# Tokenizing Single-Channel EEG with Time-Frequency Motif Learning

**Jathurshan Pradeepkumar**[1] **Xihao Piao**[2] **Zheng Chen**[2] **Jimeng Sun**[1]
[1]Siebel School of Computing and Data Science, UIUC
[2]SANKEN, Osaka University
{jp65,jimeng}@illinois.edu, {park88,chenz}@sanken.osaka-u.ac.jp

## Abstract

We introduce TFM-Tokenizer, a novel tokenization framework tailored for EEG analysis that transforms continuous, noisy brain signals into a sequence of discrete, well-represented tokens for various EEG tasks. Conventional approaches typically rely on continuous embeddings and inter-channel dependencies, which are limited in capturing inherent EEG features such as temporally unpredictable patterns and diverse oscillatory waveforms. In contrast, we hypothesize that critical time-frequency features can be effectively captured from a single channel. By learning tokens that encapsulate these intrinsic patterns within a single channel, our approach yields a scalable tokenizer adaptable across diverse EEG settings. We integrate the TFM-Tokenizer with a transformer-based TFM-Encoder, leveraging established pretraining techniques from natural language processing, such as masked token prediction, followed by downstream fine-tuning for various EEG tasks. Experiments across four EEG datasets show that TFM-Token outperforms state-of-the-art methods in single dataset settings. Comprehensive analysis shows that the learned tokens capture class-specific features, preserve frequency content, and encode interpretable time–frequency motifs. The code is available at https://github.com/Jathurshan0330/TFM-Tokenizer.

## 1 Introduction

Electroencephalograms (EEGs) captures real-time neuronal activity with millisecond precision, reflecting the responses to various event stimuli. This makes EEGs essential for fundamental research [1, 2] and diverse clinical applications[3–10]. Deep learning (DL) models have shown remarkable success in automating EEG analysis across various tasks [11–13], driven by their ability to project noisy signals into discriminative latent spaces that aligns with neurophysiological events.

Despite their success, effectively representing EEGs remains a primary challenge. Real-world EEGs vary widely due to diverse devices, channel configurations and lengths[14]. Unfortunately, most existing methods typically learn representations on a case-by-case basis with specific architectures or fixed channel settings. These methods exhibit limited generalization across tasks and poor scalability to different data formats. There is thus an urgent need to develop an EEG analysis method that serves broader research objectives.

Recently, the transformative impact of large foundation models[15, 16] has elevated EEG representation learning to new heights. Several foundation EEG models have been proposed [17–19, 14], demonstrating both enhanced performance and generalization. Researchers often *tokenize EEGs* into short-duration snapshots across different data formats and model their dependencies using powerful Transformers. However, this direction remains nascent, and several limitations remain:

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Learning from Time Series for Health.

- **Inappropriate Tokenization Representation.** One reason large language models (LLMs) succeed is their effective tokenization and similar benefits have been shown in image [20] and video [21, 22] tokenization. However, existing foundation EEG models generally do not adopt a discrete tokenization paradigm. Although some methods claim to provide an EEG "tokenizer," they typically lack a discrete approach similar to NLP. We hypothesize that EEGs consist of recurring motifs distorted by noise, scaling, and temporal warping. Discretizing these into invariant tokens reduces data complexity and simplifies downstream task learning. Empirically, we show that our fully discrete approach outperforms continuous baselines across multiple tasks, with fewer parameters.
- **Insufficient Frequency Representation.** Capturing eventful EEG features, which are characterized by distinct frequencies, is a primary focus of EEG analysis. However, tokenizing raw EEGs often lead to a loss of frequency diversity. This frequency representation collapse is a common issue in time-series modeling, as low-frequency components typically dominate the EEG data, biasing models toward lower frequencies while overlooking critical high-frequency features (e.g.spikes).
- **Scalability and Generalization.** EEG-related tasks vary in channel configurations. For example, seizure detection typically uses 16 channels, whereas sleep studies often require only 1–2 channels. However, existing models are primarily designed for multi-channel settings, heavily relying on cross-channel prediction. This design limits their scalability and adaptability to configurations with fewer or even single channels, as well as to varying acquisition setups.

Therefore, in this paper, we propose `TFM-Token`, an effective, fully discretized EEG tokenization framework that captures time-frequency motifs from single-channel EEG signals into distinct tokens. Technically, our contributions are as follows:

- `TFM-Tokenizer` **and** `TFM-Encoder`**:** We introduce a scalable discrete tokenization framework for EEG, transforming single-channel EEG into discrete token sequences akin to NLP. `TFM-Tokenizer` converts EEG into discrete tokens, and `TFM-Encoder` uses them for downstream tasks.
- **Joint Modeling of Frequency and Temporal Dynamics:** Our tokenizer integrates raw EEG patches with time-frequency representations, using frequency band and temporal masking to capture essential frequency patterns while disentangling temporal variations.
- **Scable tokenization:** Our single-channel approach enables flexible adaptation across EEG tasks and channel configurations. `TFM-Tokenizer` further enhances existing EEG models, such as LaBraM [23] (Appendix C.4).
- **Empirical Validation and Token Quality Analysis:** We evaluate our framework on four EEG downstream tasks, demonstrating state-of-the-art performance. Additionally, we analyze token quality, including token visualization, class-specific uniqueness, and frequency learning analysis

## 2 Preliminaries

**EEG Data:** Let $\mathbf{X} \in \mathbb{R}^{C\times T}$ be a multi-channel EEG. Each channel $x^c \in \mathbb{R}^T$ is segmented into raw patches $\{x_i\}_{i=1}^N$ and corresponding spectrogram windows $\{\mathbf{S}_i\}_{i=1}^N$ using STFT (window $L$, hop $H$). For simplicity, we omit the channel index and denote $x$ as a single-channel EEG.

**Problem Statement 1 (EEG Tokenization):** Given a single channel EEG $x$, we aim to learn a tokenization function $f_{\text{tokenizer}} : \mathbb{R}^T \to \mathcal{V}^{N\times D}$ that maps $x$ (or transformations) to a sequence of discrete tokens $\{v_i\}_{i=1}^N$, where each from a learnable EEG token vocabulary $\mathcal{V}$ of size $k$ and embedding size of $D$. These tokens should represent various time-frequency "*motifs*" derived from both $x_i$ and $\mathbf{S}_i$. **Remark:** We here hold several expectations for the learned motif tokens. First, these tokens are expected to reduce redundancy, noise, and complexity, providing a compact, sparse, and informative representation of EEGs. Second, these motifs should capture key neurophysiological patterns from temporal and frequency domains. Third, the tokens should generalize across EEG tasks.

**Problem Statement 2 (Multi-Channel EEG Classification):** Given EEGs $\mathbf{X}$ and a fixed, learned single-channel tokenizer $f_{\text{tokenizer}}$, we apply $f_{\text{tokenizer}}$ independently to each channel $c$ to obtain a token sequences $\left\{\{v_i^c\}_{i=1}^N\right\}_{c=1}^C$. These tokens are aggregated and mapped to output labels by:$f_{\text{classifier}} : (\mathcal{V}^D)^{N\times C} \to \mathbf{Y}$ where $Y$ is the target labels (e.g., EEG events, seizure types). Notably, $f_{\text{classifier}}$ can be any downstream model, and its training is performed separately from the EEG tokenizer $f_{\text{tokenizer}}$.

