Spontaneous Yet Predictable: Shapelet-Driven, Channel-Aware Intention Decoding from Multi-Region ECoG

Keren Cao[1,2*], **Yuhang Tian**[1*], **Kaizhong Zheng**[1,2], **Wei Xi**[1], **Xinjian Li**[3], **Liangjun Chen**[1,2†]

[1]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, National Engineering Research Center for Visual Information and Applications, Xi'an Jiaotong University

[2]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University

[3]Department of Neurology of the Second Affiliated Hospital and Interdisciplinary Institute of Neuroscience and Technology, School of Medicine, Zhejiang University

{caokr, tianyeoh, kzzheng, xiwei, liangjunchen}@stu.xjtu.edu.cn, lxjbio@zju.edu.cn

Abstract

Proactive intention decoding remains a critical yet underexplored challenge in brain–machine interfaces (BMIs), especially under naturalistic, self-initiated behavior. Existing systems rely on reactive decoding of motor cortex signals, resulting in substantial latency. To address this, we leverage the common marmoset's spontaneous vocalizations and develop a high-resolution, dual-region ECoG recording paradigm targeting the prefrontal and auditory cortices and a neural decoding framework that integrates shapelet-based temporal encoding, position-aware attention, frequency-aware channel masking, contrastive clustering and a minimum error entropy-based robust loss. Our approach achieves 91.9% accuracy up to 200 ms before vocal onset—substantially outperforming 13 competitive baselines. Our model also uncovers a functional decoupling between auditory and prefrontal regions. Furthermore, joint modeling in time and frequency domains reveals novel preparatory neural signatures preceding volitional vocal output. Together, our findings bridge the gap between foundational neuroscience and applied BMI engineering, and establish a generalizable framework for intention decoding from ecologically valid, asynchronous behaviors.

Code — https://github.com/kkkiland/Marmoset-ECoG

Introduction

In recent years, brain–machine interfaces (BMIs) have advanced significantly in enabling neural control of external devices, with applications in assistive robotics, neuroprosthetics, and communication aids (Tonin et al. 2022; Rupp et al. 2015; Meng et al. 2016). However, most existing systems adopt a *reactive* paradigm, decoding intentions from sensorimotor signals after the intention has already been formed or partially executed. As a result, system responses often lag behind user behavior, undermining the seamlessness, reliability, and safety required for real-world deployment. This limitation stems from the reliance on movement-related signals originating in the sensorimotor cortex—a downstream cortical region at the terminus of the perception–decision–action loop—which are inherently constrained by delays in neural processing and motor execution. For example, electroencephalography (EEG)- and electrocorticography (ECoG)-based BMIs for finger movement decoding typically exhibit latencies of 1.0–1.2 s (Ding et al. 2025), well beyond the threshold for real-time interaction (LaRocco and Paeng 2020; Skomrock et al. 2018). To overcome this bottleneck, a shift toward *proactive* intention prediction is essential—forecasting user intentions as early as possible before action onset. Achieving this requires moving beyond traditional motor cortices and incorporating higher-order cognitive and sensory regions, such as the prefrontal, auditory, and visual cortices, which are involved in early-stage perception, volition, and decision-making. Neural signals from these areas may encode predictive markers that precede overt motor preparation. For instance, Tsunada and Eliades simultaneously recorded multi-channel activity from the frontal and auditory cortices of common marmosets (Tsunada and Eliades 2025), demonstrating that both regions exhibit significant activation approximately 200–300 ms prior to spontaneous vocalizations. Notably, this pre-vocalization activity was predictive of subsequent acoustic features—such as spectral dynamics—thereby offering an expanded temporal window for predicting internal intentions.

Despite its critical importance, intention prediction remains a critically underexplored frontier in BMI research. The majority of existing work has centered on classifying overt behaviors under structured, trial-based paradigms (Ding et al. 2025; Nagashima et al. 2025), with only sporadic efforts have addressed early or pre-movement decoding in naturalistic settings—efforts that primarily reside within the domain of basic neuroscience rather than applied BMI engineering (Liu et al. 2022). While deep learning has considerably advanced neural decoding capabilities (Wang et al. 2024; Zheng et al. 2024), major obstacles persist. These challenges originate from limitations in neural recording, structural constraints in experimental design, and fundamental barriers in decoding algorithms and their interpretability:

Precise and real-time intention decoding. Spontaneous behaviors such as vocalizations are inherently infrequent, lack clear external cues, and unfold without structured trial

*These authors contributed equally.

†Corresponding Author

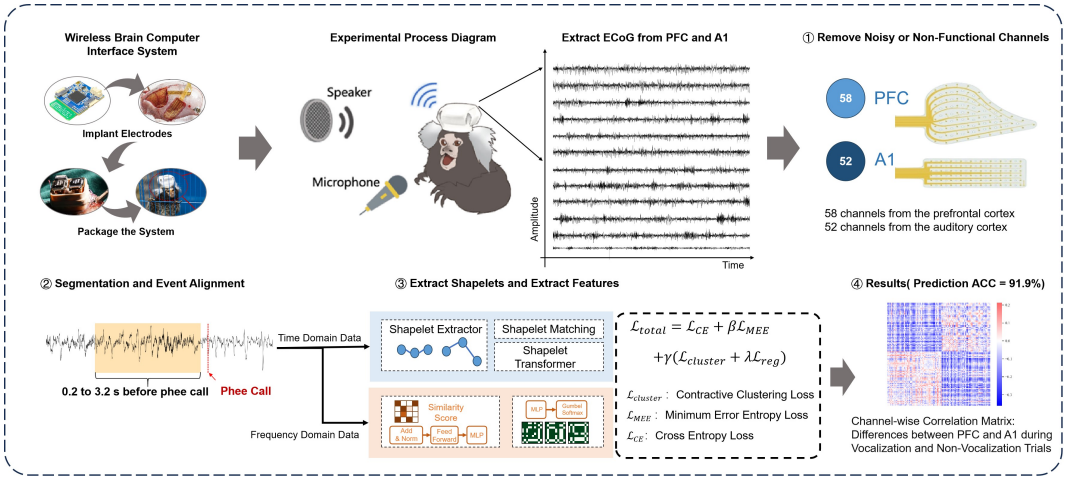

Figure 1: **Overall experimental pipeline.** Wireless BCI systems were implanted in marmosets to record ECoG signals from A1 and PFC under an auditory stimulation paradigm. The process involved bad channel removal, dataset construction, and neural decoding, achieving 91.9% accuracy for vocal onset prediction 200 ms ahead and interpretable vocalization-related patterns.

boundaries—resulting in sparse, noisy neural signals with high temporal variability. These characteristics make anticipatory decoding during the pre-movement phase extremely challenging. In particular, conventional EEG-based methods often perform at near-chance levels (40–50%) (Blankertz et al. 2003), due to limited spatial resolution and low signal-to-noise ratios. This limitation severely constrains the development of real-time BMIs capable of proactive control in ecologically valid scenarios, such as free movement or natural communication.

