INTERPRETABLE MULTI-VIEW DYNAMIC GRAPH LEARNING FOR INTENTION PREDICTION VIA INFORMATION BOTTLENECK

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting intention before behavioral onset provides valuable lead time for responsive brain-computer interfaces. However, pre-behavior neural activity exhibits nonstationary spectral dynamics, evolving functional connectivity, and redundant graph structures, posing challenges to both predictive performance and interpretability. We propose an interpretable dynamic graph learning framework for anticipatory intention prediction using ECoG recordings from non-human primates. The framework constructs a multi-view spatio-temporal-frequency representation, models the evolving coupling between neural features and functional topology through dynamic graphs, and learns dynamic predictive subgraphs with an information bottleneck motivated objective. On the vocalization prediction task, our framework achieves 95.55% accuracy and compares favorably with 20 representative baselines. Further analyses reveal distinct predictive neural structures between rest and vocal preparation and show that these structures reorganize throughout the pre-vocalization interval, with window-specific spectral dynamics providing informative predictive evidence, supporting interpretable advance prediction for responsive brain-computer interfaces. Code is available at https://anonymous.4open.science/r/iclr-anon-7f3c2a/.

1 INTRODUCTION

Anticipatory intention decoding is essential for responsive interaction between humans and intelligent systems. Brain–computer interfaces have demonstrated increasingly effective decoding of speech and locomotor behaviors from cortical activity (Willett et al., 2023; Lorach et al., 2023), but systems that respond only after overt behavioral onset leave limited time for sensing, computation, and actuation. Neural activity preceding behavior therefore provides an important opportunity for earlier prediction, potentially enabling more timely assistance, communication, and physical coordination (Gavrilov et al., 2017; Tsunada & Eliades, 2025).

Current neural decoders typically extract temporal or spatial features, or encode channel dependencies with graph structures (Lawhern et al., 2018; Schirrmeister et al., 2017). Recent methods further incorporate spatio-temporal representations and dynamic connectivity (Wang et al., 2025a; Hajisafi et al., 2024). However, anticipatory neural activity is inherently nonstationary, with temporal importance, functional organization, and spectral patterns evolving together before behavior (Khilkevich et al., 2024; Volkova et al., 2019). Modeling these views independently or at mismatched resolutions can obscure the correspondence between a window's neural representation, functional topology, and spectral context, limiting the ability to capture evolving preparatory states. We therefore treat temporal, spatial, and spectral aspects as complementary views of the same anticipatory state.

A related challenge concerns interpreting such evolving graph states (Ying et al., 2019). While dynamic graph models characterize changes in connectivity or node representations, the predictive relationship between neural representations and functional topology within each state remains underexplored (Veličković et al., 2018; Zhao & Wang, 2023). Meanwhile, graph information bottleneck (IB) methods provide a principled basis for selecting task-relevant structures (Yu et al., 2021; Miao et al., 2022), but extending this idea to dynamically evolving prediction pathways requires explanations that change with the underlying graph state. We therefore develop a framework that

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

couples cross-view graph states over time and learns window-specific predictive subgraphs directly within the decoding pathway.

To address these challenges, we develop an interpretable dynamic graph framework for anticipatory marmoset vocalization decoding from ECoG over primary auditory cortex and prefrontal cortex. Sliding windows preserve temporal evolution; electrode-aligned graph states capture window-specific functional connectivity; and complementary whole-segment and window-specific spectral branches provide multi-scale frequency information. These graph states are modeled as an identity-aligned dynamic sequence, while window-specific node and edge masks generate the subgraphs directly used for prediction. Our contributions are threefold:

(1) Multi-Scale Spatio-Temporal-Frequency Fusion. We develop a unified multi-view representation that integrates temporal windows, electrode topology, and global- and window-level spectra to better capture preparatory dynamics across complementary scales.

(2) Cross-View Coupled Dynamic Graph Modeling. By coupling each spectral representation with its corresponding functional topology, our dynamic graph formulation captures their evolving correspondence throughout the pre-behavior interval.

(3) Time-Varying IB Subgraph Learning. An information bottleneck motivated selector learns window-specific node and edge masks, reducing redundant graph information while retaining prediction-relevant structures for temporally resolved interpretation.

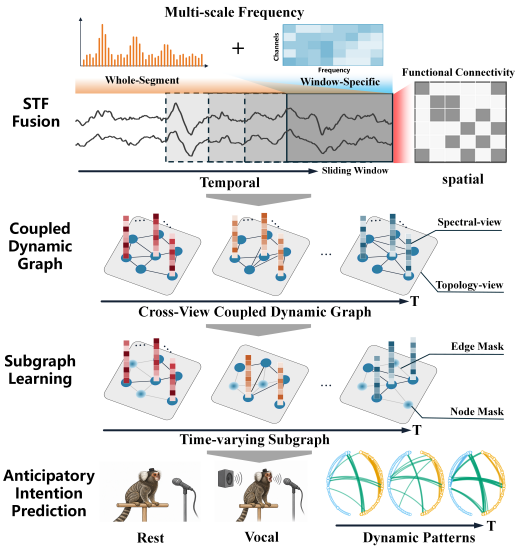

Figure 1: Motivation for interpretable dynamic graph learning. Multi-scale spectral features and window-specific functional topology are coupled in an electrode-aligned graph sequence. IB-guided node and edge selection yields time-varying predictive subgraphs for intention decoding before vocal onset.

2 Related Work

2.1 Multi-Scale Spatio-Temporal-Frequency Modeling

Neural-signal decoders combine temporal, spatial, and spectral information through convolutions, attention, or frequency-domain modeling. EEGNet couples temporal and spatial filtering (Lawhern et al., 2018); EEG-ConvMamba and Medformer capture temporal structure at different ranges or granularities (Huang et al., 2026; Wang et al., 2024). LaBraM and CBraMod learn transferable spatio-temporal representations (Jiang et al., 2024; Wang et al., 2025a); and FilterTS separates complementary spectral components (Wang et al., 2025b). These studies demonstrate the value of modeling different views of neural signals. For anticipatory decoding, however, temporal evolution, electrode-level organization, and spectral dynamics need to be considered jointly while preserving their correspondence across windows. We therefore construct a physically aligned multi-scale representation that integrates these complementary views within each graph state.

2.2 Cross-View Coupled Dynamic Graph Modeling

Graph-based neural decoders explicitly represent channel dependencies through learned adjacency, regional relations, or electrode semantics, as in DGCNN, LGGNet, and NeuroGNN (Song et al., 2020; Ding et al., 2024). Recent dynamic approaches further capture changing connectivity or evolving graph states (Hajisafi et al., 2024; Gao et al., 2026; Afzal et al., 2024). While these methods effectively model temporal changes in neural graphs, anticipatory activity requires understanding

how different aspects of a neural state change together within each time window. We therefore formulate each graph state by coupling the fused node representation with its corresponding functional topology and model the evolution of this coupled state over time.

2.3 Time-Varying IB Subgraph Learning

The graph information bottleneck provides a principled framework for identifying task-relevant structures (Yu et al., 2021). BrainIB selects discriminative brain connections, IRENE combines IB-guided graph construction with structure-aware encoding, and TIF organizes graph reductions into a multigranular tree (Zheng et al., 2025; Li et al., 2026; Yang et al., 2025). In dynamic neural decoding, predictive structures may change with the underlying graph state. We therefore learn window-specific node and edge masks to obtain temporally resolved predictive subgraphs.

3 Method

We map a pre-onset multichannel recording to a prediction through three coordinated stages. Throughout the pipeline, temporal, spatial, and spectral streams are treated as complementary views of each neural state. First, ordered sliding windows are represented by complementary whole-segment and window-specific spectra on a common physical frequency grid, preserving temporal, spatial, and frequency information. Second, within each window, the fused node representation is coupled with its corresponding functional connectivity to form a unified graph state; the resulting graph sequence captures how this cross-view coupling evolves over the pre-event interval. Third, a stochastic selector generates window-specific node and edge masks, producing time-varying predictive subgraphs that directly enter the prediction pathway and are regularized by label, dependence, and structural objectives. Figure 2 summarizes the overall framework.

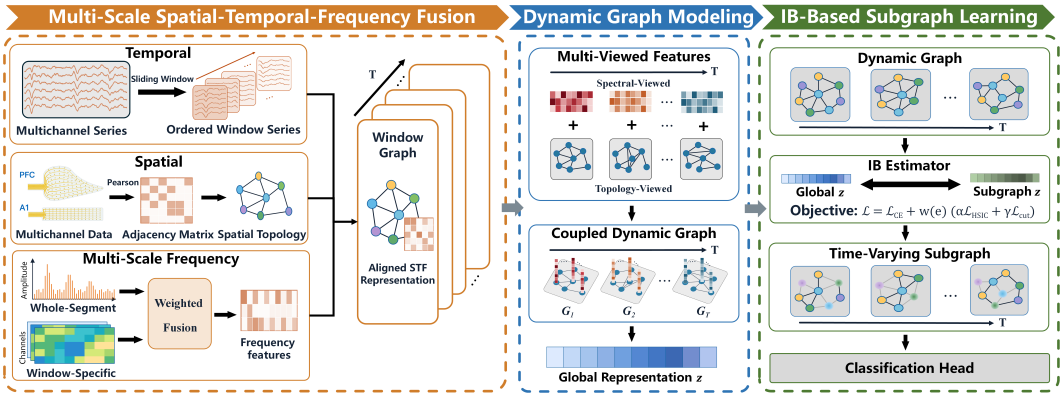

Figure 2: Overview of our method. The model consists of spatio-temporal-frequency fusion, dynamic graph modeling, and IB-based interpretable subgraph learning.

3.1 Problem Definition

We study anticipatory vocal-state decoding from pre-onset ECoG. Given a multichannel recording $X \in \mathbb{R}^{N\times L}$, the task is to predict whether the upcoming behavior is vocalization or rest, represented by $y \in \{0, 1\}$. We divide the pre-event interval into T ordered windows and formulate the decoding path as

$$X \longrightarrow \{G_t\}_{t=1}^{T} \longrightarrow \{G_t^{\mathrm{sub}}\}_{t=1}^{T} \longrightarrow z \longrightarrow \hat{y}. \tag{1}$$

Here, $G_t = (V, A_t, H_t^f)$ is the graph state for window t: V indexes ECoG channels, A_t is the window-specific connectivity, and H_t^f contains the fused spectral node features. Node identities are shared across windows, whereas connectivity and node representations may vary with t. A learned selector further induces the predictive subgraph G_t^{sub} through node and edge masks.

