International Journal of Allied Practice, Research and Review
Website: www.ijaprr.com (ISSN 2350-1294)

# NeuroTokenNet: A Foundation Model-Inspired Hybrid Framework for Adaptive EEG Tokenization, Contextual Vector Embedding, and Explainable Brain–Computer Interfaces

**Riaz Ahmed**
Associate Professor,
Department of Computer Applications,
Government Degree College, Rajouri, Jammu and Kashmir, India.

**Abstract - Electroencephalography (EEG)-based brain–computer interfaces (BCIs) remain constrained by low signal-to-noise ratio, strong inter-subject variability, and a lack of generalizable, interpretable representations. Inspired by the tokenization-and-embedding paradigm underlying large language foundation models, we propose NeuroTokenNet, a hybrid framework that adaptively segments multichannel EEG signals into variable-length "neuro-tokens," embeds them into a contextual vector space via a channel-aware transformer encoder, and produces clinically and behaviourally interpretable predictions through an integrated attention-attribution explainability module. NeuroTokenNet combines (i) an Adaptive Wavelet-Guided Tokenizer (AWGT) that segments EEG into non-uniform temporal tokens aligned with signal non-stationarity rather than fixed-length windows, (ii) a Spatio-Spectral Contextual Embedding (SSCE) module that jointly encodes electrode topology and frequency-band energy into each token embedding, and (iii) a Cross-Token Attention Transformer (CTAT) pretrained with a masked-token reconstruction objective across multiple BCI paradigms before task-specific fine-tuning. An Explainability Head built on integrated gradients and attention-rollout produces per-token and per-channel saliency maps aligned with known neurophysiological correlates. We evaluate NeuroTokenNet on motor imagery (BCI Competition IV-2a), P300 spelling (BCI Competition III), and emotion recognition (DEAP) benchmarks, achieving average classification accuracies of 84.6%, 91.3%, and 78.9% respectively, outperforming CNN, EEGNet, and standard fixed-window transformer baselines by 3.1–6.8 percentage points while reducing the number of tokens processed per trial by 41% relative to fixed-window tokenization. Cross-subject transfer experiments show a 12.4% relative improvement in zero-shot accuracy over fixed-window pretraining baselines, and explainability evaluations confirm that attributed saliency aligns with established motor-cortex and parietal activation patterns for the respective paradigms.**

**Keywords - EEG, Brain–Computer Interface, Foundation Models, Tokenization, Transformer, Explainable AI, Self-Supervised Pretraining**

## I. Introduction

Brain-computer interfaces translate brain activity into actionable commands or diagnostic labels, with electroencephalography being the leading non-invasive modality due to its low cost and high temporal resolution. Yet, after more than 3 decades of progress, EEG-based BCIs still suffer from a lack of generalization across subjects and sessions, mainly because state-of-the-art pipelines rely on fixed-length windowing, hand-crafted spectral features or convolutional architectures that implicitly assume signal stationarity within each window. EEG often violates these assumptions, with statistical features that change fast with cognitive state, artifact contamination, and electrode-specific noise. Recent foundation models in language and vision have shown that learning a discretized or continuous tokenization of raw input, followed by self-supervised contextual embedding and large-scale pretraining, leads to representations that transfer well across downstream tasks. We hypothesize that an analogous tokenize-embed-pretrain pipeline, adapted to the non-stationary, multichannel structure of EEG, can similarly improve cross-subject and cross-paradigm generalization in BCIs, if the tokenization respects EEG's temporal and spatial structure, rather than naively reusing fixed patch sizes from vision transformers. We propose NeuroTokenNet with three contributions. First, the Adaptive Wavelet-Guided Tokenizer decomposes the signal for each channel into tokens of different lengths, with the boundaries determined by wavelet-based change-point detection, rather than arbitrary fixed windows, so that each token corresponds to a segment of the signal that is quasi-stationary in a local sense. Second, the Spatio-spectrum Contextual Embedding module integrates electrode-topology positional encoding with per-band spectrum energy information into a single token embedding, allowing the downstream transformer to reason jointly across place, frequency, and time. Third, we pre-train a Cross-Token Attention Transformer with a masked-token reconstruction objective on multiple public BCI corpora covering different paradigms, and then fine-tune lightweight task heads, and combine the resulting model with an explainability head that produces clinically interpretable per-token and per-channel attributions.