Multi-region synergy mechanisms investigation. Although localized neural activity—especially in the sensorimotor cortex—has long been associated with volitional control, growing evidence suggests that intention emerges from distributed and synergistic interactions across large-scale networks, particularly involving the prefrontal and parietal cortices (Gordon-Fennell et al. 2023). However, most current decoding frameworks rely on region-specific features and trial-structured paradigms, failing to model critical cross-area dynamics. Without capturing these inter-regional dependencies, it is difficult to improve decoding performance or gain mechanistic insights into how intentions are generated and propagated throughout the brain. Addressing this gap is crucial for advancing both computational BMI design and systems-level neuroscience.

Neural patterns for specific intention extraction. Isolating neural representations that correspond to specific behavioral intentions is foundational for robust anticipatory decoding. Yet, current models struggle to capture the rapid, nonlinear evolution of neural dynamics underlying intent formation. This challenge is compounded by limitations in signal resolution, as well as the lack of frameworks capable of modeling complex, multi-scale dependencies. As a result, many intention-relevant patterns remain undercharacterized, limiting both predictive performance and biological interpretability. Developing more expressive and interpretable models is therefore essential to unlock actionable intent signatures from neural data.

To address these challenges, we leverage the marmoset—a New World primate whose rich and spontaneous vocal behavior offers an ecologically valid model for studying volitional communication. Unlike reflexive calls, marmoset vocalizations are self-initiated, context-dependent, and goal-directed (Li, Aoi, and Miller 2024), providing an ideal substrate for intention decoding research. Crucially, recent studies have shown that these vocalizations are consistently preceded by preparatory neural activity in both frontal and auditory cortices (Tsunada and Eliades 2024; Grijseels et al. 2023), offering a reliable temporal window for proactive neural decoding. Moreover, we employ a ECoG-based neural acquisition system implanted in the marmoset brain, which enables high-resolution, stable, and high-SNR recordings of cortical activity. This setup provides the high-fidelity neural data essential for decoding fine-grained and distributed dynamics underlying volitional behavior. (Schwarz et al. 2014).

Building on this, we developed a vocalization-based experimental paradigm and a multi-region available neural decoding framework for predicting spontaneous vocalizations from high-density ECoG recordings in marmosets. The experimental paradigm captures naturally occurring, asynchronous vocal behaviors without explicit task constraints, simultaneously recording neural activity from the auditory (A1) and prefrontal cortices (PFC). To extract informative neural signatures predictive of vocal onset, our framework combined shapelet-based temporal feature extraction with a position-aware attention mechanism to encode temporal dependencies. Concurrently, frequency-aware channel masking and contrastive clustering were implemented, selectively aggregating inter-channel relationships within distinct spectral bands. To enhance robustness against acquisition and neural noises, a minimum error entropy (MEE) loss derived from Rényi entropy was adopted, effectively reducing prediction uncertainty during model training (Li et al. 2021b). This framework addresses a critical gap in BMI by enabling the extraction of intention-relevant neural features from unstructured, self-initiated behaviors—an underexplored yet essential capability

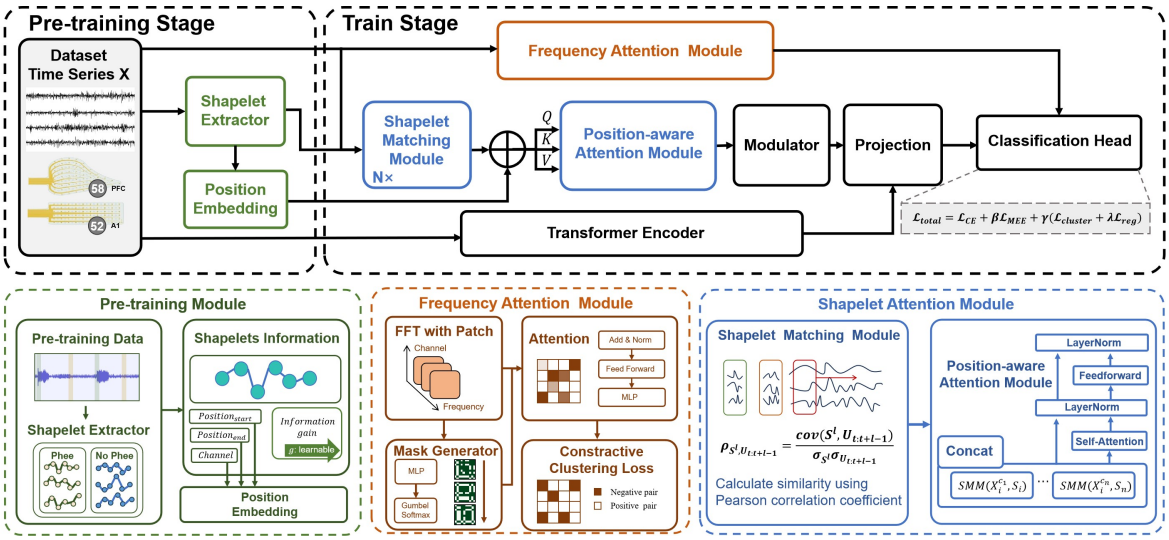

Figure 2: **Overall architecture of the decoding framework for marmoset vocalization intentions.** *In the pre-training stage*, class-relevant shapelets are extracted as informative temporal subsequences. *In the training stage*, transformer-based modules incorporate shapelet information to refine temporal representations, while spectral features are learned through channel masking attention and contrastive clustering, jointly capturing temporal and frequency-domain patterns.

for advancing naturalistic neural decoding.

