engineering Letters*, 13(3):375–390, May 2023. doi: 10.1007/s13534-023-00286-8.

Roy, S. and Wang, X. Distinct neural activities in premotor cortex during natural vocal behaviors in a new world primate, the common marmoset. *Journal of Neuroscience*, 36(48):12168–12179, 2016. doi: 10.1523/JNEUROSCI.1646-16.2016.

Seo, S., Kim, S., and Park, C. Interpretable prototype-based graph information bottleneck. *Advances in Neural Information Processing Systems*, 36:76737–76748, 2023.

Shafiul Hasan, S. M., Siddiquee, M. R., Atri, R., Ramon, R., Marquez, J. S., and Bai, O. Prediction of gait intention from pre-movement eeg signals: a feasibility study. *Journal of NeuroEngineering and Rehabilitation*, 17(1):50, 2020. doi: 10.1186/s12984-020-00675-5.

Shin, D., Watanabe, H., Kambara, H., Nambu, A., Isa, T., Nishimura, Y., and et al. Prediction of muscle activities from electrocorticograms in primary motor cortex of primates. *PLOS ONE*, 7 (10):e47992, 2012. doi: 10.1371/journal.pone.0047992.

Siegel, M., Buschman, T. J., and Miller, E. K. Cortical information flow during flexible sensorimotor decisions. *Science*, 348(6241):1352–1355, 2015. doi: 10.1126/science.aab0551.

Skomrock, N. D., Schwemmer, M. A., Ting, J. E., Trivedi, H. R., Sharma, G., Bockbrader, M. A., and Friedenberg, D. A. A characterization of brain–computer interface performance trade-offs using support vector machines and deep neural networks to decode movement intent. *Frontiers in Neuroscience*, 12:763, 2018. ISSN 1662-453X. doi: 10.3389/fnins.2018.00763. URL `https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2018.00763`.

Sun, H., Blakely, T. M., Darvas, F., Wander, J. D., Johnson, L. A., Su, D. K., Miller, K. J., Fetz, E. E., and Ojemann, J. G. Sequential activation of premotor, primary somatosensory and primary motor areas in humans during cued finger movements. *Clinical Neurophysiology*, 126(11):2150–2161, November 2015. doi: 10.1016/j.clinph.2015.01.005. Epub 2015 Jan 23. PMID: 25680948; PMCID: PMC4512926.

Tauste Campo, A., Martinez-Garcia, M., Nácher, V., Luna, R., Romo, R., and Deco, G. Task-driven intra- and interarea communications in primate cerebral cortex. *Proceedings of the National Academy of Sciences*, 112(15):4761–4766, March 2015. ISSN 1091-6490. doi: 10.1073/pnas.1503937112. URL `http://dx.doi.org/10.1073/pnas.1503937112`.

Tourville, J. A. and Guenther, F. H. The diva model: A neural theory of speech acquisition and production. *Language and Cognitive Processes*, 26(7):952–981, 2011. doi: 10.1080/01690960903498424.

Tremblay, S., Shiller, D. M., and Ostry, D. J. Somatosensory basis of speech production. *Nature*, 423 (6942):866–869, 2003. doi: 10.1038/nature01710.

Tsunada, J. and Eliades, S. J. Frontal–auditory cortical interactions and sensory prediction during vocal production in marmoset monkeys. *Current Biology*, 35(10):2307–2322.e3, May 2025. doi: 10.1016/j.cub.2025.03.077. Updated and peer-reviewed version of the original bioRxiv preprint.

Tsunada, J., Wang, X., and Eliades, S. J. Multiple processes of vocal sensory-motor interaction in primate auditory cortex. *Nature Communications*, 15:3093, 2024. doi: 10.1038/s41467-024-47510-2.

Umeda, T., Koizumi, M., Katakai, Y., Saito, R., and Seki, K. Decoding of muscle activity from the sensorimotor cortex in freely behaving monkeys. *NeuroImage*, 197:512–526, 2019. doi: 10.1016/j.neuroimage.2019.04.045.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph attention networks. In *International Conference on Learning Representations*, 2018a.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph attention networks. In *International Conference on Learning Representations*, 2018b. URL `https://openreview.net/forum?id=rJXMpikCZ`.

Volkova, K., Lebedev, M. A., Kaplan, A., and Ossadtchi, A. Decoding movement from electrocorticographic activity: A review. *Frontiers in Neuroinformatics*, 13:74, 2019. doi: 10.3389/fninf.2019.00074.

Walker, J. D., Pirschel, F., Sundiang, M., Niekrasz, M., MacLean, J. N., and Hatsopoulos, N. G. Chronic wireless neural population recordings with common marmosets. *Cell Rep.*, 36(2):109379, 2021. doi: 10.1016/j.celrep.2021.109379. Epub 2021 Jul 13.

Wang, P. T., King, C. E., McCrimmon, C. M., Lin, J. J., Sazgar, M., Hsu, F. P., Shaw, S. J., Millet, D. E., Chui, L. A., Liu, C. Y., Do, A. H., and Nenadic, Z. Comparison of decoding resolution of standard and high-density electrocorticogram electrodes. *J Neural Eng.*, 13(2):026016, 2016. doi: 10.1088/1741-2560/13/2/026016. Epub 2016 Feb 9.

Wang, Y., Yang, X., Zhang, X., et al. Implantable intracortical microelectrodes: reviewing the present with a focus on the future. *Microsyst Nanoeng*, 9:7, 2023. doi: 10.1038/s41378-022-00451-6. Received 07 May 2022; Revised 08 August 2022; Accepted 22 August 2022; Published 05 January 2023.

Wang, Y., Huang, N., Li, T., Yan, Y., and Zhang, X. Medformer: A multi-granularity patching transformer for medical time-series classification. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Wu, J., Shuman, B. R., Brunton, B. W., Steele, K. M., Olson, J. D., Rao, R. P., and Ojemann, J. G. Multistep model for predicting upper-limb 3d isometric force application from pre-movement electrocorticographic features. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1564–1567, August 2016. doi: 10.1109/EMBC.2016.7591010.

