Vector Quantization Pretraining for EEG Time Series with Random Projection and Phase Alignment

Haokun Gui [1] **Xiucheng Li** [1] **Xinyang Chen** [1]

Abstract

In this paper, we propose a BERT-style self-supervised learning model, VQ-MTM (Vector Quantization Masked Time-Series Modeling), for the EEG time series data analysis. At its core, VQ-MTM comprises a theoretically grounded random-projection quantization module and a phase-aligning module guided by the Time-Phase-Shift Equivariance of Fourier Transform, the two modules can generate well-defined semantic units (akin to words in natural language) for the corrupted and periodic time series, thus offering robust and consistent learning signals for the EEG self-supervised learning. VQ-MTM also owns low model complexity and can easily adapt to large-scale datasets. We conduct experiments on five real-world datasets including two large-scale datasets to verify the efficacy of our proposed model, the experiment results show that VQ-MTM is able to consistently surpass the existing methods by large margins on both seizure detection and classification tasks. Our code is available at https://github.com/HaokunGUI/VQ_MTM.

1. Introduction

Electroencephalogram (EEG) measures the dynamics of the electrical activity of the brain in a non-invasive way. It finds a wide spectrum of applications in diagnosing various neurological and psychiatric disorders (Schomer & Lopes da Silva, 2017; Hämäläinen et al., 1993) in medical practice. In particular, it plays a central role in seizure—the most common neurological disease affecting 50 million people worldwide (WHO, 2023)—detection and classification (Zarei et al., 2023; Guharoy et al., 2023). Clinically, EEG-based disorder diagnoses are mostly performed by a well-trained clinician who visually examines a patient's EEG observation over long periods ranging from minutes to hours (Tavares, 2020), which is extremely labor-intensive and time-consuming. Hence, it is highly appealing to design automated algorithms to accurately detect and classify the underlying disorders from the EEG signals.

Motivated by its immense practical importance, many supervised learning automated diagnosis models (Raghu et al., 2020; Gupta et al., 2021) have been developed in the past years. These methods have achieved promising results in cases where a large amount of labeled data is available. However, certain disorder types are rare in nature, which poses a serious challenge for these supervised learning models. Moreover, with the advances in data acquisition technology, an increasing amount of EEG data is being accumulated, and a large fraction of EEG data is generally normal signals (Obeid & Picone, 2016), which are simply abandoned by these methods.

The pretrain-finetune paradigm has proven remarkably effective in natural language processing (Devlin et al., 2019; Chiang et al., 2023) and computer vision (Bao et al., 2022; He et al., 2022; Liu et al., 2021). To exploit these vast quantities of unlabeled EEG data, the self-supervised learning approaches are also employed in EEG data analysis (Banville et al., 2020; Mohsenvand et al., 2020; Tang et al., 2022) and have been reported to be useful in enhancing model performance. In addition, EEG data is a special sort of time series data, and the advances in time series self-supervised learning (Wu et al., 2023; Nie et al., 2023; Dong et al., 2023) can also be applied to boost EEG data analysis.

Despite their performance gain, the present self-supervised methods for EEG often fall short concerning one of these desiderata. The successes of the self-supervised learning schemes in NLP can largely be credited to their contextual representation learning capability achieved by learning to predict or infer unseen tokens (words) in a context. However, 1) there lack of well-defined semantic units (words) in the EEG time series data. One feasible solution is to adopt the same strategy of MAE (Masked AutoEncoder) (He et al., 2022), i.e., partitioning the time series into patches and

[1]School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), China. Correspondence to: Xiucheng Li <lixiucheng@hit.edu.cn>, Xinyang Chen <chenxinyang@hit.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

learning to reconstruct a fraction of patches by using the remaining ones. But 2) unlike the images, the EEG signals are often noisy and even corrupted due to the variability during data acquisition (Wagh et al., 2022). It is prone to yielding an unstable model training process by treating corrupted data as precise reconstruction targets; moreover, the noise will also be encoded into the learned representations and is likely to degrade the self-supervised learning quality. Lastly, 3) EEG serving as a physiological signal is inherently periodic and it could be desirable to consider this periodicity for a good model design, however, it is absent from the existing self-supervised models.

To address these three issues, we introduce a BERT-style self-supervised learning method dubbed Vector Quantization Masked Time-Series Model (VQ-MTM) for the EEG data analysis. Inspired by the random projection technique in speech pretraining (Chiu et al., 2022), we design a theoretically grounded random-projection quantizer to identify semantic units from the raw EEG data. The identified semantic units not only can serve as consistent training signals (labels) to guide contextual representation learning but also are robust to various forms of data variation. In contrast to speech signals, EEG signals are inherently periodic and the random projection itself cannot guarantee to identify the phase-shifted patterns commonly arising in periodic signals. Hence, we further develop a phase-aligning module based on the Time-Phase-Shift Equivariance of the Fourier Transform to identify the variants of a pattern. The phase-aligning module enables us to better recognize semantically close words generated by phase change. We conduct experiments on five real-world EEG/ECG datasets to assess the efficacy of our proposed VQ-MTM against existing approaches. The experiments show that our proposed VQ-MTM excels in contextual representation learning and surpasses the baseline methods by large margins in both seizure detection and classification tasks.

2. Related Work

Semantic Units in Self-Supervised Learning. The BERT-style pretrain-finetune self-supervised paradigm has proven remarkably effective in natural language modeling (Devlin et al., 2019). Its core idea is to learn the contextual representations by predicting the masked tokens based on the unmasked ones. Hence, the well-defined semantic tokens are greatly important to the success of these pretraining models. Due to the lack of such tokens, the previous works (Baevski et al., 2020; Bao et al., 2022) on computer vision and speech recognition circumvent this dilemma by partitioning continuous signals into patches and learning the patches quantized representations, whereas MAE (He et al., 2022) proposes to directly reconstruct the patches. However, these methods tend to yield suboptimal solutions when applied to the corrupted time series data. Instead of learning quantizer, the works (Chiu et al., 2022; Zhang et al., 2023) propose to first map the signals by random projection, and then identify each projected vector with the index of a random vector that is close to it in terms of ℓ_2 norm. We adopt a similar strategy in the paper, however, in contrast to the ℓ_2 norm, inspired by the stochastic theory we generalize the technique to work with the cosine similarity, and consequently, the result is well supported by the stochastic theory.

Time Series Self-Supervised Learning and Analysis. Motivated by the great progress made in natural language processing and computer vision, there has been an increasing trend in developing self-supervised learning methods for time series analysis. Many of them concentrate on contrastive learning (Tonekaboni et al., 2021; Woo et al., 2022; Yue et al., 2022; Dong et al., 2023) by formulating self-supervised learning as an instance classification task. Besides, various methods (Wu et al., 2023; Nie et al., 2023; Liu et al., 2023; Wang et al., 2023; Jiang et al., 2024b; Luo & Wang, 2024; Chen et al., 2024) have been proposed to address the time series analysis tasks such as forecasting, imputation, and classification in recent years. However, none of them pays attention to the well-defined semantic units problem.

Seizure Detection and Classification. Automated seizure detection and classification have been studied extensively in literature (Covert et al., 2019; Raghu et al., 2020; Gupta et al., 2021; Iesmantas & Alzbutas, 2020) due to their immense practical value. To exploit the increasingly accumulated unlabeled data, self-supervised approaches have also been employed to analyze EEG data in recent years. The works (Banville et al., 2020; Mohsenvand et al., 2020; Kostas et al., 2021) learn the representations with contrastive loss. Xu et al. (Xu et al., 2020) proposed a pseudo-label self-supervised learning approach that assigns different labels for varying scale ratios on the original data. Tang et al. (Tang et al., 2022) propose a DCRNN-based self-supervised learning model by learning to regress the readings of future time windows. A concurrent work (Jiang et al., 2024a) also adopts the idea of vector quantization to develop a pretraining model LaBraM for EEG data based on VQ-VAE (van den Oord et al., 2017). However, there are two key distinctions. 1) The quantizer of VQ-MTM is a parameter-free function whereas the quantizer of LaBraM employs a learnable function that requires being trained first, and thus VQ-MTM is more parameter-efficient and computationally-efficient. 2) VQ-MTM is innovated with a well-designed phase alignment module and can handle more signal segment variants than LaBraM, i.e., the variants caused by time shift. Because the quantizer of LaBraM is trained by reconstructing the Fourier coefficients (amplitudes and phases), which is a time-shift-sensitive loss.

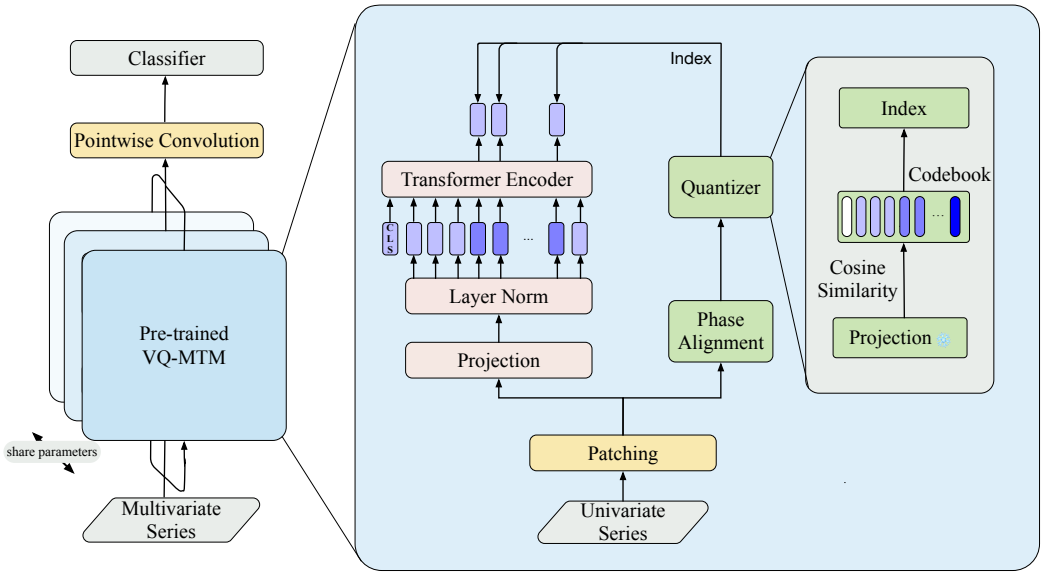

Figure 1. The fine-tuning pipeline of the model VQ-MTM. The pointwise convolution is used to fuse channel-wise representations and the [CLS] token is used to extract sequence-level representation.