3.2 Multi-Scale Spatio-Temporal-Frequency Fusion

Preparatory neural activity may evolve simultaneously in temporal organization, functional connectivity, and spectral content. We therefore construct each graph state from three aligned views: ordered windows preserve temporal evolution, electrode-level nodes maintain spatial correspondence, and complementary global and window-specific frequency branches characterize spectral information at two temporal scales.

We construct each graph state by aligning temporal evolution, functional topology, and multi-scale spectral information. The recording is decomposed into T ordered overlapping windows $X_t \in \mathbb{R}^{N\times W}$. Each ECoG channel is retained as a node, and window-specific functional connectivity is defined using Pearson correlation:

$$R_{t,ij} = \mathrm{corr}(X_{t,i,:}, X_{t,j,:}), \qquad A_{t,ij} = \mathbf{1}\{|R_{t,ij}| > \tau_A\}, \quad i \neq j, \tag{2}$$

Here, τ_A retains the strongest candidate correlations and $A_{t,ii} = 0$. Thus, node indices remain aligned while functional topology changes across windows.

For frequency encoding, we combine whole-segment spectral context with window-specific dynamics. Both branches use orthonormal real FFTs and are aligned to a common physical frequency grid defined by $\nu_f = ff_s/(2W)$. The global-context branch reconstructs the complete segment and produces $S \in \mathbb{C}^{N\times F}$ through complex embedding and smoothed frequency selection. The window-specific branch produces $D_t \in \mathbb{C}^{N\times F}$ through soft frequency selection, cross-node complex mixing, and gated residual updates. Their complete updates are given in Appendix G.

The two representations are fused with learned weights:

$$\begin{aligned} [w^s_{t,n,f}, w^d_{t,n,f}] &= \mathrm{softmax}\left(\Theta_{t,n,f,:}\right), \\ U_{t,n,f} &= w^s_{t,n,f} S_{n,f} + w^d_{t,n,f} D_{t,n,f}, \\ H^f_{t,n,:} &= \mathrm{LayerNorm}\left(\log\left(1 + |U_{t,n,:}|\right)\right). \end{aligned} \tag{3}$$

The resulting H^f_t preserves window-level spectral variation while incorporating whole-segment context, and is used by both the spatial encoder and the subgraph selector.

3.3 Cross-View Coupled Dynamic Graph Modeling

The fused node representation and functional topology describe complementary aspects of the neural state within each temporal window. Rather than modeling them as independently evolving views, we couple the window-specific node state H^f_t with its corresponding functional connectivity A_t to form a unified graph state $G_t = (V, A_t, H^f_t)$. Here, H^f_t characterizes the neural content of window t, while A_t specifies the functional interactions through which this content is propagated. As both components vary across windows, the sequence $\{G_t\}_{t=1}^T$ captures the evolving correspondence between neural states and functional topology.

For each window, a shared graph-attention encoder models this coupled state by propagating the fused node features over the corresponding functional graph. Writing $H^0_t = H^f_t$ and letting GAT_ℓ denote the multi-head graph-attention update at layer ℓ, we compute

$$H^{\ell+1}_t = \mathrm{GAT}_\ell\left(H^\ell_t, A_t\right), \qquad g_t = \mathrm{Readout}\left(H^L_t, A_t\right). \tag{4}$$

Thus, A_t directly conditions how H^f_t is spatially integrated within each window, rather than serving as an independent auxiliary view. The same spatial encoder is shared across windows, providing a consistent transformation while the node states and functional connectivity evolve over time. The masked encoder path used for prediction is defined in Section 3.4.

The resulting graph representations $\{g_1, \ldots, g_T\}$ are then aggregated across the pre-event interval. Let p_t denote a learned temporal-position vector and β_t the attention weight of window t. We compute

$$\beta_t = \mathrm{softmax}_t\left(\mathrm{MLP}_{\mathrm{time}}(g_t + p_t)\right), \quad z = W_z \sum_t \beta_t g_t + b_z, \quad \hat{y} = \mathrm{MLP}_{\mathrm{cls}}(z). \tag{5}$$

This formulation couples neural content with its corresponding functional topology within each window and models how these coupled graph states evolve throughout the pre-event interval.

3.4 TIME-VARYING IB SUBGRAPH LEARNING

We generate a predictive subgraph for each graph state, enabling the explanation to evolve with the underlying neural dynamics. For each window, the selector produces node and edge masks (M_t^v, M_t^e), yielding a sequence $\{G_t^{\text{sub}}\}_{t=1}^T$ whose structure can evolve over time and directly participates in prediction.

The selector operates on detached fused node features, with $Q_{t,n} = \text{ReLU}(W_e\ \text{sg}(H_{t,n}^f) + b_e)$. Node and symmetric edge logits are

$$s_{t,n}^v = \text{MLP}_v(Q_{t,n}), \qquad s_{t,ij}^e = \frac{(W_q Q_{t,i})^\top (W_k Q_{t,j}) + (W_q Q_{t,j})^\top (W_k Q_{t,i})}{2\sqrt{r}}. \tag{6}$$

During training, we use the Binary-Concrete relaxation (Maddison et al., 2017), defined as $B_\tau(s) = \sigma((s + \log u - \log(1-u))/\tau)$ with $u \sim \mathcal{U}(0,1)$. The window-specific masks are

$$M_t^v = \mathcal{B}_{\tau_v}(s_t^v), \qquad M_t^e = A_t \odot \text{Sym}(\mathcal{B}_{\tau_e}(s_t^e)). \tag{7}$$

At inference, stochastic relaxation is replaced by $\sigma(s)$. The node and edge masks gate the node representations, graph propagation, and readout, so the representation used for prediction is generated from the selected subgraph rather than the full candidate graph.

Mutual Information Estimation. The graph information bottleneck balances label-relevant information and input-dependent information through

$$\min_{G_t^m} \ I(G_t^m; G_t) - \beta I(G_t^m; y). \tag{8}$$

At the representation level, we use the masked embedding Z_t^m for G_t^m and the full-graph embedding Z_t for G_t. The label-information term is optimized with the variational classification objective

$$H(y) - I(Z_t^m; y) = H(y|Z_t^m) \le \mathbb{E}_{y,Z_t^m}[-\log q_\theta(y|Z_t^m)] := \mathcal{L}_{\text{CE}}, \tag{9}$$

where the constant $H(y)$ does not affect optimization. We measure the input dependence between the full-graph and masked representations with the Hilbert-Schmidt independence criterion (HSIC), using its normalized kernel-alignment form averaged over temporal windows. Combining this dependence term with the label-preserving classification loss and the structural coherence regularizer gives

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + w(e)\,(\alpha \mathcal{L}_{\text{HSIC}} + \gamma \mathcal{L}_{\text{cut}}), \qquad w(e) = \min\left\{1, \max\left[0, \frac{e-5}{5}\right]\right\}. \tag{10}$$

For $Z_t, Z_t^m \in \mathbb{R}^{B\times d}$, let K_t^c and L_t^c denote their centered RBF Gram matrices; the resulting window-averaged dependence term is

$$\mathcal{L}_{\text{HSIC}} = \frac{1}{T}\sum_{t=1}^{T} \frac{\langle K_t^c, L_t^c\rangle_F}{\max\{\|K_t^c\|_F \|L_t^c\|_F, \epsilon\}}. \tag{11}$$

The empirical HSIC construction underlying this normalized form is given in Appendix A. The full-graph representation is detached, whereas the masked branch updates the selector and shared spatial encoder. Finally, we encourage the selected nodes and edges to form a coherent structure. Let $S_t = [M_t^v, 1 - M_t^v]$, $\hat{A}_t = \text{OffDiag}(\text{Sym}(A_t \odot M_t^e))$, and $P_t = S_t^\top \hat{A}_t S_t$. With row-normalized Q_t, the structural term is

$$Q_{t,ij} = \frac{P_{t,ij}}{\max\{\sum_k P_{t,ik}, \epsilon\}}, \quad \mathcal{L}_{\text{cut}} = \frac{1}{BT}\sum_{b,t} \|Q_{b,t} - I_2\|_F^2. \tag{12}$$

The three learning signals jointly shape the window-specific masks, producing a sequence of predictive subgraphs whose node and connection patterns can vary throughout the pre-event interval.

4 EXPERIMENTS

Data and evaluation. We recorded ECoG from two awake adult common marmosets (BQ and AZ) using 128-channel arrays spanning A1 and PFC (64 channels each; 1.0-1.52 mm spacing). Sessions involved pure-tone stimulation (BQ B1-B4; 2, 4, 8, and 16 kHz), conspecific-call playback (AZ A1: chair-fixed; A2: free-moving), and antiphonal calling (BQ B5-B7; AZ A3). Signals were sampled at 500 Hz and band-pass filtered at 1-200 Hz. The benchmark comprises 2,038 samples: 1,074 from AZ and 964 from BQ. Positive samples span the one-second interval $[-1.2, -0.2)$ s relative to phee-call onset, whereas negative samples are one-second pre-stimulus segments without vocalization. Each channel is standardized over the full segment and divided into nine 200-ms windows with a 100-ms stride. We report means and standard deviations across ten-fold cross-validation folds.

Implementation and baselines. Each training split uses a 90/10 train-validation partition, and the held-out fold is evaluated at the validation-selected checkpoint. We optimize with Adam using learning rate 10^{-4}, weight decay 10^{-5}, and batch size 32. The candidate graph retains approximately 10% of off-diagonal correlations per sample. Detailed settings are given in Appendix F. Baselines include the convolutional model EEGNet; the convolution-state-space model EEG-ConvMamba; the time-series transformer Medformer; graph decoders SFE-GAT, LGGNet, DGCNN-Raw, NeuroGNN, and TIF; foundation models LaBraM and CBraMod; and interpretable or information bottleneck decoders BrainIB, IRENE, and REST-Raw. Table 1 presents representative methods; Appendix D contains the extended comparison and adaptation details.

4.1 ANTICIPATORY DECODING PERFORMANCE

Table 1 compares the proposed framework with five groups of baselines, including convolution-based decoders, sequence and time-series models, graph-based decoders, foundation models, and interpretable graph models. The proposed framework achieves the strongest overall performance across the compared methods, demonstrating its effectiveness for anticipatory vocalization decoding across diverse architectural paradigms. In particular, the improvement over both conventional temporal models and graph-based decoders suggests that effective prediction of pre-vocalization states requires jointly considering the temporal evolution, functional organization, and spectral characteristics of neural activity. Our framework combines these aspects within a unified dynamic graph formulation, while the specific contribution of each component is further examined through ablation and representation analyses below.