Wu, T., Ren, H., Li, P., and Leskovec, J. Graph information bottleneck. *Advances in Neural Information Processing Systems*, 33:20437–20448, 2020.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Yang, J., Wang, L., Liu, L., and Zheng, X. Graphpca: a fast and interpretable dimension reduction algorithm for spatial transcriptomics data. *Genome Biology*, 25:287, 2024. doi: 10.1186/s13059-024-03429-x.

Ying, R., You, J., Morris, C., Ren, X., Hamilton, W., and Leskovec, J. Hierarchical graph representation learning with differentiable pooling. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Ying, Z., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. Gnnexplainer: Generating explanations for graph neural networks. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E. B., and Garnett, R. (eds.), *NeurIPS*, pp. 9240–9251, 2019. URL `http://dblp.uni-trier.de/db/conf/nips/nips2019.html#YingBYZL19`.

Zanini, A., Dureux, A., Selvanayagam, J., and Everling, S. Ultra-high field fmri identifies an action-observation network in the common marmoset. *Communications Biology*, 6:553, 2023. doi: 10.1038/s42003-023-04942-8.

Zheng, K., Yu, S., Li, B., Jenssen, R., and Chen, B. Brainib: Interpretable brain network-based psychiatric diagnosis with graph information bottleneck. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. doi: 10.1109/TNNLS.2024.3449419.

A Use of LLMs

In this work, our usage of large language models (LLMs) was strictly limited to language polishing and editorial assistance. No LLM was used in conceiving the research, designing the methodology, writing technical derivations, generating experiment code, or interpreting results.

Specifically:

- We used an LLM (e.g. GPT-4) only to refine grammar, wording, phrasing, or clarity in certain prose sections.
- All substantive content—formulations, theorems, proofs, algorithmic design, experiments, data analysis—were authored and verified by the human authors.
- Any text generated by the LLM was carefully reviewed, edited, and fact-checked by the authors to ensure correctness and consistency with the intended meaning.

- Since the LLM usage played no role in the core scientific contributions, we do not treat it as a "contributor" under the NeurIPS LLM usage policy.

This approach ensures that all intellectual responsibility remains with the human authors, consistent with NeurIPS's policy requiring disclosure and accountability.

B Noise-Channel Detection in Marmoset BQ

We apply four complementary diagnostics to verify that ECoG-IBGT's subgraph masks isolate noisy electrodes reliably. No electrodes were manually excluded; noisy channels were retained and down-weighted by the learned node masks

(a) Time-domain Variance (Figure 5a) We compute each of the 128 electrodes' signal variance across the recording session. Light-gray bars represent all electrodes, and electrodes exceeding a variance Z-score of 1.5 (indicating abnormally high variance) are highlighted in red, marking them as potential noise channels. These elevated variance values typically correspond to artifacts such as electrical interference or poor electrode contacts.

(b) Mean Inter-Channel Correlation (Figure 5b) Next, we calculate the mean Pearson correlation coefficient of each electrode's signals with all other channels. Electrodes exhibiting mean correlations below 0.5 (marked in red) suggest significantly reduced signal consistency and synchrony with neighboring electrodes. Importantly, many of these low-correlation electrodes overlap with those identified by variance, reinforcing the identification of these channels as genuine noise sources.

(c) Power Spectral Density (PSD) Comparison (Figure 5c) To further validate channel noise, we examine the PSD of signals from flagged channels (red) and compare them with known clean electrodes (blue). Although both channel types show characteristic power-line interference at harmonics of 50 Hz (50, 100, and 150 Hz), noise channels exhibit markedly higher amplitude peaks and a pronounced elevation across a broad frequency spectrum. This broadband increase in noise floor indicates genuine noise contamination rather than typical physiological variations.

(d) Spatial Z-score Mapping (Figure 5d) Mapping the variance Z-scores spatially onto the 8×16 electrode grid allows visualization of the noise distribution. Notably, high-noise electrodes (dark red) appear spatially clustered rather than randomly scattered, suggesting a localized hardware or electrode contact problem. This pattern is critical for interpreting noise sources, as spatial clustering typically implicates technical rather than biological factors.

Centering the thresholding criterion, we define:

$$\text{noise_idx} = \left\{ i \mid |z_{\text{var},i}| > 1.5 \ \vee \ \text{mean_corr}_i < 0.5 \right\}, \tag{12}$$

where $z_{\text{var},i}$ is the variance Z-score and mean_corr_i is the mean Pearson correlation of channel i.

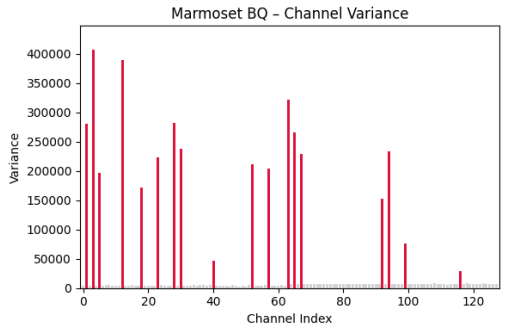
Figure 6 repeats these diagnostics for Marmoset AZ with identical thresholds and layouts. Direct comparison between Figures 5 and 6 confirms the reliability of our approach. Both subjects exhibit consistent identification of severely affected electrodes, evident in similarly exaggerated PSD profiles.

To summarize clearly and succinctly, Table 5 lists all identified noisy electrode indices for both subjects. This structured detection and validation process ensures the accuracy of downstream neural analyses by eliminating channels compromised by extraneous noise.

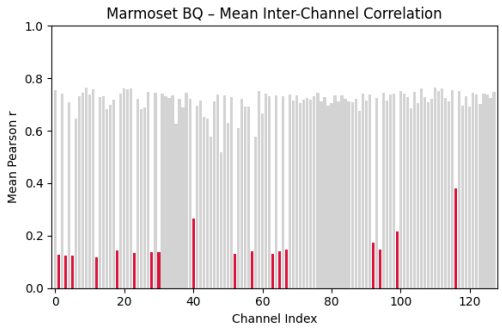
Table 5: Detected noisy channels for Marmoset BQ and AZ (1-based indices).