We evaluate NeuroTokenNet on three canonical BCI paradigms, motor imagery, P300 event-related potentials, and emotion recognition, and show consistent improvements over convolutional and fixed-window transformer baselines in accuracy and cross-subject transfer, and qualitative and quantitative evidence that the attributions of the explain ability head are neuro-physiologically plausible.

## II. Related Work

### 2.1 Deep Learning for EEG-Based BCIs

Convolution architectures such as EEGNet and shallow/deep ConvNets established strong baselines for motor imagery and ERP classification by applying compact temporal and spatial filters. While effective and parameter-efficient, these architectures process EEG at a fixed temporal resolution and offer limited capacity to model long-range dependencies across non-stationary segments.

### 2.2 Transformers and Tokenization for Physiological Signals

Transformer-based EEG models typically adopt fixed-length patching analogous to vision transformer patches, dividing each channel or multichannel window into uniform segments before applying self-attention. While this improves long-range temporal modelling relative to convolution baselines, uniform patching discards information about where genuine non-stationarity occurs in the signal, motivating content-adaptive tokenization schemes such as the one proposed here.

### 2.3 Self-Supervised Pertaining for EEG

Masked reconstruction and contrastive pretraining objectives have been applied to EEG to improve label efficiency and cross-subject transfer, generally following the masked-autoencoding recipe from vision and language pretraining. Our work differs in pretraining over adaptively tokenized, spatio-spectrally embedded inputs rather than fixed-length raw or spectrogram patches, and in pairing pretraining with a dedicated explainability head evaluated against neurophysiological priors.

### 2.4 Explain ability in BCI Systems

Saliency and attention-based explanation methods have been used to interpret CNN-based EEG classifiers, typically producing per-channel or per-timepoint heatmaps. NeuroTokenNet's token-level granularity, combined with attention-rollout across the pretrained transformer, allows attributions to be expressed at the level of semantically meaningful, variable-length signal segments rather than arbitrary fixed time bins.

## III. NeuroTokenNet Methodology

### 3.1 Overview

NeuroTokenNet processes raw multichannel EEG in four steps: (i) the Adaptive Wavelet-Guided Tokenizer (AWGT) transforms each channel into a sequence of variable-length tokens; (ii) the Spatio-Spectral Contextual Embedding (SSCE) module embeds each token into a fixed-dimensional space guided by electrode topology and spectral content; (iii) the Cross-Token Attention Transformer (CTAT) learns contextual dependencies across tokens and channels, pretrained with masked-token reconstruction and fine-tuned per task; and (iv) the Explainability Head produces per-token and per-channel attributions for each prediction.

### 3.2 Adaptive Wavelet-Guided Tokenizes

AWGT performs a multi-resolution discrete wavelet transform (DWT) on each channel, and identifies change points where the distribution of wavelet coefficients differs markedly between adjacent scales, using a cumulative-sum based change-point statistic . Token boundaries are placed at detected change points, with minimum and maximum token lengths (50-500 ms in our experiments) to control computational cost and avoid degenerate single-sample tokens. This results in a variable number of tokens per channel per trial, with representational capacity devoted to transient, information-rich segments such as event-related potentials or artifact onsets, while quasi-stationary background activity is compressed into fewer, longer tokens.

### 3.3 Spatio-Spectral Contextual Embedding

Each token is embedded by concatenating three components: a learned temporal-waveform embedding produced by a lightweight 1D convolutional encoder applied to the raw token samples, a spectral-energy embedding computed from band-power features in the delta, theta, alpha, beta, and gamma bands, and a topology-aware positional encoding derived from the electrode's 3D scalp coordinates using a graph-Laplacian eigenbasis. These three components are projected into a shared embedding dimension and summed, allowing the transformer to attend simultaneously to temporal pattern, spectral content, and spatial location.

### 3.4 Cross-Token Attention Transformer and Pretraining

The CTAT is a typical pre-norm transformer encoder acting on the sequence of token embeddings across all channels, enhanced by channel-segment embeddings to disambiguate tokens coming from separate electrodes. We pre-train the CTAT using a masked-token reconstruction objective: a random subset of tokens (15% in our trials) is masked, and the model is taught to reconstruct the waveform and spectral-energy embeddings of the masked tokens from surrounding context. We jointly pretrain on motor imagery, P300 and emotion-recognition corpora, encouraging paradigm-agnostic representations, and then finetune a lightweight classification head for each task.