Table 1: Representative baseline results (ten-fold mean and standard deviation). All metrics are percentages. Best and second-best means in each column are marked in **bold** and underlined, respectively. Raw denotes a time-domain-input adaptation.

Model	Accuracy	Prec.	Rec.	F1	ROC-AUC	PR-AUC
EEG-ConvMamba	90.21 ± 2.54	90.76 ± 3.89	89.85 ± 4.95	90.17 ± 2.61	94.16 ± 0.64	94.17 ± 0.69
EEGNet	93.32 ± 4.36	95.25 ± 3.47	91.41 ± 9.25	93.00 ± 5.16	$\mathbf{98.05 \pm 1.23}$	$\underline{97.94 \pm 1.38}$
SFE-GAT	77.67 ± 7.08	79.08 ± 9.46	76.80 ± 5.93	77.64 ± 6.11	85.55 ± 6.27	84.90 ± 7.28
Medformer	72.71 ± 3.22	74.37 ± 4.71	69.98 ± 4.96	71.93 ± 3.04	76.15 ± 5.51	76.27 ± 5.65
CBraMod	92.49 ± 3.53	93.48 ± 5.11	91.78 ± 5.70	92.44 ± 3.46	95.69 ± 2.67	96.47 ± 1.92
LGGNet	90.27 ± 3.12	92.91 ± 6.40	88.02 ± 7.35	90.01 ± 3.29	94.98 ± 2.80	95.19 ± 3.03
DGCNN	89.58 ± 3.18	92.00 ± 4.97	87.32 ± 8.29	89.22 ± 3.56	94.77 ± 2.28	94.22 ± 3.23
NeuroGNN	84.13 ± 5.59	85.17 ± 6.38	83.30 ± 8.82	83.92 ± 5.72	90.86 ± 5.14	90.33 ± 4.72
LaBraM	91.80 ± 3.86	92.05 ± 5.43	91.82 ± 4.59	91.83 ± 3.82	95.51 ± 3.08	95.87 ± 2.72
BrainIB	$\underline{94.37 \pm 2.77}$	$\underline{96.46 \pm 2.89}$	$\underline{92.17 \pm 3.84}$	$\underline{94.23 \pm 2.88}$	$\underline{97.57 \pm 2.32}$	$\mathbf{98.07 \pm 1.63}$
IRENE	90.08 ± 4.99	93.11 ± 5.78	87.08 ± 8.70	89.68 ± 5.30	93.26 ± 3.59	93.92 ± 4.11
REST	85.32 ± 6.69	85.61 ± 7.81	85.31 ± 7.78	85.32 ± 6.84	89.68 ± 6.89	89.12 ± 7.69
TIF	89.08 ± 4.81	92.25 ± 5.81	85.75 ± 7.82	88.65 ± 5.12	93.63 ± 4.49	94.50 ± 3.52
Proposed	$\mathbf{95.55 \pm 2.46}$	$\mathbf{98.04 \pm 2.90}$	$\mathbf{93.16 \pm 5.34}$	$\mathbf{95.40 \pm 2.63}$	97.26 ± 2.11	97.89 ± 1.48

4.2 GENERALIZATION ANALYSIS

To evaluate whether the learned representation captures transferable anticipatory neural patterns beyond subject-specific characteristics, we train on BQ and test on the disjoint AZ dataset. As

shown in Table 2, the proposed framework maintains competitive performance under cross-subject transfer. This suggests that the learned representation captures shared properties of pre-vocalization neural states rather than relying solely on subject-specific signal patterns.

Such transferability is supported by the electrode-aligned dynamic graph representation, which preserves consistent spatial correspondence across recordings, and by the IB-based selection mechanism, which reduces reliance on redundant subject-dependent structures. Results under the complementary LOSO protocol are provided in Appendix E.

Table 2: BQ–AZ generalization comparison. Results use ten-fold source training and a fixed, disjoint target set; all metrics are percentages. Best and second-best means in each column are marked in **bold** and underlined, respectively.

Method	Acc.	Prec.	Rec.	F1	ROC-AUC	PR-AUC
EEG-ConvMamba	64.96 ± 6.86	$\mathbf{82.30 \pm 9.81}$	36.44 ± 15.40	49.03 ± 18.26	77.02 ± 6.30	$\underline{77.34 \pm 7.84}$
SFE-GAT	68.35 ± 2.29	68.26 ± 5.04	70.32 ± 9.98	67.63 ± 3.69	74.47 ± 2.76	75.48 ± 2.94
Medformer	49.95 ± 0.94	51.10 ± 4.25	39.99 ± 15.52	42.37 ± 13.35	49.63 ± 1.58	50.35 ± 1.24
LGGNet	65.79 ± 3.45	68.12 ± 3.70	61.33 ± 15.06	63.04 ± 10.51	73.39 ± 2.26	71.95 ± 2.68
DGCNN-Raw	51.47 ± 0.89	51.73 ± 1.09	45.74 ± 10.92	47.92 ± 5.92	52.26 ± 1.47	51.20 ± 1.42
NeuroGNN	65.58 ± 2.94	65.69 ± 5.26	68.07 ± 8.72	66.29 ± 2.60	72.02 ± 3.41	71.06 ± 3.76
LaBraM	57.19 ± 2.79	58.07 ± 2.18	51.07 ± 15.72	53.22 ± 9.57	61.76 ± 2.94	58.18 ± 2.96
BrainIB	$\underline{70.45 \pm 5.51}$	69.03 ± 5.53	75.58 ± 17.76	$\underline{70.94 \pm 9.71}$	$\underline{78.31 \pm 6.10}$	76.00 ± 6.93
IRENE	62.89 ± 2.43	61.85 ± 3.09	68.78 ± 8.71	64.77 ± 3.36	67.74 ± 3.84	65.17 ± 4.78
REST-Raw	52.65 ± 3.72	52.62 ± 3.63	51.23 ± 6.17	51.86 ± 4.76	53.92 ± 5.08	54.36 ± 4.91
CBraMod	65.08 ± 5.24	61.49 ± 4.70	$\underline{81.96 \pm 11.33}$	69.92 ± 5.43	70.39 ± 8.17	66.83 ± 7.68
TIF	67.53 ± 3.34	$\underline{72.20 \pm 3.17}$	57.41 ± 9.58	63.48 ± 6.09	73.13 ± 4.60	74.83 ± 3.49
Proposed	$\mathbf{72.57 \pm 3.12}$	69.84 ± 7.00	$\mathbf{83.51 \pm 11.66}$	$\mathbf{75.07 \pm 3.29}$	$\mathbf{82.84 \pm 2.32}$	$\mathbf{82.03 \pm 2.56}$

4.3 Ablation Study

Table 3 evaluates the contribution of the major model components and optimization objectives.

Table 3: Component and objective ablations (%).

Configuration	Accuracy	F1	ROC-AUC
Multi-scale STF fusion			
Full model	95.55 ± 2.46	95.40 ± 2.63	97.26 ± 2.11
w/o frequency encoder	83.12 ± 3.31	83.00 ± 3.27	86.99 ± 4.52
w/o spatial propagation	92.50 ± 4.38	92.61 ± 4.20	95.77 ± 4.90
w/o temporal attention	94.87 ± 2.67	94.66 ± 2.83	98.23 ± 1.29
Time-varying IB subgraph learning			
Full model	95.55 ± 2.46	95.40 ± 2.63	97.26 ± 2.11
w/o explainer	92.82 ± 3.96	92.46 ± 4.34	96.39 ± 3.15
w/o MI constraint	94.53 ± 3.13	94.42 ± 3.16	97.27 ± 2.08
w/o connectivity constraint	94.35 ± 3.19	94.03 ± 3.61	97.14 ± 2.78
w/o both constraints	93.84 ± 3.10	93.53 ± 3.50	96.65 ± 2.69

The first block examines the multi-scale spatio-temporal-frequency representation. Removing individual components consistently degrades performance, demonstrating that anticipatory neural activity contains complementary information across spectral, spatial, and temporal dimensions. This supports the necessity of integrating multiple views rather than relying on a single representation. Additional branch perturbations and the controlled comparison with the original FilterTS frontend further verify the benefit of physically aligned frequency modeling (Table 5, Appendix B; Table 14, Appendix H.2).

The second block evaluates the role of IB-based subgraph learning. Removing the explainer or either auxiliary constraint weakens the learned representation, indicating that prediction benefits from identifying compact and informative graph structures. The MI constraint implemented by HSIC and the connectivity constraint guide the selector to retain task-relevant information while suppressing redundant structures, allowing the learned subgraphs to serve as both predictive representations and interpretable explanations. Additional density and objective-weight analyses are provided in Appendix H.

4.4 Analysis of Cross-view Dynamic Representation

We analyze whether the proposed framework benefits from coupling spectral representations with functional topology. Across the pre-event interval, both functional connection patterns and the spectral composition of highly selected nodes evolve over time (Figure 3A-B). These coordinated

changes suggest that neural activity and functional organization jointly characterize the evolving preparatory state.

Cross-module alignment interventions further examine the importance of this correspondence (Figure 3C). Preserving the alignment between frequency features and graph states maintains more reliable prediction, whereas disrupting the correspondence between neural content, topology, and selection leads to larger degradation. These results demonstrate that the model effectively exploits the interaction between spectral dynamics and functional topology during intention decoding. Complete intervention results are reported in Appendix C.5.

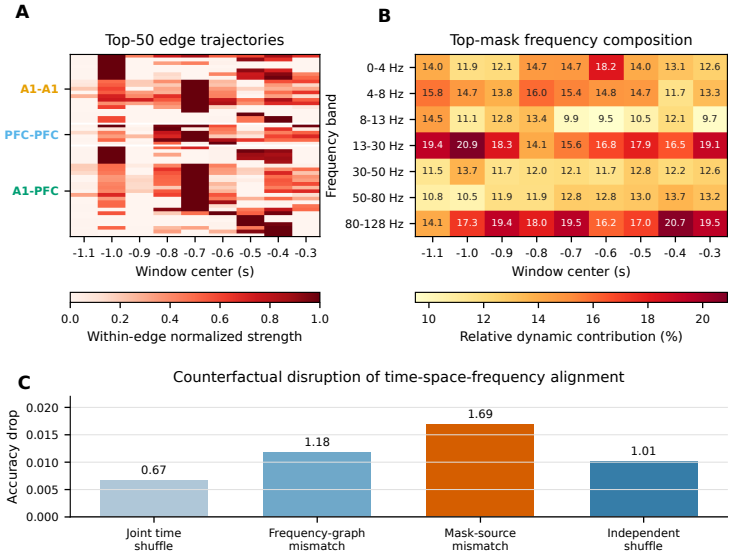

Figure 3: Evolving cross-view coupling and its predictive relevance. (A) Within-edge normalized trajectories of the top 50 connections. (B) Window-specific frequency composition of the top-32 node-mask channels. (C) Accuracy drops under interventions that preserve or disrupt cross-view correspondence.