BQ	AZ
2, 4, 6, 13, 19, 24, 29, 31, 41, 53, 58, 64, 66, 68, 93, 95, 100, 117	2, 4, 7, 29, 31, 41, 49, 66, 68, 74, 85, 87, 93, 95

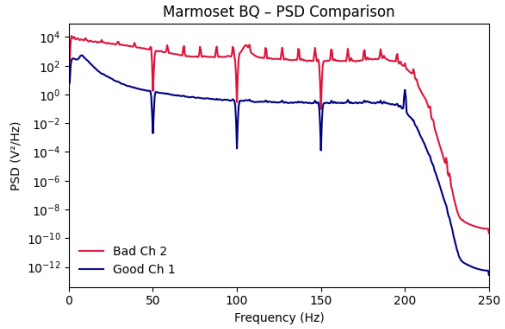
These comprehensive diagnostics (variance, correlation, PSD, and spatial analyses) collectively demonstrate our pipeline's efficacy in identifying, validating, and localizing noisy electrodes, thus ensuring that the subsequent analyses are focused exclusively on physiologically meaningful ECoG signals.



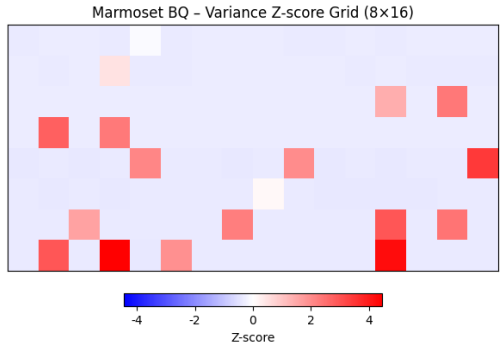
(a) Variance



(b) Correlation

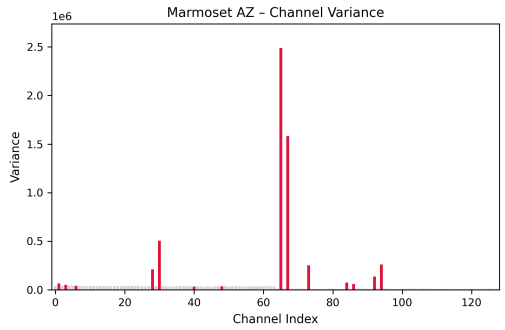


(c) Power spectral density

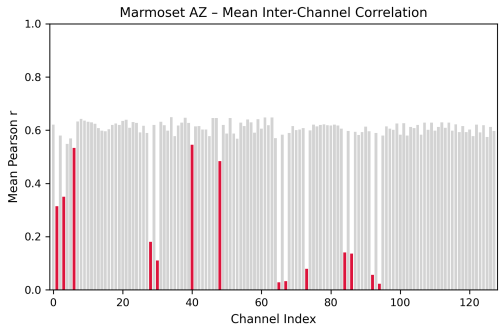


(d) Spatial Z-score grid

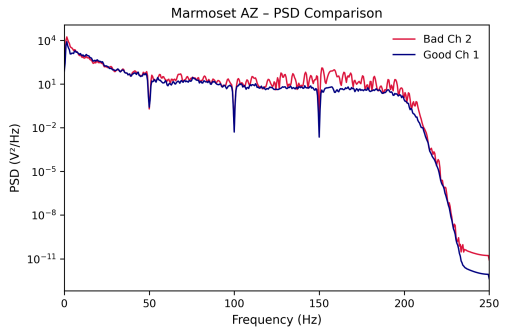
Figure 5: Noise-channel detection in Marmoset BQ: (a) variance, (b) mean inter-channel correlation, (c) power spectral density comparison, (d) spatial Z-score grid.



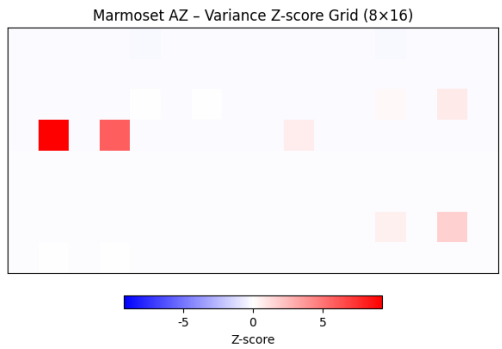
(a) Variance



(b) Correlation



(c) Power spectral density

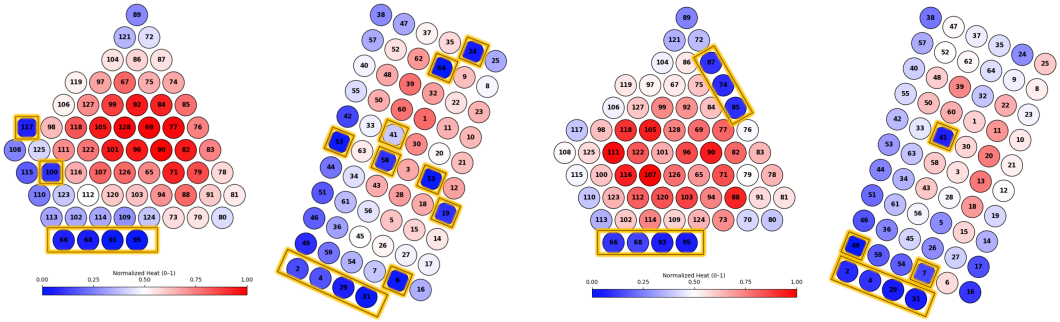


(d) Spatial Z-score grid

Figure 6: Noise-channel detection in Marmoset AZ: same diagnostics and thresholds as in Figure 5.