3. Methodology

Problem and Notation. Given a multivariate EEG time series data with M channels sampled uniformly at T time steps, $\mathbf{X} = [\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{T-1}]$ where $\mathbf{x}_t \in \mathbb{R}^M$ is the observation at step t, we aim to learn representations for $\mathbf{X}$ without the requirement of annotation such that the learned representations (with finetuning) can generalize well to the downstream disorder diagnosis such as seizure detection and fine-grained type classification, etc. We use $\mathcal{F} : \mathbb{R}^N \mapsto \mathbb{C}^N$ to represent the Fourier Transform, whose input is a signal $v_0, v_1, \ldots, v_{N-1}$ and yield the Fourier coefficients $\hat{v}_0, \hat{v}_1, \ldots, \hat{v}_{N-1}$ by following the notation convention in analysis (Folland, 2009); the function $\mathcal{F}_k : \mathbb{R}^N \mapsto \mathbb{C}$ returns the k-th Fourier coefficient of a signal; let L_τ denote the translation operator, i.e., given a function $f : \mathbb{R} \mapsto \mathbb{R}$ the translation operator transforms it as $L_\tau(f(t)) = f(t + \tau)$.

3.1. Self-supervised Pretraining

3.1.1. Preprocessing and Pipeline

Figure 1 shows the architecture of our proposed VQ-MTM. The core idea of our approach is to identify well-defined semantic units (akin to words in natural language) from potentially corrupted and periodic multivariate time series data. To this end, we need to first partition the time series into patches along the time dimension, which shares the analogous spirit as the works in computer vision (He et al., 2022) and speech recognition (Baevski et al., 2020; Chiu et al., 2022; Zhang et al., 2023). Unlike images and speech, EEG signals are often noisy and corrupted due to variability during data acquisition (Wagh et al., 2022), as a consequence, there may exist various forms of variation for a given pattern. Moreover, the physiological signals are inherently periodic and a small shift in time may also produce different variants of a fixed pattern. Hence, it is critical to identify these variants by the same genesis pattern to provide consistent learning signals for the model. That is, to be *well-defined semantic units*, we mean that the variants of a given pattern (in a patch) should be identified with the same word in the vocabulary[1]. We address this by proposing a random-projection quantizer and a phase-aligning module based on the Johnson-Lindenstrauss lemma and Time-Phase-Shift Equivariance of the Fourier Transform, respectively.

Patching and Channel Separation. Given the input $\mathbf{X} \in \mathbb{R}^{M \times T}$ and patch size N. Suppose $T \bmod N = 0$ (can be achieved by padding or truncating $\mathbf{X}$) with $L = T/N$, we can partition $\mathbf{X}$ into L disjoint patches by reshaping it into the tensor $\mathbf{X}^{\text{re}} \in \mathbb{R}^{M \times N \times L}$. Note that $\mathbf{X}^{\text{re}}$ can be treated as a multi-channel patch sequence with length L, whose ℓ-th patch can be selected as $\mathbf{X}^{\text{re}}_{:,:,\ell}$ $(0 \le \ell < L)$, and we can then denote the patch sequence by $\left[\mathbf{X}^{\text{re}}_{:,:,\ell}\right]_{\ell=0}^{L-1}$. Since each $\mathbf{X}^{\text{re}}_{:,:,\ell}$ contains M-channel observation, such a processing manner implies we directly fuse the raw features from different channels into one latent space. However, these M channels often measure M heterogeneous physiological quantities

[1]Mathematically, this can be expressed as partitioning a collection of variants into equivalence classes and identifying each equivalence class with a canonical element.

that may have quite different magnitudes, dynamics, and physiological interpretations. Simply fusing them up in the raw feature space can mess up the subsequent learning process, e.g., it is hard to normalize features coming from heterogeneous sources with different magnitudes.

For this reason, we adopt a channel separation strategy in this paper, namely, we treat each channel observation $\mathbf{Z}_m \triangleq \mathbf{X}^{\text{re}}_{m,:,:} \in \mathbb{R}^{N\times L}$ independently and encode them separately. The benefits are two-fold: 1) each channel observation has its unique physiological dynamics and magnitude (easier to do normalization) and can be embedded into a suitable representation space; 2) the channel dimension and temporal dimension are separated, which enables the model to focus on learning temporal transition that is invariant across dimensions, and thus enhancing the parameter efficiency. We will fuse the learned contextual representations to capture channel-wise information with pointwise convolution as presented in Section 3.2 and the left part of Figure 1. Therefore, in the remainder of paper we will focus on a particular channel m and drop the subscript $\mathbf{Z}_m$ as $\mathbf{Z} = [\mathbf{z}_0, \ldots, \mathbf{z}_{L-1}]$ to keep the notation uncluttered.

VQ-MTM Pipeline. Now given a patch sequence $\mathbf{z}_0, \ldots, \mathbf{z}_{L-1} \in \mathbb{R}^N$, as Figure 1 shows, we will feed them into two branches in the pretraining stage. One branch f_{encoder} encodes them to produce the contextual representations $\mathbf{z}^c_0, \ldots, \mathbf{z}^c_{L-1} \in \mathbb{R}^{D_c}$ whereas the other one f_{label} creates them pseudo-labels $y_0, \ldots, y_{L-1}$ serving as the self-supervised learning signals. The f_{encoder} is defined as follows:

$$\begin{aligned} \mathbf{z}^0_i &= \text{LayerNorm}(\mathbf{W}\mathbf{z}_i) && \in \mathbb{R}^{D_0} \\ [\mathbf{z}^c_i]_{i=0}^{L-1} &= \text{Transformer}([\mathbf{z}^0_i]_{i=0}^{L-1}) && \in \mathbb{R}^{D_c\times L}, \end{aligned} \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{D_0\times N}$ is the parameter of linear map, whereas f_{label} is composed of the phase-aligning module and random-projection quantizer, that is,

$$y_i = \text{Quantizer}(\text{PhaseAlignment}(\mathbf{z}_i)). \tag{2}$$

Next, we detail the design of the two modules. We first present the random-projection quantizer and then discuss why only it cannot guarantee to identify the variants of phase-shifted patterns frequently arising in EEG data, hence, we introduce the phase aligning module.

3.1.2. Random-Projection Quantizer

As aforementioned, the critical part of the pretraining lies in identifying the variants of the same genesis pattern from the noisy and corrupted time series so as to provide consistent and robust learning signals for the model. Now given a patch $\mathbf{z} \in \mathbb{R}^{N-1}$ obtained from the raw EEG data, how do we identify its variants effectively? To answer this question, we draw inspiration from the stochastic process theory (Shiryaev, 2016), for two real discrete-time processes $\{X_t\}, \{Y_t\}$, their correlation $\mathbb{E}[X_tY_t]$ can serve as a good similarity measure under noise and corruption, which is defined as

$$\mathbb{E}[X_tY_t] \triangleq \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{N} X_tY_t. \tag{3}$$

Therefore, two patches can be considered variants of the same genesis pattern if they are close in terms of correlation. Inspired by this and the random-projection technique in speech pretraining (Chiu et al., 2022), we propose a random-projection quantizer as follows. Let $\mathbf{C} \in \mathbb{R}^{D\times K}$ be a randomly initialized codebook with size K, $\mathbf{c}_k \in \mathbb{R}^D$ be its k-th column, and $\mathbf{A} \in \mathbb{R}^{D\times N}$ be a randomly initialized projection matrix, we identify $\mathbf{z}$ by the index of the column $\mathbf{c}_k$ that is closest to it in terms of the correlation defined in Eq. 3, that is,

$$y = \arg\max_k \left\langle \frac{\mathbf{c}_k}{\|\mathbf{c}_k\|_2}, \frac{\mathbf{A}\mathbf{z}}{\|\mathbf{A}\mathbf{z}\|_2} \right\rangle. \tag{4}$$

The projection matrix $\mathbf{A}$ first maps $\mathbf{z}$ to $\mathbb{R}^D$, then the ℓ_2-normalization operation projects both $\mathbf{c}_k$ and $\mathbf{A}\mathbf{z}$ into the sphere $\mathbb{S}^{D-1}$ and then estimates $\mathbb{E}[X_tY_t]$ on the sphere by inner product. The codebook $\mathbf{C}$ and projection matrix $\mathbf{A}$ will not be updated once after the initialization.

Since the correlation $\mathbb{E}[X_tY_t]$ can measure similarity under noise, all $\mathbf{z}'$ have a high chance of being assigned to the same pseudo-label y of $\mathbf{z}$ in $\mathbb{R}^D$ space as long as they are close to $\mathbf{z}$ in terms of correlation in $\mathbb{R}^N$. We now give a qualitative analysis based on Theorem 3.1.

Theorem 3.1. *Given n points $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n \in \mathbb{R}^d$ and $0 < \epsilon < 1$, there exists a random matrix $\mathbf{A} \in \mathbb{R}^{m\times d}$ which randomly sampled from $\mathcal{N}(0, \frac{1}{n})$ when $m \in \mathbb{Z}_+$ and $m > \frac{24\ln(n)}{\epsilon^2}$, such that*

$$\Pr[|\langle \mathbf{v}_i, \mathbf{v}_j\rangle - \langle \mathbf{u}_i, \mathbf{u}_j\rangle| \le \epsilon] \ge 1 - \frac{2}{n}, \tag{5}$$

where $\mathbf{v}_i = \mathbf{A}\mathbf{u}_i$.

Theorem 3.1 is a generalization of the Johnson-Lindenstrauss lemma from Euclidean distance to inner product, and its proof is presented in Appendix A.2. In our case, the linear mapping is implemented as the normalized random projection $\mathbf{z} \mapsto \mathbf{A}\mathbf{z}/\|\mathbf{A}\mathbf{z}\|_2$ and it does satisfy the requirement of the theorem. Thus, we can guarantee the inequality holds with a high probability by setting D (corresponding to m in Theorem 3.1) to appropriate values.

