NeuroRVQ: Multi-Scale Biosignal Tokenization for Generative Foundation Models

Konstantinos Barmpas *
Imperial College London & Cogitat

Na Lee
Imperial College London & Cogitat

Dimitrios Chalatsis
Imperial College London

William Raftery
Imperial College London

Yannis Panagakis
National and Kapodistrian University of Athens &
Archimedes Research Unit & Cogitat

Dimitrios A. Adamos
Imperial College London & Cogitat

Nikolaos Laskaris
Aristotle University of Thessaloniki & Cogitat

Alexandros Koliousis
Northeastern University London

Dario Farina
Imperial College London

Stefanos Zafeiriou
Imperial College London & Cogitat

Abstract

Biosignals such as electroencephalography (EEG), electrocardiography (ECG), and electromyography (EMG) encode physiological activity across multiple temporal and spectral scales, yielding representations that are rich but challenging for machine learning. Foundation models trained to predict masked signal tokens have shown promise in learning generalizable biosignal representations, yet their performance depends on the tokenizer's ability to preserve high-frequency dynamics and reconstruct signals with high fidelity. We introduce NEURORVQ, a modality-adaptive biosignal tokenizer family designed for high-fidelity signal reconstruction. To capture the full frequency spectrum, NEURORVQ decomposes biosignals into frequency-specific representations via multi-scale temporal convolutions, each encoded into hierarchical RVQ codebooks to preserve high-frequency detail, combined with a novel phase-aware training loss that respects the circular topology of Fourier phase. By tuning the temporal resolution, number and size of temporal kernels and RVQ depth, this design adapts to the spectro-temporal characteristics of each biosignal modality. To validate that tokenizer quality drives downstream performance, we train a simple masked-token foundation model for each modality (NEURORVQ-FM) using the corresponding NEURORVQ tokenizer. The NEURORVQ-FM family achieves competitive or superior downstream performance compared to existing modality-specific foundation models, demonstrating that high-fidelity tokenization is a critical factor for effective biosignal modeling. Code and models are available on our project page: https://neurorvq.github.io

*Correspondence to konstantinos.barmpas16@imperial.ac.uk.

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

1 Introduction

Biosignals, electrical recordings of physiological activity, are fundamental to modern healthcare and Human-Computer Interaction (HCI): Electroencephalography (EEG) captures brainwave dynamics underlying cognition, sleep and emotion; electrocardiography (ECG) records the cardiac electrical cycle critical for diagnosing arrhythmias and heart disease; and electromyography (EMG) measures muscular activation patterns essential for movement analysis and prosthetic control. Across these modalities, accurate signal decoding underpins a broad spectrum of applications, from Brain-Computer Interfaces (BCIs) for emotion recognition [Torres et al., 2020, Xu et al., 2018] and epileptic seizure detection [Alkawadri, 2019, Djoufack Nkengfack et al., 2021], to cardiac anomaly detection [Hannun et al., 2019] and gesture recognition from muscle activity [Atzori et al., 2014]. Despite their different physiological origins, EEG, ECG and EMG signals share a common characteristic: **information is distributed across multiple frequency scales.** EEG brainwaves span distinct frequency bands (e.g., delta, alpha, gamma) that capture various physiological states [Newson and Thiagarajan, 2018]. ECG signals encode both low-frequency morphological features (e.g., the QRS complex) and subtle high-frequency components indicative of abnormalities [Goldberger et al., 2017]. Similarly, EMG signals combine low-frequency motor unit recruitment patterns with high-frequency firing dynamics [De Luca, 1997]. **Any analysis framework for biosignals must capture this multiscale structure, both slow and fast dynamics, to fully preserve the information content of the signal.**

A range of techniques, from advanced signal processing to deep learning, have been employed for biosignal decoding. More recently, foundation models trained with self-supervised masked learning have emerged as a powerful paradigm: signals are first discretized into *tokens* by a tokenizer and during training, tokens are intentionally masked so that the model learns to reconstruct them. This approach has shown strong generalization capabilities without relying on task-specific data collection. However, the success of masked modeling relies heavily on tokenizer quality. Tokenizers must effectively capture spatio-temporal patterns to convert continuous signals into meaningful discrete representations while enabling faithful reconstruction. **Current tokenizers struggle to preserve complete structural information, especially high-frequency components, necessary for robust generative masked modeling.** The key to unlocking foundation-scale masked modeling for biosignals lies in the tokenizer:

> A well-designed biosignal tokenizer should not only compress continuous physiological signals into discrete neural tokens but also enable faithful reconstruction of the original waveform across all important frequency scales.

In this work, we introduce NEURORVQ, a modality-adaptive biosignal tokenizer family designed for high-fidelity signal reconstruction. The main contributions of this work are:

(1) a **modality-adaptive multiscale tokenizer architecture** that decomposes biosignals into frequency-specific representations via temporal convolutions with varying kernel sizes and encodes each frequency scale into hierarchical Residual Vector Quantization (RVQ) codebooks, enabling high-fidelity signal reconstruction across EEG, ECG, and EMG. By tuning the temporal resolution, number and size of temporal kernels, and RVQ depth, this architecture adapts to the spectro-temporal characteristics of each biosignal modality.

(2) a **phase-aware training loss** that reconstructs the Fourier spectrum through three complementary components: a log-amplitude loss for Fourier amplitude to emphasize high-frequency content, a temporal-domain regularization loss and a novel phase loss that respects the circular topology of phase angles by leveraging cosine similarity for directional alignment; and

(3) a **simple masked-token foundation model** (NEURORVQ-FM) for each modality that leverages the corresponding NEURORVQ tokenizer during pre-training and achieves competitive or superior downstream performance compared to existing modality-specific foundation models, demonstrating that high-fidelity tokenization (rather than model scale or architectural complexity) is a critical ingredient for effective biosignal modeling.

Compared to existing approaches that operate at a single scale with a single codebook, NEURORVQ with its multi-scale architecture and per-scale hierarchical RVQ codebooks achieves substantially lower reconstruction error, while the phase-aware loss delivers the largest reconstruction improvement. When used in the pre-training of NEURORVQ-FM, the high-fidelity tokenization translates directly

into strong downstream performance: NEURORVQ-EEG achieves the best mean balanced accuracy across five classification tasks, NEURORVQ-ECG doubles the next-best balanced accuracy on 43-class PTB-XL classification and NEURORVQ-EMG outperforms existing EMG foundation models across four classification benchmarks.

2 Background & Related Work

The roots of biosignal analysis can be traced back to classical methods, when handcrafted features such as Power Spectral Density (PSD) and Independent Component Analysis (ICA) were regarded standard [James and Hesse, 2004, McFarland et al., 2006]. Although these methods provided valuable insights into biosignal decoding, they often failed to generalize due to high inter-subject variability and their limited capacity to capture the full complexity of brain, heart and muscle activities.

The deep learning era marked a shift towards data-driven feature extraction: models such as EEG-Net [Lawhern et al., 2018], EEGInception [Santamaría-Vázquez et al., 2020], Brainwave-Scattering Net [Barmpas et al., 2023] and EEGConformer [Song et al., 2023] for EEG, along with deep residual networks for ECG [Ribeiro et al., 2020] and convolutional architectures for EMG [Côté-Allard et al., 2019], reduced reliance on handcrafted features by automatically learning signal representations. To address domain-specific challenges such as subject variability, researchers incorporated techniques like subject-alignment into these architectures [Bakas et al., 2022, Wei et al., 2022, Barmpas et al., 2024b]. Despite these advances, most deep models remain limited: they require carefully labeled data and extensive task-specific training, hindering their ability to generalize across diverse tasks.

More recently, the success of large foundation models in natural language processing [Brown et al., 2020, Touvron et al., 2023] and computer vision [Mizrahi et al., 2023] has motivated their adoption in biosignal decoding. While foundation models for EEG are now relatively mature, those for ECG and EMG are still emerging [Lee et al., 2025b]. Notable examples include BIOT [Yang et al., 2023], LaBraM [Jiang et al., 2024], NeuroGPT [Cui et al., 2024], EEGPT [Wang et al., 2024], CBraMod [Wang et al., 2025], BrainOmni [Xiao et al., 2025] and REVE [Ouahidi et al., 2026] for EEG; HuBERT-ECG [Coppola et al., 2024] and ECGFounder [Li et al., 2025] for ECG; and PhysioWave [Chen et al., 2025] and TinyMyo [Fasulo et al., 2026] for EMG. These models are pretrained on large unlabeled datasets using self-supervised signal reconstruction objectives and have shown promising generalization capabilities. However, they share a common bottleneck: the quality of signal tokenization, which determines the fidelity of the representations the model learns from.

Tokenization is critical for robust generative masked modeling, yet current tokenizers struggle to capture the full structure of biosignals. Some models rely on deep backbone-based tokenizers (e.g. CBraMod), while others employ transformer-based approaches (e.g. BIOT), which often learn entangled and redundant representations not optimized for reconstruction. As a result, the generated tokens lack the fidelity required for effective masked prediction. More recently, codebook-based approaches, such as LaBraM and BrainOmni, have been introduced to learn reconstructable discrete tokens using vector quantization (VQ) and residual VQ (RVQ) respectively. However, both rely on a single codebook to capture the full spectrum of biosignal activity and train their tokenizers with losses that do not always respect the signal processing properties of their reconstruction targets. The result is codebook-based tokenizers with unfaithful signal reconstruction.

3 NeuroRVQ Tokenizer Design

3.1 Model Architecture

NEURORVQ tokenizer is a specialized network designed to convert raw biosignals into a sequence of compact and informative neural tokens. This transformation reduces the inherently high-dimensional and noisy nature of biosignals into a structured lower-dimensional representation preserving essential temporal–spectral patterns. In doing so, the tokenizer provides a kind of "grammar" for the underlying physiological activity.

Figure 1 illustrates the tokenizer architecture and processing flow. The input multi-variate biosignal timeseries is first segmented into patches. These patches are encoded by the multi-scale temporal encoder, that captures features in multiple resolutions and are then combined via the transformer encoder. For each scale, RVQ codebooks discretize the embeddings into a sequence of neural tokens.

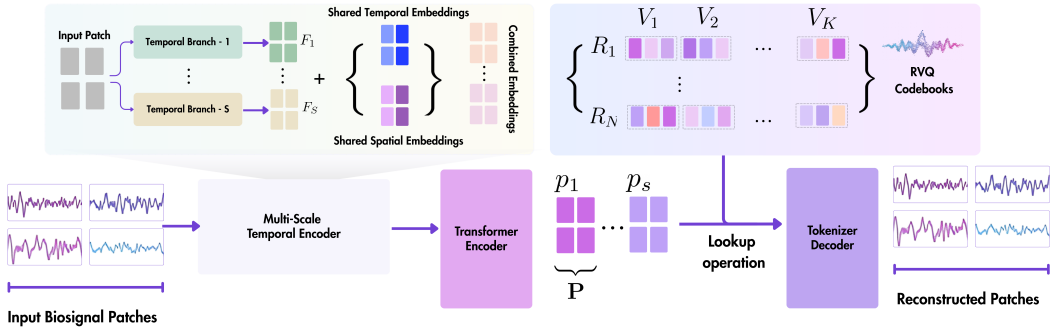

Figure 1: The NeuroRVQ tokenizer model. The multi-scale temporal encoder extracts multi-scale temporal features that are passed to the transformer encoder. RVQ codebooks per temporal branch discretize the biosignal patches into neural tokens. The tokenizer decoder reconstructs the biosignal by using the Fourier spectrum.

Finally, these tokens from all scales are combined and passed through the tokenizer decoder to reconstruct the input biosignal patches using the Fourier spectrum.

Generating patches. Let $X \in \mathbb{R}^{C \times T}$ denote the input multi-variate biosignal, where T is the number of time points and C is the number of electrodes. The signal is first segmented into temporal patches. To ensure that the tokenizer can accommodate biosignals with arbitrary numbers of channels and variable durations, we adopt the following approach: during model pre-training, each input multi-variate biosignal is represented by P patches of length w. This results in a segmented input sample $\mathbf{x} \in \mathbb{R}^{P \times w}$ which provides flexibility to handle heterogeneous biosignal recordings and facilitates robust pre-training across diverse datasets.

Multi-scale temporal encoder. The first step is to extract multi-scale temporal features from each input patch using an inception-style module with S distinct temporal scales. To achieve this, we apply 1-D temporal convolutions with varying kernel sizes $K_{\text{temporal}_1}, K_{\text{temporal}_2}, \ldots, K_{\text{temporal}_S}$. Each temporal branch consists of the following sequence of operations: a 1-D convolution, followed by group normalization, GELU activation and pooling (this sequence of operations is repeated twice). This design enables the network to capture temporal features at multiple scales resulting in S outputs $F_1, F_2, \ldots, F_S$ where $F_i \in \mathbb{R}^w$ where w is the patch time window length.

Transformer encoder. Since the tokenization of the biosignal takes place in a per-patch level, the vital temporal and spatial information is incorporated through the use of trainable temporal TE and spatial SE embeddings, each with embedding dimension w. $SE = \{se_1, ..., se_{|\mathcal{C}|}\}$ is indexed by the positions of the patches' electrodes within a list of all electrodes in the entire pre-training database $\mathcal{C}$ (e.g., 10-20 system for EEG or standard 12-lead for ECG) and $TE = \{te_1, ..., te_P\}$ since each input consists of P patches of length w. These embeddings are added to the extracted multi-scale features $F_1, F_2, \ldots, F_S$ and are shared across all S temporal branches. The combined embeddings of each temporal branch are then passed through a series of shared transformer layers [Vaswani et al., 2017] resulting in the multi-scale patch representations $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_S \in \mathbb{R}^D$. The shared transformer layers have incorporated the modification of Dehghani et al. [2023].

RVQ codebook. To efficiently discretize the multi-scale patch representations $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_S$ while preserving fine-grained information of the biosignal, we use S Residual Vector Quantization (RVQ) codebooks [Gray, 1984]. For each temporal branch, the RVQ codebook $\mathcal{R}$ is defined as:

$$\mathcal{R} = \{\mathcal{V}_i | i = 1, ..., N\} \tag{1}$$

where $\mathcal{V}_i$ is a single codebook of neural tokens defined as:

$$\mathcal{V}_i = \{\mathbf{v}_j | j = 1, ..., K\} \in \mathbb{R}^{K \times D} \tag{2}$$

where K is the number of discrete neural tokens and D is the dimensionality of each token. The discretization of the patch representations is performed iteratively. For each $\mathcal{R}$, at step 1, a quantizer maps each patch representation $\mathbf{p}^1$ to its corresponding neural token $\mathbf{v}^1 \in \mathcal{V}_1$. Specifically, for each patch, the quantizer selects the nearest neighbor $\mathbf{z_1}$ from the codebook $\mathcal{V}_1$:

$$\mathbf{z_1} = \arg\min_{\mathbf{v^1} \in \mathcal{V}_1} \left\| l_2(\mathbf{p}^1) - l_2(\mathbf{v^1}) \right\| \tag{3}$$

At each next step $i+1$, the residual patch representation $\mathbf{p}^{i+1}$ for the next codebook $\mathcal{V}_{i+1}$ is updated by subtracting the selected token $\mathbf{z}_i$:

$$\mathbf{p}^{i+1} = \mathbf{p}^i - \mathbf{z}_i \tag{4}$$

This process is repeated for all N codebooks, producing a sequence of tokens $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N$ for each temporal branch separately. The final reconstructed patch representations $\hat{\mathbf{p}}$ is obtained by summing the contributions of all selected tokens $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N$:

$$\hat{\mathbf{p}} = \sum_{i=1}^{N} \mathbf{z}_i \tag{5}$$

At the end of this module, we have $\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2, \ldots, \hat{\mathbf{p}}_S$ corresponding to the S temporal branches.