C Visualization of Noise-Channel Suppression

This appendix provides a detailed visualization illustrating how ECoG-IBGT effectively suppresses activations in channels previously identified as noisy. Figures 7a and 7b present normalized latent activation heatmaps for subjects Marmoset BQ (a) and AZ (b), respectively. Activation values are scaled between 0 and 1 to facilitate consistent comparison across all channels.



(a) Marmoset BQ heatmap

(b) Marmoset AZ heatmap

Figure 7: Latent activation heatmaps showing suppression of noise channels. Activations normalized to [0,1]; warmer colors indicate stronger activation. Previously flagged noisy channels are outlined in yellow.

Identification of Noise Channels Channels previously marked as noisy (refer to Table 5 in Appendix B) are highlighted with yellow outlines in each heatmap. This clear demarcation enables direct visual assessment of the model's suppression effectiveness relative to previously established noise locations.

Effectiveness of Suppression. Across both animals, channels identified as noisy consistently display minimal activation levels (represented by cooler colors, primarily blue). In contrast, adjacent physiological channels exhibit broader and higher activation ranges (green to red), indicating intact and meaningful signal representation. This clear distinction quantitatively demonstrates that the masking mechanism of ECoG-IBGT effectively attenuates noise-contaminated channels without adversely impacting the integrity of physiologically relevant signals.

Robustness Across Datasets. We averaged latent activations over 50 randomly selected samples per dataset. The resulting averaged heatmaps reveal a consistent pattern of suppression across multiple datasets, reinforcing the robustness and generalizability of our approach.

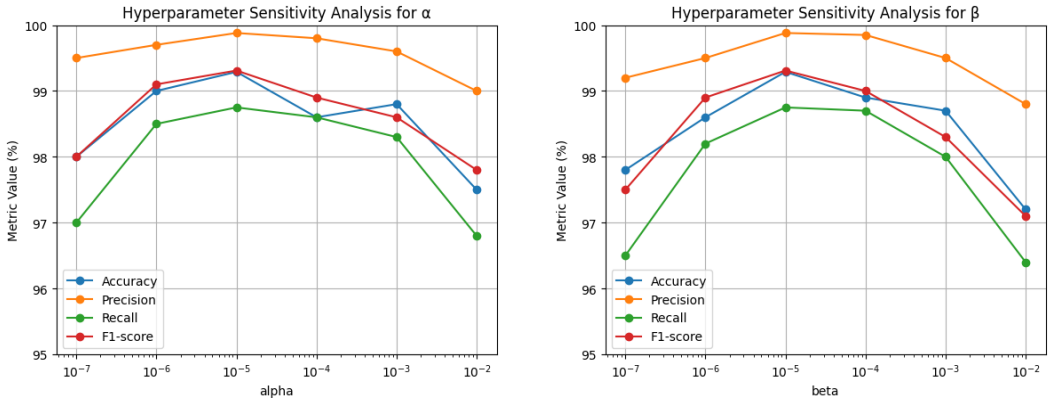
Implications for Analytical Interpretability By substantially reducing the influence of noise channels, our method significantly enhances the reliability and interpretability of downstream graph-level analyses. Consequently, this targeted suppression not only optimizes predictive accuracy but also ensures that physiological interpretations derived from subgraph analyses accurately reflect true neural dynamics, untainted by artifacts.

D Hyperparameter Sensitivity Analysis

We evaluate sensitivity to key hyperparameters. Figures 8, 9, and 10 show their effects on core metrics.

Pooling Readout. We performed comprehensive sensitivity analyses to evaluate the robustness of ECoG-IBGT concerning critical hyperparameters. Figures 8, 9, and 10 illustrate the effect of varying hyperparameters on evaluation metrics including Accuracy, Precision, Recall, F1-score, AUC-ROC, and AUC-PR.

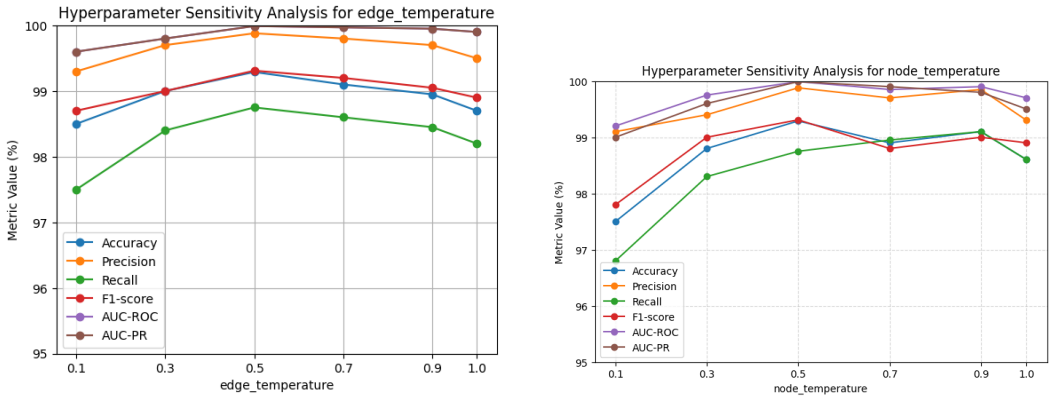
Information Bottleneck and Connectivity Loss Weights. As shown in Figure 8, both the mutual information (MI) term weight (α) and connectivity loss weight (β) have optimal settings at 10^{-5}. Deviating from this optimum negatively impacts all metrics, indicating the importance of balanced information capture and subgraph connectivity.



(a) MI term weight α. (b) Connectivity loss weight β.

Figure 8: Sensitivity of information bottleneck and connectivity loss weights.

Gumbel–Sigmoid Mask Temperatures. Figure 9 indicates optimal edge-mask and node-mask temperatures are both at 0.5. Lower or higher values result in gradual performance deterioration, demonstrating that proper mask temperature is essential for maintaining predictive subgraph quality and mask sparsity.



(a) Edge-mask temperature. (b) Node-mask temperature.

Figure 9: Sensitivity of Gumbel–Sigmoid temperatures for masks.