3.1.3. Phase Alignment

Quantizer Limitation. The random-projection quantizer is effective in handling the random noise occurring in EEG

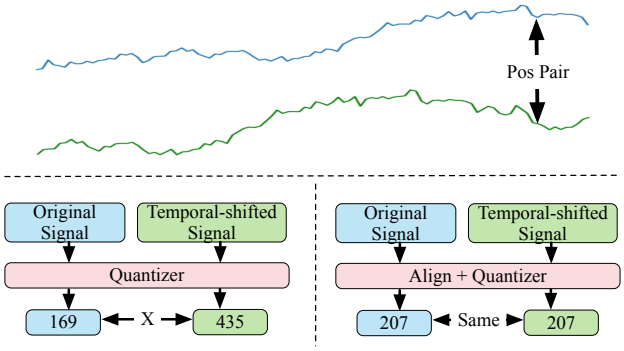

Figure 2. **Top**: A signal patch and its time-delayed counterpart generated by a minor time lay. **Bottom Left**: The pseudo-labels of the two patches produced by the Quantizer. **Bottom Right**: The pseudo-labels of the two patches produced by the Quantizer after being aligned.

during data acquisition. However, as EEG data measures the physiological signals that are often periodic, minor time lags or shifts can generate multiple different variants that may have noticeable discrepancies with their origins in terms of Eq. 3, and thus they are very likely to be treated as distinct tokens by the quantizer. Figure 2 shows an example, in which a minor time-delayed variant and its origin signal are recognized as two tokens with different semantic meanings.

To address this limitation, we introduce a phase-aligning module based on the Fourier Transform and one of its key properties, namely, the Time-Phase-Shift Equivariance. The Fourier Transform transforms a time signal $z_0, z_1, \ldots, z_{N-1}$ into its frequency representation $\hat{z}_0, \hat{z}_1, \ldots, \hat{z}_{N-1}$ through the change of basis (Fourier basis),

$$\hat{z}_k = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} z_n e^{-2\pi i \frac{n}{N} k}, \quad k = 0, 1, \ldots, N-1 \quad (6)$$

where $\hat{z}_k \in \mathbb{C}$ represents the k-th coefficient of the time series under the Fourier basis. The corresponding modulus and phase argument of $\hat{z}_k$ is denoted by $|\hat{z}_k|$ and $\arg(\hat{z}_k)$ (whose definitions are presented in Appendix A.1), respectively. As $\mathbf{z}$ and $\hat{\mathbf{z}}$ are two equivalent representations of the same vector with different bases, given its Fourier representation $\hat{\mathbf{z}}$ we can obtain $\mathbf{z}$ by the inverse Fourier Transform,

$$z_n = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \hat{z}_k e^{2\pi i \frac{k}{N} n}, \quad n = 0, 1, \ldots, N-1. \quad (7)$$

One of the appealing properties of the Fourier Transform is its Time-Phase-Shift Equivariance, which offers us a solution to address the time-shift issue of periodic signals.

Theorem 3.2. *(Time-Phase-Shift Equivariance) If a time signal $x_0, x_1, \ldots, x_{N-1}$ is shifted by τ in the time domain, then its Fourier representation $\hat{z}_k$ ($k = 0, 1, \ldots, N-1$) are shifted by $2\pi k\tau/N$ in the phase space. More precisely, let $\arg \mathcal{F}_k$ be the function composition $\arg \circ \mathcal{F}_k$, then*

$$\begin{aligned} \arg \mathcal{F}_k(L_\tau(\mathbf{x})) &= \arg(\hat{z}_k) + 2\pi k\tau/N \\ &= L_{2\pi k\tau/N}(\arg \mathcal{F}_k(\mathbf{x})). \end{aligned} \quad (8)$$

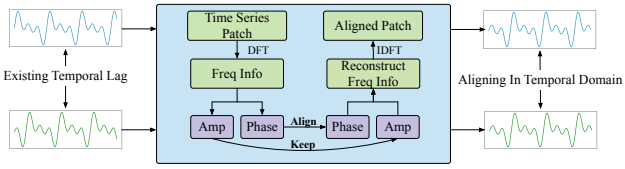

Figure 3. The structure of the phase aligning module.

The proof of the theorem can be found in Appendix A.2. Intuitively. the theorem states that shifting in the time domain is equivalent to shifting in the phase space of the Fourier representation (in the sense of equivariance). Inspired by this equivariance property, we propose to align different variants of a signal in the phase space (since it offers a natural choice for the unit shift $2\pi\tau/N$ in Eq. 8) and then reconstruct the aligned ones by the inverse Fourier Transform as follows.

$$\begin{aligned} \hat{\mathbf{z}} &= \mathrm{DFT}(\mathbf{z}), \\ \hat{\mathbf{z}}_{\mathrm{aligned}} &= \mathrm{PhaseAlign}(\hat{\mathbf{z}}), \\ \mathbf{z}_{\mathrm{aligned}} &= \mathrm{IDFT}(\hat{\mathbf{z}}_{\mathrm{aligned}}). \end{aligned} \quad (9)$$

We first compute the Fourier representation $\hat{\mathbf{z}}$ of a signal $\mathbf{z}$ and then align $\hat{\mathbf{z}}$ in the phase space, the aligned Fourier representation will be used to generate $\mathbf{z}_{\mathrm{aligned}}$ with the inverse Fourier Transform. The $\mathrm{PhaseAlign}$ is defined as follows. According to Eq. 8, we should align each $\hat{z}_k$ by preserving the modulus $|\hat{z}_k|$ and subtracting a shifted term $2\pi k\tau/N = k \cdot \arg(\hat{z}_1)$ from the phase $\arg(\hat{z}_k)$. More formally,

$$\begin{aligned} \theta_k &= \arg(\hat{z}_k) - k \cdot \arg(\hat{z}_1), \\ \hat{\mathbf{z}}_{\mathrm{aligned}} &= \left[|\hat{z}_k| e^{i\theta_k} \right]_{k=0}^{N-1}. \end{aligned} \quad (10)$$

It is also illustrated in Figure 3 where two shifted variants in time are aligned to produce an identical genesis pattern $\hat{\mathbf{z}}_{\mathrm{aligned}}$, and we will verify its efficacy in Section 4.3.

3.1.4. Self-supervised Training Loss

As shown in Figure 1, given the patch sequence of a particular channel $\mathbf{Z} = [\mathbf{z}_0, \ldots, \mathbf{z}_{L-1}]$, we create its pseudo-label sequence $y_0, \ldots, y_{L-1}$ with f_{label} and generate the contextual representation sequence $\mathbf{Z}^c = [\mathbf{z}^c_0, \ldots, \mathbf{z}^c_{L-1}]$ by feeding $\mathbf{Z}^0 = [\mathbf{z}^0_0, \ldots, \mathbf{z}^0_{L-1}]$ (calculated in Eq. 1) into the Transformer. To compute the self-supervised learning loss,

we adopt the BERT-style masked strategy, namely, randomly masking a fraction of tokens with index set $\mathcal{M}$, and replacing the masked tokens in $\mathbf{z}_0^0, \ldots, \mathbf{z}_{L-1}^0$ with one learnable vector $\mathbf{z}_{\text{masked}}$. The model is trained by maximizing the probability of predicting the pseudo-labels of masked tokens as

$$\text{maximize} \sum_{m=1}^{M} \sum_{i \in \mathcal{M}} \log p(y_{m,i} | \mathbf{z}_{m,i}^c) \tag{11}$$

where the probability $p(y_{m,i}|\mathbf{z}_{m,i}^c)$ is parameterized with the K-dimension (vocabulary size) softmax function. The learnable parameters $\mathbf{\Theta}$ include $\mathbf{W}$ and the parameters of Transformer in Eq. 1, whereas the codebook $\mathbf{C}$ and random projection matrix $\mathbf{A}$ in Eq. 4 are fixed and will not be updated once after initialization. Hence, our proposed self-supervised approach incurs no more complexity than a vanilla Transformer.

3.2. Fine-tuning on Downstream Tasks

As shown in the left of Figure 1, the pre-trained VQ-MTM (i.e., f_{encoder}) will be fine-tuned to perform the downstream tasks, and we add a [CLS] token to extract the sequence-level representation for the series-level tasks such as time-series classification. Specifically, we concatenate the pre-trained f_{encoder} with a 1×1 convolution to fuse the channel-wise representations, and then a one-hidden-layer feedforward function maps the fused representation to predict the targets, that is,

$$\begin{aligned} \mathbf{z}_{m,\text{cls}}^c &= \text{takeFirst}(f_{\text{encoder}}([\mathbf{z}_{m,i}]_{i=0}^{L-1})), \\ \mathbf{y}_{\text{pred}} &= \text{FFN}(\text{Conv}_{1\times1}([\mathbf{z}_{m,\text{cls}}^c]_{m=0}^{M-1})), \end{aligned} \tag{12}$$

where the operator takeFirst takes the first element from a sequence, namely, the contextual representation of [CLS]. The prediction $\mathbf{y}_{\text{pred}}$ will be used to fine-tune the model parameters (including f_{encoder}) with the ground-truth labels. One salient feature of our proposed approach is that it can adapt to variable-length sequences on the downstream tasks since we use the contextual representation of [CLS] to represent the sequence.

Remark. Since the model complexity of our proposed VQ-MTM is almost identical to the vanilla Transformer, it is very scalable and can easily adapt to large-scale datasets as we will show in Section 4, and its computational cost and model complexity against the existing methods are also studied in Appendix I. We also would like to highlight that VQ-MTM is not limited to the EEG data and can also benefit other corrupted and periodic time series data. Hence, we further validate its general representation learning ability by conducting more experiments on ECG data in Appendix H.

4. Experiments

We start this section by presenting the experiment settings, and then we evaluate the efficacy of VQ-MTM on two large real-world EEG datasets with the tasks of seizure detection and seizure classification. Next, we design experiments to verify the function of the channel separation strategy adopted in our self-supervised learning architecture. In the end, we examine the effectiveness of the proposed phase-aligning module by ablation study. Besides, we conduct further experiments to validate the performance and applicability of our proposed methods on more EEG and ECG data in Appendix H.

4.1. Experimental Setup

Datasets. *TUH EEG Seizure Corpus (TUSZ).* The Temple University Hospital EEG Seizure Corpus (TUSZ) v2.0.0 (Shah et al., 2018) dataset is the largest publicly available EEG dataset with a size of **79.4 GB**. The dataset contains labels for both seizure detection and seizure classification.