To summarize, our main contributions are threefold:

1. **Robust and interpretable neural decoding framework:** We formulate early vocal onset prediction as a multivariate time-series classification problem and introduce a decoding architecture, which enables the extraction of intention-relevant spatiotemporal and spectral features. Our model achieves 91.9% accuracy up to 200 ms before vocal onset, outperforming 13 state-of-the-art baselines.
2. **Neuroscientific insight into cross-regional dynamics:** We show that multi-region neural signals substantially enhance predictive performance and uncover a functional decoupling between the prefrontal and auditory cortices during spontaneous vocal behavior. These findings provide empirical evidence for top-down modulation mechanisms and offer new avenues for intention-aware BMI.
3. **Joint analysis of time- and frequency-domain patterns:** By first localizing a preparatory window via salient temporal shapelets and then identifying statistically significant frequency-specific activation patterns, our method enables fine-grained characterization of volitional dynamics. This dual-domain approach reveals previously inaccessible neural signatures preceding self-initiated vocalizations, providing a principled strategy for early-stage intention decoding in naturalistic settings.

Related Work

Experimental Paradigms for Neural Decoding ECoG offers markedly superior spatial and temporal resolution, broader frequency bandwidth, and substantially higher SNR compared to EEG, making it well suited for precise neural decoding in both primate and non-primate models (Yan et al. 2023). Non-human primates such as marmosets provide a compelling platform for investigating brain-wide dynamics due to their smaller brains, distinct cortical architectures, and more constrained behavioral repertoires.

Although studies in primates have extended to decoding complex volitional behaviors including three-dimensional movements (Hu et al. 2018) and reward-driven decision signals (Zabeh et al. 2023), decoding research in non-human primates has largely focused on externally driven or conditioned behaviors, such as simple motor tasks and reflex responses (Jiricek et al. 2021; Cho et al. 2023). Recently, increasing attention has shifted to spontaneous, internally generated behaviors (Liu et al. 2022), which offer richer opportunities to understand volitional control. However, decoding ECoG signals remains challenging due to signal sparsity, timing variability, and the need for models capable of integrating distributed brain dynamics.

Deep Learning for Neural Time Series Decoding Recent advances in deep learning have provided powerful tools for neural decoding, particularly in modeling long-range temporal dependencies using transformer-based architectures (Zhou et al. 2022). Shapelet-based models have been explored to extract discriminative subsequences reflecting local task-relevant dynamics (Li et al. 2021a; Le et al. 2024). In parallel, spectral decomposition and multi-band filtering have enabled the extraction of multi-scale neural features (Huang et al. 2025; Yi et al. 2024), while clustering and contrastive learning techniques have been used to discover structured latent patterns and improve robustness (Wu et al. 2024).

Despite these advances, most approaches focus on modeling either temporal or spectral structure in isolation and are often limited to single-region analysis, which constrains their ability to decode complex behaviors that rely on distributed neural processes. Furthermore, their outputs are frequently implicit or continuous-valued, lacking categorical interpretability crucial for linking neural activity to discrete behavioral intentions.

Method

Experimental Paradigm and Problem Definition

To investigate the neural basis of spontaneous vocal planning, we implemented an auditory stimulation paradigm in awake marmosets implanted with ECoG electrode (128 electrodes, 1.0–1.52 mm spacing) covering A1 (64 channels) and PFC (64 channels), which were seated in a primate chair and passively exposed to pure tones at four frequencies (2,000, 4,000, 8,000, and 16,000 Hz; 200 ms duration; 5 s inter-stimulus interval). Marmosets occasionally produced spontaneous phee calls in response to tones. Simultaneous ECoG recordings were acquired from a 128-channel array spanning the A1 and PFC, allowing us to examine distributed cortical dynamics associated with volitional vocal behavior.

Marmosets occasionally produced phee calls—a stereotyped long-distance vocalization typically emitted in social isolation (Zhao and Wang 2023). Pure-tone stimuli did not acoustically trigger these calls but facilitated their spontaneous occurrence, which was otherwise rare. The calls were not time-locked to stimulus onset and showed variable latencies, suggesting endogenous rather than stimulus-driven initiation. This asynchronous paradigm, where vocal onset is decoupled from stimulus timing, is essential for non-human primate studies and allowed ECoG recordings to capture cortical dynamics underlying internally driven vocal behavior.

To transform the task of predicting spontaneous vocalizations into a multivariate time series classification problem, we apply a sliding window approach to the continuous ECoG recordings. Specifically, we segment the multivariate time series using a fixed-length window of size T and a stride s. Given a full recording $X_{\text{full}} \in \mathbb{R}^{C \times T_{\text{full}}}$, where T_{full} denotes the total duration of the recording and C is the number of channels, we extract a series of overlapping windows:

$$X_i = X_{\text{full}}[:, i \cdot s : i \cdot s + T], i = 0, 1, \ldots, \left\lfloor \frac{T_{\text{full}} - T}{s} \right\rfloor, \quad (1)$$

where each segment $X_i \in \mathbb{R}^{C \times T}$ denotes an ECoG window, paired with a binary label $Y_i \in \{0, 1\}$ indicating whether it precedes a vocalization event, forming the dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{M}$, where M is the number of samples. The prediction output is defined as $\hat{Y}_i = f(X_i)$, where $f(\cdot)$ is a trainable classifier to detect vocal onset.

Methodological Details

The overall architecture of our decoding framework is shown in Fig 2, which captures both temporal and frequency-domain neural patterns underlying marmoset vocalization intentions. Firstly, pre-training stage employs a shapelet extractor to generate a series of class-relevant temporal subsequences and their weights (information gain scores g) with crucial information. g are then refined through transformer-based encoder that integrates correlation and spatiotemporal context for improved temporal representation. In parallel, spectral features are obtained via channel masking attention mechanism while contrastive clustering loss is introduced to uncover frequency-specific inter-channel patterns.