Pooling Readout Functions. As illustrated in Figure 10, max pooling consistently outperforms mean and sum pooling across all metrics. This suggests that emphasizing the strongest edge contributions yields the most discriminative and effective subgraph representations.

E Notation and Algorithmic Details

We use the following notation: N is the number of nodes; E the number of edges; F the per-node feature dimension; L TransformerConv layers; M MLP classifier layers (including output); H attention heads per layer; D the final node embedding dimension (`latent_dim[-1]` $\times H$); B batch size; C output classes; and τ_e, τ_n the Gumbel–sigmoid temperatures for edge- and node-mask sampling.

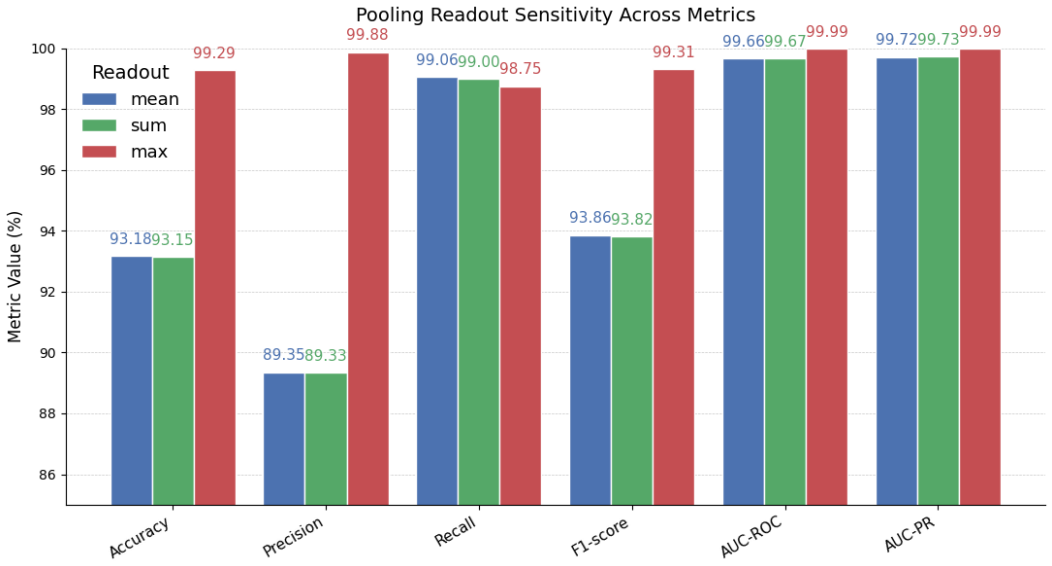


Figure 10: Comparison of pooling readout functions. Max pooling yields best performance across metrics.

Below we give pseudocode and concise explanations for the two core components: the IB–Subgraph Generator (Algorithm 1) and the Dual Structure Graph Encoder (Algorithm 2).

IB–Subgraph Generator (Algorithm 1). This module generates differentiable masks to highlight important nodes and edges for subgraph extraction. It operates by first concatenating node feature pairs for each edge to form edge-specific features, which are then passed through an edge-level multilayer perceptron (EdgeMLP). Node features independently go through a node-level MLP (NodeMLP). During training, Gumbel–sigmoid sampling ensures differentiability by introducing stochastic noise controlled by temperatures τ_e and τ_n, enabling gradient-based optimization of discrete masks. In inference, deterministic sigmoid activation is applied for mask generation.

Dual Structure Graph Encoder (Algorithm 2). This encoder leverages Transformer-based convolutional layers (TransformerConv) to learn node embeddings. The learned embeddings undergo normalization (BatchNorm) and activation (ReLU) before applying masks generated by the IB–Subgraph Generator. Graph-level embeddings are computed via global max pooling. An MLP classifier processes these embeddings, consisting of multiple linear transformations interleaved with batch normalization, ELU activations, and dropout regularization, culminating in the final prediction logits.

F Perturbation Experiments for Interpretability

To further evaluate the interpretability of our model, we conducted perturbation experiments by systematically altering the graph structure and observing the corresponding changes in classification accuracy. Specifically, we examined four scenarios:

- **base**: The original unperturbed graph.
- **rmTop**: Removing the most important edges ranked by the learned edge mask.
- **rmRnd**: Randomly removing the same number of edges as in **rmTop**.
- **addRnd**: Randomly adding new edges between unconnected node pairs.

As shown in Figure 11, the model demonstrates a clear robustness pattern. When important edges (**rmTop**) are removed, classification accuracy degrades rapidly, confirming that the identified edges are indeed critical for predictive performance. In contrast, randomly removing edges (**rmRnd**) has a

Algorithm 1 IB–Subgraph Generator

Require: Node features $X \in \mathbb{R}^{N \times F}$, edge indices `edge_index`$\in \mathbb{Z}^{2 \times E}$, edge temperature τ_e, node temperature τ_n, training flag
Ensure: Edge mask $m_e \in [0,1]^E$, node mask $m_n \in [0,1]^N$

Transfer inputs to device
for each edge $k = 1, \ldots, E$ **do**
    $(i, j) \leftarrow$ `edge_index`$[:, k]$
    expl_in$[k] \leftarrow [X[i]; X[j]]$
end for
$w_e \leftarrow \mathrm{EdgeMLP}(\textit{expl_in})$
$w_n \leftarrow \mathrm{NodeMLP}(X)$
function SAMPLE(w, τ)
    if training **then**
        Sample Gumbel noise $\varepsilon \sim \mathrm{Uniform}(10^{-5}, 1 - 10^{-5})$
        $g \leftarrow \ln(\varepsilon) - \ln(1 - \varepsilon)$
        $\ell \leftarrow (w + g)/\tau$
        return $\sigma(\ell)$
    else
        return $\sigma(w)$
    end if
end function
$m_e \leftarrow \mathrm{Sample}(w_e, \tau_e)$, $m_n \leftarrow \mathrm{Sample}(w_n, \tau_n)$
return m_e, m_n