TUH EEG Abnormal Corpus (TUAB). TUH EEG Abnormal Corpus (TUAB) is the Temple University Hospital Abnormal Corpus (TUAB) v3.0.0 (Obeid & Picone, 2016) with a size of **58.6 GB** focusing on the annotation of seizure detection. The details of the datasets can be found in Appendix B.

Preprocessing. Following the previous studies (Tang et al., 2022), we use all 19 EEG channels, and each time series (corresponding to a patient) is partitioned into 12-second and 60-second clips without overlap. The patch size N is set to 250, which results in 12 and 60 tokens for the 12-second and 60-second clips, respectively. To evaluate the generalizability of the model to unseen patients, there is no common patient in the training and testing datasets. More details on data preprocessing are available in Appendix D.

Evaluation Tasks. We assess the quality of the self-supervised learning on EEG data with two tasks, namely, seizure detection and seizure classification. The seizure detection performs a binary classification to recognize whether a given clip of EEG data contains seizure patterns, while the seizure classification steps further to classify the specific type of seizure. Since TUSZ contains labels for both seizure detection and seizure classification(4 classes), we perform both two tasks on it, whereas we only study the seizure detection task on TUAB.

Evaluation Metrics. The datasets collected from real-world medical practice often exhibit long-tailed distributions, and this is also true for the TUSZ and TUAB datasets, in which the label imbalance issue is particularly severe (because the large fraction of a time series is non-seizure). Hence, we adopt the AUROC (Area Under the Receiver Operating Characteristic curve) and Weighted F1 Score as the

Table 1. The seizure detection (AUROC) and classification (weighted F1-score) performance of different methods on the TUSZ dataset. The best results are highlighted in **bold**, and the second best results are indicated in underline.

Model	Seizure Detection AUROC		Seizure Classification Weighted F1-Score	
	12-s	60-s	12-s	60-s
DCRNN	0.836	0.753	0.603	0.478
TimesNet	0.845	0.713	0.504	0.475
MAE	0.799	0.747	0.592	<u>0.585</u>
PatchTST	<u>0.866</u>	<u>0.834</u>	<u>0.607</u>	0.554
SimMTM	0.653	0.637	0.491	0.455
VQ-MTM	**0.887**	**0.904**	**0.620**	**0.615**
Improvement(%)	2.42	8.39	2.14	5.13

criteria to evaluate model performance for seizure classification and detection, respectively, by following the previous study (Asif et al., 2020).

Baseline Methods. We evaluate our proposed VQ-MTM against the state-of-the-art time series self-supervised learning and classification approaches including, 1) DCRNN (Tang et al., 2022), 2) TimesNet (Wu et al., 2023), 3) PatchTST (Nie et al., 2023), 4) SimMTM (Dong et al., 2023), 5) MAE (He et al., 2022), 6) BIOT (Yang et al., 2023). The description of the baseline methods is presented in Appendix E.

Training Setup. For the pretraining, our proposed model is trained using the AdamW optimizer (Loshchilov & Hutter, 2017b) with a peak learning rate of 2×10^{-3}, total training epochs of 150, and a batch size of 512. We adopt the warm-up strategy followed by the cosine annealing (Loshchilov & Hutter, 2017a) with an initial learning rate 1×10^{-4} to schedule the learning rate. The Xavier initialization (Glorot & Bengio, 2010) is used for the projection matrix $\mathbf{A}$, and the codebook $\mathbf{C}$ is initialized with the standard Normal distribution, the two matrices remain fixed once after initialization. The codebook size K is set to 1024 and the dimension D is 256. The Transformer in Eq. 1 consists of 2 layers of encoders with model dimension 256 and a feed-forward network dimension of 1024. For the methods VQ-MTM, MAE, and SimMTM, the randomly masked tokens are sampled from the Normal distribution $\mathcal{N}(0, 0.01)$. More parameter settings can be found in Appendix K.

For the finetuning, we use the cosine annealing strategy with an initial learning rate of 1×10^{-3}, and a small learning rate 1×10^{-4} is used to finetune the pre-trained parts. The total number of training epochs for fine-tuning is set to 60 for both the TUSZ and TUAB datasets.

For the baseline methods, we use the officially released code and adopt the suggested training strategies and hyperparameter settings in their original papers. All experiments are conducted on the NVIDIA RTX 4090 GPUs.

Table 2. The seizure detection (AUROC) performance of different methods on the TUAB dataset.

Model	Seizure Detection AUROC	
	12-s	60-s
DCRNN	0.806	0.772
TimesNet	<u>0.863</u>	0.853
MAE	0.852	0.841
PatchTST	0.858	0.847
SimMTM	0.724	0.697
BIOT	0.849	<u>0.865</u>
VQ-MTM	**0.868**	**0.871**
Improvement(%)	0.58	0.69

4.2. Experimental Results

The seizure detection and classification performance of different methods on the TUSZ data is presented in Table 1, whereas Table 2 shows the seizure detection results on the TUAB dataset.

Analysis of Seizure Detection. It can be observed from Table 1 and 2 our proposed VQ-MTM consistently achieves the best seizure detection performance for different lengths of clips in terms of AUROC on both datasets. Notably, it surpasses the second best baseline by a large margin (0.904 vs 0.834) for the 60-second clips on the TUSZ dataset.

Among all baseline methods, PatchTST and TimesNet show the best performance. It is noted that the detection performance drops for all baseline methods when raising the clip lengths from 12s to 60s. This can be explained by that a larger clip length is more likely to include the false positive patterns (i.e., non-seizure segments) into the clip and thus mislead the models. In contrast, the AUROC of VQ-MTM grows from 0.887 to 0.904 when the clip length increases from 12s to 60s on the TUSZ dataset, which can be credited to its superior contextual representation learning ability for long sequence under corrupted and periodic time series

Table 3. Ablation study of the channel separation strategy. VQ-MTM (MIX) is a variant of VQ-MTM that fuses the channel features first before feeding to the pretraining model.

Model	Detection AUROC		Classification Weighted F1-Score	
	12-s	60-s	12-s	60-s
VQ-MTM(Mix)	0.838	0.856	0.507	0.594
VQ-MTM	**0.887**	**0.904**	**0.620**	**0.615**
Improvement(%)	5.85	5.61	22.29	3.53

Table 4. Ablation study of the phase aligning module on the TUSZ dataset.

Model	Detection AUROC		Classification Weighted F1-Score	
	12-s	60-s	12-s	60-s
VQ-MTM w/o Phase Aligning Module	0.870	0.880	0.550	0.507
VQ-MTM w Phase Aligning Module	**0.887**	**0.904**	**0.620**	**0.615**
Improvement(%)	1.95	2.73	12.73	21.30

signals. In addition, the performance gap between VQ-MTM and MAE also verifies our hypothesis that learning to reconstruct the corrupted time series directly can degrade the quality of the learned contextual representations.

Analysis of Seizure Classification. As shown in Table 1, VQ-MTM also gives rise to the best weighted F1-score on the seizure classification tasks on the TUSZ dataset. It improves the weighted F1-score over the best baseline method by 5.13% for the 60s clips. The results show that our proposed VQ-MTM is not only helpful for seizure detection but also beneficial to fine-grained seizure classification, which further validates its general contextual representation learning capability.

On both tasks and datasets, SimMTM yields the least desirable results. We hypothesize that the poor performance of SimMTM is due to its scalability issue. Its generation of various copies of positive pairs incurs significant memory burdens, and thus it cannot well fit to large datasets such as TUSZ and TUAB. To verify this hypothesis, we further conduct experiments on several relatively small datasets in Appendix H.

4.3. Ablation Study

Channel Separation. As mentioned in Section 3.1.1 we adopt the channel separation strategy in our self-supervised learning architecture, that is, we first pass each channel independently through the pretraining model to compute the contextual representation and then fuse the channel contextual representations with a pointwise convolution when fine-tuning. To verify the efficacy of this strategy, we design a variant of VQ-MTM that fuses the channel features first before feeding to the pretraining model, referred to as VQ-MTM (MIX). The results on the TUSZ dataset are presented in Table 3. As the table shows, VQ-MTM consistently achieves much better results on both seizure detection and classification tasks. This suggests that the channel separation strategy is very effective for time series with multiple channels coming from heterogeneous sources in the self-supervised learning, this is in accordance with the observation in time series forecasting (Nie et al., 2023).

Phase Alignment. To verify the effectiveness of our proposed phase-aligning module, we design experiments on the TUSZ dataset by removing it from our proposed model, and the results are shown in Table 4. Both results on the seizure detection and classification tasks demonstrate the efficacy of the proposed module. The performance gains are more evident for the seizure classification task (up to 21.30%). The experiments show that the random-projection quantization itself is not sufficient to ensure well-defined semantic units for periodic time series data, in which multiple variants caused by small time shifts frequently emerge. It is essential to take seasonality into consideration when developing self-supervised learning models for the periodic signals.

5. Conclusion

In this paper, we present a BERT-style self-supervised model VQ-MTM for the EEG data with the consideration of its inherent characteristics. Our proposed VQ-MTM is able to generate well-defined semantic units for corrupted and periodic time series, which can serve as robust and consistent learning signals for the pretraining. To this end, we design a random-projection quantization module based on the Johnson-Lindenstrauss lemma and a phase-aligning module guided by the Time-Phase-Shift Equivariance of the Fourier Transform, the resulting model is also very scalable. The experiments on two large real-world EEG data demonstrate the efficacy of the proposed modules as well as the consistent superiority of VQ-MTM over the baseline methods on both seizure detection and classification. It is worthwhile pointing out that our proposed method is not limited to EEG data and can benefit more general time series data with corruption and periodicity, hence, we also design experiments on three more datasets including two ECG datasets to verify the generalizability of our proposed model. In the future, we would like to generalize VQ-MTM to more general time series data. In addition, since the random-projection and codebook matrix are fixed during model training, we would also like to design more effective initialization strategies for them in the case of more general time series self-supervised learning.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant No. 62206074 and No. 62306085, Shenzhen College Stability Support Plan under Grant No. GXWD20220811173233001 and No. GXWD20231130151329002.

Impact Statement

We study the self-supervised learning method for EEG time series by designing inductive bias to consider its inherent characteristics. The core idea is to generate well-defined semantic units for corrupted and periodic time series. Our method is guided by the random projection technique and Fourier analysis theory. We mainly focus on scientific research and there is no obvious negative impact on the community.