Pre-training Stage To effectively predict vocalization, it is crucial to discover discriminative temporal patterns and localize when such patterns emerge in the neural signal. This motivates the incorporation of a pre-trained shapelet extractor inspired by (Le et al. 2024), which aims to identify informative subsequences in multivariate time series. Specifically, our method employs Perceptually Important Points (PIPs) (Chung et al. 2001) to extract representative shapelet candidates by preserving the essential structure of the original signal. Candidates are formed from consecutive PIP triplets based on reconstruction distance. To ensure efficiency, the number of PIPs is set to $n_{\text{pip}} = 0.5 \times T$, yielding a manageable number of candidates. Then, candidates are ranked by their optimal information gain score $g \in \mathbb{R}$, obtained by computing the Perceptual Subsequence Distance (PSD) (Le, Tran, and Huynh 2022) and identifying the optimal split point that maximizes inter-class separability. Given a shapelet S^l with length l and X_i, the PSD is defined as:

$$\text{PSD}(X_i, S^l) = \min_{j=1}^{T-l+1} \text{CID}(X[j : j + l - 1], S^l), \quad (2)$$

where CID denotes the Complexity-Invariant Distance (Batista, Wang, and Keogh 2011). The top candidates per class are retained as shapelets for downstream learning.

Formally, a shapelet is defined as a univariate temporal subsequence $S^l = [x_{\text{p}_0}, x_{\text{p}_0+1}, \ldots, x_{\text{p}_l}]$, specified by its start position p_0, end position p_l, channel index c, and score g. It is extracted from a univariate signal $X^c \in \mathbb{R}^T$, which corresponds to channel c of the input X_i. These attributes characterize the spatio-temporal localization and discriminative importance of the shapelet.

Shapelet-Driven Temporal Encoding To capture fine-grained temporal dynamics, we propose a shapelet-based encoding framework that matches input signals to learned shapelets via Pearson's correlation and aggregates the resulting features using a position-aware attention mechanism.

(1) Shapelet Matching Module (SMM): Inspired by the interchannel temporal synchrony observed in ECoG signals, we measure the similarity between shapelets and input signals based on Pearson's correlation to capture discriminative temporal patterns.

Specifically, each shapelet S^l is slid along X^c to identify local temporal motifs that best align with the shapelet. This matching process is formulated as:

$$t^* = \arg\max_t \rho\big(X^c_{t:t+l-1},\, S^l\big),$$
$$\text{SMM}(X^c, S^l) = \text{Linear}\big(X^c_{t^*:t^*+l-1}\big) - \text{Linear}(S^l), \quad (3)$$

where $\rho(\cdot, \cdot)$ denotes the Pearson's correlation between S^l and the subsequence $X^c_{t:t+l-1}$. The index t represents the starting position of a temporal window within X^c, and t^* corresponds to the window that yields the highest correlation with S^l. The most similar subsequence is then projected through a linear layer, and its representation is normalized by subtracting the projection of the shapelet itself.

(2) Position-Aware Attention Module (PAM): To effectively aggregate the temporal positional information learned

from shapelets, we apply a position-aware attention module. For input X_i, the temporal representation Z is defined as:

$$Z = \text{Concat}\left(\text{SMM}(X_i^{c_1}, S_1),\ \text{SMM}(X_i^{c_2}, S_2),\ \ldots,\ \text{SMM}(X_i^{c_n}, S_n)\right) + \text{PE}, \tag{4}$$

where $\text{PE} = \text{Emb}(p_0) + \text{Emb}(p_l) + \text{Emb}(c)$ is the learnable position encoding derived from shapelet attributes. The position-enhanced features Z are passed through a self-attention module by computing query, key, and value projections (Q, K, V). The attention output is computed as: $\text{Attention}(Z) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V$, where d is the dimension of the key vectors, and the attention output is further modulated by score g, which serves as a weighting factor for each matched segment:

$$\text{Att}(Z) = \text{LayerNorm}\left(Z + \text{Attention}(Q, K, V)\right) \times g. \tag{5}$$

Finally, the attended features are passed through a feedforward network followed by a residual connection and layer normalization, yielding the output:

$$\text{Out}(Z) = \text{LayerNorm}\left(\text{Att}(Z) + \text{FeedForward}(\text{Att}(Z))\right). \tag{6}$$

To selectively preserve salient features and suppress noise, max pooling is applied along the temporal dimension. In parallel, the output is concatenated with that of a lightweight Transformer variant designed to capture complementary global temporal dependencies, as shown in Fig 2.

Channel-Aware Spectral Attention To extract frequency domain features relevant to marmoset vocalization intentions, neural signals were segmented into spectral patches, based on the assumption that distinct frequency bands reflect different cortical processes and inter-channel interactions. A channel-aware attention mechanism, combined with a contrastive clustering loss, was then employed to capture informative spectral dependencies and enhance interpretability.

(1) Spectral Channel Attention: Firstly, each frequency-domain patch $F^p \in \mathbb{R}^{C \times L_p}$, where L_p is the patch length, is projected into query Q^p, key K^p, and value V^p. The attention scores between channels are modulated by a learnable binary mask $M^p \in \mathbb{R}^{C \times C}$, and frequency-domain features, obtained via masked attention, are defined as:

$$\text{MaskedAtt}(F^p) = \text{Softmax}\left(M^p \odot \frac{Q^p (K^p)^\top}{\sqrt{d}}\right)V^p, \tag{7}$$

where $M^p = (1 - I_C) \odot \text{GS}(\mathbf{W}F^p) + I_C$, $\mathbf{W}$ is a learnable projection, $\odot$ denotes element-wise multiplication, $\text{GS}(\cdot)$ applies Gumbel-Softmax (Jang, Gu, and Poole 2016) for differentiable binary sampling, and I_C is the identity matrix that preserves self-connections. The output is then integrated with time-domain features for final prediction.