Algorithm 2 Dual Structure Graph Encoder

Require: Node features X, edge index E, batch indices B, edge weights W_e (optional), node mask m_n (optional)
Ensure: Logits, graph embedding z_g, node embeddings Z_n

Initialize missing W_e, m_n to ones
for layer $i = 1$ to L **do**
    $X \leftarrow \mathrm{TransformerConv}_i(X, E, W_e)$
    $X \leftarrow \mathrm{BatchNorm}_i(X)$
    $X \leftarrow \mathrm{ReLU}(X)$
end for
$Z_n \leftarrow m_n \odot X$
$z_g \leftarrow \mathrm{global_max_pool}(Z_n, B)$
$h \leftarrow z_g$
for MLP layer $i = 1$ to $M - 1$ **do**
    $h \leftarrow \mathrm{Linear}_i(h)$
    $h \leftarrow \mathrm{BatchNorm}_i(h)$
    $h \leftarrow \mathrm{ELU}(h)$
    $h \leftarrow \mathrm{Dropout}(h)$
end for
$\mathrm{logits} \leftarrow \mathrm{Linear}_M(h)$
return $\mathrm{logits}, z_g, Z_n$

more gradual effect, while random edge additions (**addRnd**) lead to moderate performance declines, especially at larger perturbation scales. Notably, the **base** model maintains consistently high accuracy, serving as the upper bound.

These results indicate that the subgraph masks produced by our model successfully capture structurally meaningful connections: perturbing edges deemed important by the model has a much stronger impact than perturbing random edges. This validates the interpretability of the learned subgraphs and further highlights the model's ability to reveal functionally relevant neural circuit motifs.

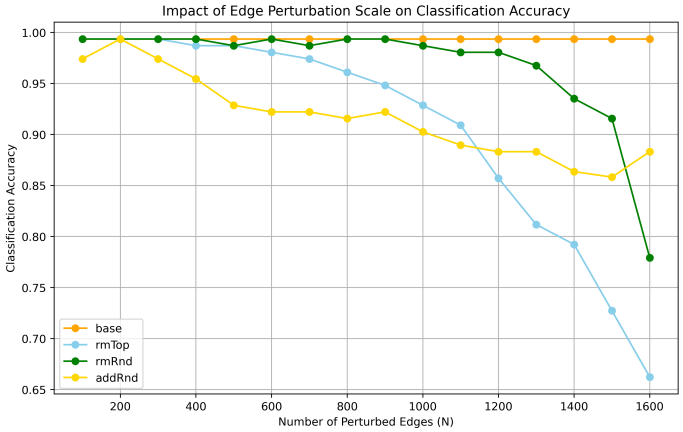


Figure 11: Impact of edge perturbation scale on classification accuracy. **base**: no perturbation; **rmTop**: removing important edges identified by the model; **rmRnd**: randomly removing the same number of edges; **addRnd**: randomly adding edges between previously unconnected nodes.

G Additional Benchmark: BCI Competition III Dataset I

Protocol. To further evaluate generalization beyond our in-house marmoset dataset, we test ECoG-IBGT on the public **BCI Competition III, Dataset I** (Lal et al., 2004). This dataset records imagined finger versus tongue movements from invasive ECoG grids (278 training samples, 100 test samples, 64 channels, 1000 Hz sampling). Unlike our main experiments, we adopt a *minimalist* configuration: (1) functional connectivity computed by Pearson correlation, (2) binary adjacency via top-50% thresholding, (3) no additional "strong" features (e.g., high-gamma power, LMP, PSD), (4) no hyperparameter tuning (default: batch size 128, epochs 300, learning rate 5×10^{-5}).

Table 6: ECoG-IBGT on BCI Competition III Dataset I.

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	AUC-PR
GIN	52.33 ± 0.47	61.06 ± 9.98	27.33 ± 17.15	32.81 ± 16.19	59.55 ± 4.74	62.27 ± 4.68
GAT	67.00 ± 4.55	76.44 ± 3.54	50.00 ± 14.14	58.96 ± 11.01	72.27 ± 2.33	72.77 ± 2.06
GCN	65.00 ± 2.16	76.92 ± 2.20	43.33 ± 8.06	54.83 ± 6.48	73.53 ± 2.97	72.28 ± 1.80
GraphSAGE	65.00 ± 9.93	68.47 ± 9.33	50.00 ± 32.12	52.01 ± 27.17	75.87 ± 7.90	74.59 ± 9.77
BrainIB	66.33 ± 3.86	78.13 ± 5.66	47.33 ± 15.69	56.92 ± 10.60	74.65 ± 3.93	75.47 ± 1.90
EEGNet	50.00 ± 1.63	48.67 ± 3.60	32.00 ± 13.37	37.34 ± 11.76	55.57 ± 2.37	54.95 ± 2.26
MedFormer	56.00 ± 2.83	56.68 ± 3.41	53.33 ± 1.89	54.83 ± 0.69	58.33 ± 0.36	57.15 ± 1.17
TopKPool	54.67 ± 3.86	77.12 ± 16.20	19.33 ± 17.91	25.55 ± 20.12	67.15 ± 0.59	65.61 ± 1.65
DiffPool	56.67 ± 6.24	56.41 ± 41.82	21.33 ± 23.46	26.29 ± 25.42	61.36 ± 4.53	60.87 ± 4.97
SAGPool	57.00 ± 2.45	62.45 ± 5.56	46.00 ± 28.89	46.42 ± 18.17	61.71 ± 3.42	62.53 ± 2.43
Cluster-GT	53.67 ± 2.49	70.20 ± 8.17	12.67 ± 8.06	20.43 ± 10.99	66.83 ± 1.04	66.65 ± 1.75
GraphPCA	59.33 ± 5.44	$\mathbf{83.07 \pm 15.02}$	34.67 ± 32.06	37.76 ± 23.36	74.73 ± 5.03	75.50 ± 6.22
GIB	56.67 ± 3.30	70.10 ± 7.11	24.67 ± 13.70	34.21 ± 13.38	63.64 ± 2.61	66.49 ± 2.56
PGIB	55.33 ± 6.18	63.41 ± 10.29	45.33 ± 28.11	43.98 ± 23.31	61.73 ± 8.33	64.08 ± 7.81
ECoG-IBGT	$\mathbf{70.67 \pm 7.41}$	72.67 ± 10.91	$\mathbf{70.67 \pm 2.49}$	$\mathbf{71.08 \pm 4.67}$	$\mathbf{77.84 \pm 7.44}$	$\mathbf{77.42 \pm 11.66}$

Observations. Results are shown in Table 6. Even under this simplified setup, our model achieves **70.7% test accuracy** with **AUC-ROC 77.8%**. This demonstrates that ECoG-IBGT can extract discriminative subgraphs without task-specific feature engineering or dataset-specific tuning. The results confirm the *positive transferability* of our framework across benchmarks.