References

Asif, U., Roy, S., Tang, J., and Harrer, S. Seizurenet: Multispectral deep feature learning for seizure type classification. In *The Medical Image Computing and Computer Assisted Interverntion Society (MICCAI)*, 2020.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Bagnall, A. J., Dau, H. A., Lines, J., Flynn, M., Large, J., Bostrom, A., Southam, P., and Keogh, E. J. The UEA multivariate time series classification archive, 2018. *CoRR*, 2018.

Banville, H. J., Chehab, O., Hyvärinen, A., Engemann, D., and Gramfort, A. Uncovering the structure of clinical EEG signals with self-supervised learning. *CoRR*, 2020.

Bao, H., Dong, L., Piao, S., and Wei, F. Beit: BERT pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*, 2022.

Chen, X., Li, X., Liu, B., and Li, Z. Biased temporal convolution graph network for time series forecasting with missing values. In *International Conference on Learning Representations (ICLR)*, 2024.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

Chiu, C., Qin, J., Zhang, Y., Yu, J., and Wu, Y. Self-supervised learning with random-projection quantizer for speech recognition. In *International Conference on Machine Learning (ICML)*, 2022.

Covert, I. C., Krishnan, B., Najm, I., Zhan, J., Shore, M., Hixson, J., and Po, M. J. Temporal graph convolutional networks for automatic seizure detection. In *Machine Learning for Healthcare Conference (MLHC)*, 2019.

Dau, H. A., Bagnall, A. J., Kamgar, K., Yeh, C. M., Zhu, Y., Gharghabi, S., Ratanamahatana, C. A., and Keogh, E. J. The UCR time series archive. *IEEE CAA J. Autom. Sinica*, 2019.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics(NAACL)*, 2019.

Dong, J., Wu, H., Zhang, H., Zhang, L., Wang, J., and Long, M. Simmtm: A simple pre-training framework for masked time-series modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Folland, G. B. *Fourier analysis and its applications*, volume 4. American Mathematical Soc., 2009.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Artificial Intelligence and Statistics (AISTATS)*, 2010.

Guharoy, R., Jana, N. D., Biswas, S., and Garg, L. Empirical analysis of different dimensionality reduction and classification techniques for epileptic seizure detection, 2023.

Gupta, S., Meena, J., and Gupta, O. P. Neural network based epileptic EEG detection and classification. *CoRR*, 2021.

Hämäläinen, M. S., Hari, R., Ilmoniemi, R. J., Knuutila, J., and Lounasmaa, O. V. Magnetoencephalography-theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics*, 1993.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. B. Masked autoencoders are scalable vision learners. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Iesmantas, T. and Alzbutas, R. Convolutional neural network for detection and classification of seizures in clinical data. *Medical Biol. Eng. Comput.*, 2020.

Jiang, W., Zhao, L., and liang Lu, B. Large brain model for learning generic representations with tremendous eeg data in bci. In *International Conference on Learning Representations (ICLR)*, 2024a.

Jiang, Y., Li, X., Chen, Y., Liu, S., Kong, W., Lentzakis, A. F., and Cong, G. A scalable adaptive graph diffusion

forecasting network for multivariate time series forecasting. In *IEEE International Conference on Data Engineering (ICDE)*, 2024b.

Kostas, D., Aroca-Ouellette, S., and Rudzicz, F. BENDR: using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *CoRR*, 2021.

Liu, S., Li, X., Cong, G., Chen, Y., and Jiang, Y. Multivariate time-series imputation with disentangled temporal representations. In *International Conference on Learning Representations (ICLR)*, 2023.

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., and Long, M. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision (ICCV)*, 2021.

Loshchilov, I. and Hutter, F. SGDR: stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017a.

Loshchilov, I. and Hutter, F. Fixing weight decay regularization in adam. *CoRR*, 2017b.

Luo, D. and Wang, X. Moderntcn: A modern pure convolution structure for general time series analysis. In *International Conference on Learning Representations (ICLR)*, 2024.

Mohsenvand, M. N., Izadi, M. R., and Maes, P. Contrastive representation learning for electroencephalogram classification. In *Machine Learning for Health Workshop, ML4H@NeurIPS*, 2020.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

Obeid, I. and Picone, J. W. The temple university hospital eeg data corpus. *Frontiers in Neuroscience*, 10, 2016.

Raghu, S., Sriraam, N., Temel, Y., Rao, S. V., and Kubben, P. L. EEG based multi-class seizure type classification using convolutional neural network and transfer learning. *Neural Networks*, 2020.

Schomer, D. L. and Lopes da Silva, F. H. *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Oxford University Press, 2017.

Shah, V., Weltin, E. V., de Diego, S. L., McHugh, J. R., Veloso, L., Golmohammadi, M., Obeid, I., and Picone, J. The temple university hospital seizure detection corpus. *Frontiers Neuroinformatics*, 2018.

Shiryaev, A. N. *Probability-1*, volume 95. Springer, 2016.

Tang, S., Dunnmon, J., Saab, K. K., Zhang, X., Huang, Q., Dubost, F., Rubin, D. L., and Lee-Messer, C. Self-supervised graph neural networks for improved electroencephalographic seizure analysis. In *International Conference on Learning Representations (ICLR)*, 2022.

Tavares, T. P. *Electroencephalography (EEG)*, pp. 1266–1269. Springer International Publishing, Cham, 2020. ISBN 978-3-319-24612-3.

Tonekaboni, S., Eytan, D., and Goldenberg, A. Unsupervised representation learning for time series with temporal neighborhood coding. In *International Conference on Learning Representations (ICLR)*, 2021.

van den Oord, A., Vinyals, O., and Kavukcuoglu, K. Neural discrete representation learning. In *Advances in Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Wagh, N., Wei, J., Rawal, S., Berry, B. M., and Varatharajah, Y. Evaluating latent space robustness and uncertainty of EEG-ML models under realistic distribution shifts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Wang, H., Peng, J., Huang, F., Wang, J., Chen, J., and Xiao, Y. Micn: Multi-scale local and global context modeling for long-term series forecasting. In *International Conference on Learning Representations (ICLR)*, 2023.

WHO. Epilepsy. https://www.who.int/news-room/fact-sheets/detail/epilepsy, February 2023.

Woo, G., Liu, C., Sahoo, D., Kumar, A., and Hoi, S. C. H. Cost: Contrastive learning of disentangled seasonal-trend representations for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2022.

Wu, H., Xu, J., Wang, J., and Long, M. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., and Long, M. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations (ICLR)*, 2023.

Xu, J., Zheng, Y., Mao, Y., Wang, R., and Zheng, W. Anomaly detection on electroencephalography with self-supervised learning. In *International Conference on Bioinformatics and Biomedicine (BIBM)*, 2020.

Yang, C., Westover, M. B., and Sun, J. BIOT: biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems (NeurIPS 2023)*, 2023.

Yue, Z., Wang, Y., Duan, J., Yang, T., Huang, C., Tong, Y., and Xu, B. Ts2vec: Towards universal representation of time series. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2022.

Zarei, A., Zhu, B., and Shoaran, M. Enhancing epileptic seizure detection with eeg feature embeddings, 2023.

Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z., Chen, N., Li, B., Axelrod, V., Wang, G., Meng, Z., Hu, K., Rosenberg, A., Prabhavalkar, R., Park, D. S., Haghani, P., Riesa, J., Perng, G., Soltau, H., Strohman, T., Ramabhadran, B., Sainath, T. N., Moreno, P. J., Chiu, C., Schalkwyk, J., Beaufays, F., and Wu, Y. Google USM: scaling automatic speech recognition beyond 100 languages. *CoRR*, 2023.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2021.

Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., and Jin, R. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning (ICML)*, 2022.

A. Background and Theorem Proof

A.1. Background on Complex Numbers

Given a complex number $\hat{z} \in \mathbb{C}$, its modulus and phase argument is defined as follows,

$$\begin{aligned} |\hat{z}| &= \sqrt{\mathrm{Re}(\hat{z})^2 + \mathrm{Im}(\hat{z})^2}, \\ \arg(\hat{z}) &= \arctan\left(\mathrm{Re}(\hat{z}) / \mathrm{Im}(\hat{z})\right) \end{aligned} \tag{13}$$

where $\mathrm{Re}(\cdot)$ and $\mathrm{Im}(\cdot)$ indicate the operations of taking real and imaginary parts of a complex variable, respectively.

A.2. Thorem Proof

Lemma A.1. *Let $\mathbf{u} \in \mathbb{R}^m$ be a vector whose entries are independently and identically distributed according to the standard normal distribution $\mathcal{N}(0,1)$. For any given $\epsilon \in (0,1)$, it holds that*

$$\Pr[\big|\|\mathbf{u}\|^2 - 1\big| \geq \epsilon] \leq 2\exp\left(-\frac{\epsilon^2 m}{8}\right). \tag{14}$$

Proof. We employ the Chernoff bound and Markov's inequality to establish the tail bounds for $\|\mathbf{u}\|^2$. Start by noting that for any $x \geq a$ and any $\lambda > 0$, it holds that $e^{\lambda x} \geq e^{\lambda a}$, which suggests

$$\Pr[x \geq a] \leq \min_{\lambda > 0} e^{-\lambda a}\mathbb{E}[e^{\lambda x}], \tag{15}$$

according to Markov's inequality.

Applying this to $\|\mathbf{u}\|^2 - 1 \geq \epsilon$, we find

$$\Pr[\|\mathbf{u}\|^2 - 1 \geq \epsilon] \leq \min_{\lambda > 0} e^{-\lambda\epsilon}\mathbb{E}[e^{\lambda(\|\mathbf{u}\|^2 - 1)}]. \tag{16}$$

Since the components of $\mathbf{u}$ are independent,

$$\mathbb{E}[e^{\lambda\|\mathbf{u}\|^2}] = \prod_{i=1}^{n} \mathbb{E}[e^{\lambda \mathbf{u}_i^2}], \tag{17}$$

where

$$\mathbb{E}[e^{\lambda \mathbf{u}_i^2}] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\mathbf{u}_i^2/2} e^{\lambda \mathbf{u}_i^2/m} du_i = \sqrt{\frac{m}{m - 2\lambda}}, \tag{18}$$

assuming $\lambda < \frac{m}{2}$.