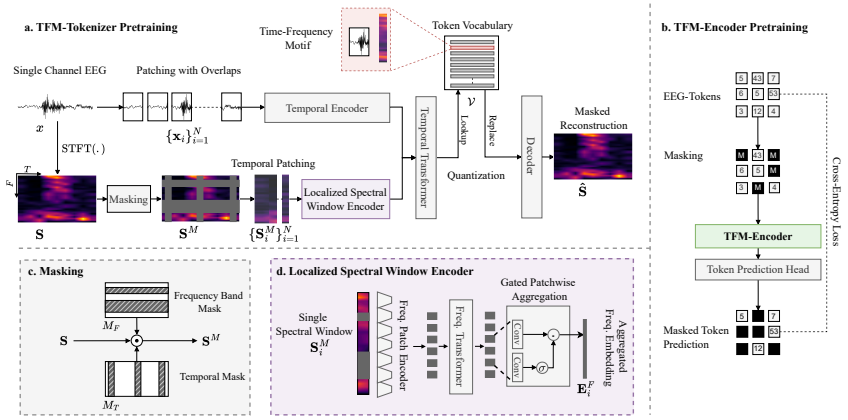


Figure 1: Overview of the TFM-Token framework. (a) TFM-Tokenizer Pretraining: through dual-path encoding and masked prediction, learns to capture time-frequency motifs into discrete tokens. (b) TFM-Encoder Pretraining: uses masked token prediction on learned tokens. (c) Masking: combination of frequency band and temporal masking. (d) Localized Spectral Window Encoder: extracts and aggregates frequency band features from spectral windows into compact embedding.

# 3 TFM-Token

TFM-Token comprises two components: (1) TFM-Tokenizer ($f_{\textbf{tokenizer}}$)**:** converts continuous EEG signals into discrete tokens, capturing key time-frequency motifs, and (2) TFM-Encoder $f_{\textbf{classifier}}$**:** leverages these tokens for downstream EEG tasks. To mitigate the quadratic complexity of standard Transformers [24], we employ a linear attention mechanism [25, 26]. To train TFM-Token, we first conduct an unsupervised pretraining of TFM-Tokenizer in a single-channel setting (Figure 1a, Sec 3.1). The tokenizer is then frozen, and TFM-Encoder undergoes masked token prediction pretraining (Figure 1b, Sec 3.2), followed by fine-tuning for downstream tasks.

## 3.1 Single Channel TFM-Tokenizer

We introduce the TFM-Tokenizer, a scalable module for tokenizing single-channel EEG signals $x$ by capturing their temporal and frequency dynamics. Our design is inspired by the Vector-Quantized Variational Autoencoder (VQ-VAE) [20], which has been widely adopted for tokenization efforts in other domains, such as video processing [22]. At a high level, TFM-Tokenizer adopts a *frequency-then-time* paradigm and comprises three components as illustrated in Figure 1a: (1) Localized Spectral Window Encoder, (2) Temporal Encoder, and (3) Temporal Transformer.

**Localized Spectral Window Encoder.** EEG signals often contain distinct oscillatory patterns (e.g., alpha, beta bands). To capture such frequency-band structures, each spectral window $\mathbf{S}_i$ is patched along the frequency axis into $P$ non-overlapping patches spanning $\Delta f$ frequency bins such that $P.\Delta f = F$ (Figure 1d). Each patch $\mathbf{S}_{(i,p)}$ is projected: $e_{(i,p)} = \text{GroupNorm}\left(\text{GeLU}\left(\mathbf{W}_p \mathbf{S}_{(i,p)}\right)\right)$, where $\mathbf{W}_p \in \mathbb{R}^{D \times \Delta f}$ is a learnable matrix. Then, a frequency transformer operates along the frequency axis to model intra-spectral window cross-frequency band dependencies. In many EEG scenarios, large portions of the frequency spectrum can be irrelevant. To emphasize task-relevant frequency patches, we apply a gated aggregation mechanism to obtain a single embedding for each $S_i$: $\mathbf{E}_i^F = \text{Concat}\left[\sigma\left(\mathbf{W_{g1}} \mathbf{e_{(i,p)}}\right) \mathbf{W_{g2}} \mathbf{e_{(i,p)}}\right]$, where $\mathbf{W_{g1}}, \mathbf{W_{g2}}$ are learnable matrices and $\sigma(\cdot)$ is the element-wise sigmoid function.

**Temporal Encoder and Temporal Transformer:** To capture temporal dynamics from the raw EEG patches $\{x_i\}_{i=1}^N$, we perform a linear projection followed by GELU[27] activation and group normalization, producing $\{\mathbf{E}_i^T\}_{i=1}^N$. We then concatenated each aggregated frequency embedding $\mathbf{E}_i^F$ with its corresponding temporal embedding $\mathbf{E}_i^T$, and input the sequence into a temporal transformer. The output is then quantized into discrete tokens $\{v_i\}_{i=1}^N$ using a learnable codebook $\mathcal{V}^k$.

**Tokenizer Codebook.** Our tokenizer captures temporal–frequency motifs by applying vector quantization along the time axis, treating each short-duration patch as a discrete unit. This contrasts with conventional visual tokenizers, which typically operate on the embedding dimension [20, 28]. As a result, each token represents a short-duration waveform segment, enabling interpretability(Section 4).

**Frequency Masking Prediction for Tokenizer Learning.** To facilitate frequency learning, we apply frequency-band and temporal masking during `TFM-Tokenizer` training. $\mathbf{S}$ is split into $N_F = \lfloor \frac{F}{\delta_f} \rfloor$ frequency groups of size $\delta_f$. We apply a random frequency mask $M_F$ and temporal mask $M_T$, combining them as $M = M_F \wedge M_T$ to produce the masked spectrogram $\mathbf{S}^M$. The masked input $\mathbf{S}^M$ and raw EEG patch $x$ are encoded, quantized, and passed through a transformer + linear decoder to reconstruct masked regions: $\mathcal{L}_{\text{rec}} = \sum_{(f,t)} \left\| \mathbf{S}(f,t) - \hat{\mathbf{S}}(f,t) \right\|_2^2$, where $\hat{\mathbf{S}}$ is the reconstructed output. Additionally, we apply the codebook and commitment losses [20].

### 3.2 Token-Wise `TFM-Encoder`

The `TFM-Encoder` aggregates EEG tokens across channels for downstream tasks. Given a multi-channel recording $\mathbf{X} \in \mathbb{R}^{C \times T}$, the pretrained `TFM-Tokenizer` produces token sequences $\left\{ \{v_i^c\}_{i=1}^N \right\}_{c=1}^C$ for each channel $c$ independently. A [CLS] token is prepended [29], and the sequence is processed by transformer layers. The [CLS] output is used for classification. We pretrain `TFM-Encoder` using masked token prediction and then it is finetuned to downstream tasks.