(2) Contrastive Clustering Loss: To promote structured frequency-domain representations, we employ a contrastive clustering loss that encourages channels with similar modulation patterns to be grouped. The loss for each patch is:

$$\mathcal{L}^p_{\text{cluster}} = -\frac{1}{C}\sum_{i=1}^{C}\log\left(\frac{\sum_{j=1}^{C} M^p_{ij}\exp\left(Sim^p_{ij}/\tau\right)}{\sum_{j=1}^{C}\exp\left(Sim^p_{ij}/\tau\right)}\right), \tag{8}$$

where M^p_{ij} indicates whether channel j is a positive neighbor of channel i, Sim^p_{ij} denotes their similarity based on masked attention, and τ is a temperature hyperparameter controlling the sharpness of the distribution. To improve interpretability, we regularize the mask via a regularization loss:

$$\mathcal{L}^p_{\text{reg}} = \frac{1}{C(C-1)}\left\|I_C - M^p\right\|_1. \tag{9}$$

Minimum Error Entropy Loss via Rényi Entropy To enhance robustness against outliers—such as those introduced by signal acquisition artifacts or unrelated neural activity—we adopt a MEE criterion based on Rényi entropy. Unlike conventional losses (e.g., cross-entropy or MSE), which typically assume Gaussian noise, MEE directly minimizes the uncertainty of the error distribution, making it less sensitive to outliers and non-Gaussian noise (Silvestrin, Yu, and Hoogendoorn 2023).

Let $e_i = \hat{y}_i - y_i$ denote the prediction error. The eigenvalues $\{\lambda_i\}_{i=1}^N$ are obtained by solving $K_{v_i} = \lambda_i v_i$, where v_i denotes the corresponding eigenvector. The Gaussian-kernel Gram matrix $K \in \mathbb{R}^{N \times N}$ is defined as $K_{ij} = \exp\left(-\|e_i - e_j\|^2/\sigma\right)$ and normalized by its trace, where the smoothing parameter σ controls the smoothness of the similarity measure between errors. The Rényi entropy of order $\alpha > 0, \alpha \neq 1$ is then approximated as:

$$\mathcal{L}_{\text{MEE}} = \frac{1}{1-\alpha}\log_2\left(\sum_{i=1}^{N}\lambda_i^\alpha\right). \tag{10}$$

The overall training objective is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \beta\mathcal{L}_{\text{MEE}} + \gamma\left(\mathcal{L}_{\text{cluster}} + \lambda\mathcal{L}_{\text{reg}}\right), \tag{11}$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss. β, γ, and λ balance the contributions of each component and are set to 0.2, 0.2, and 0.5, respectively.

Experiments

Experimental Settings

Dataset We established a spontaneous vocalization dataset from a single common marmoset (Callithrix jacchus, ID: BQ) across four sessions, totaling 142 minutes of ECoG and audio data. A total of 291 spontaneous calls were manually annotated. Positive samples were defined as the 3.2–0.2 s preceding vocal onset, with the 0.2 s cutoff chosen to allow sufficient time for downstream responses while ensuring adequate data processing and algorithmic inference. Negative samples were drawn from non-vocal periods at least 10 s away from any call. In total, we obtained 430 samples with a 1:1 ratio of positive to negative examples. Due to issues with channel contact, 110 valid ECoG channels (52 A1, 58 PFC) with 500 Hz sampling were retained and band-pass filtered from 1–200 Hz. All samples were aligned using synchronized neural and dual-channel audio recordings.

Models	Accuracy	Precision	F1 score	AUROC	AUPRC
SVM	75.1±2.8	70.6±6.7	75.4±1.9	81.8±2.5	76.9±3.8
Minirocket	75.2±1.2	86.8±0.8	70.5±1.0	74.9±1.1	71.9±0.9
EEGNet	72.8±2.0	73.6±1.8	72.0±1.9	69.3±1.5	48.2±2.1
TCAN	80.7±1.1	75.1±1.3	77.8±1.4	88.1±0.9	84.2±1.2
ST-CCNet	65.0±2.5	81.9±2.2	54.7±3.0	65.6±1.8	63.8±2.1
SparseDGCNN	66.0±5.7	66.1±5.9	65.8±5.8	71.1±3.1	63.7±4.4
Transformer	70.6±1.3	70.8±1.1	70.5±1.0	69.3±0.9	69.0±1.0
Crossformer	84.2±0.7	86.4±0.6	83.5±0.7	87.8±0.5	88.9±0.6
Shapeformer	85.2±1.0	85.0±1.1	85.1±1.0	88.3±0.8	81.5±1.2
Autoformer	77.4±1.5	77.4±1.4	77.3±1.3	89.7±1.0	89.5±1.1
Informer	66.0±2.0	65.7±2.2	65.6±2.1	62.9±1.8	61.7±2.0
iTransformer	79.6±0.9	80.8±1.0	79.6±0.8	87.6±0.7	88.9±0.9
Medformer	68.3±1.8	69.1±1.9	66.6±2.0	79.2±1.6	79.8±1.5
Ours	**91.9**±0.6	**92.1**±0.5	**91.9**±0.6	**96.0**±0.9	**93.1**±2.6

Table 1: Performance comparison between the proposed model and competing methods in %. **Bold**: best.

Metric	Backbone	+Shapelet	+FAM		+MEE
			CD	+MG	(**Ours**)
Accuracy	79.1±1.5	86.0±1.0	89.5±0.8	90.7±0.7	**91.9**±0.6
Precision	82.2±1.6	87.2±0.9	89.5±0.6	90.7±0.7	**92.1**±0.5
F1 score	78.2±1.4	85.8±0.9	89.5±0.8	90.7±0.7	**91.9**±0.6
AUROC	86.5±1.2	89.0±0.9	90.8±0.8	92.8±0.8	**96.0**±0.9
AUPRC	84.7±1.4	88.5±1.0	88.7±1.1	90.3±1.4	**93.1**±2.6

Table 2: Ablation study with stepwise addition of modules on the test set in %. **Bold**: best.

Implementation We evaluate all models using 5 standard metrics: accuracy, precision, F1 score, area under the receiver operating characteristic curve (AUROC), and area under the precision–recall curve (AUPRC). Model training is performed using a fixed data split of 70%, 10%, and 20% for training, validation, and testing, respectively. To initialize shapelet extraction in the pretraining stage, 100 samples are randomly selected from the training set. All experiments are conducted on a single NVIDIA RTX 4090 GPU. Optimization is performed using the Adam optimizer with a learning rate of 5×10^{-5} and a weight decay of 5×10^{-4}. The batch size is set to 16, and training proceeds for 100 epochs. All reported results represent the average performance over five runs with different random seeds on the test set.