Tokenizer decoder. The purpose of the decoder module (a series of shared transformer layers followed by prediction heads) is to reconstruct the original signal information based on the learned codebook tokens. Directly reconstructing raw biosignals has been shown to yield unstable training and poor performance, due to the inherently noisy and high-dimensional nature of these signals. Therefore, we leverage the Fourier spectrum (more specifically the phase ϕ and spectral amplitude A) as a more structured target for reconstruction.

Modality-adaptive design parameters. NEURORVQ tokenizer architecture exposes four principled parameters that can be tuned per modality to accommodate biosignals with fundamentally different spectro-temporal characteristics: (*i*) the patch length w, which controls the temporal context per patch; (*ii*) the number of temporal branches S, which determines how many frequency scales the signal is decomposed into; (*iii*) the temporal convolution kernel sizes $K_{\text{temporal}_1}, \ldots, K_{\text{temporal}_S}$, which align each branch's receptive field with the characteristic time scales of the target modality; and (*iv*) the number of RVQ codebooks N per branch, which dictates the quantization depth.

3.2 Phase-Aware Training Loss

However, architecture alone does not account for NEURORVQ's reconstruction gains: the training loss is equally critical. The NEURORVQ tokenizer loss operates in the Fourier domain (phase and amplitude) and consists of three complementary loss components.

Phase loss. Reconstructing the phase ϕ directly with mean squared error (MSE) is suboptimal due to the periodic nature of the Fourier phase (see Appendix A). To address this, the tokenizer decoder replaces raw phase reconstruction with a sine/cosine phase representation and predicts two values $\hat{c}$ and $\hat{s}$, corresponding to estimates of $\cos\phi$ and $\sin\phi$ respectively, which are smooth, continuous and better suited as loss terms. Since both predictions $\hat{c}$ and $\hat{s}$ refer to the same phase angle, we optimize them jointly rather than independently. We represent each phase as a 2D vector and compute the cosine similarity between the predicted vector $(\hat{c}, \hat{s})$ and the target $(\cos\phi, \sin\phi)$. This formulation naturally respects the circular topology of phase: the phases are represented as vectors in 2D space, making the errors between them geometrically meaningful.

Amplitude loss. For the amplitude component, rather than reconstructing raw spectral magnitudes A, we operate in the log-amplitude domain $\log(1+A)$. This logarithmic scaling compresses the large dynamic range of signal spectra and places relatively greater emphasis on reconstructing high-frequency components. Therefore, the tokenizer decoder has three prediction heads responsible for reconstructing $\log(1+\hat{A})$, $\hat{c}$ and $\hat{s}$, respectively.

Temporal loss. Finally, although reconstructing raw biosignal X waveforms might lead to instability issues [Jiang et al., 2024], we found that incorporating a temporal-domain reconstruction loss provides additional guidance. Specifically, we transform the predicted log-amplitude and phase back to the time domain (the predicted log-amplitudes are exponentiated and rescaled to recover the original amplitude spectrum before the inverse Fourier transformation) and compute a temporal MSE regularization term.

The NEURORVQ tokenizer is thus trained end-to-end with the objective:

$$\mathcal{L}_T = \sum_i \underbrace{\left\| \log(1+\hat{A}_i) - \log(1+A_i) \right\|_2^2}_{\text{Log-Amplitude Loss}} + \underbrace{1 - \sum_i \frac{\hat{c}_i \cos\phi_i + \hat{s}_i \sin\phi_i}{\sqrt{\hat{c}_i^2 + \hat{s}_i^2}}}_{\text{Phase Loss}} + \underbrace{\left\| \hat{X}_i - X_i \right\|_2^2}_{\text{Temporal Loss}} + \mathcal{L}_Q \tag{6}$$

where $\hat{A}$, $\hat{c}$, and $\hat{s}$ denote reconstructed Fourier components, $\hat{X}$ denotes reconstructed biosignal in the time domain and $\mathcal{L}_Q$ is the commitment loss.

As shown in Appendix B, the phase-aware loss is a critical component. The principled combination of multi-scale architecture (Section 3.1) and its novel training loss (Section 3.2) is what enables NEURORVQ's high-fidelity reconstruction.

4 Modality-Specific Configuration

The NEURORVQ tokenizer architecture is modality-agnostic by design, but its parameters (namely patch length, sampling rate, number of temporal branches and RVQ codebooks) must be adapted to the distinct spectro-temporal characteristics of each biosignal. In this section, we describe the signal properties and the scaling analysis that motivate our configuration choices.

4.1 Signal Characteristics

EEG. Neural oscillations of interest span distinct frequency bands, namely delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (> 30 Hz). Each band is linked to specific cognitive states, with clinically and cognitively relevant content concentrated below 100 Hz [Newson and Thiagarajan, 2018]. We resample all EEG recordings to 200 Hz, which satisfies the Nyquist criterion for frequencies up to 100 Hz. We set the patch length to $w = 1sec$, offering a practical trade-off between temporal localization and spectral resolution for downstream tokenization.

ECG. The cardiac electrical cycle is characterized by a sequence of morphologically distinct waveforms whose shapes and duration carry critical diagnostic information: the P wave ($\sim 60 - 100\,ms$), the QRS complex ($\sim 80 - 120\,ms$) and the T wave (typically $\sim 160\,ms$) [Sattar and Chhabra, 2023]. The clinically relevant frequency content of ECG signals is largely concentrated below 100 Hz, with high-frequency components above this range primarily reflecting noise [Kligfield et al., 2007]. Accordingly, we resample ECG recordings to 200 Hz. We set the patch length to $w = 200ms$, a window that is sufficiently long to capture the QRS complex while remaining short enough to isolate individual waveform components without mixing adjacent cardiac features.

EMG. EMG signals arise from the superposition of Motor Unit Action Potentials (MUAPs), which are brief electrical discharges lasting approximately 5–30 ms [Dumitru et al., 1997]. The resulting surface EMG signal contains significant spectral energy extending up to 500 Hz, with the dominant power typically concentrated between 20 and 250 Hz [De Luca, 1997]. To adequately capture these high-frequency dynamics, we resample EMG recordings to 1000 Hz (substantially higher than for EEG and ECG). We set the patch length to $w = 200ms$, which is long enough to encompass multiple MUAP events while remaining short enough to track rapid changes in muscle activation.

Although NEURORVQ design is flexible, these modality-specific signal-level choices also influence the design of the temporal encoder, particularly the choice of kernel sizes (see Appendix C).

4.2 Scaling Analysis

To determine the optimal tokenizer configuration for each modality, we conduct a systematic scaling analysis over two key architectural hyperparameters [1]: (*i*) the number of temporal branches $S \in \{1, 2, 4\}$ and (*ii*) the number of RVQ codebooks per branch $N \in \{1, 2, 4, 8, 16, 32\}$. For each modality, we select a compact representative dataset, namely PhysioNet MI [Schalk et al., 2004] for EEG, PTB [Bousseljot et al., 2004] for ECG, and the small version of emg2pose [Salter et al.,

[1] Larger configurations were not explored due to computational budget constraints.

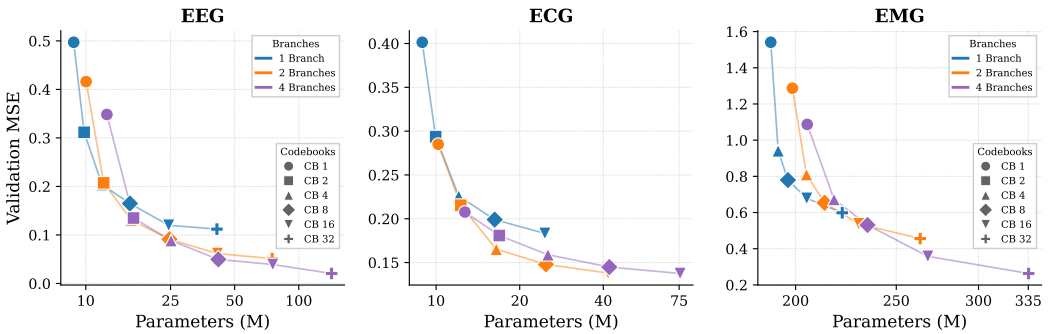

Figure 2: Scaling Analysis. Mean squared error (MSE) reconstruction performance and number of trainable parameters for NEURORVQ with various numbers of temporal branches and RVQ codebooks (CB) per branch across three modalities (EEG, ECG, EMG).

2024] for EMG. We train the NEURORVQ tokenizer for 50 epochs using the settings described in Appendix C. We report validation MSE (in an unseen set) averaged over the last 10 epochs.

General trends. As seen from the results in Figure 2 (separate plots for parameter counts and reconstruction MSE as a function of codebook depth / number of branches are provided in Appendix D), two consistent patterns emerge across all modalities: (*i*) increasing the number of codebooks per branch yields large reductions in reconstruction error, confirming that deeper residual quantization refines the discrete approximation of continuous signal features. (*ii*) adding temporal branches provides complementary gains (particularly when the number of codebooks is small) by adding specific priors and enabling the model to decompose the signal into frequency-specific representations that are each easier to quantize. When both dimensions are scaled jointly, the improvements accumulate.

Modality-specific observations. For modalities sampled at the same rate (200Hz), signal complexity dictates the saturation behavior: less noisy ECG saturates earlier, whereas EEG with richer spectral structure demands finer-grained quantization, continues to benefit from scaling the number of codebooks. For EMG, which is sampled at a substantially higher rate (1000Hz) to capture its broadband spectral content, the steepest improvements come from increasing codebook depth, suggesting that the wider frequency range requires greater representational capacity (codebook utilization analysis is provided in Appendix E). These trends validate a core design principle of NEURORVQ: the multi-scale branch structure and hierarchical RVQ codebooks are complementary mechanisms whose optimal balance is modality-dependent.

5 Evaluation

5.1 Signal Reconstruction Evaluation

Tokenizer training. For each biosignal modality, a NEURORVQ tokenizer was trained using the hyperparameters described in Appendix C on the datasets listed in Appendix F. All datasets were resampled to the modality-specific sampling frequency defined in Section 4.1. Based on the scaling analysis in Section 4.2 and available computational resources, we select $S = 4$ temporal branches for all modalities with $N = 8$ codebooks per branch for EEG and ECG and $N = 16$ for EMG, reflecting the greater representational demands of broadband EMG signals. All experiments were conducted on an NVIDIA DGX system with 4 Tesla V100 GPUs.

Reconstruction results. As shown in Figure 3, NEURORVQ achieves *high-fidelity signal reconstruction* across all modalities. ECG exhibits the fastest and sharpest convergence consistent with its stereotyped morphology, EEG converges to a stable plateau reflecting its richer spectral structure and EMG converges more gradually, consistent with its broadband nature. For ECG and EMG, no unimodal codebook-based tokenizers currently exist; NEURORVQ is the first to demonstrate high-fidelity codebook-based reconstruction for these modalities (reconstruction examples are provided in Appendix G). For EEG, we compare NEURORVQ against two pretrained open-source

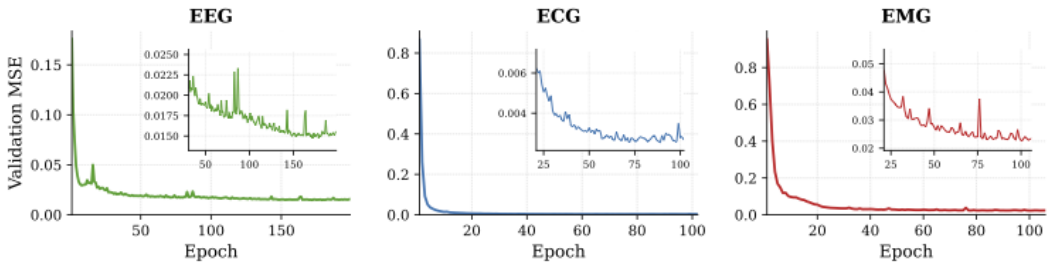

Figure 3: Validation mean squared error (MSE) during NEURORVQ tokenizer training across all three modalities. Note that absolute MSE values are not directly comparable across modalities.

codebook-based tokenizers: LaBraM and BrainOmni. NEURORVQ consistently outperforms both across evaluation setups (see Appendix H for detailed comparisons).

5.2 Downstream Task Performance

To demonstrate that high-fidelity tokenization translates into strong downstream performance, we evaluate fine-tuned NEURORVQ foundation models (NEURORVQ-FMs) across all three biosignal modalities using the tokenizers of Section 5.1.

Foundation model architecture and training. NEURORVQ-FM operates on the tokenized representation produced by the NEURORVQ tokenizer. The architecture consists of the multi-scale temporal encoder (Section 3.1), followed by transformer layers that model long-range dependencies and token prediction heads that reconstruct masked tokens using the learned codebooks $\mathcal{R}$ (see Figure 4). The model is pre-trained with a masked-token prediction objective using a symmetric masking strategy [Jiang et al., 2024]. For downstream tasks, the pre-trained backbone is extended with task-specific classification heads. We deliberately keep the foundation model architecture simple to isolate the contribution of the tokenizer: improvements in downstream performance can be attributed to the quality of the learned token representations rather than architectural complexity. The hyperparameters for the pre-training and fine-tuning of NEURORVQ foundation model for each biosignal modality are described in Appendix J on the datasets listed in Appendix F. All datasets were resampled to the modality-specific sampling frequency defined in Section 4.1.

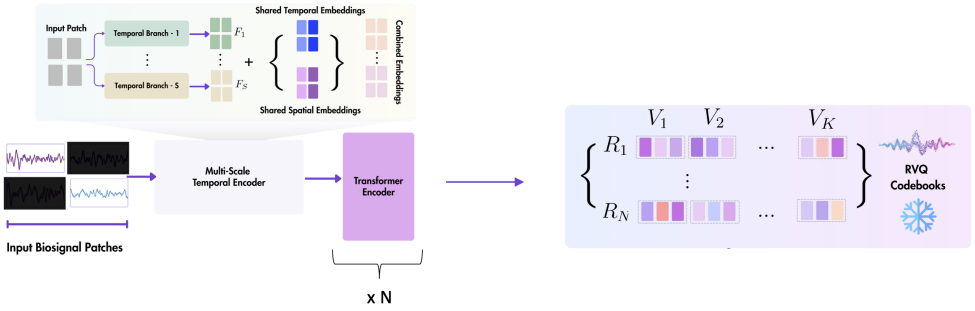

Figure 4: The NeuroRVQ-FM model. The multi-scale temporal encoder extracts multi-scale temporal features that are passed to a series of transformer encoder layers. The training objective is to reconstruct masked tokens using the learned codebooks $\mathcal{R}$.

Downstream task evaluation. We evaluate fine-tuned NEURORVQ-FMs across all three modalities. For EEG, we compare against nine foundation models on five downstream classification tasks from the benchmark of Lee et al. [2025a]. These models include two models that use codebook-based tokenizers like NEURORVQ, namely LaBraM and BrainOmni and seven common foundation model

Table 1: Downstream task performance across three biosignal modalities. **Bold** and underlined values indicate best and second-best performance per column, respectively. All evaluations are subject-independent. EEG: Balanced accuracy (BAcc). ECG: Accuracy and Balanced accuracy (BAcc). EMG: Accuracy, F1, Balanced Accuracy (BAcc) and CLER (Classification Error Rate).