## 4 Experiments and Results

Table 1: EEG classification performance on TUEV and TUAB datasets under single dataset settings (Results on CHB-MIT and IIIC Seizure are provided in Table 4 and 5 in Appendix C.1 ).

| Models | Number of Params | TUEV (event type classification) | | | TUAB (abnormal detection) | | |
|---|---|---|---|---|---|---|---|
| | | Balanced Acc. | Cohen's Kappa | Weighted F1 | Balanced Acc. | AUC-PR | AUROC |
| SPaRCNet[30] | 0.79M | 0.4161 ± 0.0262 | 0.4233 ± 0.0181 | 0.7024 ± 0.0104 | 0.7896 ± 0.0018 | 0.8414 ± 0.0018 | 0.8676 ± 0.0012 |
| ContraWR[4] | 1.6M | 0.4384 ± 0.0349 | 0.3912 ± 0.0237 | 0.6893 ± 0.0136 | 0.7746 ± 0.0041 | 0.8421 ± 0.0104 | 0.8456 ± 0.0074 |
| CNN-Transformer[31] | 3.2M | 0.4087 ± 0.0161 | 0.3815 ± 0.0134 | 0.6854 ± 0.0293 | 0.7777 ± 0.0022 | 0.8433 ± 0.0039 | 0.8461 ± 0.0013 |
| FFCL[32] | 2.4M | 0.3979 ± 0.0104 | 0.3732 ± 0.0188 | 0.6783 ± 0.0120 | 0.7848 ± 0.0038 | 0.8448 ± 0.0065 | 0.8569 ± 0.0051 |
| ST-Transformer[33] | 3.5M | 0.3984 ± 0.0228 | 0.3765 ± 0.0306 | 0.6823 ± 0.0190 | 0.7966 ± 0.0023 | 0.8521 ± 0.0026 | 0.8707 ± 0.0019 |
| Vanilla BIOT[14] | 3.2M | 0.4682 ± 0.0125 | 0.4482 ± 0.0285 | 0.7085 ± 0.0184 | 0.7959 ± 0.0057 | 0.8792 ± 0.0023 | 0.8815 ± 0.0043 |
| BIOT*[14] | 3.2M | 0.4679 ± 0.0354 | 0.4890 ± 0.0407 | 0.7352 ± 0.0236 | 0.7955 ± 0.0047 | 0.8819 ± 0.0046 | 0.8834 ± 0.0041 |
| LaBraM-Base*[23] | 5.8M | 0.4682 ± 0.0856 | 0.5067 ± 0.0413 | 0.7466 ± 0.0202 | 0.7720 ± 0.0046 | 0.8498 ± 0.0036 | 0.8534 ± 0.0027 |
| TFM-Token-R | 1.8M | <u>0.4898</u> ± 0.0105 | 0.5194 ± 0.0195 | 0.7518 ± 0.0095 | <u>0.8033</u> ± 0.0021 | <u>0.8908</u> ± 0.0027 | <u>0.8849</u> ± 0.0024 |
| TFM-Token-S | 1.9M | 0.4708 ± 0.0339 | <u>0.5275</u> ± 0.0314 | <u>0.7538</u> ± 0.0152 | 0.7927 ± 0.0044 | 0.8814 ± 0.0095 | 0.8836 ± 0.0052 |
| **TFM-Token** | 1.9M | **0.4943 ± 0.0516** | **0.5337 ± 0.0306** | **0.7570 ± 0.0163** | **0.8152 ± 0.0014** | **0.8946 ± 0.0008** | **0.8897 ± 0.0008** |

1. Best results are **bolded**, second-best are <u>underlined</u>. 2. LaBraM's parameter count includes only the classifier. The size of their neural tokenizer was 8.6M. 3. TFM-Token-R and S use only raw EEG or STFT as inputs. 4. * indicates single dataset setting

**Performance comparison:** We evaluate on four EEG datasets, including: (1)TUEV[34, 35], (2)TUAB[36], (3)IIIC Seizure sourced from [30, 37] and (4)CHB-MIT[38]. Full experimental details are provided in Appendix B. Table 1 presents EEG event classification results on TUEV and abnormal detection performance on TUAB. Our `TFM-Token` consistently outperforms baselines on all datasets and metrics in the single-dataset setting.

`TFM-Token` achieves better performance with fewer parameters, 3× smaller than LaBraM (5.8M → 1.9M) and 1.5× smaller than BIOT (3.2M → 1.9M). Empirically, this reduction can be attributed to the discrete tokenization approach, which compresses the EEG into a token sequence, reducing data complexity. Additional results and token quality analysis are provided in Appendix C.

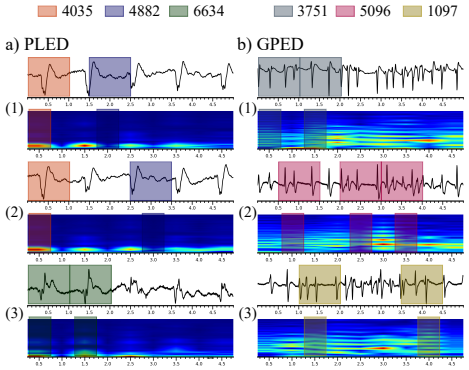


Figure 2: Motifs captured by `TFM-Tokenizer` on TUEV: (a) shows three samples from the PLED class and (b) shows three samples from the GPED class.

**Interpretability of Learned Tokens:** We visually examine whether `TFM-Tokenizer` captures meaningful time–frequency motifs. Figure 2 shows some representative tokens learned by `TFM-Tokenizer` on the TUEV. Each token corresponds to a 1s EEG patch (0.5s overlap) and its spectral window. For clarity, we highlight the most frequent tokens per class. The results shows that `TFM-Tokenizer` encodes class-specific patterns into discrete tokens. For instance, token 4035 in the PLED class consistently captures a characteristic drop followed by a rise waveform, maintaining its structure across different samples despite variations in noise, amplitude, and minor shifts within the window.

## 5 Conclusion

We introduced `TFM-Token`, a fully discrete tokenization framework consisting of `TFM-Tokenizer` and `TFM-Encoder` modules. Comprehensive evaluations across multiple datasets demonstrate that `TFM-Token` outperforms existing baselines with fewer parameters in single dataset settings.

## References


[1] Erin C Conrad, Samuel B Tomlinson, Jeremy N Wong, Kelly F Oechsel, Russell T Shinohara, Brian Litt, Kathryn A Davis, and Eric D Marsh. Spatial distribution of interictal spikes fluctuates over time and localizes seizure onset. *Brain*, pages 554–569, 2019.

[2] Adam Li, Chester Huynh, Zachary Fitzgerald, Iahn Cajigas, Damian Brusko, Jonathan Jagid, Angel Claudio, Andres Kanner, Jennifer Hopp, Stephanie Chen, Jennifer Haagensen, Emily Johnson, William Anderson, Nathan Crone, Sara Inati, Kareem Zaghloul, Juan Bulacio, Jorge Gonzalez-Martinez, and Sridevi Sarma. Neural fragility as an eeg marker of the seizure onset zone. *Nature Neuroscience*, 24:1–10, 2021.

[3] Huy Phan, Kaare Mikkelsen, Oliver Y Chén, Philipp Koch, Alfred Mertins, and Maarten De Vos. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering*, 69(8):2456–2467, 2022.

[4] Chaoqi Yang, Danica Xiao, M Brandon Westover, and Jimeng Sun. Self-supervised eeg representation learning for automatic sleep staging. *JMIR AI*, page e46769, 2023.