H Additional Generalization Evaluation

To further assess generalization across distinct experimental dataset, we conducted **cross-dataset decoding experiments** using different datasets as training and test sets. Specifically:

- **Dataset-3** $\rightarrow$ **Dataset-4**: model trained on Dataset-3 (antiphonal calling, 1012 samples) and tested on Dataset-4 (mixed playback sessions, 1012 samples);
- **Dataset-4** $\rightarrow$ **Dataset-3**: reciprocal configuration with Dataset-4 as training and Dataset-3 as testing;
- **Dataset-5** $\rightarrow$ **Dataset-6**: model trained on Dataset-5 (free-behavior session, 222 samples) and tested on Dataset-6 (chair-restrained session, 222 samples);
- **Dataset-6** $\rightarrow$ **Dataset-5**: reciprocal configuration with Dataset-6 as training and Dataset-5 as testing.

Table 7 summarizes the cross-dataset validation and testing accuracy. We note that, despite the distributional shifts across recording conditions (antiphonal calling vs. pure-tone/conspecific playback; free-behavior vs. chair-restrained), our method maintains robust performance, achieving up to **95.98%** cross-dataset accuracy.

Table 7: Generalization results of cross-dataset.

Training $\rightarrow$ Testing	Validation Acc.	Cross-Dataset Acc.
Dataset-3 $\rightarrow$ Dataset-4	98.91%	80.08%
Dataset-4 $\rightarrow$ Dataset-3	98.37%	61.83%
Dataset-5 $\rightarrow$ Dataset-6	100.00%	95.98%
Dataset-6 $\rightarrow$ Dataset-5	100.00%	81.10%

Furthermore, we further use leave-session-out cross validation to validate generalization. Table 8 summarizes the performance of leave-session-out cross validation. These results suggest that the model is minimally affected by variations in experimental conditions, highlighting its strong robustness and generalization capability.

Table 8: Generalization results of leave-session-out cross validation. Test accuracy, precision, recall, F1 score, AUC-ROC, and AUC-PR are reported (% ± SD).

Held-out Session	Test Acc (%)	Precision (%)	Recall (%)	F1 (%)	AUC-ROC (%)	AUC-PR (%)
B1	81.93 ± 10.70	77.28 ± 12.21	96.49 ± 4.96	85.00 ± 7.09	95.69 ± 1.50	93.77 ± 3.10
B2	82.05 ± 7.49	98.05 ± 1.60	65.64 ± 16.00	77.29 ± 12.13	95.87 ± 1.14	96.67 ± 0.68
B3	97.30 ± 1.25	97.56 ± 1.83	97.06 ± 1.20	97.30 ± 1.24	99.41 ± 0.26	99.42 ± 0.28
B4	99.74 ± 0.36	99.49 ± 0.71	100.00 ± 0.00	99.75 ± 0.36	100.00 ± 0.00	100.00 ± 0.00
B5	88.33 ± 0.59	86.34 ± 1.04	91.11 ± 1.57	88.65 ± 0.61	93.19 ± 0.75	93.75 ± 0.20
B6	75.76 ± 5.67	83.33 ± 16.36	75.76 ± 18.68	75.48 ± 3.85	73.83 ± 10.68	77.81 ± 13.26
B7	91.09 ± 1.47	93.84 ± 5.19	88.51 ± 4.53	90.86 ± 1.43	95.59 ± 0.70	96.62 ± 0.55
Average	88.17 ± 4.71	90.57 ± 6.14	87.37 ± 7.71	86.76 ± 5.20	93.93 ± 2.56	92.99 ± 3.46

To better analyze the types of intentions, we introduced a three-class classification task in our experiment, including resting state, stimulus-induced vocalizations, and responses to vocalizations from conspecifics. In this three-class experiment, we performed data balancing, ensuring that each category contained 189 samples, for a total of 567 samples. ECoG-IBGT achieved the best performance across all evaluation metrics, demonstrating the strong applicability and effectiveness of our proposed method (see Table 9).

I Additional Analysis on Graph Construction Sensitivity

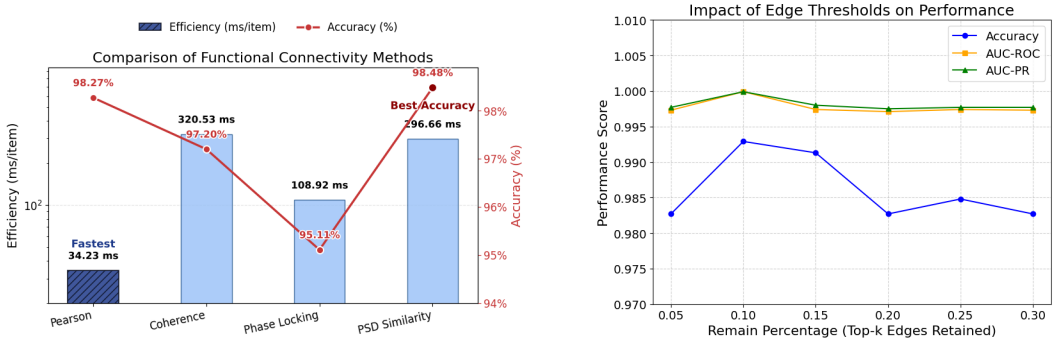
Our graph construction strategy is based on Pearson correlation with a fixed top-k threshold (10%). We note that this choice—of threshold, the use of absolute values, and the length of the time window—may affect downstream performance. To assess robustness, we performed two complementary experiments.