(A) EEG — FIVE CLASSIFICATION TASKS [LEE ET AL., 2025A]

Model (Backbone Params)	Motor	ERP	Memory	Sleep	Eyes
BIOT [Yang et al., 2023] (3.2M)	44.3±7.9	50.0±0.0	51.0±1.8	–	76.3±4.9
MIRepNet [Liu et al., 2025] (5.2M)	68.9±8.6	–	–	–	–
CBraMod [Wang et al., 2025] (4.9M)	61.4±10.4	77.7±5.2	57.4±3.8	63.5±4.1	83.9±4.1
EEGPT [Wang et al., 2024] (25.7M)	31.3±3.5	66.8±14.6	52.0±1.7	63.4±4.4	79.7±3.7
NeuroGPT [Cui et al., 2024] (79.5M)	68.2±8.3	75.7±4.8	**59.7±2.9**	67.4±3.3	82.7±3.6
CSBrain [Zhou et al., 2025] (4.9M)	67.1±11.1	76.5 ±5.7	57.0±3.5	67.2±3.8	81.9±4.0
REVE [Ouahidi et al., 2026] (69.2M)	69.4±10.5	82.6±4.3	59.3±3.2	62.9±3.5	79.9±3.4
BrainOmni† [Xiao et al., 2025] (37.8M)	58.5±5.8	72.3±4.1	51.8±2.3	57.6±5.5	85.2±2.2
LaBraM† [Jiang et al., 2024] (5.8M)	63.0±7.6	82.2±4.0	52.6±2.6	65.2±3.7	79.9±4.7
NEURORVQ-EEG† (5.9M)	**70.0±7.3**	**87.6±3.3**	57.4±2.7	**72.8±2.8**	**86.9±2.6**

† Models that use codebook-based tokenizers

(B) ECG — PTB-XL CLASSIFICATION

Model (Backbone Params)	5-class Accuracy	5-class BACC	43-class Accuracy	43-class BACC
HuBERT-ECG [Coppola et al., 2024] (30.2M)	72.60	60.23	62.49	20.71
ECGFounder [Li et al., 2025] (30.7M)	**76.55**	**65.39**	65.51	28.96
NEURORVQ-ECG (0.3M)	70.19	64.50	**79.17**	**58.33**

(C) EMG — FOUR CLASSIFICATION TASKS

Model (Backbone Params)	Discrete Gestures BAcc ↑	Discrete Gestures CLER ↓	EPN-612 Acc	EPN-612 F1	NinaPro DB5 Acc	NinaPro DB5 F1	UCI-EMG Acc	UCI-EMG F1
PhysioWave [Chen et al., 2025] (4.8M)	54.70±0.0	64.20±0.9	90.30±0.6	90.35±0.3	24.91±2.5	22.95±2.5	56.52±2.4	55.76±2.8
TinyMyo [Fasulo et al., 2026] (3.6M)	39.70	64.20	84.68±1.9	84.68±1.9	25.26±4.5	23.29±4.9	85.99±0.4	85.66±0.4
NEURORVQ-EMG (5.9M)	**70.80±0.6**	**27.60±0.8**	**94.65±0.4**	**94.66±0.5**	**41.36±2.8**	**38.76±3.7**	**89.43±1.1**	**89.28±1.1**

baselines, namely BIOT[2], MIRepNet[3], CBraMod, EEGPT, NeuroGPT, CSBrain and REVE. For ECG, we compare against two foundation models, HuBERT-ECG and ECGFounder, on the PTB-XL benchmark [Wagner et al., 2020]. For EMG, we compare against two foundation models, PhysioWave and TinyMyo, on four classification datasets (Discrete Gestures [Kaifosh et al., 2025], EPN-612 [Benalcazar et al.], NinaPro DB5 [Pizzolato et al.], UCI-EMG [Krilova et al., 2018]). Appendix K provides fine-tuning, implementational details and further analysis. Appendix M presents the head-only fine-tuning results for all NEURORVQ-FMs and other foundation model baselines.

As shown in Table 1, NEURORVQ achieves strong downstream performance across all three biosignal modalities. NEURORVQ-EEG achieves the best balanced accuracy on four of five tasks and the highest mean balanced accuracy overall. NEURORVQ-ECG achieves high balanced accuracy on both PTB-XL settings, doubling the balanced accuracy of the next-best model on the fine-grained 43-class task. ECGFounder leads on the 5-class setting with NEURORVQ in the second-best bacc while in the 43-class setting NEURORVQ outperforms all baselines. NEURORVQ-EMG outperforms PhysioWave and TinyMyo across all four classification tasks by a wide margin. Across all three modalities, NEURORVQ's consistent improvements with a compact and simple foundation model architecture confirm the central claim of this work: *high-fidelity tokenization is the critical ingredient for effective biosignal foundation models* and its benefits transfer across physiologically diverse signal types. To isolate the effect of high-fidelity tokenization in downstream task performance, we conducted a tokenizer-swap ablation study (Appendix N). The results show that tokenization quality translates into downstream performance: balanced accuracy improves with tokenizer fidelity and the effect is clearest under full fine-tuning.

[2]BIOT could not be tested on Sleep since the benchmark electrodes are missing from its pre-trained model. Its pre-trained bipolar channels were mapped to standard locations.

[3]As a motor-specific foundation model, MIRepNet was tested only on the Motor task.

6 Discussion

The NEURORVQ tokenizer demonstrates state-of-the-art reconstruction capabilities (Section 5.1) on out-of-distribution and in-distribution signals (Appendix H and Appendix I). Importantly, these results cannot be attributed solely to the number of codebooks (single or multiple), their size or their type (e.g., RVQ-based). Instead, they arise from the inductive biases embedded in the tokenizer design (Section 4) which, together with the newly introduced phase-aware training loss (Section 3.2 and Appendix B), are particularly well-suited for capturing the complex nature of biosignals.

Since biosignals differ fundamentally in their spectro-temporal characteristics, we train separate tokenizers in this work, each optimally configured for its target modality while the architecture remains identical: only signal-level hyperparameters change (see Section 4). Whether a shared cross-modal tokenizer can match this per-modality performance is an interesting future direction.

A natural concern is whether pursuing high-fidelity reconstruction is warranted for inherently noisy signals. However, high-fidelity reconstruction does not mean memorizing noise. It ensures sufficient representational capacity to preserve all signal components, including subtle discriminative features that overlap spectrally with noise, letting downstream foundation models learn which are task-relevant.

The main contribution of this work is the NEURORVQ tokenizer, which can be integrated into any biosignal foundation model. Further improvements can come from larger foundation architectures, cross-modal pre-training and incorporating causal reasoning-inspired training principles such as more targeted temporal and spatial masking strategies [Barmpas et al., 2024a].

7 Conclusion

In this work, we introduce NEURORVQ, a modality-adaptive biosignal tokenizer family designed for high-fidelity signal reconstruction across EEG, ECG and EMG. Our main contribution is the design and training of the tokenizer, which captures low- and high-frequency components of biosignals, achieves state-of-the-art reconstruction performance and facilitates the development of more effective biosignal foundation models. We showed that when a simple masked-token foundation model is trained for each modality using NEURORVQ tokenizers, it outperforms existing pre-trained open-source modality-specific foundation models in various downstream tasks. In summary, NEURORVQ advances the development of efficient large biosignal foundation models by providing a high-fidelity biosignal tokenizer, paving the way for broader use of foundation models in the biosignal space.

Acknowledgments

K. Barmpas was responsible for the development of NeuroRVQ, including dataset preprocessing, tokenizer analysis and models' training. K. Barmpas and S. Zafeiriou jointly designed NeuroRVQ's theoretical formulation. N. Lee conducted the EEG and ECG benchmarking analysis. D. Chalatsis and W. Raftery conducted the EMG benchmarking analysis. Y. Panagakis, D. Adamos, N. Laskaris, A. Koliousis and D. Farina provided valuable guidance and feedback throughout the project.

This work was supported by the EPSRC Turing AI Fellowship (Grant Ref: EP/Z534699/1), *Generative Machine Learning Models for Data of Arbitrary Underlying Geometry (MAGAL)*.

References

Rafeed Alkawadri. Brain–Computer Interface (BCI) Applications in Mapping of Epileptic Brain Networks Based on Intracranial-EEG: An Update. *Frontiers in Neuroscience*, 13:191, 2019. ISSN 1662-453X. doi: 10.3389/fnins.2019.00191.

Manfredo Atzori, Arjan Gijsberts, Claudio Castellini, et al. Electromyography data for non-invasive naturally-controlled robotic hand prostheses. *Scientific Data*, 1:140053, 2014.

Stylianos Bakas, Siegfried Ludwig, Konstantinos Barmpas, Mehdi Bahri, Yannis Panagakis, Nikolaos Laskaris, Dimitrios A. Adamos, and Stefanos Zafeiriou. Team Cogitat at NeurIPS 2021: Benchmarks for EEG Transfer Learning Competition, 2022.

Konstantinos Barmpas, Yannis Panagakis, Dimitrios A Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. BrainWave-Scattering Net: a lightweight network for EEG-based motor imagery recognition. *Journal of Neural Engineering*, 20(5):056014, September 2023. ISSN 1741-2552. doi: 10.1088/1741-2552/acf78a.

Konstantinos Barmpas, Yannis Panagakis, Dimitrios Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. A Causal Perspective in Brainwave Foundation Models. In *Causality and Large Models @NeurIPS 2024*, 2024a. doi: 10.1088/1741-2552/ad3eb5.

Konstantinos Barmpas, Yannis Panagakis, Georgios Zoumpourlis, Dimitrios A Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. A causal perspective on brainwave modeling for brain–computer interfaces. *Journal of Neural Engineering*, 21(3):036001, may 2024b. doi: 10.1088/1741-2552/ad3eb5.

Marco E. Benalcazar, Lorena Barona, Leonardo Valdivieso, Xavier Aguas, and Jonathan Zea. EMG-EPN-612 Dataset.

Benjamin Blankertz, Guido Dornhege, Matthias Krauledat, Klaus-Robert Müller, and Gabriel Curio. The non-invasive Berlin Brain–Computer Interface: Fast acquisition of effective performance in untrained subjects. *NeuroImage*, 37(2):539–550, 2007. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2007.01.051.

Ralf-Dieter Bousseljot, D Kreiseler, and A Schnabel. The PTB diagnostic ECG database, 2004.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, 2020.

Grace Buckwalter, Sidney Chhin, Sabariah Rahman, Iyad Obeid, and Joseph Picone. Recent Advances in the TUH EEG Corpus: Improving the Interrater Agreement for Artifacts and Epileptiform Events. In *2021 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–3, 2021. doi: 10.1109/SPMB52430.2021.9672302.

Yanlong Chen, Mattia Orlandi, Pierangelo Maria Rapa, Simone Benatti, Luca Benini, and Yawei Li. Physiowave: A multi-scale wavelet-transformer for physiological signal representation. *arXiv preprint arXiv:2506.10351*, 2025.

Edoardo Coppola, Mattia Savardi, Mauro Massussi, Marianna Adamo, Marco Metra, and Alberto Signoroni. HuBERT-ECG: a self-supervised foundation model for broad and scalable cardiac applications. *medRxiv preprint doi:10.1101/2024.11.14.24317328*, 2024. doi: 10.1101/2024.11.14.24317328.

Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A. Joshi, and Richard M. Leahy. Neuro-GPT: Towards A Foundation Model for EEG, 2024.

Ulysse Côté-Allard, Cheikh Latyr Fall, Alexandre Drouin, Alexandre Campeau-Lecours, Clément Gosselin, Kyrre Glette, François Laviolette, and Benoit Gosselin. Deep learning for electromyographic hand gesture signal classification using transfer learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(4):760–771, 2019. doi: 10.1109/TNSRE.2019.2896269.

Carlo J De Luca. The use of surface electromyography in biomechanics. *Journal of Applied Biomechanics*, 13(2):135–163, 1997.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. Featured Certification, Reproducibility Certification.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin F. Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Patrick Collier, Alexey Gritsenko, Vighnesh Birodkar, Cristina Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetić, Dustin Tran, Thomas Kipf, Mario Lučić, Xiaohua Zhai, Daniel Keysers, Jeremiah Harmsen, and Neil Houlsby. Scaling Vision Transformers to 22 Billion Parameters, 2023.

Paolo Detti, Giampaolo Vatti, and Garazi Zabalo Manrqiue de Lara. EEG Synchronization Analysis for Seizure Prediction: A Study on Data of Noninvasive Recordings. *Processes*, 8:846, 07 2020. doi: 10.3390/pr8070846.

Laurent Chanel Djoufack Nkengfack, Daniel Tchiotsop, Romain Atangana, Valérie Louis-Door, and Didier Wolf. Classification of EEG signals for epileptic seizures detection and eye states identification using Jacobi polynomial transforms-based measures of complexity and least-square support vector machine. *Informatics in Medicine Unlocked*, 23:100536, 2021. ISSN 2352-9148. doi: 10.1016/j.imu.2021.100536.

Daniel Dumitru, John C. King, and Sanjeev D. Nandedkar. Motor unit action potential duration recorded by monopolar and concentric needle electrodes: Physiologic implications. *American Journal of Physical Medicine & Rehabilitation*, 76(6):488–493, November 1997. ISSN 0894-9115. doi: 10.1097/00002060-199711000-00011.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming Transformers for High-Resolution Image Synthesis, 2020.

Matteo Fasulo, Giusy Spacone, Thorir Mar Ingolfsson, Yawei Li, Luca Benini, and Andrea Cossettini. Tinymyo: a tiny foundation model for flexible emg signal processing at the edge, 2026.

Ary L Goldberger, Zachary D Goldberger, and Alexei Shvilkin. *Goldberger's Clinical Electrocardiography: A Simplified Approach*. Elsevier, 9th edition, 2017.

Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W Waks, Parastou Eslami, Tanner Carbonati, Ashish Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset. *PhysioNet*, September 2023. doi: 10.13026/4nqg-sb35. Version 1.0.

Robert M. Gray. Vector quantization. *IEEE ASSP Magazine*, 1:4–29, 1984.

Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25:65–69, 2019.

Christopher J James and Christian W Hesse. Independent component analysis for biomedical signals. *Physiological Measurement*, 26(1):R15, dec 2004. doi: 10.1088/0967-3334/26/1/R02.

Weibang Jiang, Liming Zhao, and Bao liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024.

Patrick Kaifosh, Thomas R Reardon, and CTRL-labs at Reality Labs. A generic non-invasive neuromotor interface for human-computer interaction. *Nature*, 645(8081):702–711, September 2025.

Bastiaan Kemp, Aeilko H. Zwinderman, Bert Tuk, Hilbert A. C. Kamphuisen, and Josefien J. L. Oberyé. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000. doi: 10.1109/10.867928.

Hong-Kyung Kim, John Williamson, Min-Ho Lee, O-Yeon Kwon, Seong-Whan Lee, Siamac Fazli, Yong-Jeong Kim, and Young-Eun Lee. Supporting data for "EEG Dataset and OpenBMI Toolbox for Three BCI Paradigms: An Investigation into BCI Illiteracy", 2019.

Paul Kligfield, Leonard S. Gettes, James J. Bailey, and et al. Recommendations for the standardization and interpretation of the electrocardiogram: Part i: The electrocardiogram and its technology. *Circulation*, 115(10):1306–1324, 2007. doi: 10.1161/CIRCULATIONAHA.106.180200.

Louis Korczowski, Martine Cederhout, Anton Andreev, Grégoire Cattan, Pedro Rodrigues, Violette Gautheret, and Marco Congedo. Brain Invaders calibration-less P300-based BCI with modulation of flash duration Dataset (bi2015a), 07 2019.