Baselines We benchmarked our model against 13 baselines, including a classical machine learning method (feature extracted within frequency domain), a feature-based method, an EEG-specific model, CNN-based models, a GNN-based model, and Transformer-based architectures: SVM, MiniRocket (Dempster, Schmidt, and Webb 2021), EEGNet (Lawhern et al. 2018), TCAN (Hao et al. 2020), ST-CCNet (Zhang et al. 2021b), SparseDGCNN (Zhang et al. 2021a), Transformer (Vaswani et al. 2017), Crossformer (Zhang and Yan 2023), Shapeformer (Le et al. 2024), Autoformer (Wu et al. 2021), Informer (Zhou et al. 2021), iTransformer (Liu et al. 2023), and Medformer (Wang et al. 2024).

Quantitative Results

Comparison Study Across all evaluation metrics, our method consistently outperforms existing approaches, achieving the highest AUROC of **96.0%**. Competitive baselines such as Shapeformer (85.2% accuracy), which captures task-relevant temporal subsequences, and Crossformer (84.2% accuracy), which models cross-dimensional dependencies, also perform relatively well. Nevertheless, our framework surpasses them with a substantial improvement, reaching **91.9%** accuracy—representing a **6.7%** gain over the best-performing baseline, indicating its superior prediction ability.

Multi-brain Region Analysis To evaluate the contribution of distinct cortical areas to vocal intention decoding, we compare 3 input configurations: dual-region input (A1 and PFC), PFC-only, and A1-only. As shown in Fig 3a, the PFC-only model performs better than the A1-only model (AUROC 0.930 vs. 0.899), indicating that prefrontal activity carries more intention-relevant information. Notably, the dual-region model consistently outperforms the single-region variants across all evaluation metrics, suggesting that sensory and executive cortices provide complementary information for spontaneous vocal planning.

Ablation Study We conduct a step-by-step ablation study to assess the contribution of each component by incrementally integrating them into the Transformer-based backbone (Table 2). Adding the shapelet-driven module, which significantly improves performance, demonstrating its effectiveness in capturing fine-grained, temporally discriminative local patterns. Next, we incorporate the Frequency Attention Module (FAM) with a contrastive loss to extract informative and distinctive spectral representations by modeling inter-channel interactions, and further compare the proposed mask generator (MG) with the traditional channel dependency (CD) modeling. The results demonstrate the superior performance for above components. Finally, adding the MEE loss yields the best overall results by enhancing the model's robustness to non-Gaussian noise and signal artifacts.

Interpretable Analysis

Temporal Analysis To assess the interpretability of learned temporal representations, we visualized the shapelets extracted during pretraining stage (Fig.3b). These shapelets were predominantly localized to the PFC and clustered within the 0.8–0.4 s window preceding vocal onset, highlighting a critical preparatory period. Time–frequency analysis during this interval revealed significant modulation in the α, θ, and γ bands (Fig.4), with α and θ changes most pronounced in PFC channels—consistent with the spatial distribution of shapelets (FDR-corrected). Prior studies have reported suppressed α and θ activity in the PFC during vocal preparation (Tsunada and Eliades 2024), lending biological support to our findings. These results indicate that the model captures task-relevant spatiotemporal dynamics in a neurobiologically meaningful manner.

Functional Decoupling Between Frontal and Auditory Cortices To further investigate inter-regional and intra-regional coordination patterns associated with marmoset vo-

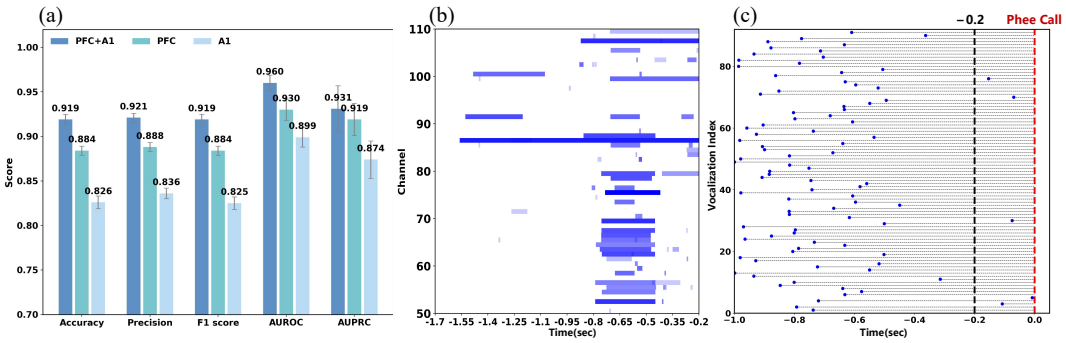

Figure 3: (a) Performance comparison: dual-region (A1+PFC), PFC-only, and A1-only. (b) Visualization of extracted shapelets and their temporal distribution relative to vocal onset in PFC channels (53–110). (c) Pseudo-online vocalization prediction.

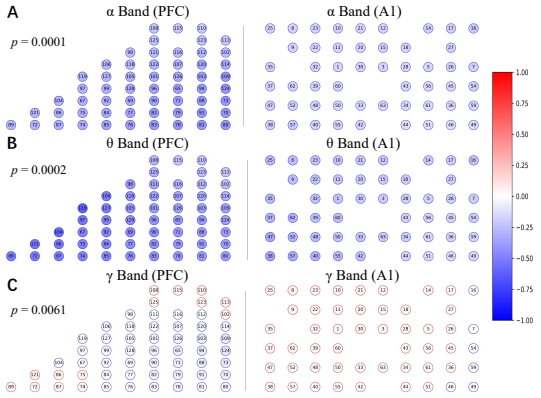

Figure 4: Increases and decreases in power across three frequency bands during 0.8–0.4 s before vocal onset.

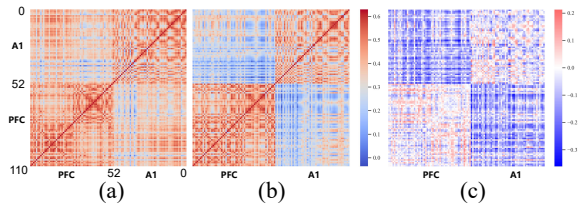

Figure 5: Channel-wise correlation matrix: (a) non-vocalization trials; (b) vocalization trials; (c) Difference between vocalization and non-vocalization trials.

cal behavior, we analyzed the channel-wise correlation structure based on the frequency representations derived from the FAM (Fig. 5). For each sample, we computed a correlation matrix across channels, then separately averaged these matrices across positive (pre-vocalization) and negative (non-vocalization) samples. Subtracting the negative from the positive sample average yielded a differential correlation map. Notably, we observed an **increased intra-regional correlation** within both PFC and A1, alongside a **decrease in inter-regional correlation** between these two areas during vocalization trials. This pattern suggests that during the pre-vocal phase, local coordination within each region is enhanced, possibly reflecting region-specific processing such as motor planning in the PFC and sensory modulation or prediction in A1. Conversely, the reduced cross-area correlation may reflect a functional decoupling between PFC and A1, consistent with top-down suppressive signaling from PFC to A1 observed in previous neurophysiological studies (Tsunada and Eliades 2024).