4.5 Interpretability Analysis

State-dependent spatial patterns and predictive relevance. We first evaluate whether the learned subgraphs capture structures that are directly relevant to the decoding process.

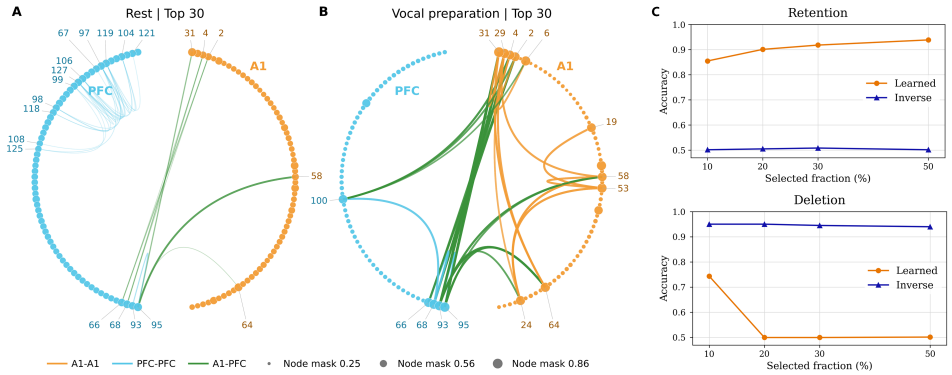

Figure 4: State-dependent predictive subgraphs and faithfulness. (A, B) Node-mask and connectivity patterns for rest and vocal preparation. (C) Accuracy under progressive retention and targeted deletion of the selected structure.

At matched checkpoints, masked inference achieves better performance than the full-graph pathway (Table 7), showing that the selected structures preserve discriminative information while reducing

unnecessary graph components. Furthermore, targeted removal of highly selected nodes and candidate edges substantially affects prediction, whereas retaining the selected structures maintains decoding capability (Figure 4C).

These results demonstrate that the learned subgraphs are integrated into the prediction pathway and contribute directly to decision making. The state-dependent node and connection patterns shown in Figure 4A-B further reveal that different neural structures are emphasized during rest and vocal preparation, suggesting that the selector identifies behavior-related graph patterns. Full inference-path and intervention results appear in Appendix C.

Temporal evolution of selected neural structures. We further examine how the selected predictive structures change throughout the pre-event interval. Adjacent-window comparisons reveal temporal variations in both selected nodes and connections (Table 8), indicating that predictive structures evolve with the underlying neural state. Figure 5 illustrates the transition of predictive subgraphs across early, peak-attention, and late preparation stages, which follows the temporal profile of model attention.

These findings show that the learned selector provides a time-resolved representation of predictive structures. The resulting dynamic subgraphs characterize the evolving neural organization associated with intention formation before behavioral onset.

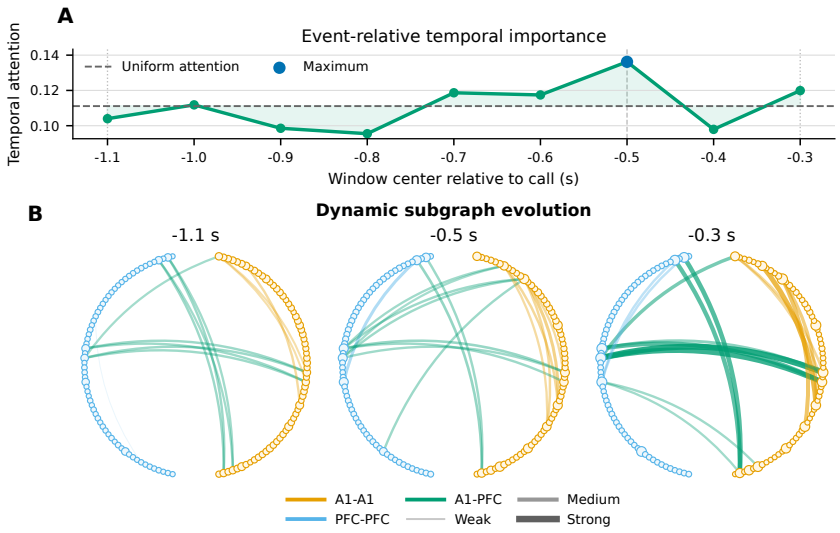

Figure 5: Temporal attention and evolution of predictive subgraphs before vocal onset. (A) Mean temporal attention over vocalization OOF samples. (B) Mean masked subgraphs for the OOF samples at early, peak-attention, and late windows.

5 Conclusion

We presented a framework for anticipatory decoding from multi-region ECoG. The model combines global-context and window-specific frequency encoding with HSIC-regularized soft subgraph learning and temporal attention. It achieves 95.55% accuracy on pre-vocalization decoding, with consistent gains from joint auditory-prefrontal input. Component ablations, state-dependent patterns, and targeted perturbations connect the learned representations and spatial selections to prediction, supporting interpretable multi-region neural anticipation. Although the evaluation now includes two animals and cross-animal transfer, the number of animals and recording sessions remains limited; larger cohorts and strictly nested model selection are needed to establish broader generalization.

AI USE STATEMENT

AI tools were used to assist with manuscript writing and language polishing, and to support literature retrieval and the identification of relevant prior work, and the implementation and debugging of experimental code

REFERENCES

Arshia Afzal, Grigorios Chrysos, Volkan Cevher, and Mahsa Shoaran. REST: Efficient and accelerated EEG seizure analysis through residual state updates. In *International Conference on Machine Learning*, pp. 271–290, 2024.

Yi Ding, Neethu Robinson, Chengxuan Tong, Qiuhao Zeng, and Cuntai Guan. LGGNet: Learning from local-global-graph representations for brain–computer interface. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):9773–9786, 2024. doi: 10.1109/TNNLS.2023.3236635.

Xin Gao, Guohua Cao, and Guoqing Ma. SFE-GAT: Structure-feature evolution graph attention network for motor imagery decoding. *Sensors*, 26(5):1730, 2026. doi: 10.3390/s26051730.

Natalja Gavrilov, Steffen R. Hage, and Andreas Nieder. Functional specialization of the primate frontal lobe during cognitive control of vocalizations. *Cell Reports*, 21(9):2393–2406, 2017. doi: 10.1016/j.celrep.2017.10.107.

Arash Hajisafi, Haowen Lin, Yao-Yi Chiang, and Cyrus Shahabi. Dynamic GNNs for precise seizure detection and classification from EEG data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 207–220, 2024.

Jianxi Huang, Yinghui Chang, Sen Yang, Jigang Tong, Kai Zhang, and Shengzhi Du. EEG-ConvMamba: Motor imagery EEG decoding and visualization via convolutional neural networks and Mamba. *Biomedical Signal Processing and Control*, 121:110357, 2026. doi: 10.1016/j.bspc.2026.110357.

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

Andrei Khilkevich, Michael Lohse, Ryan Low, Ivana Oršolić, Tadej Božič, Paige Windmill, and Thomas D. Mrsic-Flogel. Brain-wide dynamics linking sensation to action during decision-making. *Nature*, 634:890–900, 2024. doi: 10.1038/s41586-024-07908-w.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529, 2019.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. 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.

Lincan Li, Rikuto Kotoge, Xihao Piao, Zheng Chen, and Yushun Dong. Optimizing EEG graph structure for seizure detection: An information bottleneck and self-supervised learning approach. *arXiv preprint arXiv:2604.01595*, 2026. doi: 10.48550/arXiv.2604.01595.

Henri Lorach, Andrea Galvez, Valeria Spagnolo, et al. Walking naturally after spinal cord injury using a brain–spine interface. *Nature*, 618:126–133, 2023. doi: 10.1038/s41586-023-06094-5.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.

Siqi Miao, Mia Liu, and Pan Li. Interpretable and generalizable graph learning via stochastic attention mechanism. In *International Conference on Machine Learning*, pp. 15524–15543, 2022.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017. doi: 10.1002/hbm.23730.

Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3): 532–541, 2020. doi: 10.1109/TAFFC.2018.2817622.

Joji Tsunada and Steven J. Eliades. Frontal-auditory cortical interactions and sensory prediction during vocal production in marmoset monkeys. *Current Biology*, 35(10):2307–2322.e3, 2025. doi: 10.1016/j.cub.2025.03.077.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

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

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

Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A multi-granularity patching transformer for medical time-series classification. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yulong Wang, Yushuo Liu, Xiaoyi Duan, and Kai Wang. FilterTS: Comprehensive frequency filtering for multivariate time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2025b. doi: 10.1609/aaai.v39i20.35438.

Francis R. Willett, Erin M. Kunz, Chaofei Fan, Donald T. Avansino, Guy H. Wilson, Eun Young Choi, Foram Kamdar, Matthew F. Glasser, Leigh R. Hochberg, Shaul Druckmann, Krishna V. Shenoy, and Jaimie M. Henderson. A high-performance speech neuroprosthesis. *Nature*, 620: 1031–1036, 2023. doi: 10.1038/s41586-023-06377-x.

Jie Yang, Yuwen Wang, Kaixuan Chen, Tongya Zheng, Yihe Zhou, Zhenbang Xiao, Ji Cao, Mingli Song, and Shunyu Liu. From GNNs to trees: Multi-granular interpretability for graph neural networks. In *International Conference on Learning Representations*, 2025.

Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. GNNExplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Junchi Yu, Tingyang Xu, Yu Rong, Yatao Bian, Junzhou Huang, and Ran He. Graph information bottleneck for subgraph recognition. In *International Conference on Learning Representations*, 2021.

Lingyun Zhao and Xiaoqin Wang. Frontal cortex activity during the production of diverse social communication calls in marmoset monkeys. *Nature Communications*, 14:6634, 2023. doi: 10.1038/s41467-023-42052-5.

Kaizhong Zheng, Shujian Yu, Baojuan Li, Robert Jenssen, and Badong Chen. BrainIB: Interpretable brain network-based psychiatric diagnosis with graph information bottleneck. *IEEE Transactions on Neural Networks and Learning Systems*, 36(7):13066–13079, 2025. doi: 10.1109/TNNLS.2024.3449419.