[5] Jathurshan Pradeepkumar, Mithunjha Anandakumar, Vinith Kugathasan, Dhinesh Suntharalingham, Simon L Kappel, Anjula C De Silva, and Chamira US Edussooriya. Towards interpretable sleep stage classification using cross-modal transformers. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2024.

[6] Rikuto Kotoge, Zheng Chen, Tasuku Kimura, Yasuko Matsubara, Takufumi Yanagisawa, Haruhiko Kishima, and Yasushi Sakurai. Splitsee: A splittable self-supervised framework for single-channel eeg representation learning. *arXiv preprint arXiv:2410.11200*, 2024.

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

[8] Doina Precup and Philip Bachman. Improved estimation in time varying models. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1735–1742, 2012.

[9] Lukas Wolf, Ard Kastrati, Martyna Beata Plomecka, Jie-Ming Li, Dustin Klebe, Alexander Veicht, Roger Wattenhofer, and Nicolas Langer. A deep learning approach for the segmentation of electroencephalography data in eye tracking applications. In *International Conference on Machine Learning, ICML*, pages 23912–23932, 2022.

[10] Jingyuan Li, Leo Scholl, Trung Le, Pavithra Rajeswaran, Amy Orsborn, and Eli Shlizerman. Amag: Additive, multiplicative and adaptive graph neural network for forecasting neuron activity. In *Advances in Neural Information Processing Systems*, pages 8988–9014. Curran Associates, Inc., 2023.

[11] Lakhan Dev Sharma, Vijay Kumar Bohat, Maria Habib, Al-Zoubi Ala'M, Hossam Faris, and Ibrahim Aljarah. Evolutionary inspired approach for mental stress detection using eeg signal. *Expert systems with applications*, 197:116634, 2022.

[12] Siyi Tang, Jared Dunnmon, Khaled Kamal Saab, Xuan Zhang, Qianying Huang, Florian Dubost, Daniel Rubin, and Christopher Lee-Messer. Self-supervised graph neural networks for improved electroencephalographic seizure analysis. In *International Conference on Learning Representations*, 2022.

[13] Zheng Chen, Ziwei Yang, Lingwei Zhu, Wei Chen, Toshiyo Tamura, Naoaki Ono, Md Altaf-Ul-Amin, Shigehiko Kanaya, and Ming Huang. Automated sleep staging via parallel frequency-cut attention. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, pages 1974–1985, 2023.

[14] Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36, 2024.

[15] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[16] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[17] Yiqun Duan, Charles Zhou, Zhen Wang, Yu-Kai Wang, and Chin-teng Lin. Dewave: Discrete encoding of eeg waves for eeg to text translation. In *Thirty-seventh Conference on Neural Information Processing Systems*, pages 9907 – 9918, 2023.

[18] Ke Yi, Yansen Wang, Kan Ren, and Dongsheng Li. Learning topology-agnostic eeg representations with geometry-aware modeling. *Advances in Neural Information Processing Systems*, 36, 2024.

[19] Navid Mohammadi Foumani, Geoffrey Mackellar, Soheila Ghane, Saad Irtza, Nam Nguyen, and Mahsa Salehi. Eeg2rep: enhancing self-supervised eeg representation through informative masked inputs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5544–5555, 2024.

[20] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[21] Hanyu Wang, Saksham Suri, Yixuan Ren, Hao Chen, and Abhinav Shrivastava. Larp: Tokenizing videos with a learned autoregressive generative prior. *arXiv preprint arXiv:2410.21264*, 2024.

[22] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[23] 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.

[24] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[25] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR, 2020.

[26] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

[27] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[28] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020.

[29] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[30] Jin Jing, Wendong Ge, Shenda Hong, Marta Bento Fernandes, Zhen Lin, Chaoqi Yang, Sungtae An, Aaron F Struck, Aline Herlopian, Ioannis Karakis, et al. Development of expert-level classification of seizures and rhythmic and periodic patterns during eeg interpretation. *Neurology*, 100(17):e1750–e1762, 2023.

[31] Wei Yan Peh, Yuanyuan Yao, and Justin Dauwels. Transformer convolutional neural networks for automated artifact detection in scalp eeg. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3599–3602. IEEE, 2022.

[32] Hongli Li, Man Ding, Ronghua Zhang, and Chunbo Xiu. Motor imagery eeg classification algorithm based on cnn-lstm feature fusion network. *Biomedical signal processing and control*, 72:103342, 2022.

[33] Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie. Transformer-based spatial-temporal feature learning for eeg decoding. *arXiv preprint arXiv:2106.11170*, 2021.

[34] Amir Harati, Meysam Golmohammadi, Silvia Lopez, Iyad Obeid, and Joseph Picone. Improved eeg event classification using differential energy. In *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–4. IEEE, 2015.

[35] Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in neuroscience*, 10:196, 2016.

[36] Sebas Lopez, G Suarez, D Jungreis, I Obeid, and Joseph Picone. Automated identification of abnormal adult eegs. In *2015 IEEE signal processing in medicine and biology symposium (SPMB)*, pages 1–5. IEEE, 2015.

[37] Wendong Ge, Jin Jing, Sungtae An, Aline Herlopian, Marcus Ng, Aaron F Struck, Brian Appavu, Emily L Johnson, Gamaleldin Osman, Hiba A Haider, et al. Deep active learning for interictal ictal injury continuum eeg patterns. *Journal of neuroscience methods*, 351:108966, 2021.

[38] Ali Hossam Shoeb. *Application of machine learning to epileptic seizure onset detection and treatment*. PhD thesis, Massachusetts Institute of Technology, 2009.

[39] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*, 2021.

[40] Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in Neural Information Processing Systems*, 35:3988–4003, 2022.

[41] Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

[42] Hsiang-Yun Sherry Chien, Hanlin Goh, Christopher M Sandino, and Joseph Y Cheng. Maeeg: Masked auto-encoder for eeg representation learning. *arXiv preprint arXiv:2211.02625*, 2022.

[43] Hanwen Liu, Daniel Hajialigol, Benny Antony, Aiguo Han, and Xuan Wang. Eeg2text: Open vocabulary eeg-to-text decoding with eeg pre-training and multi-view transformer. *arXiv preprint arXiv:2405.02165*, 2024.

[44] Jiaqi Wang, Zhenxi Song, Zhengyu Ma, Xipeng Qiu, Min Zhang, and Zhiguo Zhang. Enhancing eeg-to-text decoding through transferable representations from pre-trained contrastive eeg-text masked autoencoder. *arXiv preprint arXiv:2402.17433*, 2024.

[45] Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. *Advances in Neural Information Processing Systems*, 2024.

[46] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

[47] Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Biloš, Hena Ghonia, Nadhir Hassen, Anderson Schneider, et al. Lag-llama: Towards foundation models for time series forecasting. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023.

[48] Maxwell A Xu, Alexander Moreno, Hui Wei, Benjamin M Marlin, and James M Rehg. Rebar: Retrieval-based reconstruction for time-series contrastive learning. *arXiv preprint arXiv:2311.00519*, 2023.