Pseudo-online Prediction on Unseen Session To assess the applicability of our model in real-time scenarios, we conduct a pseudo-online prediction analysis using a held-out session containing 95 spontaneous vocal onsets. We simulate a streaming setting by continuously feeding the model with 3 s windows, shifted at 500 ms intervals. This setup generates a time-resolved vocal prediction, mimicking the temporal dynamics of real-time decoding. Our model correctly detected 91 out of 95 vocalizations and achieves an average inference time of 7.41 ms per sample, a throughput of 134.95 samples/s, and a model size of 14.58 MB, demonstrating its efficiency for real-time deployment. As shown in Fig. 3c, given the presence of temporally clustered vocalizations in certain periods, only the 1 s pre-onset segments were visualized for each vocal event. Notably, the average lead time for accurate detections was 710.4 ms, demonstrating the capacity of our model to anticipate vocal events. These results highlight the high temporal precision of the model in intention prediction and its potential for real-time decoding applications.

Conclusion

In this work, we propose a robust and interpretable neural decoding framework for predicting spontaneous vocal intentions from high-density dual-region ECoG recordings in marmosets. By integrating shapelet-based temporal encoding, spectral clustering, and entropy-regularized learning, our model achieves 91.9% accuracy up to 200 ms before vocalization—outperforming 13 baseline methods. Beyond strong predictive accuracy, the model reveals interpretable time–frequency neural patterns and functional decoupling between auditory and prefrontal cortices, offering insights into the distributed cortical dynamics underlying volitional behavior. Despite these promising results, the current study was conducted on a single marmoset, limiting the generalizability of the findings across individuals and behavioral variability. These efforts lay a critical foundation for the development of intention-aware brain–machine interfaces and truly seamless human–machine interaction and also advance our understanding of how goal-directed behavior is encoded across distributed cortical networks.

Acknowledgments

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 (NSFC, Grant Nos. 62088102, 62436005, and 62473303).

Ethics Statement

All procedures were approved by institutional animal ethics committees.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Batista, G. E.; Wang, X.; and Keogh, E. J. 2011. A complexity-invariant distance measure for time series. In *Proceedings of the 2011 SIAM international conference on data mining*, 699–710. SIAM.

Blankertz, B.; Dornhege, G.; Schäfer, C.; Curio, G.; and Müller, K.-R. 2003. Boosting bit rates and error detection for the classification of fast-paced motor commands based on single-trial EEG analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2): 127–131.

Cho, Y. K.; Koh, C. S.; Lee, Y.; Park, M.; Kim, T. J.; Jung, H. H.; Chang, J. W.; and Jun, S. B. 2023. Somatosensory ECoG-based brain–machine interface with electrical stimulation on medial forebrain bundle. *Biomedical Engineering Letters*, 13(1): 85–95.

Chung, F. L. K.; Fu, T.-C.; Luk, W. P. R.; and Ng, V. T. Y. 2001. Flexible time series pattern matching based on perceptually important points. In *Workshop on learning from temporal and spatial data in international joint conference on artificial intelligence*.

Dempster, A.; Schmidt, D. F.; and Webb, G. I. 2021. Minirocket: A very fast (almost) deterministic transform for time series classification. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 248–257.

Ding, Y.; Udompanyawit, C.; Zhang, Y.; and He, B. 2025. EEG-based brain-computer interface enables real-time robotic hand control at individual finger level. *Nature Communications*, 16(1): 1–20.

Gordon-Fennell, A.; Barbakh, J. M.; Utley, M. T.; Singh, S.; Bazzino, P.; Gowrishankar, R.; Bruchas, M. R.; Roitman, M. F.; and Stuber, G. D. 2023. An open-source platform for head-fixed operant and consummatory behavior. *eLife*, 12: e86183.

Grijseels, D. M.; Prendergast, B. J.; Gorman, J. C.; and Miller, C. T. 2023. The neurobiology of vocal communication in marmosets. *Annals of the New York Academy of Sciences*, 1528(1): 13–28.

Hao, H.; Wang, Y.; Xue, S.; Xia, Y.; Zhao, J.; and Shen, F. 2020. Temporal convolutional attention-based network for sequence modeling. *arXiv preprint arXiv:2002.12530*.

Hu, K.; Jamali, M.; Moses, Z. B.; Ortega, C. A.; Friedman, G. N.; Xu, W.; and Williams, Z. M. 2018. Decoding unconstrained arm movements in primates using high-density electrocorticography signals for brain-machine interface use. *Scientific reports*, 8(1): 10583.

Huang, S.; Zhao, Z.; Li, C.; and Bai, L. 2025. TimeKAN: KAN-based Frequency Decomposition Learning Architecture for Long-term Time Series Forecasting. *arXiv preprint arXiv:2502.06910*.

Jang, E.; Gu, S.; and Poole, B. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Jiricek, S.; Koudelka, V.; Lacik, J.; Vejmola, C.; Kuratko, D.; Wójcik, D. K.; Raida, Z.; Hlinka, J.; and Palenicek, T. 2021. Electrical source imaging in freely moving rats: Evaluation of a 12-electrode cortical electroencephalography system. *Frontiers in neuroinformatics*, 14: 589228.

LaRocco, J.; and Paeng, D.-G. 2020. Optimizing Computer–Brain Interface Parameters for Non-invasive Brain-to-Brain Interface. *Frontiers in Neuroinformatics*, 14: 1. Epub 2020 Feb 7.