### 3.5 Explain ability Head

The Explain ability Head produces a per-token attribution score by combining integrated gradients over token embeddings and attention-rollout over the self-attention layers of CTAT, which is subsequently transposed to per-channel and per-time-interval saliency according to the known channel and time span of each token. This allows for interpretable visualizations directly comparable to standard EEG topographic maps and qualitative validation against known neurophysiological correlations for each BCI paradigm.

## IV. Experimental Setup

### 4.1 Datasets

• BCI Competition IV-2a: 22-channel motor imagery, 4 classes (left hand, right hand, feet, tongue), 9 subjects. • BCI Competition III (dataset II): 64-channel P300 speller paradigm, 2 subjects with large trial repeats. • DEAP: 32-Channel emotion recognition with valence/arousal self-report labels, 32 individuals, median-binarized for classification.

### 4.2 Baselines

- EEGNet: compact CNN baseline widely used in BCI literature.
- Deep ConvNet: deeper convolutional baseline for comparison on raw-signal classification.
- Fixed-Window Transformer (FWT): a transformer encoder identical to CTAT but using uniform fixed-length patching instead of AWGT, to isolate the effect of adaptive tokenization.
- NeuroTokenNet (no pretraining): NeuroTokenNet trained from scratch per task, to isolate the effect of masked-token pretraining.

### 4.3 Training Protocol

All models are trained with subject-wise 5-fold cross validation for within-subject evaluation and leave-subjects-out cross validation for cross-subject transfer evaluation. Variants of NeuroTokenNet are pre-trained for 200 epochs on the combined multi-paradigm corpus and then fine-tuned for 100 epochs for each downstream task, using the AdamW optimizer with cosine learning rate decay.

### 4.4 Metrics

• Classification accuracy and macro-F1 score for each paradigm. • Tokens per Trial: mean number of tokens processed per trial, used as a proxy for computing cost. • Cross-Subject Transfer Accuracy: zero-shot accuracy on unseen subjects without fine-tuning • Attribution Alignment Score: overlap between explainability-head saliency and paradigm-specific regions of interest specified from past neurophysiological literature (e.g., central electrodes C3/C4 for motor imagery).

## V. Experimental Results

### 5.1 Overall Classification Performance

Table 1 reports within-subject classification accuracy across the three benchmark paradigms.

| Model | Motor Imagery (%) | P300 Speller (%) | DEAP Emotion (%) |
|---|---|---|---|
| EEGNet | 77.8 | 85.1 | 72.4 |
| Deep ConvNet | 79.2 | 86.7 | 73.8 |
| Fixed-Window Transformer | 80.5 | 87.9 | 75.1 |
| NeuroTokenNet (no pretrain) | 82.1 | 89.0 | 76.6 |
| NeuroTokenNet (full) | 84.6 | 91.3 | 78.9 |

**Table 1. Within-subject classification accuracy across BCI paradigms.**

### 5.2 Tokenization Efficiency

Table 2 compares the average number of tokens per trial and inference latency between fixed-window patching and AWGT, on the motor imagery dataset.

| Tokenization | Avg. Tokens/Trial | Inference Latency (ms) | Accuracy (%) |
|---|---|---|---|
| Fixed-Window (200 ms) | 20.0 | 14.2 | 80.5 |
| AWGT (Adaptive) | 11.8 | 9.1 | 84.6 |

**Table 2. Tokenization efficiency comparison, motor imagery dataset, single-trial inference.**

### 5.3 Cross-Subject Transfer

Table 3 reports zero-shot cross-subject accuracy (leave-subjects-out, no fine-tuning) on motor imagery, comparing pretrained and non-pretrained variants.

| Model | Zero-Shot Accuracy (%) | Relative Improvement |
|---|---|---|
| Fixed-Window Transformer (pretrained) | 61.3 | — |
| NeuroTokenNet (no pretrain) | 58.9 | −3.9% |
| NeuroTokenNet (pretrained, full) | 68.9 | +12.4% |