N. Krilova, I. Kastalskiy, V. Kazantsev, V.A. Makarov, and S. Lobov. EMG Data for Gestures. UCI Machine Learning Repository, 2018. DOI: https://doi.org/10.24432/C5ZP5C.

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, July 2018. ISSN 1741-2552. doi: 10.1088/1741-2552/aace8c.

Na Lee, Stylianos Bakas, Konstantinos Barmpas, Yannis Panagakis, Dimitrios Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. Assessing the capabilities of large brainwave foundation models. In *2025 IEEE 35th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 01–06, 2025a. doi: 10.1109/MLSP62443.2025.11204282.

Na Lee, Konstantinos Barmpas, Alexandros Koliousis, Yannis Panagakis, Dimitrios Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. A comprehensive review of biosignal foundation models. November 2025b. doi: 10.36227/techrxiv.176369849.97173246/v1.

Jun Li, Aaron D. Aguirre, Valdery Moura Junior, Jiarui Jin, Che Liu, Lanhai Zhong, Chenxi Sun, Gari Clifford, M. Brandon Westover, and Shenda Hong. An Electrocardiogram Foundation Model Built on over 10 Million Recordings. *NEJM AI*, 2(7):AIoa2401033, 2025. doi: 10.1056/AIoa2401033.

Dingkun Liu, Zhu Chen, Jingwei Luo, Shijie Lian, and Dongrui Wu. Mirepnet: A pipeline and foundation model for eeg-based motor imagery classification, 2025.

Matthew Luciw, Ewa Jarocka, and Benoni Edin. Multi-channel EEG recordings during 3,936 grasp and lift trials with varying weight and friction. *Scientific data*, 1:140047, 11 2014. doi: 10.1038/sdata.2014.47.

Perrin Margaux, Emmanuel Maby, Sébastien Daligault, Olivier Bertrand, and Jérémie Mattout. Objective and Subjective Evaluation of Online Error Correction during P300-Based Spelling. *Advances in Human-Computer Interaction*, 2012, 01 2012. doi: 10.1155/2012/578295.

Dennis J. McFarland, Charles W. Anderson, Klaus-Robert Müller, Alois Schlögl, and Dean J. Krusienski. BCI meeting 2005-workshop on BCI signal processing: feature extraction and translation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(2):135–138, 2006. doi: 10.1109/TNSRE.2006.875637.

David Mizrahi, Roman Bachmann, Oguzhan Fatih Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4M: Massively Multimodal Masked Modeling. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Jennifer J Newson and Tara C Thiagarajan. EEG frequency bands in psychiatric disorders: A review of resting state studies. *Front. Hum. Neurosci.*, 12:521, 2018.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG - adapting to any setup with large-scale pretraining on 25,000 subjects. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. doi: 10.48550/arXiv.2510.21585.

Yuri G. Pavlov, Dauren Kasanov, Alexandra I. Kosachenko, and Alexander I. Kotyusov. "EEG, pupillometry, ECG and photoplethysmography, and behavioral data in the digit span task and rest", 2022.

Stefano Pizzolato, Luca Tagliapietra, Matteo Cognolato, Monica Reggiani, Henning Müller, and Manfredo Atzori. Comparison of six electromyography acquisition setups on hand movement classification tasks. 12(10):e0186132. ISSN 1932-6203. doi: 10.1371/journal.pone.0186132.

Antônio H Ribeiro, Manoel Horta Ribeiro, Gabriela M M Paixão, Derick M Oliveira, Paulo R Gomes, Jéssica A Canazart, Milton P S Ferreira, Carl R Andersson, Peter W Macfarlane, Wagner Meira, Jr, Thomas B Schön, and Antonio Luiz P Ribeiro. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat. Commun.*, 11(1):1760, April 2020.

Sasha Salter, Richard Warren, Collin Schlager, Adrian Spurr, Shangchen Han, Rohin Bhasin, Yujun Cai, Peter Walkington, Anuoluwapo Bolarinwa, Robert Wang, et al. emg2pose: A large and diverse benchmark for surface electromyographic hand pose estimation. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, Fernando Vaquerizo-Villar, and Roberto Hornero. EEG-Inception: A Novel Deep Convolutional Neural Network for Assistive ERP-Based Brain-Computer Interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(12):2773–2782, 2020. doi: 10.1109/TNSRE.2020.3048106.

Yasar Sattar and Lovely Chhabra. Electrocardiogram. In *StatPearls [Internet]*. StatPearls Publishing, Treasure Island (FL), Jun 2023. Updated Jun 5, 2023.

Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. BCI2000: a general-purpose brain-computer interface (BCI) system. *IEEE Trans. Biomed. Eng.*, 51(6):1034–1043, June 2004. doi: 10.1109/TBME.2004.827072.

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.

Vinit Shah, Eva von Weltin, Silvia López, James Riley McHugh, Lily Veloso, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. The Temple University Hospital Seizure Detection Corpus, 2018.

Viswanath Sivakumar, Jeffrey Seely, Alan Du, Sean R Bittner, Adam Berenzweig, Anuoluwapo Bolarinwa, Alexandre Gramfort, and Michael I Mandel. emg2qwerty: A large dataset with baselines for touch typing using surface electromyography. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 91373–91389. Curran Associates, Inc., 2024. doi: 10.52202/079017-2899.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG Conformer: Convolutional Transformer for EEG Decoding and Visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023. doi: 10.1109/TNSRE.2022.3230250.

Mastaneh Torkamani-Azar, Sumeyra Demir Kanik, Serap Aydin, and Mujdat Cetin. Prediction of Reaction Time and Vigilance Variability from Spatiospectral Features of Resting-State EEG in a Long Sustained Attention Task, 2019.

Edgar P. Torres, Edgar A. Torres, Myriam Hernández-Álvarez, and Sang Guun Yoo. EEG-Based BCI Emotion Recognition: A Survey. *Sensors*, 20(18), 2020. ISSN 1424-8220. doi: 10.3390/s20185083.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, 2023.

Logan Trujillo. Raw EEG Data, 2020.

Logan T. Trujillo, Candice T. Stanfield, and Ruben D. Vela. The Effect of Electroencephalogram (EEG) Reference Choice on Information-Theoretic Measures of the Complexity and Integration of EEG Signals. *Frontiers in Neuroscience*, Volume 11 - 2017, 2017. ISSN 1662-453X. doi: 10.3389/fnins.2017.00425.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, 2017.

Lily Veloso, James R. McHugh, Eva von Weltin, Silvia López, Iyad Obeid, and Joseph Picone. Big data resources for EEGs: Enabling deep learning research. In *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–3, 2017. doi: 10.1109/SPMB.2017.8257044.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Sci. Data*, 7(1):154, May 2020.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained Transformer for Universal and Reliable Representation of EEG Signals. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

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 *The Thirteenth International Conference on Learning Representations*, 2025. doi: 10.48550/arXiv.2412.07236.

Xiaoxi Wei, A Aldo Faisal, Moritz Grosse-Wentrup, Alexandre Gramfort, Sylvain Chevallier, Vinay Jayaram, Stylianos Bakas Camille Jeunet, Siegfried Ludwig, Konstantinos Barmpas, Mehdi Bahri, Yannis Panagakis, Nikolaos Laskaris, Dimitrios A Adamos, Stefanos Zafeiriou, William C Duong, Stephen M Gordon, Vernon J Lawhern, Maciej Śliwowski, Vincent Rouanne, and Piotr Tempczyk. 2021 BEETL Competition: Advancing Transfer Learning for Subject Independence & Heterogenous EEG Data Sets, 2022.

Qinfan Xiao, Ziyun Cui, Chi Zhang, SiQi Chen, Wen Wu, Andrew Thwaites, Alexandra Woolgar, Bowen Zhou, and Chao Zhang. Brainomni: A brain foundation model for unified EEG and MEG signals. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Tao Xu, Yun Zhou, Zi Wang, and Yixin Peng. Learning Emotions EEG-based Recognition and Brain Activity: A Survey Study on BCI for Intelligent Tutoring System. *Procedia Computer Science*, 130:376–382, 2018. ISSN 1877-0509. doi: j.procs.2018.04.056. The 9th International Conference on Ambient Systems, Networks and Technologies (ANT 2018) / The 8th International Conference on Sustainable Energy Information Technology (SEIT-2018) / Affiliated Workshops.

Chaoqi Yang, M Westover, and Jimeng Sun. BIOT: Biosignal Transformer for Cross-data Learning in the Wild. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 78240–78260. Curran Associates, Inc., 2023.

Jianwei Zheng, Hangyuan Guo, and Huimin Chu. A large scale 12-lead electrocardiogram database for arrhythmia study. *PhysioNet*, August 2022. doi: 10.13026/wgex-er52. Version 1.0.0.

Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. CSBrain: A cross-scale spatiotemporal brain foundation model for EEG decoding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

A Design of the Phase Term in the Phase-Aware Loss

We describe the benefits of using a sine-cosine representation of the Fourier phase ϕ in the loss function, instead of the raw phase values.

Other tokenizer architectures (e.g LaBraM [Jiang et al., 2024]) compute phase loss directly as the squared error between predicted $\hat{\phi}$ and true ϕ:

$$\mathcal{L}_\phi = ||\hat{\phi} - \phi||^2. \tag{7}$$

This formulation, however, is discontinuous at the boundaries of the phase domain $[-\pi, \pi]$. For instance, with $\hat{\phi} = \pi - \epsilon$ and $\phi = -\pi + \epsilon$ for small $\epsilon > 0$,

$$\mathcal{L}_\phi = ||\hat{\phi} - \phi||^2 = (2\pi - 2\epsilon)^2. \tag{8}$$

The loss approaches $4\pi^2$, even though the two angles are nearly identical. Thus, small phase shifts near the $\pm\pi$ boundary (e.g., 2° between 179° and -179°) result in large discontinuous jumps in the loss, leading to unstable gradients and poor convergence.

To address this issue, we map phase ϕ to its sine and cosine components, a point on the unit circle:

$$r(\phi) = (\sin(\phi), \cos(\phi)) \in \mathbb{R}^2, \tag{9}$$

The corresponding loss is:

$$\mathcal{L}_{r(\phi)} = ||\sin(\hat{\phi}) - \sin(\phi)||^2 + ||\cos(\hat{\phi}) - \cos(\phi)||^2 = 2 - 2\cos(\hat{\phi} - \phi) \tag{10}$$

Our formulation is continuous across the $\pm\pi$ boundary. Revisiting the earlier example with $\hat{\phi} = \pi - \epsilon$ and $\phi = -\pi + \epsilon$,

$$\mathcal{L}_{r(\phi)} = 2 - 2\cos(2\epsilon) \approx 4\epsilon^2 \tag{11}$$

which smoothly approaches 0 as $\epsilon \to 0$.

While $\mathcal{L}_{r(\phi)}$ resolves the discontinuity problem, it creates a gradient issue when the prediction is maximally wrong on the unit circle, i.e $\hat{\phi} \approx \phi \pm \pi$. Consider the gradient with respect to $\hat{\phi}$:

$$\frac{\partial}{\partial \hat{\phi}} \mathcal{L}_{r(\phi)} = 2\sin(\hat{\phi} - \phi) \tag{12}$$

When $\hat{\phi} \approx \phi \pm \pi$, $\sin(\hat{\phi} - \phi) \approx 0$. This means the worst predictions receive weak learning update.

Instead, our architecture predicts two values $\hat{c}$ and $\hat{s}$ (see Section 3.2), corresponding to estimates of $\cos\phi$ and $\sin\phi$ respectively. To avoid the issues relating to $\mathcal{L}_{r(\phi)}$ but keeping its benefits, we introduce a joint optimization of $\hat{c}$ and $\hat{s}$ with the phase-aware loss term $\mathcal{L}_{\text{phase}}$, since both values refer to the same underlying phase angle. $\mathcal{L}_{\text{phase}}$ optimizes the directional alignment between the predicted vector $(\hat{c}, \hat{s})$ and the target $(\cos\phi, \sin\phi)$ via cosine similarity. Since the target is a unit vector by definition ($\cos^2\phi + \sin^2\phi = 1$), the loss simplifies to:

$$\mathcal{L}_{\text{phase}} = 1 - \sum_i \frac{\hat{c}_i \cos\phi_i + \hat{s}_i \sin\phi_i}{\sqrt{\hat{c}_i^2 + \hat{s}_i^2}} \tag{13}$$

The gradient of $\mathcal{L}_{\text{phase}}$ with respect to the predicted vector is well-behaved across the full range of phase errors: it is proportional to the component of the predicted vector orthogonal to the target direction, scaled by the inverse of the predicted norm. This means that (*i*) maximally wrong predictions produce large gradients and (*ii*) the loss jointly optimizes direction rather than treating $\hat{c}$ and $\hat{s}$ as independent quantities.

Although cosine similarity is scale-invariant and does not explicitly constrain $(\hat{c}, \hat{s})$ to lie on the unit circle (see Figure 5), the predicted norms tend to remain close to unity in practice. Let $r = \sqrt{\hat{c}^2 + \hat{s}^2}$ denote the predicted norm. The gradient of $\mathcal{L}_{\text{phase}}$ with respect to $\hat{c}$ is:

$$\frac{\partial \mathcal{L}_{\text{phase}}}{\partial \hat{c}} = -\frac{1}{r}\left(\cos\phi - \frac{\hat{c}\cos\phi + \hat{s}\sin\phi}{r^2}\hat{c}\right) \tag{14}$$



(a) Prediction off the unit circle.

(b) Prediction on the unit circle.

Figure 5: Visualization of phase predictions off (a) and on (b) the unit circle.

The $1/r$ scaling optimizes direction without creating incentives to grow or shrink the predicted norm. In practice, predicted norms remain close to unity throughout training. This is primarily because the decoder's $\tanh$ output activation bounds $\hat{c}$ and $\hat{s}$ to $[-1, 1]$, constraining $r \leq \sqrt{2}$ and preventing norm divergence. As the predictions converge toward the targets $(\cos\phi, \sin\phi)$, the norm naturally approaches 1. Nonetheless, an optional regularization term can be added to explicitly penalize deviations from unit circle:

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{circle}} \cdot \sum_{i} \left(\hat{c}_i^2 + \hat{s}_i^2 - 1\right)^2 \tag{15}$$

However, this introduces a trade-off: the regularization gradient is radial (adjusting magnitude), while the cosine similarity gradient is tangential (adjusting direction). Any nonzero λ_{circle} diverts part of each optimization step toward magnitude correction rather than directional alignment, slightly reducing the efficiency of phase learning. Our sensitivity analysis (Appendix B) confirms this: $\lambda_{\text{circle}} = 0.0$ achieves the lowest reconstruction MSE, with performance degrading as λ_{circle} increases. In our main results, however, we retain $\lambda_{\text{circle}} = 0.1$ as a conservative choice that provides an explicit geometric guarantee at negligible cost, ensuring that predicted phase vectors remain valid across the diverse signal distributions encountered when training on multiple datasets and modalities. In summary, our approach combines the strengths of all formulations while avoiding their respective weaknesses:

	Continuous at $\pm\pi$	No vanishing gradients	Joint optimization	Unit-circle
$\mathcal{L}_{\phi}$	✗	✓	–	–
$\mathcal{L}_{r(\phi)}$	✓	✗	✗	✓
$\mathcal{L}_{\text{phase}}$	✓	✓	✓	$\approx^{*}$
$\mathcal{L}_{\text{phase}} + \mathcal{L}_{\text{reg}}$	✓	✓	✓	✓*

*Approximate unit-circle enforcement via $\tanh$ activation.