A HSIC-REGULARIZED SUBGRAPH LEARNING

A.1 FROM EMPIRICAL HSIC TO NORMALIZED KERNEL DEPENDENCE

Let $Z_t, Z_t^m \in \mathbb{R}^{B\times d}$ denote the full and masked spatial representations at window t. Their RBF kernels are

$$[K_t]_{ij} = \exp\left(-\frac{\|Z_{t,i} - Z_{t,j}\|_2^2}{2\rho_t}\right), \qquad [L_t]_{ij} = \exp\left(-\frac{\|Z_{t,i}^m - Z_{t,j}^m\|_2^2}{2\rho_t^m}\right), \tag{13}$$

where ρ_t and ρ_t^m are the medians of the nonzero squared pairwise distances in their respective representations. With $H_B = I_B - B^{-1}\mathbf{1}\mathbf{1}^\top$, empirical HSIC is

$$\widehat{\mathrm{HSIC}}(Z_t, Z_t^m) = \frac{\mathrm{tr}(K_t H_B L_t H_B)}{(B-1)^2}. \tag{14}$$

Since $H_B^\top = H_B$, $H_B^2 = H_B$, and both kernels are symmetric,

$$\mathrm{tr}(K_t H_B L_t H_B) = \langle H_B K_t H_B, H_B L_t H_B\rangle_F. \tag{15}$$

Normalizing by the two self-dependence terms gives the centered kernel alignment form (Kornblith et al., 2019):

$$\begin{aligned}\mathrm{NHSIC}_t &= \frac{\widehat{\mathrm{HSIC}}(Z_t, Z_t^m)}{\sqrt{\widehat{\mathrm{HSIC}}(Z_t, Z_t)\widehat{\mathrm{HSIC}}(Z_t^m, Z_t^m)}} \\ &= \frac{\langle K_t^c, L_t^c\rangle_F}{\|K_t^c\|_F \|L_t^c\|_F}.\end{aligned} \tag{16}$$

The equality holds for nondegenerate centered kernels. The implementation floors the denominator by a positive constant and clips the result to $[0, 1]$. Averaging over windows yields Equation 11. The factor $(B-1)^{-2}$ cancels, but the self-dependence normalization depends on the representations. Thus, the normalized objective measures the alignment of sample similarities rather than the unnormalized magnitude of dependence. We use it as a dependence surrogate motivated by information bottleneck learning.

A.2 JOINT OPTIMIZATION AND GRADIENT PATHS

Write the frequency encoder, selector, and spatial encoder parameters as ψ, ϕ, and θ. Let $F = E_\psi(X)$ and $\bar{F} = \mathrm{sg}(F)$. The selector generates M from $\bar{F}, A$, and the auxiliary branches compute

$$Z_t = \mathrm{sg}\big(g_\theta(\bar{F}_t, A_t; \mathbf{1})\big), \qquad Z_t^m = g_\theta(\bar{F}_t, A_t; M_t). \tag{17}$$

The full-graph representation is held constant within each backward pass. The two branches share a spatial encoder, whose parameters change during training. Kernel bandwidths are computed from the current batch and detached. Dropout is disabled for auxiliary spatial computations, while the training-mode selector still performs Binary-Concrete sampling. Table 4 summarizes the direct gradient paths. Cross-entropy uses non-detached frequency features and trains all prediction modules.

Table 4: Modules directly updated by each objective.

Objective	Frequency	Selector	Spatial/readout	Temporal/classifier
Cross-entropy	Yes	Yes	Yes	Yes
Spatial NHSIC	No	Yes	Yes	No
Low-cut	No	Yes	No	No

A.3 BINARY RELAXATION AND SYMMETRIC EDGE MASKS

For a node logit s, let $p = \sigma(s)$. The retained coordinate of a two-class Gumbel-Softmax sample can be written as

$$\frac{\exp((\log p + g_1)/\tau)}{\exp((\log p + g_1)/\tau) + \exp((\log(1-p) + g_0)/\tau)} = \sigma\left(\frac{s + g_1 - g_0}{\tau}\right), \tag{18}$$

where g_0, g_1 are independent standard Gumbel variables. Their difference has a Logistic distribution, giving the Binary-Concrete expression in Equation 7 (Maddison et al., 2017). For edges, we first symmetrize the logits, independently relax the directional entries, and then average the two masks. The final undirected weight is therefore a symmetric continuous random variable. Normalized HSIC is evaluated on the resulting graph embeddings and does not require an explicit density for this averaged mask. Inference uses deterministic sigmoid weights.

B Additional Frequency Interpretability Analyses

B.1 Stability of the Global-Context Frequency Profile

We average each sample's global-context soft mask over all 128 channels and rank the frequency coordinates. We compare this ranking with the mean ranking of its fold using top-10 Jaccard similarity, $J(U, V) = |U \cap V|/|U \cup V|$, and full-ranking Spearman correlation. Mean sample-to-fold Jaccard similarity is 0.8182, and mean Spearman correlation is 0.8740. Global-context frequency selection is therefore largely consistent across samples.

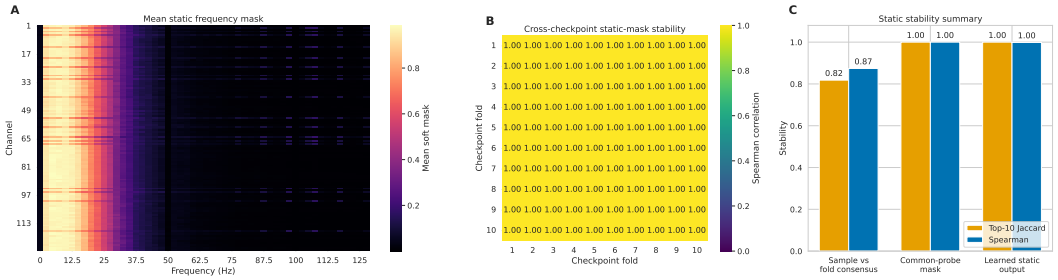

Figure 6: Global-context frequency stability. Sample-specific rankings are compared with fold consensus, including overlap among the highest-weight frequency coordinates. The global-context mask is generated from the input spectrum with a fixed quantile rule; learned gates and complex embeddings determine its contribution to the representation.

B.2 Window-Specific Frequency Patterns and Branch Contributions

Figure 7 shows class-averaged Layer 2 window-specific masks as time–frequency maps. Rest and vocalization use a shared color scale; their difference is centered at zero. The maps reveal changes in frequency selection across windows and states.

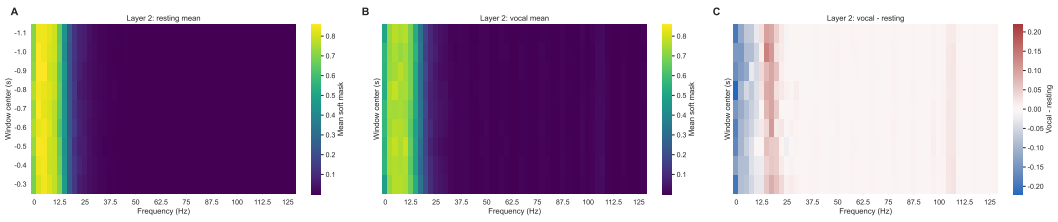

Figure 7: State-dependent window-specific frequency selection. The three panels show Layer 2 masks for rest, vocalization, and vocalization minus rest.

Although learned fusion weights are close to an equal allocation, branch magnitudes can differ. We quantify the window-specific branch's magnitude-weighted contribution as

$$c^d_{t,n,f} = \frac{w^d_{t,n,f}|D_{t,n,f}|}{w^s_{t,n,f}|S_{n,f}| + w^d_{t,n,f}|D_{t,n,f}| + \epsilon}. \tag{19}$$

Its mean over all channels is 0.7309. This statistic describes relative magnitudes entering fusion; phase interactions and subsequent normalization also affect the resulting features. We therefore assess predictive dependence through direct interventions.

B.3 Branch Perturbations

We intervene before complex branch fusion and then recompute fusion, normalization, masks, and predictions. Zeroing removes a branch representation; cross-sample shuffling breaks its correspondence with the sample; temporal shuffling changes the order of window states; phase randomization changes phase while preserving magnitude. Each condition is paired with the original prediction.

Table 5: Pre-fusion branch interventions. Reference pooled accuracy is 95.05%; intervals describe accuracy drops. Accuracy is in percent; drops and intervals are in percentage points.

Intervention	Accuracy	Drop	95% CI
Global-context zeroing	91.13	3.92	[2.38, 5.63]
Window-specific zeroing	59.90	35.15	[30.89, 39.08]
Global-context sample shuffle	83.45	11.60	[8.87, 14.33]
Window-specific sample shuffle	55.63	39.42	[35.32, 43.86]
Window-specific time shuffle	94.37	0.68	[0.17, 1.37]
Phase randomization	91.98	3.07	[1.71, 4.61]

Zeroing the window-specific branch reduces accuracy by 35.15 percentage points, and shuffling it across samples reduces accuracy by 39.42 points. Both effects exceed their global-context counterparts. Local spectral representations carry most of the sample-specific predictive information, with the whole-segment branch providing additional support.

B.4 Frequency-Band Occlusion

We zero the global-context and window-specific complex features within one predefined band at a time and recompute downstream predictions. Table 6 orders these interventions by physical frequency.

Table 6: Independent frequency-band occlusions and paired bootstrap confidence intervals for accuracy drops. Accuracy is in percent; drops and intervals are in percentage points.

Occluded band (Hz)	Accuracy	Drop	95% CI
0–4	79.52	15.53	[12.29, 18.60]
4–8	74.57	20.48	[17.06, 24.40]
8–13	55.46	39.59	[35.15, 43.86]
13–30	67.58	27.47	[23.55, 31.57]
30–50	53.07	41.98	[37.54, 46.59]
50–80	90.78	4.27	[2.56, 6.14]
80–128	52.56	42.49	[38.23, 46.76]

Occlusion has the largest effects in the 80–128, 30–50, and 8–13 Hz bands, while 50–80 Hz produces the smallest drop. These interventions measure end-to-end sensitivity to the corresponding spectral features, including the response of normalization and subgraph selection to each occlusion.