Assembling the above results,

$$\Pr[\|\mathbf{u}\|^2 - 1 \geq \epsilon] \leq \min_{\lambda > 0} e^{-\lambda(\epsilon+1)} \left(\frac{m}{m - 2\lambda}\right)^{m/2}. \tag{19}$$

Minimizing over λ gives $\lambda = \frac{m\epsilon}{2(\epsilon+1)}$, leading to

$$\Pr[\|\mathbf{u}\|^2 - 1 \geq \epsilon] \leq \exp\left(-\frac{m\epsilon^2}{8}\right). \tag{20}$$

Similarly,

$$\Pr[1 - \|\mathbf{u}\|^2 \geq \epsilon] \leq \exp\left(-\frac{m\epsilon^2}{8}\right). \tag{21}$$

Finally, by the union bound,

$$\Pr[\big|\|\mathbf{u}\|^2 - 1\big| \geq \epsilon] \leq 2\exp\left(-\frac{m\epsilon^2}{8}\right). \tag{22}$$

□

Lemma A.2. *(Johnson–Lindenstrauss lemma) Given n points $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n \in \mathbb{R}^d$ and $0 < \epsilon < 1$, there exists a random matrix $\mathbf{A} \in \mathbb{R}^{m \times d}$ which randomly sampled from $\mathcal{N}(0, \frac{1}{n})$ when $m \in \mathbb{Z}_+$ and $m > \frac{24 \ln(n)}{\epsilon^2}$, such that*

$$\Pr\left[\exists(i,j) : \left| \left\| \frac{\mathbf{v}_i - \mathbf{v}_j}{\mathbf{u}_i - \mathbf{u}_j} \right\|^2 - 1 \right| \geq \epsilon \right] \leq \frac{1}{n} \tag{23}$$

where $\mathbf{v}_i = \mathbf{A}\mathbf{u}_i$.

Proof. If $\mathbf{u}$ is a unit vector, and $\mathbf{A}$ is randomly sampled from $\mathcal{N}(0, \frac{1}{n})$, then each component of $\mathbf{A}\mathbf{u}$ satisfies $\mathcal{N}(0, \frac{1}{n})$.

Now assume $\mathbf{u} = \frac{\mathbf{u}_i - \mathbf{u}_j}{\|\mathbf{u}_i - \mathbf{u}_j\|}$, by using lemma A.1, we can get

$$\Pr\left[\left| \left\| \frac{\mathbf{A}(\mathbf{u}_i - \mathbf{u}_j)}{\|\mathbf{u}_j - \mathbf{u}_j\|} \right\|^2 - 1 \right| \geq \epsilon \right] \leq 2 \exp\left(-\frac{\epsilon^2 m}{8} \right) \tag{24}$$

when $i \neq j$.

Thus

$$\Pr\left[\exists(i,j) : \left| \left\| \frac{\mathbf{A}(\mathbf{u}_i - \mathbf{u}_j)}{\|\mathbf{u}_i - \mathbf{u}_j\|} \right\|^2 - 1 \right| \geq \epsilon \right] \leq 2\binom{n}{2} \exp\left(-\frac{\epsilon^2 m}{8} \right) \tag{25}$$

Considering $m > \frac{24 \ln(n)}{\epsilon^2}$,

$$2\binom{n}{2} \exp\left(-\frac{\epsilon^2 m}{8} \right) < \frac{1}{n} \tag{26}$$

Thus,

$$\Pr\left[\exists(i,j) : \left| \left\| \frac{\mathbf{v}_i - \mathbf{v}_j}{\mathbf{u}_i - \mathbf{u}_j} \right\|^2 - 1 \right| \geq \epsilon \right] \leq \frac{1}{n} \tag{27}$$

□

Theorem A.3. *Given n points $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n \in \mathbb{R}^d$ and $0 < \epsilon < 1$, there exists a random matrix $\mathbf{A} \in \mathbb{R}^{m \times d}$ which randomly sampled from $\mathcal{N}(0, \frac{1}{n})$ when $m \in \mathbb{Z}_+$ and $m > \frac{24 \ln(n)}{\epsilon^2}$, such that*

$$\Pr\left[|\langle \mathbf{v}_i, \mathbf{v}_j \rangle - \langle \mathbf{u}_i, \mathbf{u}_j \rangle| \leq \epsilon \right] \geq 1 - \frac{2}{n}, \tag{28}$$

where $\mathbf{v}_i = \mathbf{A}\mathbf{u}_i$.

Proof. As the Johnson–Lindenstrauss lemma demonstrates,

$$\Pr\left[\exists(i,j) : \left| \left\| \frac{\mathbf{v}_i - \mathbf{v}_j}{\mathbf{u}_i - \mathbf{u}_j} \right\|^2 - 1 \right| \geq \epsilon \right] \leq \frac{1}{n}. \tag{29}$$

Considering the two events

$$\Pr\left[\exists(i,j) : \left| \left\| \frac{\mathbf{v}_i - \mathbf{v}_j}{\mathbf{u}_i - \mathbf{u}_j} \right\|^2 - 1 \right| \geq \epsilon \right] \leq \frac{1}{n}, \tag{30}$$

$$\Pr\left[\exists(i,j) : \left| \left\| \frac{\mathbf{v}_i + \mathbf{v}_j}{\mathbf{u}_i + \mathbf{u}_j} \right\|^2 - 1 \right| \geq \epsilon \right] \leq \frac{1}{n}. \tag{31}$$

The second inequality holds due to $-\mathbf{u}_j \in \mathbb{R}^d$ and $-\mathbf{v}_j = \mathbf{A}(-\mathbf{u}_i)$. Thus, the probability of the event that at least one of them occurs (the union of the two events) should be less than $\frac{2}{n}$. Consequently, we obtain the following inequalities

$$(1-\epsilon)||\mathbf{u}_i - \mathbf{u}_j||^2 \leq ||\mathbf{v}_i - \mathbf{v}_j||^2 \leq (1+\epsilon)||\mathbf{u}_i - \mathbf{u}_j||^2 \tag{32}$$

$$(1-\epsilon)||\mathbf{u}_i + \mathbf{u}_j||^2 \leq ||\mathbf{v}_i + \mathbf{v}_j||^2 \leq (1+\epsilon)||\mathbf{u}_i + \mathbf{u}_j||^2. \tag{33}$$

By multiplying -1 to the first inequality and adding it to the second one, we obtain

$$4\langle \mathbf{u}_i, \mathbf{u}_j\rangle - 2\epsilon(||\mathbf{u}_i||^2 + ||\mathbf{u}_j||^2) \leq 4\langle \mathbf{v}_i, \mathbf{v}_j\rangle \leq 4\langle \mathbf{u}_i, \mathbf{u}_j\rangle + 2\epsilon(||\mathbf{u}_i||^2 + ||\mathbf{u}_j||^2) \tag{34}$$

As $\mathbf{u}_i$ is the normalized vector, we derive the result

$$\Pr\left[|\langle \mathbf{v}_i, \mathbf{v}_j\rangle - \langle \mathbf{u}_i, \mathbf{u}_j\rangle| \leq \epsilon\right] \geq 1 - \frac{2}{n}. \tag{35}$$

□

Theorem A.4. *If a time signal $x_0, x_1, \ldots, x_{N-1}$ is shifted by τ in the time domain, then its Fourier representation $\hat{z}_k$ ($k = 0, 1, \ldots, N-1$) are shifted by $2\pi k\tau/N$ in the phase space. More precisely, let $\arg \mathcal{F}_k$ be the function composition $\arg \circ \mathcal{F}_k$, then*

$$\begin{aligned} \arg \mathcal{F}_k(L_\tau(\mathbf{x})) &= \arg(\hat{z}_k) + 2\pi k\tau/N \\ &= L_{2\pi k\tau/N}(\arg \mathcal{F}_k(\mathbf{x})). \end{aligned} \tag{36}$$

Proof. For ease of argument, we prove the theorem for the continuous signal $x : \mathbb{R} \mapsto \mathbb{R}$ and the conclusion generalizes to the discrete signals. Recall that the Fourier Transform of a real-valued function $x(t)$ on $\mathbb{R}$ can be described as

$$\hat{x}(\omega) = \int_{-\infty}^{\infty} x(t)e^{-i\omega t}dt. \tag{37}$$

Hence, we have

$$\widehat{L_\tau(x)}(\omega) = \int_{-\infty}^{\infty} L_\tau(x(t))e^{-i\omega t}dt \tag{38}$$

$$= \int_{-\infty}^{\infty} x(t+\tau)e^{-i\omega t}dt \tag{39}$$

$$= \int_{-\infty}^{\infty} x(t+\tau)e^{-i\omega t}dt \tag{40}$$

$$= \int_{-\infty}^{\infty} x(t)e^{-i\omega(t-\tau)}dt \quad (t \triangleq t-\omega) \tag{41}$$

$$= e^{i\omega\tau}\int_{-\infty}^{\infty} x(t)e^{-i\omega t}dt \tag{42}$$

$$= e^{i\omega\tau}\hat{x}(\omega). \tag{43}$$

where in the fourth line we use the change of variable. Thus, $\arg(\widehat{L_\tau(x)}(\omega)) = \omega\tau + \arg(\hat{x}(\omega)) = L_{\omega\tau}(\arg(\hat{x}(\omega)))$.

For a discrete signal of length N, the frequence ω becomes $2\pi k/N$ for $k = 0, 1, \ldots, N-1$, and the proof is completed. □

B. Dataset Details

Temple University EEG Seizure Corpus[2]. TUSZ is a subset of the TUH EEG Corpus and contains sessions that are known to contain seizure events. It comprises 7,337 files, with a cumulative recording length exceeding 5,312,996 seconds, spanning approximately 79.4 GB. Annotations are available in event-based, term-based, and bi-class formats. The channel configurations within the dataset vary from 10 to 40 channels. Detailed information is presented in Table 5.

Temple University EEG Abnormal Corpus[3]. TUAB is a subset of the TUH EEG Corpus, containing EEG records classified as either clinically normal or abnormal. The dataset comprises 1,385 normal subjects and 998 abnormal subjects, with a total size of 58.6 GB. Detailed information is presented in Table 6.

UCR & UEA[4]. To further illustrate the performance of the VQ-MTM on smaller-scale datasets, three bioelectricity-related datasets are selected from the UCR & UEA datasets, namely Epilepsy, ECG200, and ECG5000. Detailed information is provided in Table 7.