Lawhern, V. J.; Solon, A. J.; Waytowich, N. R.; Gordon, S. M.; Hung, C. P.; and Lance, B. J. 2018. EEGNet: A Compact Convolutional Network for EEG-based Brain-Computer Interfaces. *Journal of Neural Engineering*, 15(5): 056013.1–056013.17.

Le, X.-M.; Luo, L.; Aickelin, U.; and Tran, M.-T. 2024. Shapeformer: Shapelet transformer for multivariate time series classification. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1484–1494.

Le, X.-M.; Tran, M.-T.; and Huynh, V.-N. 2022. Learning perceptual position-aware shapelets for time series classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 53–69. Springer.

Li, G.; Choi, B.; Xu, J.; Bhowmick, S. S.; Chun, K.-P.; and Wong, G. L.-H. 2021a. Shapenet: A shapelet-neural network approach for multivariate time series classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 8375–8383.

Li, J.; Aoi, M. C.; and Miller, C. T. 2024. Representing the dynamics of natural marmoset vocal behaviors in frontal cortex. *Neuron*, 112(21): 3542–3550.

Li, Y.; Chen, B.; Yoshimura, N.; and Koike, Y. 2021b. Restricted minimum error entropy criterion for robust classification. *IEEE Transactions on Neural Networks and Learning Systems*, 33(11): 6599–6612.

Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; and Long, M. 2023. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. *arXiv preprint arXiv:2310.06625*.

Liu, Y.; Nour, M. M.; Schuck, N. W.; Behrens, T. E.; and Dolan, R. J. 2022. Decoding cognition from spontaneous neural activity. *Nature Reviews Neuroscience*, 23(4): 204–214.

Meng, J.; Zhang, S.; Bekyo, A.; Olsoe, J.; Baxter, B.; and He, B. 2016. Noninvasive electroencephalogram based control of a robotic arm for reach and grasp tasks. *Scientific reports*, 6(1): 38565.

Nagashima, S.; Kaneda, K.; Wada, Y.; Iida, T.; Taguchi, M.; Hirata, M.; and Sugiura, K. 2025. ECoG Dual Context Network for Brain Machine Interfaces. *IEEE Access*.

Rupp, R.; Rohm, M.; Schneiders, M.; Kreilinger, A.; and Müller-Putz, G. R. 2015. Functional rehabilitation of the paralyzed upper extremity after spinal cord injury by noninvasive hybrid neuroprostheses. *Proceedings of the IEEE*, 103(6): 954–968.

Schwarz, D. A.; Lebedev, M. A.; Hanson, T. L.; Dimitrov, D. F.; Lehew, G.; Meloy, J.; Rajangam, S.; Subramanian, V.; Ifft, P. J.; Li, Z.; et al. 2014. Chronic, wireless recordings of large-scale brain activity in freely moving rhesus monkeys. *Nature methods*, 11(6): 670–676.

Silvestrin, L. P.; Yu, S.; and Hoogendoorn, M. 2023. Revisiting the Robustness of the Minimum Error Entropy Criterion: A Transfer Learning Case Study. In *ECAI 2023*, 2146–2153. IOS Press.

Skomrock, N. D.; Schwemmer, M. A.; Ting, J. E.; Trivedi, H. R.; Sharma, G.; Bockbrader, M. A.; and Friedenberg, D. A. 2018. 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.

Tonin, L.; Perdikis, S.; Kuzu, T. D.; Pardo, J.; Orset, B.; Lee, K.; Aach, M.; Schildhauer, T. A.; Martínez-Olivera, R.; and Millán, J. d. R. 2022. Learning to control a BMI-driven wheelchair for people with severe tetraplegia. *Iscience*, 25(12).

Tsunada, J.; and Eliades, S. J. 2024. Frontal-auditory cortical interactions and sensory prediction during vocal production in marmoset monkeys. *bioRxiv*.

Tsunada, J.; and Eliades, S. J. 2025. Frontal–Auditory Cortical Interactions and Sensory Prediction During Vocal Production in Marmoset Monkeys. *Current Biology*, 35(10): 2307–2322.e3. Updated and peer-reviewed version of the original bioRxiv preprint.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Y.; Huang, N.; Li, T.; Yan, Y.; and Zhang, X. 2024. Medformer: A multi-granularity patching transformer for medical time-series classification. *arXiv preprint arXiv:2405.19363*.

Wu, H.; Xu, J.; Wang, J.; and Long, M. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34: 22419–22430.

Wu, X.; Qiu, X.; Li, Z.; Wang, Y.; Hu, J.; Guo, C.; Xiong, H.; and Yang, B. 2024. Catch: Channel-aware multivariate time series anomaly detection via frequency patching. *arXiv preprint arXiv:2410.12261*.

Yan, T.; Suzuki, K.; Kameda, S.; Maeda, M.; Mihara, T.; and Hirata, M. 2023. Chronic subdural electrocorticography in nonhuman primates by an implantable wireless device for brain-machine interfaces. *Frontiers in neuroscience*, 17: 1260675.

Yi, K.; Fei, J.; Zhang, Q.; He, H.; Hao, S.; Lian, D.; and Fan, W. 2024. Filternet: Harnessing frequency filters for time series forecasting. *Advances in Neural Information Processing Systems*, 37: 55115–55140.

Zabeh, E.; Foley, N. C.; Jacobs, J.; and Gottlieb, J. P. 2023. Beta traveling waves in monkey frontal and parietal areas encode recent reward history. *Nature Communications*, 14(1): 5428.

Zhang, G.; Yu, M.; Liu, Y.-J.; Zhao, G.; Zhang, D.; and Zheng, W. 2021a. SparseDGCNN: Recognizing emotion from multichannel EEG signals. *IEEE Transactions on Affective Computing*, 14(1): 537–548.

Zhang, L.; Na, J.; Zhu, J.; Shi, Z.; Zou, C.; and Yang, L. 2021b. Spatiotemporal causal convolutional network for forecasting hourly PM2. 5 concentrations in Beijing, China. *Computers & Geosciences*, 155: 104869.

Zhang, Y.; and Yan, J. 2023. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*.

Zhao, L.; and Wang, X. 2023. Frontal cortex activity during the production of diverse social communication calls in marmoset monkeys. *Nature communications*, 14(1): 6634.

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

Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 11106–11115.

Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; and Jin, R. 2022. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, 27268–27286. PMLR.