# Appendix

## Contents



## A Related work

**EEG Representation Learning.** To learn general representations and address issues of label scarcity in EEG data, self-supervised learning (SSL) has emerged as a prominent paradigm, and existing works can be categorized into two main approaches: contrastive learning and self-prediction. Contrastive learning methods [39, 40, 6], leverage augmentation or transformation of EEG inputs to learn consistent representations. In contrast, self-prediction methods[41, 42, 14, 19] aim to accurately reconstruct masked or corrupted input. However, their learning objectives heavily rely on cross-channel prediction to focus on spatial characteristics. In contrast, our method emphasizes inherent time-frequency features within a single-channel setting and can later adapt to any channel configuration.

**Foundation EEG Models.** Inspired by the success of foundation models in NLP, recent efforts have sought to develop foundation models for EEG analysis. These models can be categorized into decoding and encoder-based methods. Decoding-only methods focus on generative tasks like EEG-to-text translation, with representative works including DeWave [17], EEG2Text [43], and E2T-PTR [44]. In contrast, encoder-only methods concentrate on fundamental EEG classification tasks and representation learning. Notable models include LaBraM [23], BIOT [14], BRANT [45], and MMM [18]. Our work aligns with this latter category, focusing on enhancing the representation quality to improve classification performance.

**EEG Tokenization.** Tokenization has been instrumental in NLP, where discrete subword units have proven to reduce data complexity and improve model performance and interoperability. Although time-series tokenization methods have shown promise [46, 47], they do not scale well to EEGs' higher sampling rates and other artifacts. Existing attempts for EEGs include patch-based continuous tokenization, such as BIOT [14] and BRANT [45], and vector quantization (VQ)-based methods like DeWave [17]. Patch-based methods do not involve encoding or quantization, leading to unbounded and continuous representations that lack distinctiveness and interpretability. In contrast, VQ-based tokenizers, traditionally successful in tokenizing continuous images [28], have recently been adapted

for EEG by LaBraM [23], However, LaBraM employs its neural tokenizer only during pretraining but relies on raw EEG signals during inference. Conceptually, its primary role is to pre-train classification layers, rather than encoding inputs and reducing data complexity. Here, our method is explicitly VQ-based, treating the codebook as a real tokenizer for EEG data. Moreover, we enforce each token to capture time-frequency motifs[48] in EEG inputs, ensuring a more structured and interpretable representation.

# B Experiment Setup

## B.1 Datasets:

- **TUH EEG Events (TUEV)** [34]: TUEV is a subset of the TUH EEG Corpus [35], which comprises clinical EEG recordings collected at Temple University Hospital between 2002 and 2017. The dataset is annotated for six EEG event types: spike and sharp wave (SPSW), generalized periodic epileptiform discharges (GPED), periodic lateralized epileptiform discharges (PLED), eye movement (EYEM), artifact (ARTF), and background (BCKG).
- **TUH Abnormal EEG Corpus (TUAB)** [36]: TUAB comprises EEG recordings collected at Temple University Hospital, which are labeled for normal and abnormal EEG activity.
- **IIIC Seizure** [30, 37]: The IIIC Seizure dataset is curated for the detection of six distinct ictal–interictal–injury continuum (IIIC) patterns and is sourced from [30, 37]. The annotations include: (1) others (OTH), (2) seizure types (ESZ), (3) lateralized periodic discharge (LPD), (4) generalized periodic discharge (GPD), (5) lateralized rhythmic delta activity (LRDA), and (6) generalized rhythmic delta activity (GRDA).
- **CHB-MIT** [38]: The CHB-MIT dataset is a widely used benchmark for epilepsy seizure detection. It comprises EEG recordings from 23 pediatric subjects with intractable seizures.

## B.2 Dataset Statistics and Splits

Table 2: Dataset Summary

| Dataset | # of Recordings | # of Samples | Duration (s) | Task |
|---|---|---|---|---|
| TUEV | 11,914 | 112,491 | 5 | EEG Event Classification |
| IIIC Seizure | 2,689 | 135,096 | 10 | Seizure Type Classification |
| CHB-MIT | 686 | 326,993 | 10 | Seizure Detection |
| TUAB | 2,339 | 409,455 | 10 | Abnoral EEG Detection |

This section provides detailed information on the datasets used in our experiments and their respective splits. Table 2 summarizes key statistics, including the number of recordings, the total number of samples after preprocessing, their duration, and the corresponding downstream tasks. For TUEV and TUAB, we utilized the official training and test splits provided by the dataset and further divided the training splits into 80% training and 20% validation sets. We performed a subject-wise split into 60% training, 20% validation, and 20% test on the IIIC Seizure dataset. In the CHB-MIT dataset, we used 1-19 subjects for training, 20-21 for validation, and 22-23 for testing.

## B.3 Preprocessing:

We follow the preprocessing setup of BIOT [14]. Unlike LaBraM [23], which utilized 23 channels in the TUEV and TUAB datasets, we adhere to the 16-channel bipolar montage from the international 10–20 system, as used in [14]. All EEG recordings are resampled to 200 Hz. For TUEV and TUAB, we apply a bandpass filter (0.1–75 Hz) and a notch filter (50 Hz), following the preprocessing pipeline of LaBraM [23]. STFT computation of the signals is performed using PyTorch, with detailed parameters provided in Appendix B.4. For training, validation, and test splits, we follow the recommendations from [14]. Additional details on dataset statistics and splits are provided in Appendix B.2.

### B.4 STFT parameters

To extract frequency-domain representations of the EEG, we utilized the STFT function from PyTorch. The recommendations of [14] guided our parameter selection and empirical analysis of different configurations to optimize the time-frequency resolution tradeoff. The final parameters are as follows:

Table 3: STFT parameters

| Parameter | Value | Description |
|---|---|---|
| FFT size ($n_{\text{fft}}, L$) | 200 | Number of frequency bins (equal to resampling rate) |
| Hop length $H$ | 100 | Step size for sliding window ($50\%$ overlap) |
| Window type | Hann | A smoothing window function to reduce spectral leakage |
| Output representation | Magnitude | Only the absolute values of the STFT are retained |
| Centering | False | The STFT is computed without implicit zero-padding |
| One-sided output | True | Only the positive frequency components are kept |

### B.5 Baselines and Metrics:

We evaluated our approach against the baselines from [14] as well as the current state-of-the-art methods, including BIOT [14] and LaBraM [23]. BIOT and LaBraM were reproduced using their respective open-source GitHub repositories. For other baselines we use the reported best results from [14]. To ensure a fair comparison, our experiments follow a single-dataset settings, where we reproduced BIOT and LaBraM. Specifically for BIOT, we conducted their proposed unsupervised pretraining followed by fine-tuning on the same dataset. Similarly, for LaBraM, we used their base model and conducted neural tokenizer training, masked EEG modeling, and fine-tuning within the same dataset. For performance evaluation, we used balanced accuracy, Cohen's Kappa coefficient, and weighted-F1 score for multi-class classification tasks, while balanced accuracy, AUC-PR, and AUROC were used for binary classification tasks. For TUAB, we used binary cross-entropy loss for fine-tuning, while the cross-entropy loss was applied to the TUEV and IIIC datasets. Given the class imbalance in the CHB-MIT dataset, we employed focal loss for all experiments. All experiments were conducted using five different random seeds, and we report the mean and standard deviation for each metric. Also, we used Cohen's Kappa and AUROC as monitoring metric for multiclass and binary classification tasks respectively.