[2] https://isip.piconepress.com/projects/tuh_eeg/downloads/tuh_eeg_seizure
[3] https://isip.piconepress.com/projects/tuh_eeg/downloads/tuh_eeg_abnormal
[4] https://www.timeseriesclassification.com/dataset.php

Table 5. Summary of data in dataset of TUSZ v2.0.0.

	EEG FILES(% SEIZURE)	PATIENTS(% SEIZURE)	TOTAL DURATION(%SEIZURE)
TRAIN SET	4,664(18.70%)	579(35.92%)	54,620.48 MIN(5.34%)
DEV SET	1,832(17.69%)	53(84.91%)	26,132.87 MIN(5.09%)
EVAL SET	881(22.13%)	43(79.10%)	7,796.58 MIN(5.82%)

Table 6. Summary of data in dataset of TUAB v3.0.0.

	EEG FILES	SESSIONS	SUBJECTS
ABNORMAL	1,472(49.18%)	1,472(49.18%)	998(41.88%)
NORMAL	1,521(50.82%)	1,521(50.82%)	1,385(58.12%)
TOTAL	2,993(100%)	2,993(100%)	2,383(100%)

C. Dataset Visualization

To better illustrate the differences between seizure clips and non-seizure clips in the TUH EEG Corpus, we randomly select one patient to showcase the characteristics of seizure events. As depicted in the top section of Figure 4, the seizure part (annotated in red) exhibits a larger standard variance compared to normal segments. When aligning the seizure and non-seizure clips along the timeline, as depicted at the bottom of Figure 4, we can have a more intuitive impression on it.

D. Data Preprocessing

Temple University EEG Seizure Corpus. As the data derived from the Temple University EEG Seizure Corpus (TUSZ) uses a variety of channel configurations, we adopt a channel selection approach consistent with the prior EEG preprocessing methodology as detailed in Tang's work (Tang et al., 2022). We employ a total of 19 channels, and systematically exclude the recorded EEG signals that do not encompass the entirety of these 19 channels. This culling process results in the exclusion of a total of 265 files during the preprocessing. Additionally, we implement the subsequent configurations to derive sampled EEG clips along with their corresponding labels.

In contrast to previous approaches (Tang et al., 2022) that primarily focus on the frequency information of EEG signals, our work only focuses on the temporal domain. Specifically, in the context of self-supervised learning, we extract relevant EEG clips using 12-second (60-second) segments as input. Moreover, considering the self-supervised learning objective of DCRNN, which involves predicting future signals, we designate the 12-second clips contiguous to the input as the pretraining target. Throughout this process, clips that do not cover the full length of the combined input and target clips are excluded.

Regarding the seizure detection task, the related EEG clips are acquired through the utilization of non-overlapping 12-second (60-second) sliding windows applied to the EEG signals. Clips not satisfying the length of 12 seconds (60 seconds) are excluded from datasets. The labeling schema assigns a value of 1 for the seizure type and 0 for the non-seizure type. The detailed informaton of the preprocessed dataset for the seizure detection task are elucidated in Table 8.

In the context of seizure classification tasks, we also conduct the experiments on 12-second (60-second) clips. The 12-second (60-second) clip is acquired preceding the annotated seizure time of 2 seconds. This configuration aligns with prior research by Tang et al. (Tang et al., 2022), wherein the 2-second offset serves as a tolerance for annotation. Additionally, to avoid the occurrence of two seizure types within a single clip, the clips are truncated before the second seizure type manifests. After applying the Z-Score Normalization, the truncated clips are then padded with zeros to ensure the same clip length across all instances.

Moreover, aligning with the data preprocessing methodology in DCRNN (Tang et al., 2022), we merge the simple partial (SP) seizures, focal non-specific (FN) seizures, and complex partial (CP) seizures into a newly defined class termed combined focal (CF) seizures. In addition, tonic-clonic seizures are merged with tonic seizures, forming a class labeled combined tonic (CT) seizures. As a result, seizures are categorized into four distinct types: CF, GN, AB, and CT. Each clip is consequently assigned a label $y \in 0, 1, 2, 3$, representing combined focal (CF) seizures, absence (AB) seizures, generalized non-specific

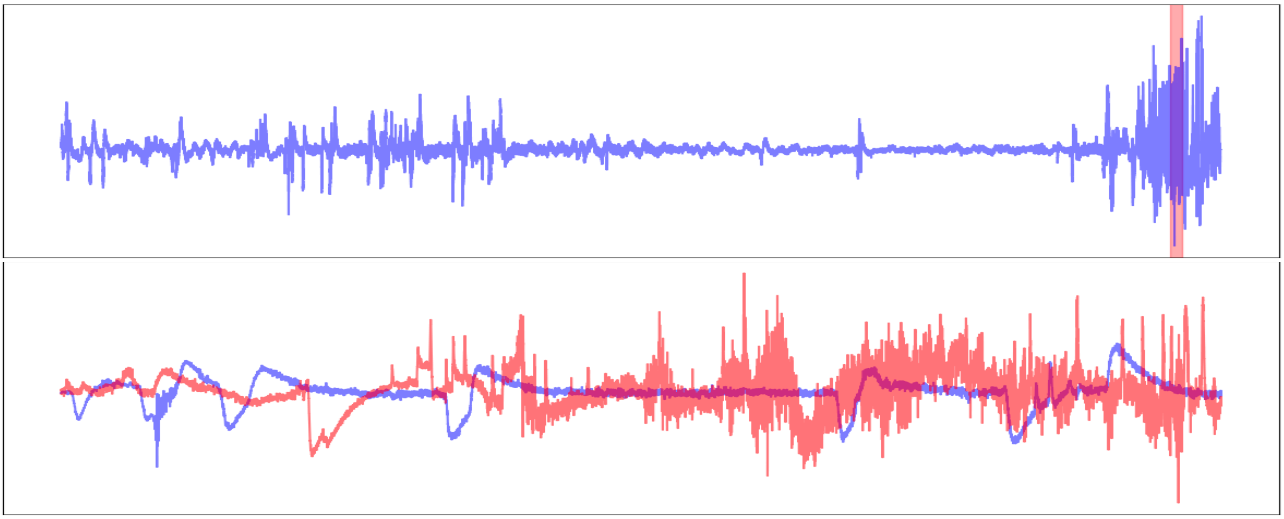

Figure 4. **Top**: One of the EEG channels of a patient experiencing a seizure, where the clip annotated in red is labeled as the seizure. **Bottom**: Comparison between the seizure clip and the non-seizure clip. The clip in blue represents the non-seizure clip, while the clip in red represents the seizure clip.

Table 7. Additional dataset descriptions. The datasets' size is organized in (Train, Test).

DATASET	DIM	DATASET SIZE	LENGTH	CLASSES
EPILEPSY	3	(137, 138)	206	4
ECG200	1	(100, 100)	96	2
ECG5000	1	(500, 4500)	140	2

(GN) seizures, and combined tonic (CT) seizures. A detailed analysis of the preprocessed dataset for the seizure classification task is outlined in Table 9.

Temple University EEG Abnormal Corpus. Regarding the data collected in the Temple University EEG Abnormal Corpus, we follow the same preprocessing procedures as outlined for the TUSZ dataset. Specifically, we utilize 60-second clips for self-supervised training, and any clips with a duration of less than 60 seconds are excluded to mitigate the introduction of extraneous noise. We conduct experiments exclusively on both 12-second and 60-second clips. Detailed information can be found in Table 10.

E. Details of Baseline Models

The brief introduction of the baseline models is summarized as follows. Considering that Informer, Autoformer, and iTransformer are only used in the Additional Experiments in Appendix H, due to scalability issues, they are not able to effectively complete the downstream tasks on TUH EEG Corpus. Additionally, DCRNN is only set as a baseline on TUH EEG Corpus. Since the additional datasets are mostly univariate, it's not beneficial to utilize the GNN-based model on them; thus, we abandon the DCRNN in additional experiments.

- DCRNN[5]. A GNN-based model incorporates a self-supervised pretraining strategy, which is used to capture non-Euclidean features across EEG signal channels.
- TimesNet[6]. A CNN-based model, by transforming the 1-D time series into folds, aims to capture the periodic information inherent in the original time series.

[5]https://github.com/tsy935/eeg-gnn-ssl
[6]https://github.com/thuml/TimesNet

Table 8. Detailed information of the TUSZ dataset on seizure detection task.

DATASET	SEIZURE CLIPS	NON-SEIZURE CLIPS
TRAIN SET(12S)	16,383(6.25%)	245,537(93.75%)
TRAIN SET(60S)	4,282(8.31%)	47,277(91.69%)
DEV SET(12S)	4,852(6.61%)	68,535(93.39%)
DEV SET(60S)	1,397(9.67%)	13,048(90.33%)
EVAL SET(12S)	2,709(6.98%)	36,088(93.02%)
EVAL SET(60S)	805(10.56%)	6,815(89.44%)

Table 9. Detailed information of the TUSZ dataset on seizure classification task.

DATASET	COMBINED FOCAL(CF)	GENERALIZED NON-SPECIFIC(GN)	ABSENCE(AB)	COMBINED TONIC(CT)
TRAIN SET(12S)	11,279(71.06%)	4,289(27.02%)	72(0.45%)	232(1.46%)
TRAIN SET(60S)	3,138(72.91%)	1,052(24.44%)	50(1.16%)	64(1.49%)
DEV SET(12S)	1,761(37.72%)	2,858(61.23%)	18(0.39%)	31(0.66%)
DEV SET(60S)	504(37.95%)	802(60.39%)	15(1.13%)	7(0.53%)
EVAL SET(12S)	1,624(63.17%)	838(32.59%)	49(1.91%)	60(2.33%)
EVAL SET(60S)	451(59.34%)	243(31.97%)	49(6.45%)	17(2.24%)

- PatchTST[7]. A Transformer-based model, wherein the local semantic information serves as tokens to facilitate attention across an extended historical context.
- MAE. A pretraining paradigm employing the method of Masked Autoencoders. The implementation of the MAE is adapted from the codebase[8]. The model employs a 1-D Convolution layer to extract local temporal features and utilizes non-overlapping temporal data as tokens on timestamps.
- SimMTM[9]. A Self-supervised framework trained by assembling the neighbor outside the manifold to generate promising representation. As it needs to generate the positive pairs for further aggregation, it will result in several times extra GPU memory usage, which will degrade the model's performance when facing large-scale dataset.
- BIOT[10]. A model comprising two stages, pre-training and fine-tuning, employs a biosignal tokenization module to generate a meticulously crafted representation of EEG signals.
- Informer[11]. A Transformer-based model that utilizes the ProbSparse Attention Module to reduce time complexity and memory usage.
- Autoformer[12]. A Transformer-based model utilizing frequency-enhanced decomposed Transformer together with seasonal-trend decomposition to better grasp global properties of time series.
- iTransformer[13]. A Transformer-based model regards independent time series as tokens to capture multivariate correlations by Transformer blocks.