Table 9: Three-class classification performance comparison.

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	AUC-PR
GIN	76.90 ± 7.99	80.27 ± 5.14	76.01 ± 8.42	76.27 ± 8.71	89.65 ± 7.10	81.95 ± 13.04
GAT	52.34 ± 2.98	42.00 ± 1.56	55.53 ± 4.01	44.30 ± 3.51	79.82 ± 2.64	66.09 ± 6.35
GCN	62.57 ± 9.13	54.48 ± 16.29	64.47 ± 7.26	56.64 ± 13.18	80.23 ± 0.62	65.98 ± 2.54
GraphSAGE	67.84 ± 8.90	65.22 ± 15.61	69.17 ± 6.36	65.08 ± 12.27	84.05 ± 0.23	75.67 ± 1.67
BrainIB	79.53 ± 2.30	79.81 ± 2.94	79.53 ± 2.13	79.18 ± 2.42	91.28 ± 2.15	87.15 ± 2.65
EEGNet	81.12 ± 3.01	82.32 ± 1.72	81.38 ± 2.65	81.10 ± 2.82	91.37 ± 1.29	84.04 ± 2.22
MedFormer	73.45 ± 0.72	74.51 ± 0.32	72.92 ± 0.62	73.38 ± 0.48	89.60 ± 0.89	82.45 ± 1.90
TopKPool	68.71 ± 8.05	71.49 ± 9.49	69.06 ± 6.61	68.21 ± 8.31	83.54 ± 1.95	74.60 ± 4.52
DiffPool	76.32 ± 0.72	76.95 ± 1.28	76.07 ± 0.66	76.31 ± 0.90	87.02 ± 0.79	79.95 ± 1.58
SAGPool	75.15 ± 2.30	77.40 ± 0.82	75.16 ± 2.92	75.10 ± 2.33	83.14 ± 1.66	73.73 ± 3.58
Cluster-GT	61.11 ± 3.68	59.10 ± 11.76	60.24 ± 4.06	55.97 ± 7.47	81.70 ± 2.89	70.02 ± 5.88
GraphPCA	62.87 ± 5.79	63.57 ± 14.36	64.37 ± 3.68	60.50 ± 9.25	84.19 ± 1.30	75.43 ± 1.93
ECoG-IBGT	$\mathbf{90.64 \pm 4.32}$	$\mathbf{91.51 \pm 3.64}$	$\mathbf{90.86 \pm 3.93}$	$\mathbf{90.69 \pm 4.10}$	$\mathbf{98.52 \pm 1.18}$	$\mathbf{97.17 \pm 2.30}$

(1) Comparing different functional connectivity measures. We evaluated whether the choice of connectivity metric substantially affects performance or computational efficiency. We compared Pearson correlation, coherence, phase-locking-value, and power spectral density (PSD) similarity, each retaining the top 30% of edges. Results are summarized in Figure 12. While accuracy is broadly similar (94%–98.5% across methods), runtime per sample differs significantly. Pearson correlation achieves both strong accuracy (96.65%) and the fastest efficiency (34.23 ms per sample), whereas coherence and PSD similarity provide slightly higher accuracy (97.2%–97.5%) at the cost of nearly an order of magnitude higher computation time. Based on this analysis, we adopt Pearson correlation for subsequent experiments.

(2) Varying edge-retention thresholds. We next assessed how performance changes when varying the proportion of retained edges under the Pearson correlation metric. Specifically, we varied the remaining percentage of top-k edges from 5% to 30% while keeping other parameters fixed (absolute value, window size). As shown in Figure 12, classification performance remains highly stable across thresholds. Both AUC-ROC and AUC-PR consistently remain above 0.997, while accuracy fluctuates only within ±0.01. This suggests our model is robust to the specific choice of edge-retention threshold.



(a) Comparison of functional-connectivity measures (top 30% edges retained).

(b) Effect of varying edge-retention thresholds under Pearson correlation.

Figure 12: (a) Performance vs. efficiency for various connectivity measures. (b) Sensitivity of classification performance to different edge-retention thresholds under Pearson correlation.

Summary. Together, these analyses show that (i) while alternative connectivity measures may yield marginal accuracy gains, Pearson correlation provides the best trade-off between predictive accuracy and computational efficiency; and (ii) under this metric, model performance is stable across a range of edge-retention thresholds. These findings strengthen our conclusions about the robustness of graph-construction choices.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our Abstract and Introduction summarize the three main contributions listed in Section 1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We include limitations at the end of Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We detail the information bottleneck theoretical framework, its assumptions, and derivations in Section 3.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide dataset splits, model architectures, training hyperparameters, random seed settings, and evaluation protocols in Section 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our code will be publicly released upon acceptance at [URL Blinded for Review], and the data supporting this study's findings are available from the authors upon request, subject to approval by [Institution Blinded for Review].

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Section 4.1 details the split ratios, graph transformer configuration, optimizer settings, and hyperparameter values.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Experiments were repeated with different random seeds; Table 2 and Table 3 present the average results, and Fig 3(b) includes error bars denoting ±1 standard deviation. Due to space constraints, the full set of standard deviations is provided in Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide hardware specifications in Section 4 to enable the reproduction of our results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All experimental procedures were approved by Animal Use and Care Committee at [Institution Blinded for Review] and in accordance with the National Institutes of Health Guidelines. The protocol number is (blinded for review).

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our work provides insights into the principles underlying biological activity and contributes to the development of real-time responsive brain-computer interfaces.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The citation has been provided in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]All new assets include full documentation.

    Justification: The GitHub repository provides a README, API guide, and data schema.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used in the core methodology.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.