## C More Experiment Results

### C.1 Performance on CHB-MIT and IIIC Seizure

Table 4 and 5 presents the performance comparison of `TFM-Token` with baselines on seizure detection (CHB-MIT) and seizure type classification (IIIC Seizure) tasks. `TFM-Token` outperforms all baselines across all metrics in both datasets. On the CHB-MIT dataset with a highly imbalanced binary classification task, BIOT is the only baseline with an AUC-PR above $0.25$. However, `TFM-Token` surpasses BIOT, achieving an $8\%$ improvement in AUC-PR ($0.3127 \rightarrow 0.3379$) and a $4.5\%$ increase in AUROC ($0.8456 \rightarrow 0.8839$), demonstrating better robustness to class imbalance. For the IIIC Seizure dataset, where the task is to classify 10-second, 16-channel EEG segments into six classes, `TFM-Token` improves Cohen's Kappa by $9.5\%$ ($0.4549 \rightarrow 0.4985$) and Weighted F1 by $8.5\%$ ($0.5387 \rightarrow 0.5847$) over ContraWR, which achieves second best results.

The superior performance of `TFM-Token` across four EEG datasets shows the promise of a fully discretized framework that has the potential to enhance future EEG foundation models. These results also underscore the importance of capturing both temporal and frequency information, highlighting the critical role of frequency learning in EEG analysis.

### C.2 Importance of Joint Frequency and Temporal Modeling:

To evaluate the importance of joint frequency-temporal modeling, we conducted an ablation study comparing three tokenization variants: (1) `TFM-Token`-Raw Signal Only (`TFM-Token`-R), which uses only raw EEG patches $\{x_i\}_{i=1}^{N}$ to predict the spectrum $\mathbf{S}$, (2) `TFM-Token`-STFT Only (`TFM-Token`-S), and (3) `TFM-Token`, which jointly models both temporal and frequency features. Masked modeling

Table 4: Seizure detection performance comparison on the CHB-MIT dataset

| Models | Number of Params | CHB-MIT (seizure detection) | | |
|---|---|---|---|---|
| | | Balanced Acc. | AUC-PR | AUROC |
| SPaRCNet[30] | 0.79M | $0.5876 \pm 0.0191$ | $0.1247 \pm 0.0119$ | $0.8143 \pm 0.0148$ |
| ContraWR[4] | 1.6M | $0.6344 \pm 0.0002$ | $0.2264 \pm 0.0174$ | $0.8097 \pm 0.0114$ |
| CNN-Transformer[31] | 3.2M | $0.6389 \pm 0.0067$ | $0.2479 \pm 0.0227$ | $0.8662 \pm 0.0082$ |
| FFCL[32] | 2.4M | $0.6262 \pm 0.0104$ | $0.2049 \pm 0.0346$ | $0.8271 \pm 0.0051$ |
| ST-Transformer[33] | 3.5M | $0.5915 \pm 0.0195$ | $0.1422 \pm 0.0094$ | $0.8237 \pm 0.0491$ |
| Vanilla BIOT[14] | 3.2M | $0.6640 \pm 0.0037$ | $0.2573 \pm 0.0088$ | $0.8646 \pm 0.0030$ |
| BIOT*[14] | 3.2M | $0.6582 \pm 0.0896$ | $0.3127 \pm 0.0890$ | $0.8456 \pm 0.0333$ |
| LaBraM-Base*[23] | 5.8M | $0.5035 \pm 0.0078$ | $0.0959 \pm 0.0742$ | $0.6624 \pm 0.1050$ |
| **TFM-Token** | 1.9M | $\mathbf{0.6750 \pm 0.0392}$ | $\mathbf{0.3379 \pm 0.0515}$ | $\mathbf{0.8839 \pm 0.0173}$ |

1. Best results are **bolded**. 2. * indicates single dataset setting

Table 5: Seizure type classification performance comparison on the IIIC Seizure dataset

| Models | Number of Params | IIIC Seizure (seizure type classification) | | |
|---|---|---|---|---|
| | | Balanced Acc. | Cohen's Kappa | Weighted F1 |
| SPaRCNet[30] | 0.79M | $0.5546 \pm 0.0161$ | $0.4679 \pm 0.0228$ | $0.5569 \pm 0.0184$ |
| ContraWR[4] | 1.6M | $0.5519 \pm 0.0058$ | $0.4623 \pm 0.0148$ | $0.5486 \pm 0.0137$ |
| CNN-Transformer[31] | 3.2M | $0.5476 \pm 0.0103$ | $0.4481 \pm 0.0139$ | $0.5346 \pm 0.0127$ |
| FFCL[32] | 2.4M | $0.5617 \pm 0.0117$ | $0.4704 \pm 0.0130$ | $0.5617 \pm 0.0171$ |
| ST-Transformer[33] | 3.5M | $0.5423 \pm 0.0056$ | $0.4492 \pm 0.0056$ | $0.5440 \pm 0.0014$ |
| Vanilla BIOT[14] | 3.2M | $0.5762 \pm 0.0034$ | $0.4932 \pm 0.0046$ | $0.5773 \pm 0.0031$ |
| BIOT*[14] | 3.2M | $0.4458 \pm 0.0183$ | $0.3418 \pm 0.0228$ | $0.4511 \pm 0.0207$ |
| LaBraM-Base*[23] | 5.8M | $0.4736 \pm 0.0101$ | $0.3716 \pm 0.0128$ | $0.4765 \pm 0.0097$ |
| **TFM-Token (Ours - Single Dataset)** | 1.9M | $\mathbf{0.5775 \pm 0.0042}$ | $\mathbf{0.4985 \pm 0.0039}$ | $\mathbf{0.5847 \pm 0.0050}$ |

1. Best results are **bolded**. 2. * indicates single dataset setting

was applied for token learning in the latter two, with consistent `TFM-Encoder` training across all variants. Results are shown in Table 6. In event classification, `TFM-Token-S` improves Cohen's Kappa over `TFM-Token-R` ($0.5194 \rightarrow 0.5275$). However, in abnormal detection, `TFM-Token-R` achieves a higher AUC-PR ($0.8814 \rightarrow 0.8908$). These results indicate that different EEG tasks rely on distinct feature domains, underscoring the necessity of joint modeling. The primary `TFM-Token` consistently outperforms both single-domain approaches across all settings, further underscoring the importance of joint modeling.

### C.3 EEG Token Quality Analysis and Frequency Learning

We study the quality of the EEG tokens learned by our `TFM-Tokenizer` by analyzing four key aspects: (1) token utilization, (2) class-specific distinctiveness, (3) similar class retrieval, and (4) frequency learning capability. We conducted our analysis using all three `TFM-Tokenizer` variants and the neural tokenizer from LaBraM [23], testing them on the test splits of both the TUEV and IIIC datasets, which have multiple classes. All tokenizers employed a fixed vocabulary size of $8,192$ tokens for consistency and fair comparison.