$\mathcal{L}_{\text{reg}}$ provides exact enforcement but introduces a convergence trade-off (Appendix B).

B Architectural and Training Components Analysis

B.1 Individual Component Contributions

As described in Section 3.2, the NEURORVQ training objective consists of multiple components. To quantify the individual contribution of each architectural and training component, we conduct the following ablation studies using the datasets described in Section 4.2. We start from a simple baseline: a single-branch, single-codebook architecture that naively reconstructs the Fourier amplitude and phase using MSE (as discussed in Appendix A). We progressively introduce to this baseline each NEURORVQ training objective component.

Table 2 reports the validation MSE for each configuration across all three modalities. Each component contributes to reconstruction quality, with consistent trends across EEG, ECG, and EMG:

Table 2: Individual training component contribution analysis. Starting from a simple baseline, each NEURORVQ training component is added incrementally. Val MSE for each modality is averaged over the last 10 training epochs.

Row	Configuration	EEG	ECG	EMG
1.	1-Branch, 1CB, MSE Phase + MSE A	1.509	1.956	1.996
2.	4-Branch, 8CB, MSE Phase + MSE A	0.946	1.417	1.651
3.	4-Branch, 8CB, MSE Phase + $\log(A)$ + Temporal	0.603	0.843	1.000
4.	4-Branch, 8CB, Phase Loss (without $\mathcal{L}_{\text{reg}}$) + MSE A	0.167	0.287	0.903
5.	4-Branch, 8CB, Phase Loss (without $\mathcal{L}_{\text{reg}}$) + $\log(A)$ + Temporal	0.035	0.115	0.447

(1) **Multi-scale architecture** (rows 1→2): Scaling from 1 branch with 1 codebook to 4 branches with 8 codebooks per branch reduces MSE by 37% for EEG (1.509→0.946), 28% for ECG (1.956→1.417) and 17% for EMG (1.996→1.651), confirming that decomposing the signal into frequency-specific representations with deeper residual quantization leads to more faithful discrete approximations across modalities.

(2) **Log-amplitude and temporal losses** (rows 2→3): Replacing raw amplitude MSE with the log-amplitude formulation $\log(1 + A)$ and adding the temporal reconstruction term reduces EEG MSE from 0.946 to 0.603, ECG from 1.417 to 0.843 and EMG from 1.651 to 1.000. The log transform compresses the dynamic range and places greater emphasis on high-frequency components, while the temporal loss provides complementary time-domain guidance.

(3) **Phase loss alone** (rows 2→4): Replacing the independent MSE phase loss with the cosine-similarity-based $\mathcal{L}_{\text{phase}}$ (see Appendix A), without any temporal or log-amplitude changes, substantially reduces MSE for EEG (0.946→0.167), ECG (1.417→0.287) and EMG (1.651→0.903), demonstrating that respecting the circular topology of phase is critical for reconstruction quality.

(4) **Full combination. Phase-Aware Loss** (rows 3→5): Combining all components produces the largest single-step improvement across all modalities: EEG drops from 0.603 to 0.035 (17×), ECG from 0.843 to 0.115 (7.3×) and EMG from 1.000 to 0.447 (2.2×). This dramatic gain demonstrates a synergy between the phase and the temporal terms: the temporal term provides stable time-domain gradients that enable the phase term to fully exploit its directional optimization.

In total, the full NEURORVQ configuration achieves a **43**× reduction in EEG reconstruction error (1.509→0.035), **17**× for ECG (1.956→0.115) and **4.5**× for EMG (1.996→0.447) relative to the baseline. These results demonstrate that architectural contributions and training loss design are both essential and synergistic: while the multi-scale structure provides consistent gains, the largest improvements emerge from the phase-aware loss, underscoring that principled loss design is as important as model architecture for high-fidelity biosignal tokenization.

B.2 Sensitivity Analysis

The phase-aware loss (see Section 3.2) follows a principled design grounded in signal processing. As shown in Appendix A, raw phase MSE is fundamentally flawed. We address this by representing phase as 2D vectors $(\cos\phi, \sin\phi)$ and use the cosine similarity to jointly optimise the predicted values $\hat{c}$ and $\hat{s}$. Although cosine similarity is scale-invariant and does not explicitly constrain $(\hat{c}, \hat{s})$ to lie on the unit circle, the predicted norms tend to remain close to unity in practice (see Appendix A). Nonetheless, as discussed, an optional regularization term can be added to explicitly penalize deviations from unit circle (see Appendix A). To determine the sensitivity of reconstruction quality to λ_{circle}, we conduct an ablation across all three modalities with the full NEURORVQ architecture (4 branches, 8 codebooks per branch), varying λ_{circle} from 0 (cosine similarity only, no unit-circle penalty) to 1.0 and using the datasets described in Section 4.2. Table 3 reports the validation MSE for each setting.

Even without the unit-circle penalty ($\lambda_{\text{circle}} = 0$), the cosine similarity alone achieves the lowest validation MSE across all modalities, already far below the MSE phase baselines (Table 2), confirming that the core improvement comes from replacing independent phase MSE with directional alignment

Table 3: Sensitivity analysis of the unit-circle penalty weight λ_{circle}. Val MSE is averaged over the last 10 training epochs.

λ_{circle}	EEG	ECG	EMG
0.0	0.035	0.115	0.447
0.1	0.050	0.144	0.530
0.4	0.106	0.144	0.534
0.8	0.107	0.180	0.528
1.0	0.111	0.143	0.525

via cosine similarity. For EEG, increasing λ_{circle} degrades reconstruction quality monotonically. For ECG and EMG, performance is relatively stable across $\lambda_{\text{circle}} \in [0.1, 1.0]$ but slightly worse than $\lambda_{\text{circle}} = 0.0$, indicating that stronger penalty values over-constrain the optimization, as discussed in Appendix A. In our main results, however, we retain $\lambda_{\text{circle}} = 0.1$ as a conservative choice that provides an explicit geometric guarantee at small cost.

C Temporal Encoder Configuration & Tokenizer Hyperparameter Settings

This section details: (*i*) the configuration of NEURORVQ's multi-scale temporal encoder (§C.1) for each modality, and (*ii*) the hyperparameter settings used for training the tokenizer (§C.2).

C.1 Multi-Scale Temporal Encoder Architecture

Our temporal encoder comprises four temporal branches. Each branch applies the following sequence of operations twice in succession (Figure 1): a 1-D convolution, followed by group normalization with $G = 4$ groups and $C = 8$ channels, GELU activation, and pooling. As described in Section 4, the signals for each modality are resampled to different sampling frequencies to maintain the necessary signal properties. Therefore, the temporal convolution kernels are chosen to span receptive fields aligned with the characteristic of each modality, with larger kernels broader morphological features and narrower kernels sensitive to capture the fast transients. Tables 4 and 5 list the filter counts, kernel sizes, padding and pooling parameters for the two sequences, where $x \to y$ denotes the configuration value(s) used in the first and second sequence, respectively. Since both EEG and ECG modalities are resampled at the same rate (see Section 4), same kernel sizes are used.

Table 4: Configuration of NEURORVQ's multi-scale temporal encoder for EEG and ECG modalities. Here, $x \to y$ indicates values for the first (x) and second (y) sequence.

Branch №	Filters	Kernel	Padding	Pooling
1	$8 \to 8$	$(1, 21) \to (1, 9)$	$(0, 10) \to (0, 4)$	$(1, 2) \to (1, 4)$
2	$8 \to 8$	$(1, 15) \to (1, 7)$	$(0, 7) \to (0, 3)$	$(1, 2) \to (1, 4)$
3	$8 \to 8$	$(1, 9) \to (1, 5)$	$(0, 4) \to (0, 2)$	$(1, 2) \to (1, 4)$
4	$8 \to 8$	$(1, 5) \to (1, 3)$	$(0, 2) \to (0, 1)$	$(1, 2) \to (1, 4)$

Table 5: Configuration of NEURORVQ's multi-scale temporal encoder for EMG modality. Here, $x \to y$ indicates values for the first (x) and second (y) sequence.

Branch №	Filters	Kernel	Padding	Pooling
1	$8 \to 8$	$(1, 51) \to (1, 25)$	$(0, 25) \to (0, 12)$	$(1, 2) \to (1, 4)$
2	$8 \to 8$	$(1, 17) \to (1, 9)$	$(0, 8) \to (0, 4)$	$(1, 2) \to (1, 4)$
3	$8 \to 8$	$(1, 8) \to (1, 4)$	$(0, 4) \to (0, 2)$	$(1, 2) \to (1, 4)$
4	$8 \to 8$	$(1, 5) \to (1, 3)$	$(0, 2) \to (0, 1)$	$(1, 2) \to (1, 4)$

C.2 Tokenizer Hyperparameter Settings

Table 6 summarises the hyperparameters used for training the tokenizer as described in Section 3.2. Certain choices (e.g., batch size) were constrained by hardware, while others were guided by ablation studies and scaling analyses (see Section 4.2).

Table 6: NEURORVQ Tokenizer Hyperparameters

Hyperparameter	EEG	ECG	EMG
Number of tokens in $\mathcal{V}_i$	8192	8192	8192
Token Dimension	128	128	128
Unit-Circle Penalty λ_{circle}	0.1	0.1	0.1
Temporal Branches S	4	4	4
Number of codebooks N per branch	8	8	16
Batch size	256	256	256
Learning rate scheduler	Cosine	Cosine	Cosine
Base learning rate	5e-5	5e-5	5e-5
Min learning rate	1e-5	1e-5	1e-5
Total epochs	100	100	100
Warmup epochs	10	10	10
Optimizer	AdamW	AdamW	AdamW
Weight decay	1e-4	1e-4	1e-4
Adam β	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)
Gradient clipping	3	3	3
Layer scale init	0.001	0.001	0.001
Encoder depth	12	12	12
Decoder depth	3	3	3
Hidden dimension	200	40	200
No. Attention heads	10	10	10

D Scaling Analysis

Figures 6 and 7 provide the individual scaling analysis plots that are combined in Figure 2 in Section 4.

Figure 6 shows the validation MSE as a function of codebook depth for each branch configuration, and Figure 7 shows the corresponding number of trainable parameters. Together, these plots illustrate that reconstruction quality improves with both codebook depth and number of branches, while the parameter cost scales linearly with codebook depth and proportionally with the number of branches.

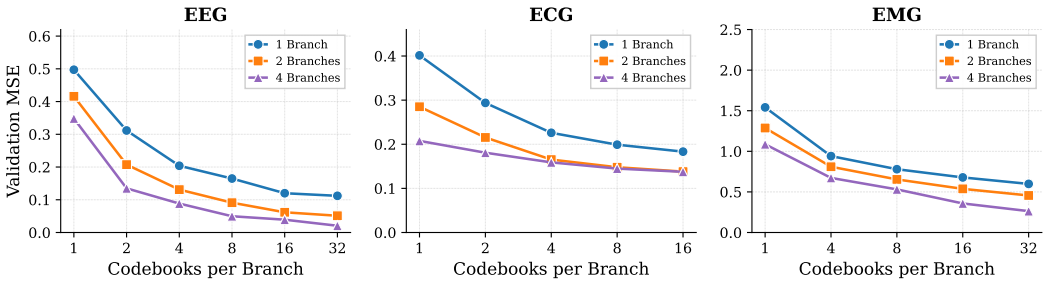

Figure 6: Mean squared error (MSE) reconstruction performance for NEURORVQ with various numbers of temporal branches and RVQ codebooks (CB) per branch across three modalities (EEG, ECG, EMG).

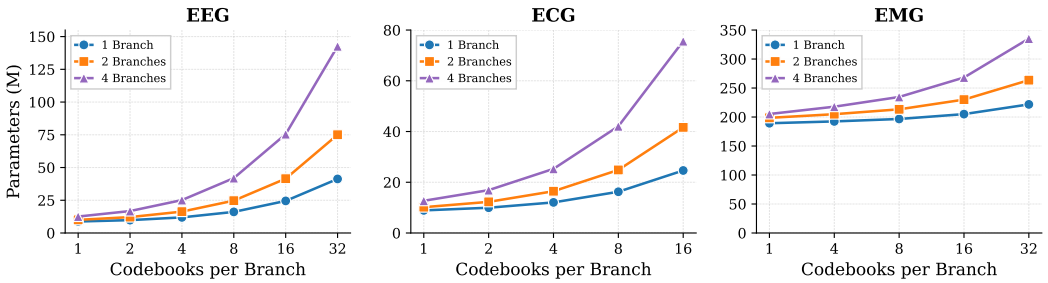

Figure 7: Number of trainable parameters for NEURORVQ with various numbers of temporal branches and RVQ codebooks (CB) per branch across three modalities (EEG, ECG, EMG).

E Codebook Utilization Analysis

Codebook utilization is an important property for an RVQ-based tokenizer. Using the held-out validation sets (Section 4.2) per modality (EEG, ECG, EMG), the following metrics across all branches (4 for all modalities) and RVQ levels (8 for EEG/ECG and 16 for EMG) were computed: *(i)* utilization, *(ii)* perplexity, *(iii)* dead codes, *(iv)* top-1% mass (selection probability captured by the 1% most-used codes). The results are presented in Tables 7, 9 and 8. NEURORVQ-tokenizer shows efficient code usage with no collapse: EEG and EMG reach 96.9% and 99.2% mean utilization with near-uniform usage and only 257 and 62 dead codes per 8192-entry codebook respectively. Within each modality the coarse first RVQ level has the lowest utilization while deeper residual levels saturate toward 100%, with perplexity increasing and top-code concentration decreasing monotonically with depth. This behavior is a residual hierarchy: first levels produce coarse details and successive levels encode progressively finer detail spread across the full codebook. ECG shows lower utilization: this is expected and consistent with our scaling analysis (in Section 4.2), which shows that ECG's stereotyped spectrally simple morphology saturates with far fewer codebooks than EEG or EMG.

Table 7: Codebook utilization for EEG

RVQ level	Utilization	Perplexity	Dead codes	Top-1% mass
1	80.9%	6099	1566	0.039
2	97.1%	7562	234	0.029
3	97.7%	7668	188	0.026
4	99.5%	7815	40	0.025
5	100.0%	7846	3	0.024
6	100.0%	7802	1	0.022
7	99.9%	7748	6	0.021
8	99.8%	7740	18	0.018
Mean	**96.9%**	**7535**	**257**	**0.026**

Table 8: Codebook utilization for ECG

RVQ level	Utilization	Perplexity	Dead codes	Top-1% mass
1	29.2%	2060	5798	0.123
2	67.9%	4708	2631	0.068
3	54.3%	3846	3746	0.101
4	47.8%	3408	4280	0.100
5	53.0%	3805	3848	0.085
6	47.0%	3465	4342	0.091
7	47.0%	3544	4338	0.079
8	50.2%	3850	4076	0.060
Mean	**49.6%**	**3586**	**4132**	**0.089**

Table 9: Codebook utilization for EMG

RVQ level	Utilization	Perplexity	Dead codes	Top-1% mass
1	88.1%	6442	978	0.053
2	100.0%	7331	1	0.049
3	99.9%	7355	5	0.048
4	100.0%	7388	0	0.046
5	100.0%	7418	2	0.044
6	100.0%	7454	0	0.042
7	100.0%	7488	1	0.039
8	100.0%	7517	0	0.037
9	100.0%	7541	0	0.036
10	100.0%	7568	0	0.035
11	100.0%	7618	0	0.034
12	100.0%	7632	0	0.034
13	100.0%	7733	0	0.032
14	100.0%	7812	0	0.030
15	100.0%	7844	0	0.028
16	100.0%	7913	0	0.027
Mean	**99.2%**	**7503**	**62**	**0.038**

F Datasets

We use a collection of publicly available datasets to pre-train NEURORVQ for each biosignal modality. Tables 10, 11, 12 summarize the datasets, their associated BCI paradigms and the number of sensors. The collection spans multiple experimental settings, including hand pose estimation, touch typing, cardiac arrhythmia detection, motor imagery, event-related potentials (ERP), resting-state activity, information-integration and rule-based cognitive categorization (II-RB) and clinical data for epilepsy and seizures.