**Table 3. Cross-subject zero-shot transfer accuracy, motor imagery, relative to the fixed-window pretrained baseline.**

### 5.4 Ablation Study

We undertake ablation studies on motor imagery for each component of NeuroTokenNet to attribute the performance gains. Removing AWGT (reverting to fixed-window tokenization) causes a 4.1 pt drop in accuracy and a 70% rise in tokens-per-trial. deleting the spectrum-energy branch of SSCE (just waveform and positional encoding) reduces accuracy by 2.3 points, while deleting the topology-aware positional encoding (only spectral and waveform) reduces accuracy by 1.8 points. We show that pretraining is the main factor for cross-subject generalization, as removal of masked-token pretraining reduces accuracy by 2.5 points and zero-shot transfer accuracy by 10.0 points.

| Configuration | Accuracy (%) | Tokens/Trial | Zero-Shot Accuracy (%) |
|---|---|---|---|
| Full NeuroTokenNet | 84.6 | 11.8 | 68.9 |
| − AWGT (fixed-window) | 80.5 | 20.0 | 64.1 |
| − Spectral embedding | 82.3 | 11.8 | 66.0 |
| − Topology positional enc. | 82.8 | 11.8 | 65.4 |
| − Pretraining | 82.1 | 11.8 | 58.9 |

**Table 4. Ablation of NeuroTokenNet components, motor imagery dataset.**

### 5.5 Explain ability Evaluation

We compute the Attribution Alignment Score between the saliency maps of the explainability head and paradigm-specific ROIs. Attributions peak over central electrodes C3, Cz and C4 consistent with activation of the sensorimotor cortex, with an alignment score of 0.79 (cosine similarity between normalized saliency and a binary region of interest mask) in motor imagery. For the P300 paradigm, attributions are maximal over parietal and centro-parietal electrodes (Pz, CPz) in the 250-500 ms post-stimulus window, consistent with known P300 generators, with a score of 0.82. These results indicate that the token level attributions of NeuroTokenNet are not only locally coherent but also consistent with independently validated neurophysiological priors, providing qualitative support for the model interpretability assertions.

## VI. Discussion

Our results indicate that adaptive, content-aware tokenization is an important direction for improving EEG transformers beyond just expanding model capacity or pretraining data, as AWGT alone (without pretraining) already increases accuracy and efficiency over fixed-window patching. The largest single-component contribution to cross-subject generalization is from masked-token pretraining, consistent with findings in the broader self-supervised representation-learning literature, suggesting that scaling multi-paradigm EEG pretraining corpora is a promising direction for future work. The explainability head also aligns with known neurophysiological regions of interest, indicating that the variable-length, semantically coherent tokens may allow for more clinically interpretable token-level attribution than fixed-bin saliency maps; however, further work is needed to formally validate this with expert neurologist annotations.

### 6.1 Limitations

Our evaluation is based on publicly accessible benchmark datasets that are standard in the BCI research but modest in scope compared to foundation-model pretraining corpora in other domains, and results on bigger, more diversified EEG corpora may differ. The Attribution Alignment Score depends on coarse, literature-based regions of interest instead of subject-specific functional localization and should be considered as a plausibility check rather than a clinical validation.

## VII. Conclusion

Here, we propose NeuroTokenNet, a framework inspired by foundation models for EEG-based brain-computer interfaces, which integrates adaptive wavelet-guided tokenization, spatio-spectral contextual embedding, masked-token pretraining, and an integrated explainability head. NeuroTokenNet consistently outperforms convolution and fixed-window transformer baselines on motor imagery, P300 and emotion-recognition benchmarks in terms of accuracy, computational efficiency and cross-subject transfer, while producing token-level attributions that are consistent with established neurophysiological correlates. Our results indicate that adaptive tokenization and multi-paradigm self-supervised pretraining are potential paths for constructing more generalizable and interpretable BCI systems.

## VIII. References

**[1] Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., & Ball, T. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. Human Brain Mapping.**

**[2] Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. Journal of Neural Engineering.**

**[3] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL.**

**[4] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., &Girshick, R. (2022). Masked autoencoders are scalable vision learners. CVPR.**

**[5] Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., & Patras, I. (2012). DEAP: A database for emotion analysis using physiological signals. IEEE Transactions on Affective Computing.**

**[6] Tangermann, M., Müller, K.-R., Aertsen, A., et al. (2012). Review of the BCI competition IV. Frontiers in Neuroscience.**

**[7] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. ICML.**

**[8] Abnar, S., & Zuidema, W. (2020). Quantifying attention flow in transformers. ACL.**