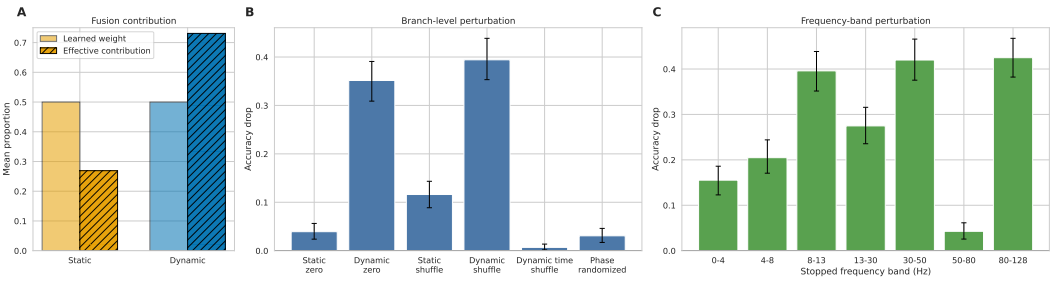

Figure 8: Frequency contributions and perturbations. Relative global-context and window-specific branch magnitudes are shown alongside changes in prediction after branch and band interventions.

C Additional Spatial and Temporal Interpretability Analyses

C.1 Masked and Full-Graph Inference

We compare masked prediction with a full-graph path that bypasses node and edge masks at the same trained checkpoints. Encoder and classifier parameters remain fixed.

Table 7: Masked and full-graph inference at the same trained checkpoints. Predictive metrics are percentages.

Inference path	Accuracy	F1	ROC-AUC	PR-AUC
Masked graph	95.05	94.89	97.25	97.79
Full graph	92.49	92.11	95.95	96.96

Masked inference improves pooled accuracy by 2.56 percentage points, and the two paths agree on 96.42% of predictions. Node masks have mean and standard deviation 0.3946 ± 0.2678; candidate-edge masks have 0.0817 ± 0.1936. The 90th percentile of node weights is 0.8404, and the 99th percentile of edge weights is 0.9618. The distributions contain a relatively small high-weight subset alongside many low-weight locations.

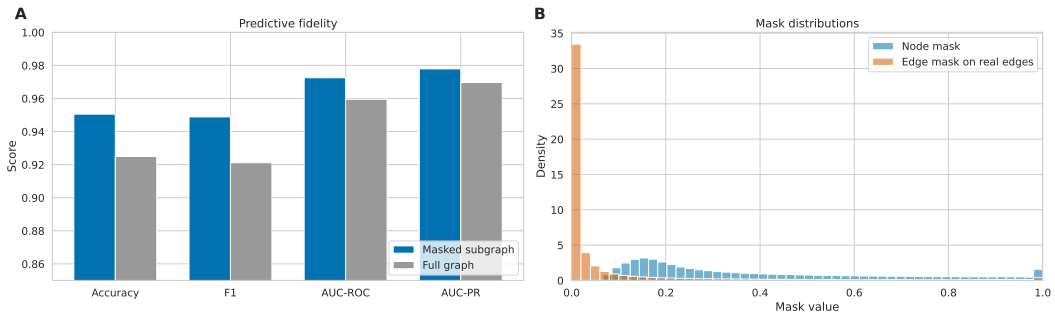

Figure 9: Spatial mask distributions and predictive fidelity. Soft node and edge weights are shown together with full-graph and masked-graph inference at fixed checkpoints.

C.2 Temporal Changes in Selected Nodes and Edges

For each of the vocalization samples, we rank nodes and off-diagonal candidate edges within each window and compare top-K sets between adjacent windows. We use $K \in \{16, 32, 64\}$ for nodes and $K \in \{25, 50, 100\}$ for edges. Statistics are averaged over samples and then over adjacent-window pairs.

Table 8: Mean adjacent-window Jaccard similarity at different selection sizes.

Element	Top-K	Jaccard
Nodes	16	0.5693
Nodes	32	0.3846
Nodes	64	0.4510
Edges	25	0.2260
Edges	50	0.1499
Edges	100	0.1373

The top-50 edge Jaccard similarity is 0.1499, and the top-32 node similarity is 0.3846. High-weight node and edge sets thus change across the pre-event interval. Edge-set turnover reflects both changing candidate graphs and changing mask rankings; comparisons on shared candidate edges further characterize temporal changes in selection.

Figure 11 connects these changes to temporal attention and mask weighting. Panel A shows higher mean node masks in high-attention windows across all 128 channels, with different effect sizes

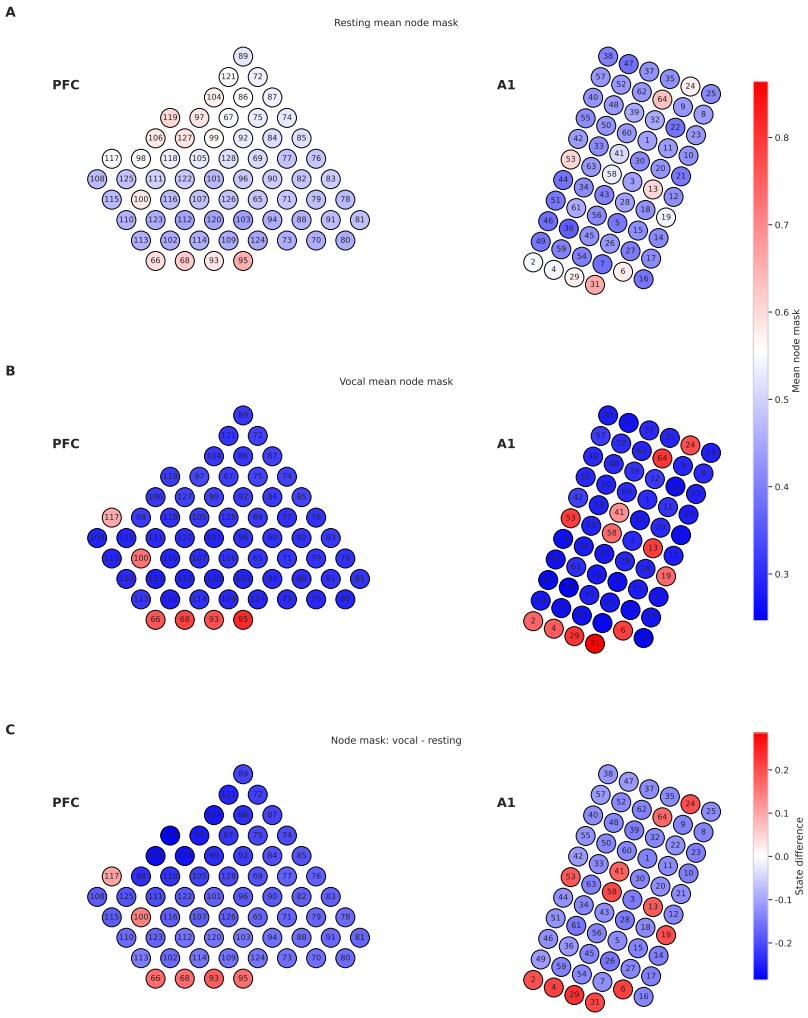

Figure 10: State-dependent node masks. Rest, vocalization, and their difference are displayed over all 128 channels.

across electrodes. Temporal emphasis is thus accompanied by nonuniform changes in node weighting. Panel B shows that node-mask rankings vary across windows while retaining some shared structure. Panels C and D provide complementary views of edge dynamics: effective subgraph similarity captures changes in the weighted graph, whereas similarity on common candidate edges tracks how the selector reweights connections available in both windows. Adjacent windows are more similar on average than nonadjacent windows in all three matrices, indicating local temporal continuity within an evolving spatial representation. Together, these patterns show that temporal attention, node weighting, and edge selection vary in a coordinated way across the pre-vocalization interval.

C.3 Attention-Conditioned Node and Frequency Profiles

For each vocalization sample, we identify its highest- and lowest-attention windows and compare their node masks and node-mask-weighted dynamic spectra. Paired sign-flip tests are followed by Benjamini–Hochberg false discovery rate correction separately over nodes and frequencies. Differences reach $q < 0.05$ for 123 of 128 nodes and 38 of 52 frequency coordinates.

We also permute window order within each sample to test the across-sample mean temporal pattern. The permutation tests yield $p = 0.0002$ for node patterns and $p = 0.0006$ for normalized weighted

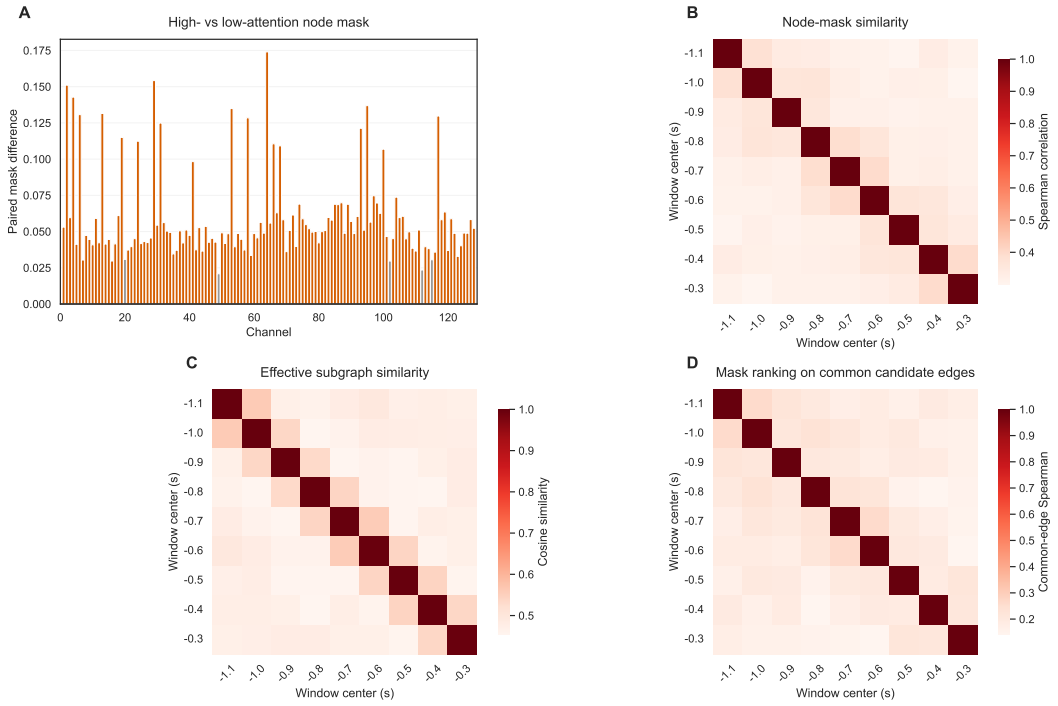

Figure 11: Temporal changes in nodes and edges. (A) Paired node-mask differences between the highest- and lowest-attention windows; orange denotes FDR-adjusted $q < 0.05$ and gray denotes the remaining channels. (B) Within-sample node-mask Spearman similarity averaged across samples. (C) Effective subgraph cosine similarity. (D) Edge-mask Spearman similarity restricted to candidate edges shared between windows. Time axes report window centers relative to vocal onset. Each similarity panel uses its own observed range.

spectra. Temporal weighting is therefore accompanied by systematic changes in spatial selection and spectral profiles.