Table 10: Datasets used for pre-training NEURORVQ in EMG.

Datasets	Task	# channels
emg2pose [Salter et al., 2024]	Hand Pose Estimation	16
emg2qwerty [Sivakumar et al., 2024]	Touch Typing	16

Table 11: Datasets used for pre-training NEURORVQ in ECG.

Datasets	Task	# leads
PTB [Bousseljot et al., 2004]	Diagnostic Classification	15
MIMIC-IV-ECG [Gow et al., 2023]	Diagnostic Classification	12
Chapman-Shaoxing [Zheng et al., 2022]	Arrhythmia Classification	12

Table 12: Datasets used for pre-training NEURORVQ in EEG.

Datasets	BCI paradigm	# channels
BCI Competition IV-1 [Blankertz et al., 2007]	Motor	64
Grasp and Lift [Luciw et al., 2014]	Motor	32
Physionet MI [Schalk et al., 2004]	Motor	64
bi2015a [Korczowski et al., 2019]	ERP	32
Inria BCI Challenge [Margaux et al., 2012]	ERP	56
Trujillo [2020]	II-RB	64
Trujillo et al. [2017]	Resting	64
SPIS Resting [Torkamani-Azar et al., 2019]	Resting	64
TUAR [Buckwalter et al., 2021]	Artifacts (e.g., eye movement)	31
Siena Scalp [Detti et al., 2020]	Epilepsy	31
TUEP [Veloso et al., 2017]	Epilepsy	31
TUSZ [Shah et al., 2018]	Seizure	31
Self-Collected Dataset	Motor	29

In the EEG modality, we additionally include a proprietary dataset comprising approximately 235 hours of motor imagery data recorded with 29 channels. The data collection was approved by the Research Governance and Integrity Team of Imperial College London (approval number 22IC7505). Recordings used non-invasive scalp EEG. All participants received the study information below and gave informed consent before taking part:

> "We will record your brain waves while you engage in simple tasks. These include real and imagined movements of the extremities, and a game which uses movements as controls. The collected data will be used to train algorithms that can be used for many things, such as facilitating navigation in virtual environments, seamless user interaction with objects in computer gaming or even rehabilitation training of post-stroke patients towards regaining the ability to control their limbs. The project received approval from the Research Governance and Integrity Team. If you want to help out our research team you can register to be a participant in our experiment! There is a small gift of £30 to every participant."

G NEURORVQ Reconstruction Examples

In this section, we present a few reconstruction examples using the NEURORVQ tokenizer across all three modalities (EEG, ECG and EMG).

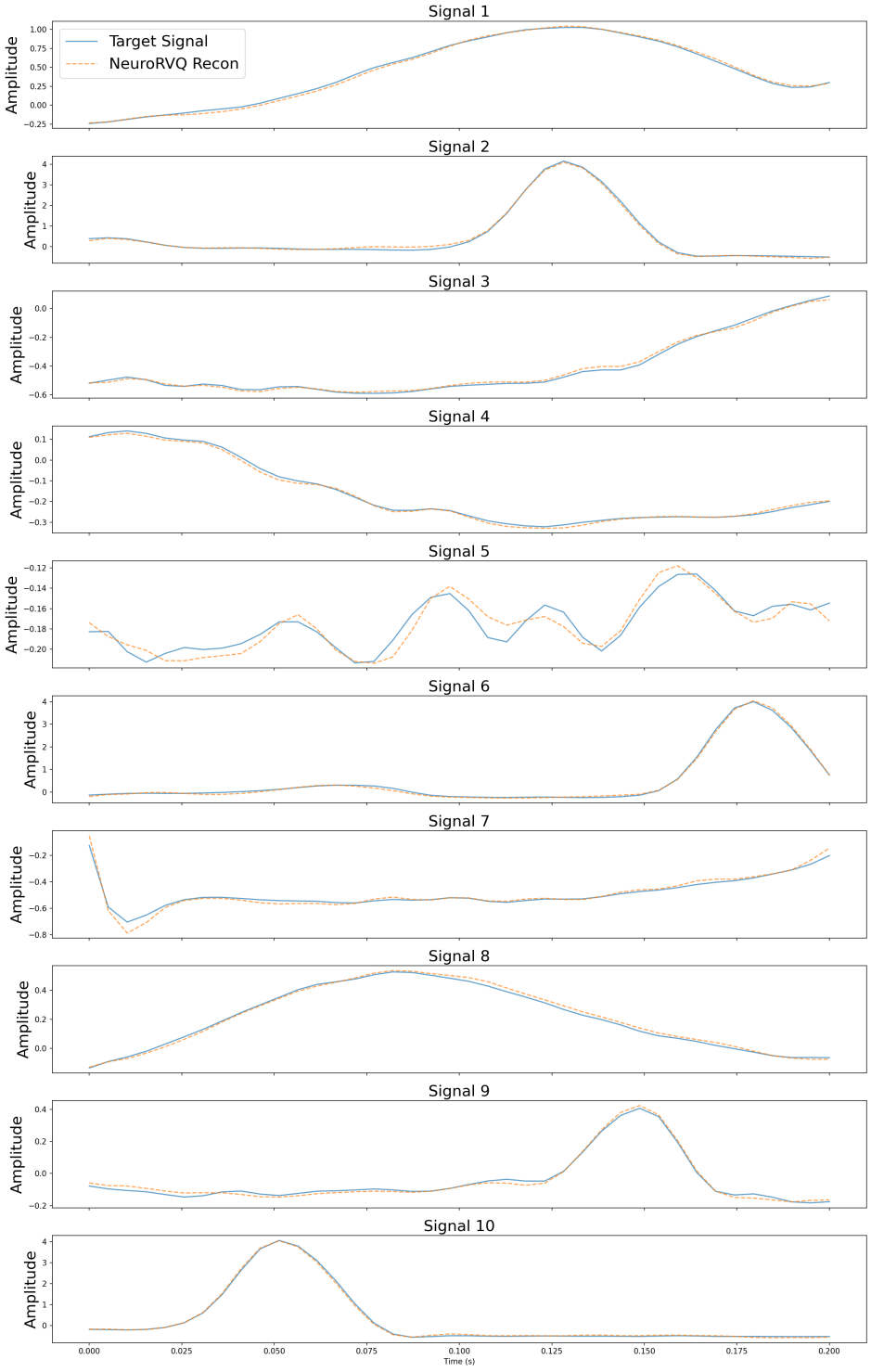

Figure 8: Reconstructed ECG signals (from the validation set) using NEURORVQ codebook-based tokenizer. Blue lines denote the input ECG signal and orange the reconstructed ECG signal.

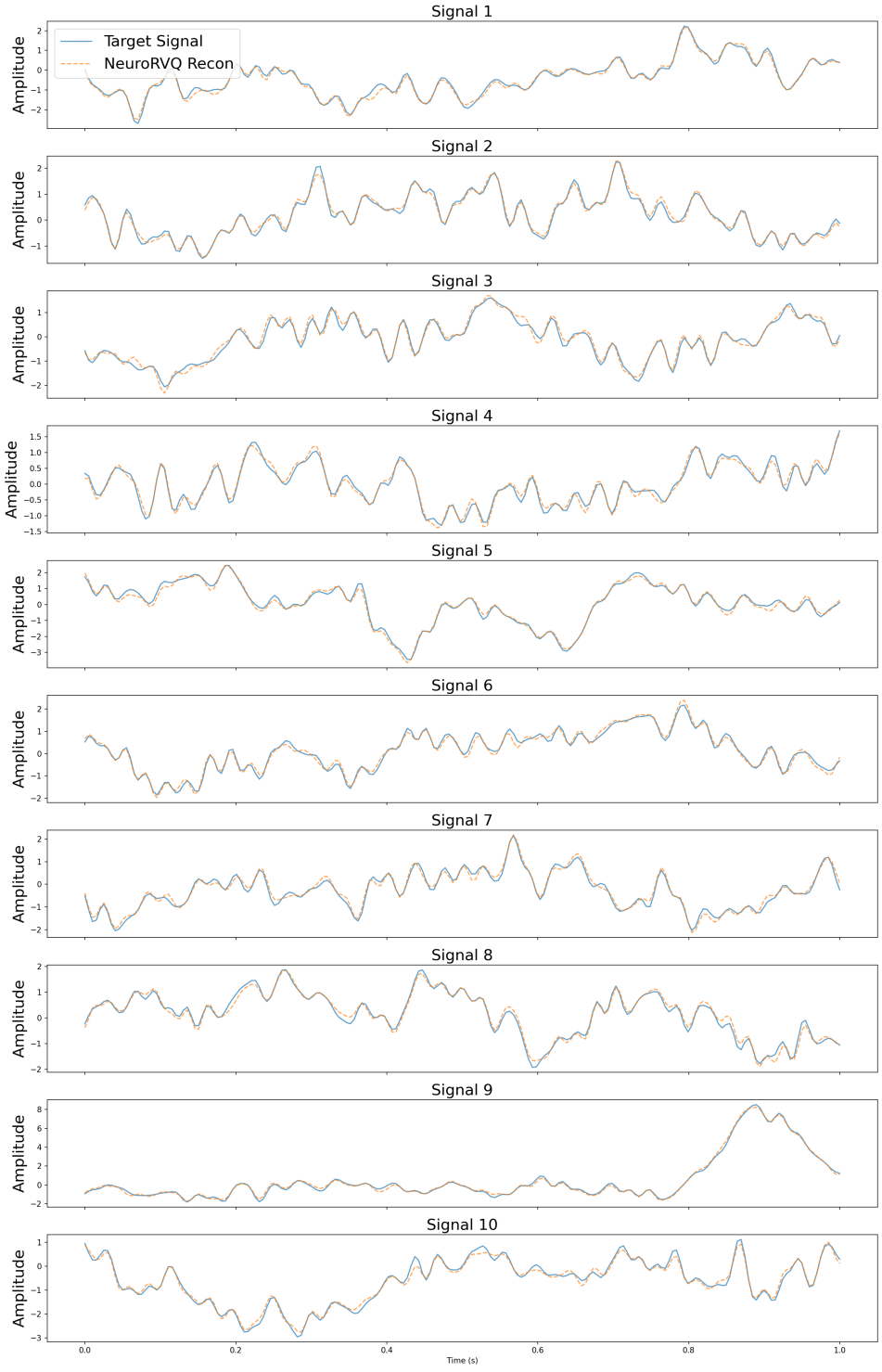

Figure 9: Reconstructed EEG signals (from the validation set) using NEURORVQ codebook-based tokenizer. Blue lines denote the input EEG signal and orange the reconstructed EEG signal.

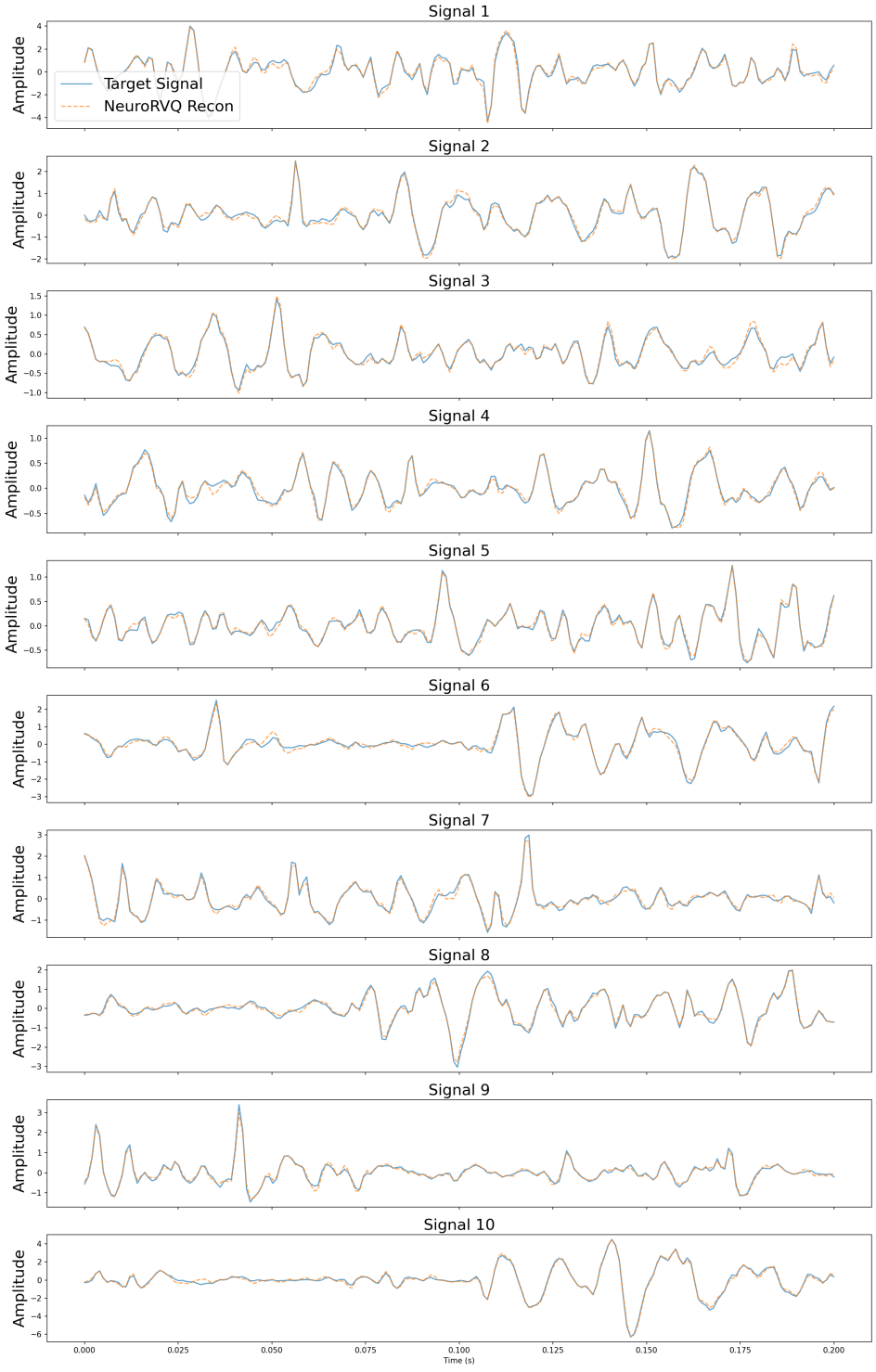

Figure 10: Reconstructed EMG signals (from the validation set) using NEURORVQ codebook-based tokenizer. Blue lines denote the input EMG signal and orange the reconstructed EMG signal.