**Token utilization and Class uniqueness:** Token utilization (%) score was calculated as the percentage of unique tokens activated from the total available vocabulary size. To quantify whether the tokenizers capture class-distinctive representations, we introduce the Class-Token Uniqueness Score, defined as:

$$\text{Class-Token Uniqueness } \% = \frac{\text{\# Unique Tokens in Class}}{\text{\# Tokens Utilized by Class}} \times 100$$

Figure 3a visualizes the class-token uniqueness scores for each class in both datasets. A robust tokenizer should capture class-distinctive tokens across all dataset classes through unsupervised pretraining. To assess this, we computed the geometric mean (GM) of class-token uniqueness scores, as shown in Table 7. Our `TFM-Tokenizer` reduces token utilization by more than two-

Table 6: Ablation study on input representation to TFM-Tokenizer

| Models | Number of Params | TUEV (event type classification) | | | TUAB (abnormal detection) | | |
|---|---|---|---|---|---|---|---|
| | | Balanced Acc. | Cohen's Kappa | Weighted F1 | Balanced Acc. | AUC-PR | AUROC |
| TFM-Token-**R** | 1.8M | <u>0.4898</u> ± 0.0105 | 0.5194 ± 0.0195 | 0.7518 ± 0.0095 | <u>0.8033</u> ± 0.0021 | <u>0.8908</u> ± 0.0027 | <u>0.8849</u> ± 0.0024 |
| TFM-Token-**S** | 1.9M | 0.4708 ± 0.0339 | <u>0.5275</u> ± 0.0314 | <u>0.7538</u> ± 0.0152 | 0.7927 ± 0.0044 | 0.8814 ± 0.0095 | 0.8836 ± 0.0052 |
| **TFM-Token** | 1.9M | **0.4943 ± 0.0516** | **0.5337 ± 0.0306** | **0.7570 ± 0.0163** | **0.8152 ± 0.0014** | **0.8946 ± 0.0008** | **0.8897 ± 0.0008** |

1. The best results are **bolded**, while the second-best are <u>underlined</u>.

Table 7: Token Utilization and class-token uniqueness comparison

| Tokenization Method | # Params | Utilization % | | Class-Token Uniqueness (GM) % | |
|---|---|---|---|---|---|
| | | TUEV | IIIC | TUEV | IIIC |
| Neural Tokenizer (LaBraM) | 8.6M | 21.13 | 15.25 | 0.034 | 0.000 |
| TFM-Tokenizer-R | 1.1M | 5.29 | 7.87 | 0.000 | 0.000 |
| TFM-Tokenizer-S | 1.1M | 13.93 | 11.04 | 0.004 | 0.619 |
| TFM-Tokenizer | 1.2M | 9.78 | 8.26 | 2.14 | 1.429 |

fold compared to the neural tokenizer on TUEV ($21.13\% \rightarrow 9.78\%$) and nearly two-fold on IIIC ($15.25\% \rightarrow 8.26\%$). It also significantly improves learning of class-unique tokens compared to neural tokenizer ($0.034\% \rightarrow 2.14\%$on TUEV, $0.0\% \rightarrow 1.429\%$ on IIIC). These results demonstrate that the TFM-Tokenizer captures more compact and useful tokens than the neural tokenizer. Additionally, TFM-Tokenizer achieves a higher class-token uniqueness score across all classes compared to TFM-Tokenizer-R ($0.0\% \rightarrow 1.429\%$ on IIIC) and TFM-Tokenizer-S ($0.619\% \rightarrow 1.429\%$ on IIIC), as depicted in Figure 3a. This further validates joint frequency-temporal modeling in EEG analysis.

**Tokens for Similar-Class Sample Mining:** We conducted an EEG signal mining experiment based on similar-class sample retrieval. Given a multi-channel EEG sample, we first obtain its discrete token representation. Using the Jaccard similarity score, we then retrieve the top $K$ most similar samples from the dataset and compute the precision score for correctly retrieving samples of the same class. For this study, we constructed a balanced subset from the IIIC and TUEV datasets and tested all four tokenization methods. The retrieval performance, illustrated in Figure 3b, shows that all TFM-Tokenizer variants significantly outperform neural tokenizer. Notably, TFM-Tokenizer-S and TFM-Tokenizer achieve nearly $60\%$ precision on the TUEV for $K = 1$. While the Jaccard similarity measure demonstrates initial feasibility, further research is needed to identify optimal metrics for token-based EEG retrieval.

**Evaluating the Frequency Learning of TFM-Tokenizer Tokens:** In this experiment, we compare the frequency and temporal-domain encoders of the TFM-Tokenizer to evaluate their ability to capture diverse frequency features in EEG signals. Specifically, we arrange all tokens in temporal order and perform a discrete Fourier transform on the token sequence. This process decomposes the tokens into frequencies, where each frequency reflects the degree of change between tokens at various scales. Larger changes indicate more diverse token representations. Then, we compute spectral entropy, defined as the normalized Shannon entropy of the amplitude values, to quantify how energy is distributed across the spectrum. Higher spectral entropy means that the model has learned a broader range of frequency features, capturing differences from both large-scale trends and fine details. Figure 4 shows that on the TUEV, TUAB, and CHBMIT datasets, the frequency encoder produces tokens with significantly higher spectral entropy than the temporal encoder. For example, on the TUEV dataset, the frequency encoder achieved an average spectral entropy of 0.26, while the temporal encoder reached only 0.14. This multi-scale sensitivity benefits downstream tasks such as classification, where learning detailed differences in EEG tokens can improve performance.

### C.4 Does TFM-Tokenizer Enhance LaBraM?

To assess the scalability of TFM-Tokenizer, we investigated its ability to enhance an existing EEG foundation model. We selected LaBraM [23], which employs a neural tokenizer solely for pretraining. This setup makes it an ideal candidate for this study. We replaced LaBraM neural tokenizer with

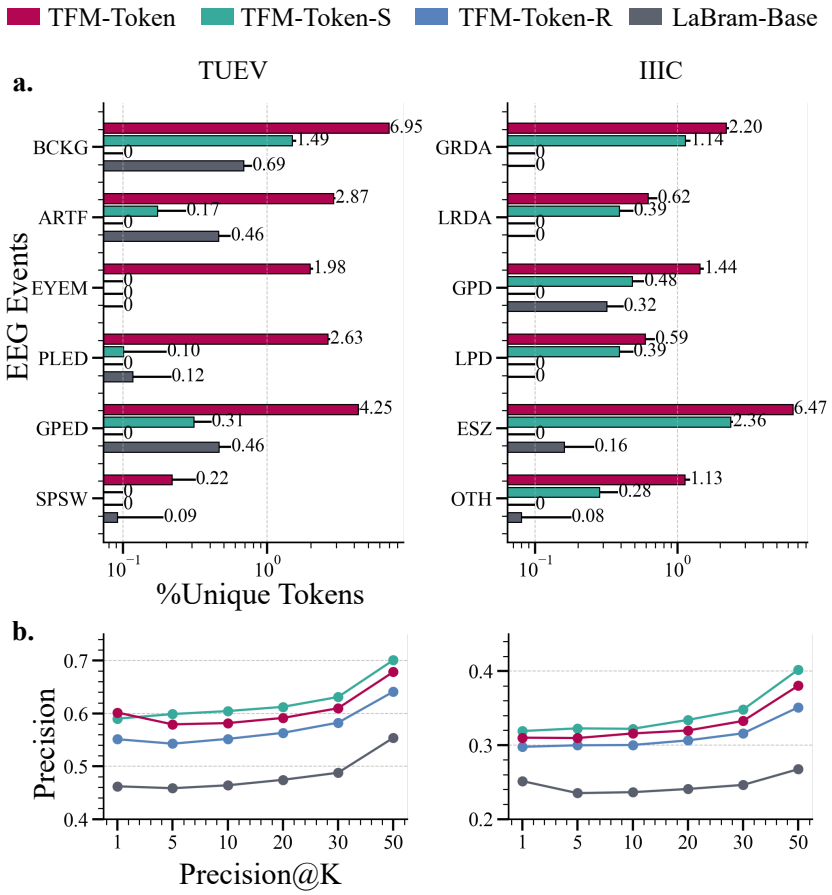


Figure 3: Analysis of token quality across three `TFM-Tokenizer` variants and the neural tokenizer. (a) Comparison of class-token uniqueness scores across all classes. (b) Retrieval performance comparison of tokenizers in a similar-class sample mining task.