C.4 Complete Retention and Deletion Results

For each sample and window, we rank nodes and off-diagonal edges in descending mask order (learned), ascending order (inverse), or random order. Retention first keeps a fraction p of nodes, then retains a fraction p of candidate edges among those nodes. Existing self-connections are retained with their nodes. Deletion removes a fraction p of nodes and a fraction p of edges according to their original candidate sets. Node deletion also removes incident propagation paths. All interventions operate after frequency encoding and keep frequency features fixed.

Table 9: Complete subgraph intervention accuracies. Fractions follow the node/edge selection procedure in the text. Predictive metrics are percentages.

Operation	Fraction	Learned	Inverse	Random
Retain	10%	85.49	50.17	86.01
Retain	20%	90.10	50.51	90.44
Retain	30%	91.81	50.85	92.83
Retain	50%	93.86	50.17	93.34
Delete	10%	74.40	95.05	95.05
Delete	20%	50.00	95.05	94.54
Delete	30%	50.00	94.54	94.20
Delete	50%	50.17	94.03	94.20

Deleting the learned top 10% gives 0.7440 accuracy, compared with 0.9505 for random deletion. At 20% deletion, learned accuracy reaches 0.5000. Retention accuracy increases from 0.8549 to 0.9386

for learned ranking and from 0.8601 to 0.9334 for random ranking. High-mask deletion is strongly disruptive, while retaining a small high-mask subset still loses predictive performance.

C.5 Cross-Module Alignment Interventions

Joint time shuffling permutes frequency and graph windows together. Frequency–graph mismatch permutes frequency windows alone and regenerates masks. Mask-source mismatch generates masks from misaligned frequency windows and applies them to the original features and adjacency. Independent shuffling uses separate permutations for frequency features and graphs.

Table 10: Cross-module interventions. Accuracy is in percent; drops are differences from the reported aligned accuracy of 95.55%, in percentage points.

Intervention	Accuracy	Drop
Aligned	95.55	–
Joint time shuffle	94.88	0.67
Frequency-graph mismatch	94.37	1.18
Mask-source mismatch	93.86	1.69
Independent shuffle	94.54	1.01

Mask-source mismatch has the largest accuracy difference from the reported aligned reference. Joint temporal shuffling has a smaller effect, consistent with a model that relies mainly on window content and its weighted aggregation.

D Extended Baseline Comparison

Table 11 reports additional adapted baselines that are not listed in the main comparison. MedSpaformer represents sparse medical time-series transformers; DGCNN, RGNN, and PGCN represent dynamic, regularized, and pyramidal EEG graph networks; REST and EvoBrain model residual or explicitly time-evolving graph states; and GIB, Cluster-GT, and MaxCutPool provide information-bottleneck, cluster-wise, or feature-aware graph reduction.

Table 11: Complete baseline comparison (ten-fold means and standard deviations). Predictive metrics are percentages.

Model	Accuracy	F1	ROC-AUC	PR-AUC
EEG-ConvMamba	90.21 ± 2.54	90.17 ± 2.61	94.16 ± 0.64	94.17 ± 0.69
EEGNet	93.32 ± 4.36	93.00 ± 5.16	98.05 ± 1.23	97.94 ± 1.38
SFE-GAT	77.67 ± 7.08	77.64 ± 6.11	85.55 ± 6.27	84.90 ± 7.28
Medformer	72.71 ± 3.22	71.93 ± 3.04	76.15 ± 5.51	76.27 ± 5.65
MedSpaformer	77.14 ± 5.85	76.86 ± 5.69	80.24 ± 7.94	80.84 ± 8.00
LGGNet	90.27 ± 3.12	90.01 ± 3.29	94.98 ± 2.80	95.19 ± 3.03
DGCNN	89.58 ± 3.18	89.22 ± 3.56	94.77 ± 2.28	94.22 ± 3.23
RGNN	88.75 ± 3.59	88.84 ± 3.33	93.72 ± 3.70	93.25 ± 4.76
PGCN	76.79 ± 3.99	76.77 ± 4.62	79.48 ± 5.71	79.71 ± 5.64
REST	80.04 ± 5.39	80.55 ± 4.54	83.61 ± 4.92	82.98 ± 4.20
NeuroGNN	84.13 ± 5.59	83.92 ± 5.72	90.86 ± 5.14	90.33 ± 4.72
EvoBrain	84.13 ± 5.15	84.31 ± 5.16	89.14 ± 5.60	88.90 ± 5.55
LaBraM	91.80 ± 3.86	91.83 ± 3.82	95.51 ± 3.08	95.87 ± 2.72
CBraMod	92.49 ± 3.53	92.44 ± 3.46	95.69 ± 2.67	96.47 ± 1.92
BrainIB	94.37 ± 2.77	94.23 ± 2.88	97.57 ± 2.32	98.07 ± 1.63
IRENE	90.08 ± 4.99	89.68 ± 5.30	93.26 ± 3.59	93.92 ± 4.11
GIB	88.40 ± 4.23	87.68 ± 4.75	91.86 ± 3.84	92.89 ± 4.13
Cluster-GT	86.86 ± 4.23	86.88 ± 4.34	92.39 ± 4.87	91.61 ± 6.68
MaxCutPool	85.84 ± 5.64	85.33 ± 5.98	88.69 ± 6.83	88.10 ± 7.33
TIF	89.08 ± 4.81	88.65 ± 5.12	93.63 ± 4.49	94.50 ± 3.52
Proposed	95.55 ± 2.46	95.40 ± 2.63	97.26 ± 2.11	97.89 ± 1.48

The extended comparison therefore complements the main baselines with broader coverage of temporal, graph-structural, dynamic, and interpretable modeling strategies.

E Leave-One-Session-Out Generalization

We evaluate leave-one-session-out decoding across BQ sessions B1–B7, training on six sessions and evaluating on the remaining session in each run. Within the six training sessions, 90% of the samples are used for optimization and 10% for internal validation. The checkpoint with the highest validation accuracy is selected, with ROC-AUC used to break ties. The held-out session is kept completely separate and evaluated only at the selected checkpoint. As shown in Table 12, the proposed framework achieves a mean accuracy of $80.13\% \pm 13.81\%$ and a mean ROC-AUC of $82.88\% \pm 13.89\%$ across the seven sessions. Accuracy reaches 99.57% on B1 and 97.39% on B4. B6 yields lower accuracy (63.64%) and F1 (50.00%) with only 22 available samples.

Table 12: Leave-one-session-out analysis on BQ sessions B1–B7. All metrics are percentages; values are reported as mean $\pm$ standard deviation.

Held-out	Acc.	Prec.	Rec.	F1	ROC-AUC	PR-AUC
B1	99.57 ± 0.26	99.38 ± 0.35	98.63 ± 0.14	99.26 ± 0.29	98.39 ± 0.32	98.82 ± 0.22
B2	74.18 ± 5.48	60.00 ± 7.37	83.08 ± 5.84	69.68 ± 6.55	79.23 ± 5.24	67.84 ± 4.12
B3	81.96 ± 4.12	75.38 ± 5.12	72.06 ± 4.89	73.68 ± 5.23	87.44 ± 3.12	82.91 ± 4.23
B4	97.39 ± 0.18	95.59 ± 0.21	98.93 ± 0.00	97.74 ± 0.15	99.46 ± 0.03	99.52 ± 0.02
B5	68.33 ± 4.32	66.67 ± 3.84	73.33 ± 4.05	69.84 ± 3.65	72.89 ± 3.63	73.11 ± 2.98
B6	63.64 ± 9.82	80.00 ± 7.49	36.36 ± 8.58	50.00 ± 9.04	60.33 ± 8.37	70.44 ± 8.73
B7	75.86 ± 4.17	82.61 ± 3.16	65.52 ± 5.43	73.08 ± 4.89	82.40 ± 3.21	84.83 ± 2.91
Mean $\pm$ SD	80.13 ± 13.81	79.95 ± 14.29	75.42 ± 21.56	76.18 ± 17.23	82.88 ± 13.89	82.50 ± 12.98

F Training Configuration

Experiments were run on Ubuntu 22.04.3 LTS with an Intel Xeon Gold 6230 CPU using 8 cores, 64 GB of RAM, and an NVIDIA GeForce RTX 4090 GPU. The software environment used Python 3.12.3, PyTorch 2.8.0, and CUDA 12.8.

We use Adam with learning rate 10^{-4}, weight decay 10^{-5}, batch size 32, and 160 epochs. The split seed is 42. GAT feature dropout, attention dropout, and classifier dropout are 0.1, 0.2, and 0.3. The classifier contains two 128-dimensional hidden layers with ELU. The selector hidden dimension is 64, the edge rank is 32, and both Binary-Concrete temperatures are 1.

Auxiliary objectives are disabled for the first five epochs and linearly increased to their target weights over the next five.

G Complete Frequency and Spatial Updates

G.1 Frequency Grid and Soft Selection

The short-window transform is

$$\widehat{X}_t = \mathrm{rFFT}(X_t; n = 2W, \mathrm{norm} = \mathrm{ortho}), \qquad \nu_f = \frac{f f_s}{2W}, \quad f = 0, \ldots, W. \tag{20}$$

We retain coordinates up to 128 Hz. Whole-segment spectra are obtained by reconstructing the signal through overlap averaging, applying a $2L$-point orthonormal FFT, and separately interpolating the real and imaginary parts to the same grid. No Hann window is used.

For a complex spectrum V, define a soft selection operator along the frequency dimension:

$$\mathcal{M}_q(V) = \sigma\left(\frac{\log(1+|V|) - \mathrm{Quantile}_q[\log(1+|V|)]}{\eta \max\{\mathrm{Std}_f[\log(1+|V|)], \epsilon\}}\right). \tag{21}$$

The standard deviation uses the population convention, and $\eta = 0.1$.