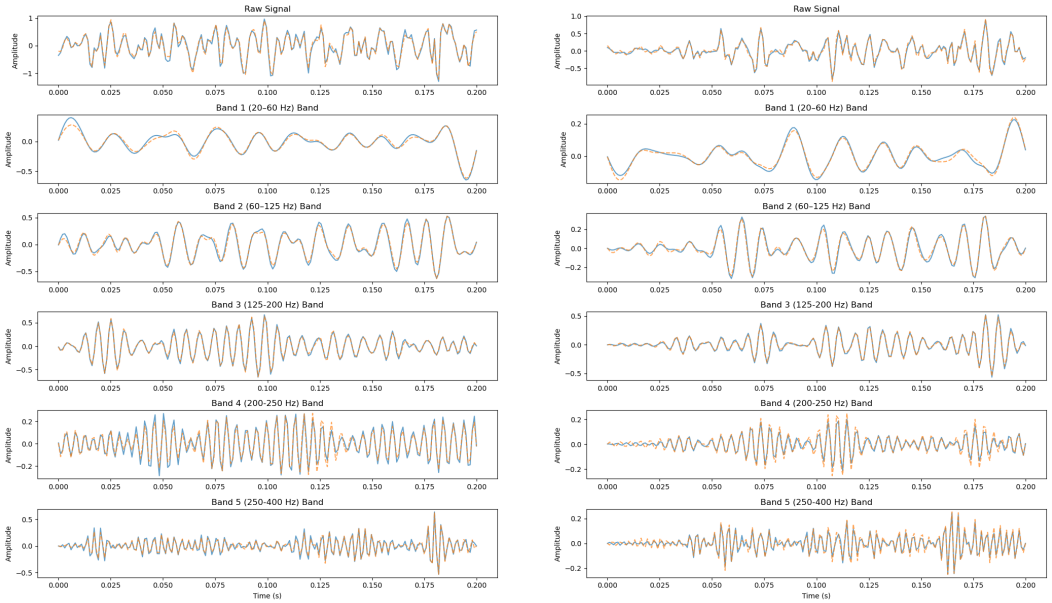

Figure 11: Per-band analysis of two sample reconstructed EMG signal (from the validation set) from NEURORVQ codebook-based tokenizer. Blue lines denote the input EMG signal and orange the reconstructed EMG signal.

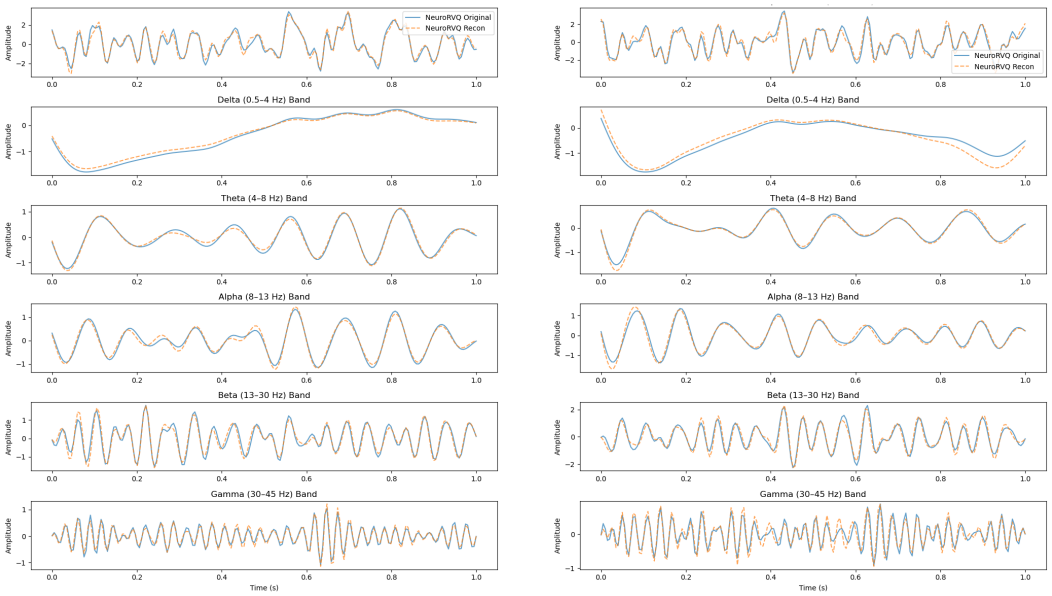

Figure 12: Per-band analysis of two sample reconstructed EEG signal (from the validation set) from NEURORVQ codebook-based tokenizer. Blue lines denote the input EEG signal and orange the reconstructed EEG signal.

H NEURORVQ Comparison with Pre-Trained Codebook-Based Tokenizers

For EEG, we compare the signal reconstruction capabilities of NEURORVQ against two pretrained open-source state-of-the-art codebook-based tokenizers: LaBraM [Jiang et al., 2024] and BrainOmni [Xiao et al., 2025]. For ECG and EMG, no unimodal codebook-based tokenizers currently exist; NEURORVQ is the first to demonstrate high-fidelity codebook-based reconstruction for these modalities, as shown in the validation curves in Figure 3 and in Appendix G.

Comparison with LaBraM. LaBraM employs a VQ-VAE-inspired tokenizer [Esser et al., 2020] with a single codebook. We evaluate both tokenizers on an unseen dataset during training: the High Gamma motor dataset [Schirrmeister et al., 2017]. To compare both reconstructions, both signal are brought to the time domain and z-scored per trial together with the input before computing MSE, so the comparison is scale-matched. However, LaBraM is pre-trained between 0.1–75 Hz (LaBraM-Setup) while NEURORVQ is pre-trained between 0.5–45 Hz (NEURORVQ-Setup). We performed both setup comparisons in the motor-trials for fair comparison. As shown in Table 13, NEURORVQ consistently outperforms LaBraM across both setups, demonstrating superior reconstruction generalization.

Table 13: Out-of-distribution MSE reconstruction performance for NEURORVQ and LaBraM tokenizers with both setups. NEURORVQ-Setup: 0.5-45Hz. LaBraM-Setup: 0.1-75Hz.

	LaBraM-Setup	NEURORVQ-Setup
LaBraM	0.598	0.855
NEURORVQ	**0.072**	**0.058**

To verify that these gains stem from architectural differences rather than training data, we also conducted an in-distribution comparison in which LaBraM and NEURORVQ were trained on the same datasets, with identical train-validation splits and training regime (see Appendix I). The results confirm that NEURORVQ's advantage is consistent regardless of the training data used.

Comparison with BrainOmni. BrainOmni adopts a single residual quantizer inspired by EnCodec [Défossez et al., 2023], whereas NEURORVQ employs multiple RVQ codebooks, one per frequency scale. BrainOmni is pre-trained between 0.1–96 Hz (BrainOmni-Setup) and uses different signal normalization strategy than NEURORVQ. Therefore, we evaluate on BrainOmni's publicly available sample data, which we use because their model requires device-specific layout information and we report results under both setups to ensure a fair comparison. As shown in Table 14 and Figure 13, NEURORVQ consistently outperforms BrainOmni under both normalization conditions.

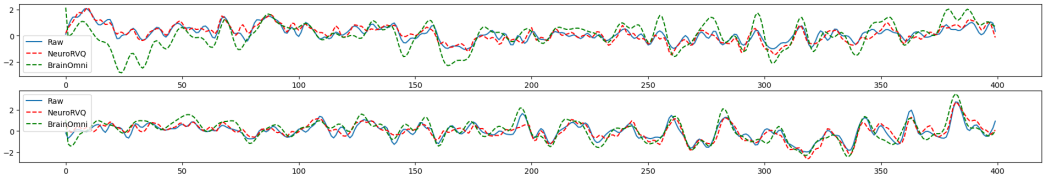

Figure 13: Two sample reconstructed EEG signals from BrainOmni and NEURORVQ codebook-based tokenizers. Blue lines denote the input EEG signal, green the reconstructed EEG signal for BrainOmni and red the reconstructed EEG signal for NEURORVQ.

Table 14: Out-of-distribution MSE reconstruction performance for NEURORVQ and BrainOmni tokenizers with both setups. NEURORVQ-Setup: 0.5-45Hz and NEURORVQ's normalization scheme. BrainOmni-Setup: 0.1-96Hz and BrainOmni's normalization scheme.

	NEURORVQ-Setup	BrainOmni-Setup
BrainOmni	0.350	0.413
NEURORVQ	**0.219**	**0.324**

I NEURORVQ vs. LaBraM: Reconstruction Performance

We train NEURORVQ and LaBraM tokenizers on **the same datasets with the same train–validation split** and an identical training regime (including preprocessing and training steps), as described in Section 5.1. As shown in Table 15 and Figure 14, NEURORVQ achieves orders of magnitude lower in-distribution reconstruction mean squared error (MSE) compared to LaBraM.

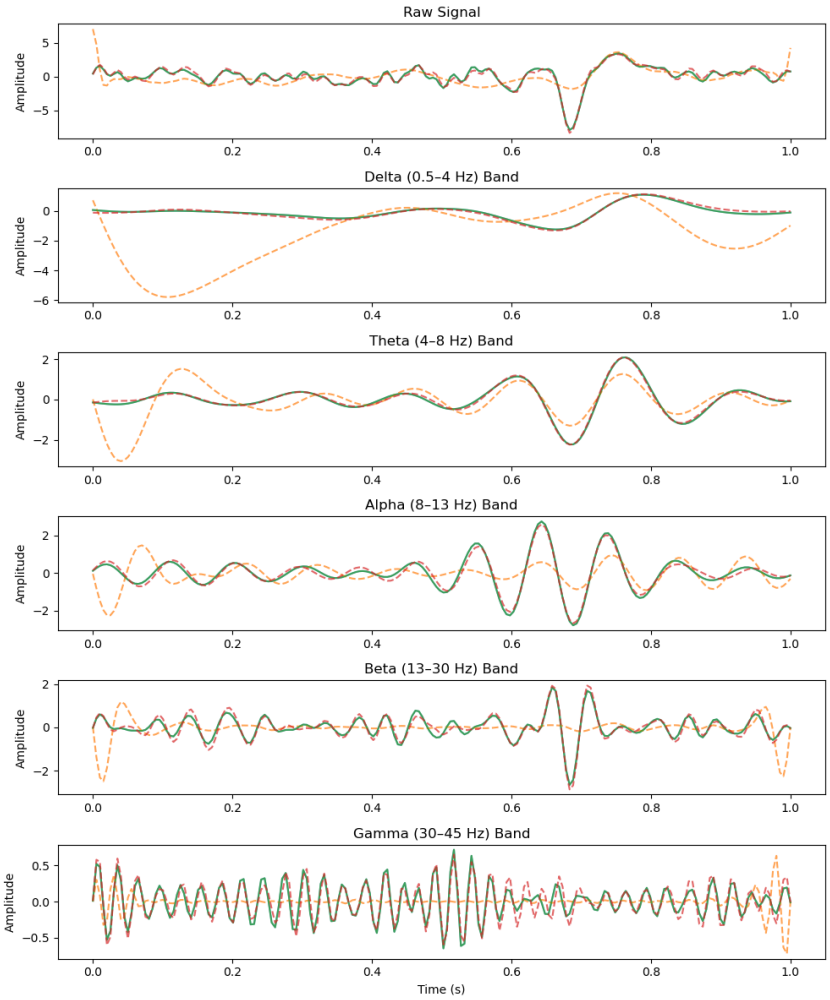

Figure 14: Per-band analysis of a sample reconstructed EEG signal (from the validation set) from LaBraM and NEURORVQ codebook-based tokenizers. Green lines denote the input EEG signal, orange the reconstructed EEG signal for LaBraM (ours) and red the reconstructed EEG signal for NEURORVQ.

Table 15: In-distribution validation mean squared error (MSE) reconstruction performance across frequency bands for NEURORVQ and LaBraM tokenizers trained on the same datasets.

	Raw Signal	Delta	Theta	Alpha	Beta	Gamma
LaBraM (ours)	1.071	1.561	0.184	0.099	0.122	0.020
NEURORVQ	**0.016**	**0.006**	**0.002**	**0.002**	**0.005**	**0.002**

J Foundation Model Configuration & Hyperparameter Settings

This section details the hyperparameter settings used for pre-training the foundation model as well as its fine-tuning on downstream tasks. NEURORVQ foundation model used the multi-scale temporal encoder configurations described in (§C.1).

Table 16: NEURORVQ-FM Hyperparameters. The rightmost column reports the fine-tuning protocol used for our 3-seed runs on Discrete Gestures and Wrist Velocity Regression.

Hyperparameter	Pre-training FM (EEG/ECG/EMG)	Finetuning (EEG/ECG)	Finetuning (EMG)	Finetuning (EMG: [Kaifosh et al., 2025])
Number of tokens in $\mathcal{V}_i$	8192	8192	8192	8192
Token Dimension	128	128	128	128
Unit-Circle Penalty λ_{circle}	0.1	0.1	0.1	0.1
Temporal Branches S	4	4	4	4
Number of codebooks N	8	8	16	16
Batch size	64	64	128	64
Learning rate scheduler	Cosine	Linear	Cosine	Cosine
Base learning rate	5e-4	5e-4	1e-3	5e-4
Min learning rate	1e-5	-	1e-6	0
Warmup lr start-end factors	-	(0.1, 1)	(0.01, 1.0)	(0.1, 1.0)
Lr start-end factors	-	(1, 0.1)	(1, 0.001)	-
Total epochs	50	20	100	15 (Gestures) / 20 (Wrist)
Warmup epochs	5	4	5	3
Optimizer	AdamW	AdamW	AdamW	AdamW
Weight decay	0.05	0.01	0.001	0.01
Adam β	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.95)	(0.9, 0.999)
Layer lr decay	-	0.975	-	-
Masking Ratio	0.5	-	-	-
Layer scale init	0.001	-	-	-
Encoder depth	12	12	12	12
Hidden dimension	200/40/200	200/40	200	200
No. Attention heads	10	10	10	10

Table 16 summarises the hyperparameters used for pre-training and fine-tuning the foundation model. Certain choices (e.g., batch size) were constrained by hardware, while others were guided by ablation studies (see Section 4.2).

K Downstream Task Fine-Tuning Analysis

In this section, we describe the baseline foundation models, datasets / benchmarks and implementation details for each modality. Full hyperparameter configurations are provided in Appendix J.

K.1 EEG downstream evaluation.

For EEG, we compare against nine foundation models on five downstream classification tasks from the benchmark of [Lee et al., 2025a]. These models include two models that use codebook-based tokenizer like NEURORVQ, namely LaBraM [Jiang et al., 2024] and BrainOmni [Xiao et al., 2025] and seven common foundation model baselines, namely NeuroGPT[Cui et al., 2024], CBraMod [Wang et al., 2025], EEGPT [Wang et al., 2024], BIOT [Yang et al., 2023], MIRepNet [Liu et al., 2025], CSBrain [Zhou et al., 2025], REVE [Ouahidi et al., 2026]. The five downstream classification tasks include: (*a*) Motor, four-class movement classification on High Gamma [Schirrmeister et al., 2017]; (*b*) ERP, binary event-related potential classification on Korea University [Kim et al., 2019]; (*c*) Memory, binary working memory classification on Pavlov et al. [2022]; (*d*) Sleep, six-class sleep stage classification on Sleep-EDF [Kemp et al., 2000]; and (*e*) Eyes, eyes-open vs. closed classification on PhysioNet Motor [Schalk et al., 2004]. This benchmark was selected for its diverse BCI paradigms and minimal spurious artifacts. All models used their original pre-processing

settings (for NEURORVQ-EEG 0.5-45Hz band pass filter), were fully fine-tuned for 20 epochs (early stopping of patience 5 epochs was employed to prevent overfitting), evaluated via 10-fold subject-independent cross-validation and compared using balanced accuracy. As shown in Table 1 (Section 5.2), NEURORVQ-EEG achieves the best balanced accuracy on four of five tasks and the highest mean balanced accuracy overall. Notably, NEURORVQ uses only 5.9M backbone parameters, substantially fewer than other models, e.g. NeuroGPT (79.5M). This combination of strong generalization and compact size underscores that the tokenizer, not model scale, is the primary driver of performance.