F. Effectiveness of the Pretraining

To assess the effectiveness of pretraining on the TUSZ dataset, we compare the performance of VQ-MTM with and without self-supervised pretraining for both seizure detection and seizure classification. As depicted in Table 11, VQ-MTM with pretraining exhibits significantly better performance than its counterpart without pretraining. This suggests that

[7] https://github.com/PatchTST/PatchTST
[8] https://github.com/facebookresearch/mae
[9] https://github.com/thuml/SimMTM
[10] https://github.com/ycq091044/BIOT
[11] https://github.com/zhouhaoyi/Informer2020
[12] https://github.com/MAZiqing/FEDformer
[13] https://github.com/thuml/iTransformer

Table 10. Detailed information of the TUAB dataset on seizure detection task.

DATASET	SEIZURE CLIPS	NON-SEIZURE CLIPS
TRAIN SET(12S)	139,914(50.51%)	137,072(49.49%)
TRAIN SET(60S)	27,957(50.29%)	27,637(49.71%)
DEV SET(12S)	19,555(55.83%)	15,470(44.17%)
DEV SET(60S)	3,278(52.95%)	2,913(47.05%)
EVAL SET(12S)	15,475(45.12%)	18,821(54.88%)
EVAL SET(60S)	2,799(46.10%)	3,272(53.90%)

Table 11. Effectiveness of the pretraining. The proposed VQ-MTM's results are highlighted in **bold**.

MODEL	SEIZURE DETECTION AUROC		SEIZURE CLASSIFICATION WEIGHTED F1-SCORE	
	12-S	60-S	12-S	60-S
VQ-MTM W/O PRETRAINING	0.866	0.834	0.607	0.554
VQ-MTM W PRETRAINING	**0.887**	**0.904**	**0.620**	**0.615**
IMPROVEMENT(%)	2.42	8.39	2.14	11.01

the pre-trained model can provide a more promising representation than the randomly initialized one, aligning with our expectations.

G. Linear Probes on TUSZ Corpus

We conduct an evaluation of linear probes on self-supervised models for downstream tasks on the TUSZ Corpus, including the seizure detection and the seizure classification. We compare VQ-MTM's effectiveness with the models MAE, DCRNN, and SimMTM, as previously discussed. For the implementation of VQ-MTM, a trainable convolution layer is employed to aggregate features from different channels. All training is performed using the AdamW Optimizer with a learning rate of 1×10^{-3}, weight decay of 1×10^{-4}, global batch size of 512, and the execution of linear probes for 60 epochs using the cosine annealing strategy.

As illustrated in Table 12, we conduct a linear probe evaluation on the TUSZ dataset to assess the representations learned by the self-supervised models. Overall, VQ-MTM outperforms other self-supervised models in linear probing, indicating that VQ-MTM excels at extracting well-defined features for subsequent downstream tasks.

H. Additional Experimental Results

Considering the large scale of the previous datasets, namely TUSZ and TUAB, which may pose challenges for training the majority of proposed models in time series analysis (Wu et al., 2021; Zhou et al., 2022; 2021; Liu et al., 2024; Vaswani et al., 2017). In order to assess the performance of the VQ-MTM relative to these methods, we conduct supplementary experiments on subsets of the UEA datasets (Bagnall et al., 2018) and UCR datasets (Dau et al., 2019) (Epilepsy, ECG200,

Table 12. Linear Probes on the TUSZ dataset. The best results are highlighted in **bold**.

MODEL	DETECTION AUROC		CLASSIFICATION WEIGHTED F1-SCORE	
	12-S	60-S	12-S	60-S
VQ-MTM	**0.832**	**0.767**	**0.593**	0.459
MAE	0.642	0.640	0.513	**0.479**
DCRNN	0.830	0.752	0.530	0.459
SIMMTM	0.627	0.617	0.489	0.441
IMPROVEMENT(%)	0.24	1.99	11.89	/

Table 13. Additional experimental results for comparison with the previous models. The best results are highlighted in **bold**, and the second best results are indicated in underline.

		Informer	Autoformer	FEDformer	iTransformer	PatchTST	TimesNet	MAE	SimMTM	VQ-MTM
Epilepsy	Acc	0.862	0.746	0.833	0.710	0.768	0.826	0.819	0.936	**0.964**
	F1	0.863	0.742	0.831	0.692	0.770	0.819	0.790	0.889	**0.963**
ECG200	Acc	0.850	0.840	**0.930**	0.840	0.850	0.850	0.830	0.830	0.860
	F1	0.847	0.836	**0.930**	0.840	0.848	0.850	0.830	0.817	0.857
ECG5000	Acc	0.938	0.933	0.939	0.941	**0.942**	0.939	0.927	0.936	**0.942**
	F1	0.928	0.924	0.931	0.932	0.935	0.933	0.897	0.479	**0.938**

and ECG5000 datasets, with detailed information listed in Table 7). However, due to the univariate nature of some datasets, the DCRNN baseline is omitted from the additional experiments.

The datasets we choose focusing on EEG/ECG signal classification. The outcomes of these experiments reveal that VQ-MTM demonstrates promising performance across both large and small datasets. Moreover, these experiments are conducted within the framework of the Time-Series-Library (Wu et al., 2023), while SimMTM is implemented using its original codebase (Dong et al., 2023). Futhermore, as the majority of the selected datasets involve multiclass classification, we exclusively report the pertinent Weighted F1 Score and the corresponding Accuracy metrics. The reported results can be found in Table 13.

I. Analysis on Model's Computational Costs

To more effectively demonstrate the efficiency of our model during the pre-training and supervised learning stages, we conducted a comparison of computational costs with self-supervised baseline models, as delineated in Table 14 and Table 15. Despite possessing fewer parameters, our model surpasses most models that have a larger number of parameters in terms of performance.

Table 14. The computational cost of pre-training.

MODEL	SIMMTM	BIOT	MAE	VQ-MTM
FLOPS	1.78G	7.84G	0.15G	0.12G
PARAMS	0.53M	1.74M	2.50M	0.20M

Table 15. The computational cost of supervised learning.

MODEL	SIMMTM	BIOT	MAE	TIMESNET	VQ-MTM
FLOPS	0.313G	0.883G	0.014G	1.980G	0.014G
PARAM	0.457M	1.604M	1.217M	0.269M	0.065M

J. Analysis of the Distribution of the Codebooks

To more effectively depict the word frequency of raw EEG signals, the logarithmic frequency of the word frequency is presented in Appendix J. It is readily apparent that the majority of the words in the codebooks are utilized.

K. Hyperparameters Settings

For a more comprehensive understanding of the hyperparameter settings employed during the pretraining phase, the intricate configurations are delineated in Table 16 and Table 17.

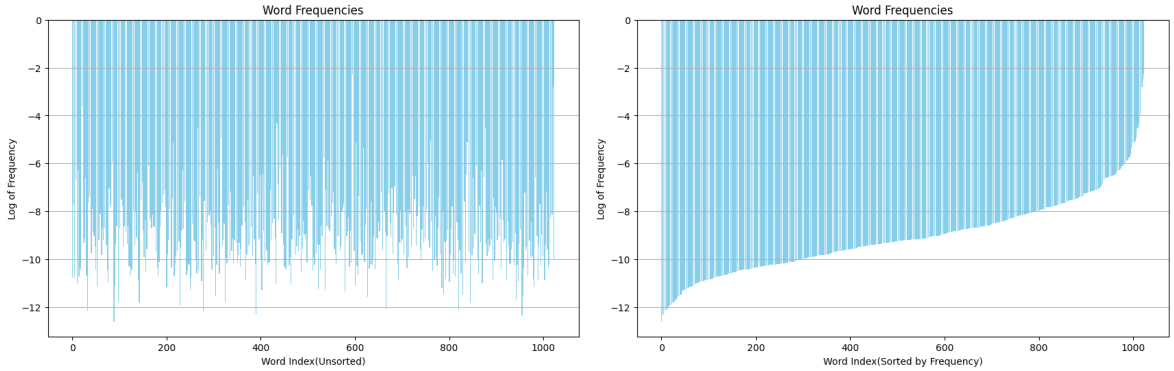

Figure 5. **Left**: The distribution of words in EEG signals (unsorted). **Right**: The distribution of words in EEG signals (sorted).

Table 16. Hyperparameters for pretraining on TUSZ dataset.

HYPERPARAMETERS	TUSZ
LAYERS	2
HIDDEN SIZE	256
ATTENTION HEADS	8
PATCH SIZE	250
ACTIVATION	GELU
TRAINING EPOCH	150
BATCH SIZE	512
WEIGHT DECAY	1×10^{-4}
PEAK LEARNING RATE	2×10^{-3}
INITIAL LEARNING RATE	1×10^{-3}
LEARNING RATE SCHEDULE	COSINE
WARMUP EPOCHES	20
MASK RATIO	0.3
DROPOUT	0.3
CODEBOOK SIZE	1024
CODEBOOK DIM	256

Table 17. Hyperparameters for pretraining on TUAB dataset.

HYPERPARAMETERS	TUAB
LAYERS	2
HIDDEN SIZE	256
ATTENTION HEADS	8
PATCH SIZE	250
ACTIVATION	GELU
TRAINING EPOCH	100
BATCH SIZE	512
WEIGHT DECAY	1×10^{-4}
PEAK LEARNING RATE	2×10^{-3}
INITIAL LEARNING RATE	1×10^{-3}
LEARNING RATE SCHEDULE	COSINE
WARMUP EPOCHES	20
MASK RATIO	0.3
DROPOUT	0.3
CODEBOOK SIZE	1024
CODEBOOK DIM	256