G.2 Global-Context Branch

Let $\widehat{S} \in \mathbb{C}^{N \times F}$ denote the aligned whole-segment spectrum. A learned embedding gives $E^s = a^s \odot \widehat{S} + b^s$, where a^s is real-valued and b^s complex-valued. The branch uses $q_s = 1 - 10/F$ and smooths its mask with a Gaussian kernel G of bandwidth 1 Hz:

$$\overline{M}^s = \frac{G * \mathcal{M}_{q_s}(\widehat{S})}{\max(G * \mathbf{1}, \epsilon)}, \qquad S = (1 - g^s) \odot E^s + g^s \odot (E^s \odot \overline{M}^s). \tag{22}$$

Convolution is over frequency, and $g^s_n = \sigma(\mathrm{Re}(r^s_n))$ is a learned channel gate. The resulting background representation is shared across windows.

G.3 Window-Specific Branch

Each window is embedded as $D^0_t = a^d \odot \widehat{X}_t + b^d$. At layer ℓ, the branch selects frequencies and mixes complex features across nodes at each frequency:

$$M^{d,\ell}_t = \mathcal{M}_{0.9}(D^{\ell-1}_t), \qquad C^\ell_{t,i,f} = \sum_j \widetilde{\Omega}^\ell_{ij} \left(M^{d,\ell}_{t,j,f} D^{\ell-1}_{t,j,f}\right)^*. \tag{23}$$

Here * denotes complex conjugation. The learned complex mixing matrix is transformed by applying ReLU separately to its real and imaginary components, normalizing magnitudes with softmax, and preserving the resulting phases. The layer updates and outer skip connection are

$$D^\ell_t = (1 - g^{d,\ell}) \odot D^{\ell-1}_t + g^{d,\ell} \odot \left(a^\ell \odot D^{\ell-1}_t + C^\ell_t\right), \quad \ell = 1, 2, \tag{24}$$

$$D_t = D^2_t + D^0_t. \tag{25}$$

The scales a^ℓ are real-valued and specific to nodes and frequencies; $g^{d,\ell}_n = \sigma(m^\ell_n)$ is a channel gate. Equation 3 combines these window-specific features with the global-context branch. Dense mixing supplies multichannel spectral context before graph-based selection.

Table 13: Candidate functional-graph density ablation. Only the fraction of strongest absolute Pearson edges retained per sample is changed. Metrics are percentages; differences are in percentage points.

Retained	Accuracy	F1	ROC-AUC	Δ Accuracy
10%	$\mathbf{95.55 \pm 2.46}$	$\mathbf{95.40 \pm 2.63}$	$\mathbf{97.26 \pm 2.11}$	–
20%	94.70 ± 3.09	94.53 ± 3.22	97.20 ± 2.39	-0.85
30%	94.02 ± 2.90	93.81 ± 3.11	97.23 ± 2.00	-1.53
40%	93.85 ± 2.79	93.71 ± 2.95	97.10 ± 2.18	-1.71
50%	92.65 ± 2.32	92.55 ± 2.50	97.03 ± 1.84	-2.90
60%	92.82 ± 2.64	92.62 ± 2.83	96.84 ± 2.04	-2.73
70%	91.79 ± 3.51	91.75 ± 3.67	96.79 ± 1.99	-3.76
80%	91.80 ± 3.66	91.71 ± 3.75	96.35 ± 2.93	-3.76
90%	88.74 ± 5.17	87.86 ± 5.80	93.37 ± 4.91	-6.81

G.4 Residual Graph Layers and Node Readout

Let O_t^ℓ concatenate the three attention heads at layer ℓ. Starting from $H_t^0 = M_t^v \odot H_t^f$, the two residual updates are

$$H_t^\ell = M_t^v \odot \mathrm{ReLU}\left(P^\ell H_t^{\ell-1} + \mathrm{LayerNorm}(O_t^\ell)\right), \qquad \ell = 1, 2. \tag{26}$$

The first residual projection maps 52 to 96 dimensions; the second is the identity. Feature dropout precedes attention computation. Final node embeddings $H_t^g = H_t^2$ receive scores $c_{t,n} = \mathrm{MLP}_{\mathrm{read}}(H_{t,n}^g)$. With $c_t^{\max} = \max_n c_{t,n}$,

$$a_{t,n}^v = \frac{\exp(c_{t,n} - c_t^{\max}) M_{t,n}^v}{\max\{\sum_j \exp(c_{t,j} - c_t^{\max}) M_{t,j}^v, \epsilon\}}, \qquad g_t = \sum_n a_{t,n}^v H_{t,n}^g. \tag{27}$$

Node masks therefore act at the input, after each residual layer, and in readout. Each window yields a 96-dimensional graph embedding for temporal aggregation.

H Additional Component and Hyperparameter Results

H.1 Candidate Functional-Graph Density

For each sample, we pool absolute Pearson correlations over all nine windows and retain the requested fraction of strongest entries before applying the learned selector. Only this retained fraction changes; model dimensions, objectives, folds, and training schedule remain fixed. Table 13 reports the complete ten-fold sweep. The 10% candidate graph gives the highest accuracy, with a general decline as weaker connections are admitted.

H.2 Controlled Frequency-Encoder Comparison

Physical frequency calibration substantially improves the FilterTS frontend under the current spatial–temporal backend (Table 14). The reference frontend follows the original FilterTS implementation, extended with a time-window dimension to process our windowed inputs. We replace it with our physically calibrated frontend while retaining the same backend. On 1-s inputs, accuracy increases from 71.83% to 95.55%, with corresponding improvements in F1, ROC-AUC, and PR-AUC. These results demonstrate the benefit of adapting the frequency representation to short-window ECoG through a common physical frequency grid and frequency cutoff in Hz, beyond extending the original architecture to windowed inputs.

H.3 Complete Component and Loss Ablations

Tables 15 and 16 supplement the main-text ablations with all recorded metrics. Removing the frequency module bypasses the encoder; removing the spatial module replaces graph propagation and readout with a node-mask-weighted mean. Removing learned temporal aggregation replaces temporal attention with a temporal mean and dimension-alignment layer. Removing the selector also disables its mask-dependent auxiliary losses.

Table 14: Original and physically calibrated FilterTS frontends with the current spatial–temporal backend on 1-s inputs. The original frontend is extended with a time-window dimension. Values are fold means and standard deviations; gains are absolute differences. Metrics are percentages; differences are in percentage points.

Frequency frontend	Accuracy	F1	ROC-AUC	PR-AUC
Original FilterTS	71.83 ± 4.88	70.35 ± 7.32	74.03 ± 6.91	73.17 ± 7.41
Physically calibrated frontend	$\mathbf{95.55 \pm 2.46}$	$\mathbf{95.40 \pm 2.63}$	$\mathbf{97.26 \pm 2.11}$	$\mathbf{97.89 \pm 1.48}$
Gain	$+23.73$	$+25.05$	$+23.23$	$+24.73$

Table 15: Complete component-ablation results. Metrics are percentages; differences are in percentage points.

Configuration	Accuracy	F1	ROC-AUC	Δ Acc.
Full model	95.55 ± 2.46	95.40 ± 2.63	97.26 ± 2.11	–
w/o explainer	92.82 ± 3.96	92.46 ± 4.34	96.39 ± 3.15	-2.73
w/o frequency encoder	83.12 ± 3.31	83.00 ± 3.27	86.99 ± 4.52	-12.44
w/o spatial propagation	92.50 ± 4.38	92.61 ± 4.20	95.77 ± 4.90	-3.06
w/o temporal attention	94.87 ± 2.67	94.66 ± 2.83	98.23 ± 1.29	-0.69

H.4 Brain-Region Ablation

The A1-only and PFC-only variants retain the corresponding 64 channels and crop the candidate adjacency accordingly. Model widths, objectives, and training settings are kept fixed. Both single-region models achieve approximately 91% accuracy, while the joint model performs better in every fold.

H.5 Auxiliary-Weight Search

We vary α and γ together from 0 to 1 in increments of 0.1. The highest accuracy occurs at 0.20, while the highest ROC-AUC occurs at 0.10. Larger auxiliary weights generally reduce predictive performance.

Table 16: Complete loss-ablation metrics and mean mask weights. Predictive metrics are percentages; mask means use the original scale.

Objective	Accuracy	F1	ROC-AUC	Node mean	Edge mean
CE only	93.84 ± 3.10	93.53 ± 3.50	96.65 ± 2.69	0.9527	0.5003
CE + HSIC	94.35 ± 3.19	94.03 ± 3.61	97.14 ± 2.78	0.7342	0.4973
CE + low-cut	94.53 ± 3.13	94.42 ± 3.16	97.27 ± 2.08	0.4423	0.0960
CE + HSIC + low-cut	95.55 ± 2.46	95.40 ± 2.63	97.26 ± 2.11	0.4257	0.0932

Table 17: Brain-region ablation with identical model widths and training settings. Metrics are percentages; differences are in percentage points.

Region	Accuracy	F1	ROC-AUC	Δ Acc.
A1 + PFC	95.55 ± 2.46	95.40 ± 2.63	97.26 ± 2.11	–
A1 only	91.12 ± 2.41	90.90 ± 2.60	95.13 ± 2.90	-4.43
PFC only	90.95 ± 3.24	90.42 ± 3.51	96.19 ± 2.48	-4.60

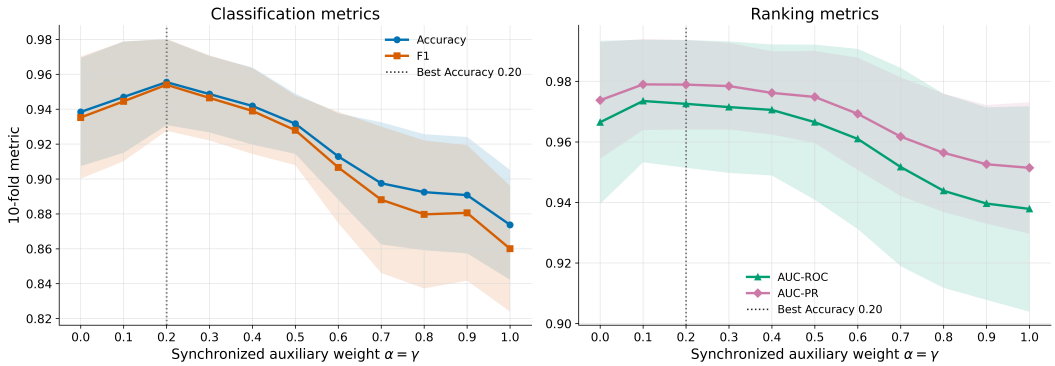

Figure 12: Joint auxiliary-weight sweep. Equal HSIC and low-cut weights are varied over the same grid, and predictive metrics are compared across the resulting models.