Ablation studies. To ensure that our downstream gains are attributable to tokenization differences rather dataset selection or model architecture, we conduct two ablation studies: *(i)* we follow the same controlled comparison protocol used for the tokenizer evaluation (Appendix I) where both NEURORVQ and LaBraM foundation models are trained on identical datasets with the same train–validation splits, preprocessing and training regime and *(ii)* a tokenizer swap ablation. The results, reported in Appendix L and Appendix N, confirms that high fidelity tokenizer translates to better downstream task performance under matched conditions.

K.2 ECG downstream evaluation.

For ECG, we compare against two foundation models, HuBERT-ECG [Coppola et al., 2024] and ECGFounder [Li et al., 2025], on the PTB-XL benchmark [Wagner et al., 2020]. PTB-XL is the most popular downstream dataset for ECG foundation models. It includes various labels covering Diagnostic, Form and Rhythm statements, but most models choose to use only a subset of these for multi-label classification. Following the structure of the original dataset Wagner et al. [2020], we evaluated models on two established ECG classification tasks covering different levels of diagnostic granularity: (i) a 5-class task based on the coarse diagnostic superclasses and (ii) a 43-class task using the full set of fine-grained diagnostic statements. We adopted the dataset's recommended 10-fold, patient-stratified splitting scheme, using the designated nine folds for model training and validation and the reserved tenth fold for final testing. In this way, we followed a unified benchmarking protocol that allows fair performance comparisons across all evaluated ECG foundation models. All models were fully fine-tuned for 20 epochs (with early stopping to prevent overfitting). Table 1 reports metrics for each classification task for open-source ECG foundation models. NEURORVQ achieves high balanced accuracy, doubling the BAcc of the next-best model on the fine-grained 43-class task, but ECGFounder leads on 5-class task.

K.3 EMG downstream evaluation.

We compared the performance of NEURORVQ-EMG against two foundation models, PhysioWave [Chen et al., 2025] and TinyMyo [Fasulo et al., 2026], on four downstream hand gesture classification tasks: i) Discrete Gestures Kaifosh et al. [2025]: a 9-class gesture classification task using 2s EMG signals recorded using a 16 channel wristband at 2000 Hz ii) EPN-612 [Benalcazar et al.]: a 6-class classification task from 5s EMG signals recorded with an 8-channel EMG armband iii) NinaPro DB5 [Pizzolato et al.]: A 52-class classification task from 5s EMG windows, recorded with two, 8-channel sEMG bands placed laterally adjacent on the forearm, iv) UCI-EMG [Krilova et al., 2018]: a 6-class classification from 1s windows of sEMG recorded using an 8-channel sEMG armband.

Unlike previous sEMG foundation model evaluations, we exclusively split datasets at the subject level into train, validation and test sets with ratios 7:1:2 (for the [Kaifosh et al., 2025] dataset, we used a 7:1.5:1.5). This difference contributes to a lower performance across models, especially on the NinaPro DB5 dataset (Table 1), compared to other works which train and test on different gesture repetitions within the same recording session. Another key difference on the NinaPro DB5 dataset is that existing evaluations included the recordings between gestures as a "no gesture" label, despite this causing large unresolved class imbalance as well as label leakage through padding operations and a lower distribution of recording duration.

Recordings were resampled to the expected sampling frequency of each model and preprocessed and normalized according to the steps outlined in each model paper: (*i*) NEURORVQ-EMG: 20-90 Hz band pass filtered; (*ii*) TinyMyo: 20-90 Hz band pass filter (Butterworth, 4th order) followed by a 50 Hz notch filter and per-channel z-score normalization; (*iii*) PhysioWave: No filtering, per-channel z-score normalization. The three models output an embedding per patch. For a fair comparison, we

applied a uniform aggregation strategy across all three models: patch embeddings were concatenated across the channel dimension and averaged across time before being passed to the classification head. We added a single linear layer classification head and fine-tuned each pretrained model on the four downstream datasets.

From scratch ablation study. To quantify the impact of the pretraining phase on model performance, we also trained NEURORVQ-EMG from scratch with randomly initialized weights on each of the four downstream datasets, using the same subject level train/test split as the finetuning evaluation. The results in Table 17 demonstrate that pretraining consistently improved performance across all datasets. The NinaPro DB5 dataset exhibits the largest improvement, showing that pretraining the NEURORVQ-EMG foundation model is most valuable when labeled data is limited relative to task complexity, exactly the setting foundation models are designed to address.

Table 17: NEURORVQ-EMG ablations. **Bold** values indicate best performance per column. All evaluations are subject-independent. Metrics: Accuracy, F1, Balanced Accuracy (BAcc) and CLER (Classification Error Rate).

	Discrete Gestures		**EPN-612**		**NinaPro DB5**		**UCI-EMG**	
Tuning	BAcc ↑	CLER ↓	Acc	F1	Acc	F1	Acc	F1
From Scratch	65.9 ± 0.6	34.3 ± 1.2	89.6 ± 0.2	89.7 ± 0.2	32.8 ± 1.8	30.6 ± 1.7	86.6 ± 1.9	86.5 ± 1.9
Full Finetune	$\mathbf{70.8 \pm 0.6}$	$\mathbf{27.6 \pm 0.8}$	$\mathbf{94.65 \pm 0.4}$	$\mathbf{94.7 \pm 0.5}$	$\mathbf{41.4 \pm 2.8}$	$\mathbf{38.8 \pm 3.7}$	$\mathbf{89.4 \pm 1.1}$	$\mathbf{89.3 \pm 1.1}$

L NEURORVQ vs. LaBraM: Downstream Task Performance

We compare the downstream task performance of NEURORVQ with LaBraM when both models are **pre-trained and fine-tuned on the same datasets** and with an identical training regime. We use a 10-fold subject-independent cross-validation to evaluate models on five classification benchmarks (Motor, Memory, Sleep, Eyes, ERP) as described in §K. Table 18 shows that, when all training settings are equal, NEURORVQ achieves higher balanced accuracy than LaBraM on all tasks (up to 13%), demonstrating the effectiveness of our tokenizer in training codebook-based models.

Table 18: Comparison of classification balanced accuracy between NEURORVQ and LaBraM, both pre-trained and finetuned on the same datasets using identical settings.

Model	Motor	ERP	Memory	Sleep	Eyes	Mean
LaBraM (ours)	0.570	0.819	0.565	0.715	0.805	0.695
NEURORVQ-EEG	**0.700**	**0.876**	**0.573**	**0.728**	**0.869**	**0.749**

M Frozen Backbone Downstream Task Performance

For each foundation model baseline, we use the original architecture implementation and pretrained weights with the backbone frozen, adding only a trainable classification head to the model's output embeddings / tokens according to their original implementations. CBraMod [Wang et al., 2025] and NeuroGPT [Cui et al., 2024] employ a 3-layer MLP as their classification heads, while the remaining EEG and ECG models use a single layer. In EMG, we used 1-second input window with trainable classification heads. For NEURORVQ, we use all tokens produced by the multi-branch encoder as flattened input to a linear classification layer. The results of classification-head-only training are reported in Tables 19, 20 and 21 using the evaluation schemes described in Section 5.2 and Appendix K. For EMG, we evaluated on 1-second windows.

Table 19: Frozen-backbone EEG classification results. Results are reported as mean balanced accuracy ± standard deviation. **Bold** and underlined values indicate best and second-best performance per column, respectively.

Model	Motor	Memory	Sleep	Eyes	ERP
BIOT	0.335 ± 0.039	0.528 ± 0.024	—	0.748 ± 0.046	0.500 ± 0.000
CBraMod	0.440 ± 0.077	0.537 ± 0.016	0.628 ± 0.044	0.800 ± 0.034	0.500 ± 0.000
EEGPT	0.331 ± 0.041	0.541 ± 0.030	**0.634 ± 0.043**	0.815 ± 0.031	0.523 ± 0.008
NeuroGPT	0.452 ± 0.047	0.500 ± 0.000	0.304 ± 0.010	0.656 ± 0.048	0.501 ± 0.002
REVE	0.574 ± 0.076	0.568 ± 0.030	0.616 ± 0.029	0.761 ± 0.053	0.728 ± 0.040
LaBraM	0.327 ± 0.029	0.500 ± 0.000	0.303 ± 0.008	0.733 ± 0.051	0.500 ± 0.000
NEURORVQ-EEG	**0.599 ± 0.044**	**0.590 ± 0.033**	0.624 ± 0.028	**0.823 ± 0.035**	**0.802 ± 0.038**

Table 20: Frozen-backbone ECG classification results. Results are reported as accuracy and balanced accuracy. **Bold** and underlined values indicate best and second-best performance per column, respectively.

Model	5-Class	
	Acc.	Bacc
HuBERT-ECG	54.94	35.55
ECGFounder	**75.06**	**63.66**
NEURORVQ-ECG	54.57	42.13

Table 21: Frozen-backbone EMG classification results. Results are reported as mean accuracy/F1-score ± standard deviation. **Bold** and underlined values indicate best and second-best performance per column, respectively.

Model	EPN-612		NinaPro DB5		UCI-EMG	
	Acc.	F1	Acc.	F1	Acc.	F1
PhysioWave	89.36 ± 0.16	89.43 ± 0.16	18.07 ± 0.91	16.71 ± 1.42	84.21 ± 1.57	84.01 ± 1.58
TinyMyo	75.28 ± 0.07	75.01 ± 0.09	21.60 ± 0.51	16.61 ± 0.49	83.67 ± 0.83	83.48 ± 0.93
NEURORVQ-EMG	**90.38 ± 0.04**	**90.41 ± 0.06**	**25.54 ± 0.68**	**21.91 ± 0.81**	**88.45 ± 1.06**	**88.45 ± 1.05**

N Tokenizer Swap Ablation

We examine the effect of same backbone, parameter budget, pretraining data, normalization and downstream protocol but different tokenizers. We use two tokenizers from §B, namely Row 2 in Table 2 (4-Branch, 8CB, MSE Phase + MSE A) and Row 5 in Table 2 (4-Branch, 8CB, Phase Loss + log(A) + Temporal), which share the same architecture and differ only in the training loss (val MSE 0.946 and 0.035 respectively). Using these two tokenizers, we pre-trained two NEURORVQ-EEG models on PhysioNet-MI [Schalk et al., 2004] and benchmarked both on the High Gamma Motor dataset [Schirrmeister et al., 2017]. We chose Motor because the tokenizers and consequently the foundation models are pre-trained on PhysioNet-MI (a motor dataset), and Motor is the downstream task most closely matched to that pretraining domain. Evaluation is under head-only (10-fold) and full-finetuning (5-fold) subject-wise cross-validation. In other words, **backbone architecture, parameter budget, pretraining data, normalization and downstream protocol are held identical. Only the tokenizer varies.** The results show that tokenization quality translates into downstream performance: balanced accuracy improves with tokenizer fidelity and the effect is clearest under full fine-tuning.

Table 22: Comparison of classification balanced accuracy between NEURORVQ-EEG on the Motor task pre-trained using various tokenizers.

Tokenizer used to train same backbone model (corresponding val MSE)	Head-Only	Full-Fine Tuning
4-Branch, 8CB, MSE Phase + MSE A (0.946)	56.5 (8.59)	58.5 (3.19)
4-Branch, 8CB, Phase Loss + log(A) + Temporal (0.035)	**58.9 (5.67)**	**64.0 (1.39)**

We extended this ablation study to other tokenizers from Section §4.2 in the EEG modality. To examine the relationship between number of codebooks and downstream performance, we used the tokenizers with (a) 4-Branch, 1CB (b) 4-Branch, 2CB and (c) 4-Branch, 8CB, which keeps the same architecture and training loss but scales only the number of codebooks.

Table 23: Comparison of classification balanced accuracy between NEURORVQ-EEG on the Motor task pre-trained using various tokenizers.

Tokenizer used to train same backbone model (corresponding val MSE)	Head-Only	Full-Fine Tuning
4-Branch, 1CB (0.35)	59.0 (4.91)	62.7 (1.61)
4-Branch, 2CB (0.12)	59.1 (6.09)	63.5 (2.23)
4-Branch, 8CB (0.035)	58.9 (5.67)	64.0 (1.39)

We notice that once the tokenizer faithfully reconstructs task-relevant structure, further fidelity captures detail less relevant to the task-label with diminishing returns: a 10× MSE reduction (0.35→0.035) yields modest gain (saturation), whereas the low vs very good fidelity loss swap (0.946→0.035) gives a large 58.5→64.0.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract clearly outlines the core contributions presented in the main paper, particularly the introduction of the biosignal tokenizer family. The claims made are consistent with the scope and findings detailed throughout the work.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The discussion section describes how the architecture needs to be configured for each modality and how a unified (cross-modal) model can be challenging. In addition, it also outlines potential directions for future research, one of which being a better foundation model, a factor that might currently limit downstream task performance.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: Appendix A describes the theoretical assumptions and a complete and mathematically sound analysis of our proposed training loss.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: The method presented in the paper is fully reproducible. All model parameters, datasets used and preprocessing / training steps are described in great detail. While one of the pre-training datasets used is self-collected, it constitutes only a minor portion of the data and does not significantly impact the reproducibility of the paper's main claims or conclusions.

Guidelines:

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

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [No]

   Justification: We provide the main code and some of our models open-source. While all model details and all settings are thoroughly documented in the main paper and supplementary material, we do not provide full open access to all datasets. Specifically, one of the datasets is our own self-collected closed-access dataset. However, this dataset represents only a minor part of the pre-training data and does not materially affect the reproducibility of the main experimental results.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The paper provides comprehensive details on data preprocessing, hyperparameters and network settings. All relevant training and testing configurations are thoroughly described in the main paper and supplementary material.

   Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: The paper reports statistical significance through cross-validation experiments, providing the mean and std across folds. These metrics are appropriate and sufficient to support the paper's experimental claims.

Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Yes, these details are provided in Section 5 and Appendices.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: All research was conducted responsibly and it fully adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The primary contribution of the paper is a biosignal tokenizer family. As discussed in the Introduction, Discussion and Conclusion, biosignal foundation models offer substantial application potential in neuroscience research, Brain-Computer Interfaces and healthcare. Therefore, NeuroRVQ advances the development of efficient large biosignal foundation models by providing a high-fidelity biosignal tokenizer, paving the way for broader use of foundation models in the biosignal space.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: We believe the method and data presented in the paper do not pose significant risks of misuse.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All external assets used in this work are properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: We release the NeuroRVQ tokenizers, foundation models and code. The self-collected EEG dataset is not released due to intellectual property constraints. Its collection was approved by Imperial College London's Research Governance and Integrity Team (approval number 22IC7505).

    Guidelines:

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

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

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

    Answer: [Yes]

    Justification: Yes, the full text of the information provided to participants is included in Appendix F. Each participant received a £30 gift.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

    Answer: [Yes]

    Justification: The collection of the self-collected EEG dataset was approved by the Research Governance and Integrity Team of Imperial College London (approval number 22IC7505), the institutional equivalent of an IRB. The study involved non-invasive scalp EEG recordings during real and imagined movements and a movement-controlled game, which poses minimal risk. Participants were informed about the procedures and the use of their data before taking part and gave informed consent (Appendix F).

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs were used for writing and editing purposes

    Guidelines:

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