Table 8: Performance Comparison of LaBraM with their neural tokenizer vs TFM-Tokenizer

| **Dataset** | **Tokenizer** | **Performance Metrics** | | |
|---|---|---|---|---|
| | | **Balanced Acc.** | **Cohen's Kappa** | **Weighted F1** |
| TUEV | Neural Tokenizer | $0.4682 \pm 0.0856$ | $0.5067 \pm 0.0413$ | $0.7466 \pm 0.0202$ |
| | `TFM-Tokenizer` | $\mathbf{0.5147 \pm 0.0174}\uparrow$ | $\mathbf{0.5220 \pm 0.0153}\uparrow$ | $\mathbf{0.7533 \pm 0.0094}\uparrow$ |
| | | **Balanced Acc.** | **AUC-PR** | **AUROC** |
| TUAB | Neural Tokenizer | $0.7720 \pm 0.0046$ | $0.8498 \pm 0.0036$ | $0.8534 \pm 0.0027$ |
| | `TFM-Tokenizer` | $\mathbf{0.7765 \pm 0.0016}\uparrow$ | $\mathbf{0.8518 \pm 0.0051}\uparrow$ | $\mathbf{0.8584 \pm 0.0022}\uparrow$ |

`TFM-Tokenizer` during the masked EEG modeling stage and evaluated its performance on TUEV and TUAB, presented in Table 8. On TUEV, LaBraM with `TFM-Tokenizer` achieves a 9% increase in balanced accuracy ($0.4682 \rightarrow 0.5147$) and a 3% increase in Cohen's Kappa ($0.5067 \rightarrow 0.5220$). On TUAB, `TFM-Tokenizer` consistently outperforms the neural tokenizer. These results confirm the capability of `TFM-Token` in enhancing the performance of EEG foundation models. The increase in balanced accuracy suggests that our tokenizer learns more class-discriminative tokens than the neural tokenizer.

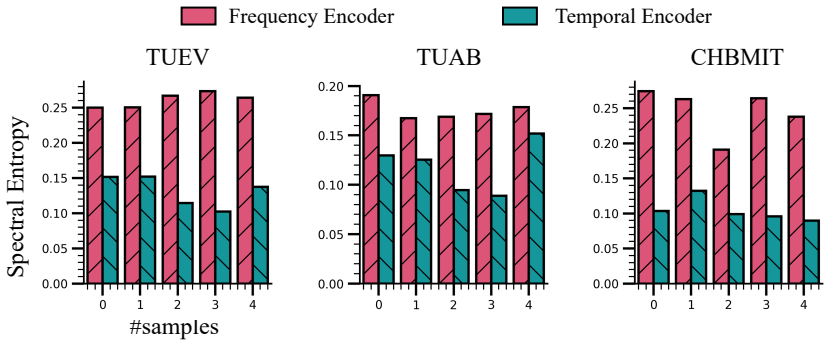


Figure 4: An analysis of how the proposed frequency and temporal-domain encoders capture frequency features, by using the spectral entropy of the learned token sequences from randomly selected samples. Higher values indicate that the tokens contain richer frequency information.

### C.5 Ablation on Masking

Table 9: Ablation on masking used for the pretraining of `TFM-Tokenizer` on TUEV Dataset

| Masking Strategy | Balanced Acc. | Cohen's Kappa | Weighted F1 |
|---|---|---|---|
| Random Masking | $0.4351 \pm 0.0462$ | $0.4772 \pm 0.0140$ | $0.7296 \pm 0.0076$ |
| Frequency Bin Masking | $0.4673 \pm 0.0540$ | $0.5193 \pm 0.0243$ | $0.7536 \pm 0.0125$ |
| Frequency Bin + Temporal Masking | $\mathbf{0.4946 \pm 0.0392}$ | $0.5045 \pm 0.0221$ | $0.7462 \pm 0.0116$ |
| Frequency Bin + Temporal Masking + Symmetric Masking | $0.4943 \pm 0.0516$ | $\mathbf{0.5337 \pm 0.0306}$ | $\mathbf{0.7570 \pm 0.0163}$ |

We conducted an ablation study on masking strategies during `TFM-Tokenizer` pretraining to assess their impact on performance. Results shown in Table 9 indicate that random masking on the spectrogram $S$ performs poorly compared to other strategies, underscoring the need for effective masking to capture frequency and temporal features from EEG. Frequency bin masking significantly improves performance over random masking, with an $8\%$ increase in Cohen's Kappa ($0.4772 \rightarrow 0.5193$) and a $7\%$ increase in balanced accuracy ($0.4351 \rightarrow 0.4673$), highlighting the importance of modeling frequency band dynamics. The addition of temporal masking further boosts balanced accuracy by $5\%$ ($0.4673 \rightarrow 0.4946$), underscoring the importance of joint temporal-frequency modeling. However, temporal masking results in a decline in Cohen's Kappa and Weighted F1, which is then resolved by introducing symmetric masking, achieving the overall best performance.

## D `TFM-Token` Implementation and Hyperparameter Tuning

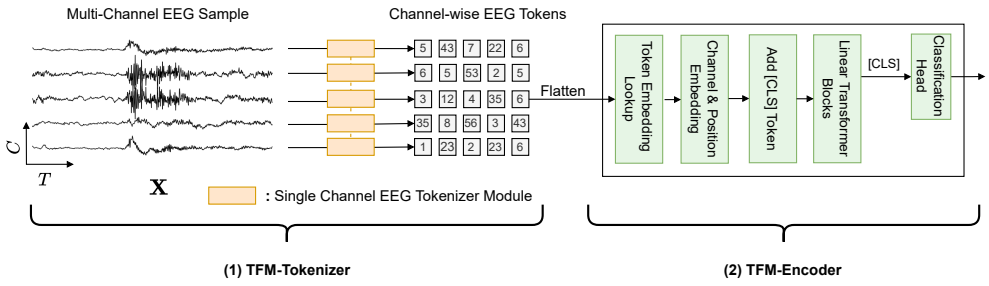


Figure 5: `TFM-Token` Overview

Figure 5 presents an overview of TFM-Token during inference. This section provides additional details on the implementation and training of the framework.

### D.1 Training Pipeline:

For all experiments, we follow a single-dataset setting, where all processes in each experiment are conducted within the same dataset. The training process of our framework is as follows: (1) TFM-Tokenizer unsupervised pretraining, (2) TFM-Encoder pretraining using masked token prediction, and finally (3) fine-tuning on the same dataset for downstream tasks.