NEUROCODEX: BEYOND HOMOGENIZATION IN EEG JOINT PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning effective EEG representations remains challenging when downstream tasks have limited labeled data. Joint pretraining on diverse large-scale EEG datasets offers a promising solution, but substantial differences in electrode montages, channel configurations, and sampling rates make it difficult to directly combine datasets. Existing approaches typically address these discrepancies by enforcing a unified input structure through signal resampling, channel padding, masking, or fixed channel selection, thereby suppressing dataset-specific structural characteristics and limiting the ability of joint pretraining to exploit complementary information across datasets. In this work, we propose NeuroCodeX, a structure-aware framework that moves beyond homogenization by explicitly modeling structural differences during EEG joint pretraining. NeuroCodeX consists of two components: a Hierarchical Topographic Encoder that captures anatomy-aware spatial dependencies through hierarchical region-level aggregation, and a Multi-Scale Quantizer that provides rhythm-aware discrete supervision across different sampling rates. We evaluate NeuroCodeX on four EEG datasets spanning diverse acquisition settings, channel configurations, and downstream tasks. Experimental results demonstrate consistent improvements over state-of-the-art baselines, with particularly strong benefits under data-scarce conditions. These findings suggest that effective EEG joint pretraining depends not simply on combining more data, but on appropriately preserving and modeling the structural characteristics of individual datasets. Code is available at: https://anonymous.4open.science/r/NeuroCodeX.

1 INTRODUCTION

Electroencephalography (EEG) is a fundamental modality for brain–computer interfaces, neurological diagnosis, and cognitive state monitoring (McFarland & Wolpaw, 2017; Dirican & Göktürk, 2011; Jiang et al., 2024). Despite the increasing availability of large-scale EEG datasets, many downstream applications remain data-scarce, where collecting sufficient task-specific annotations is expensive or impractical (Kindermans et al., 2018). Joint pretraining on existing EEG datasets therefore offers a promising opportunity to leverage the growing availability of large-scale EEG resources for improving downstream learning under limited supervision. However, these datasets are collected under diverse acquisition protocols (Yi et al., 2023; Yang et al., 2023) and exhibit substantial differences in electrode montages, channel configurations, and sampling rates, making it non-trivial to effectively combine them during joint pretraining.

Existing approaches typically facilitate joint pretraining by reducing structural discrepancies through preprocessing, such as resampling signals to a unified sampling rate or standardizing electrode configurations through channel padding, masking, or selection. Other approaches process channels independently without explicitly modeling their spatial topology (Yang et al., 2023; Jiang et al., 2024; Wang et al., 2024). While such preprocessing simplifies the integration of datasets collected under different acquisition settings, it can also suppress dataset-specific structural characteristics and limit the ability of joint pretraining to exploit complementary spatial and spectral information across datasets. More recently, several studies have begun to explicitly model specific aspects of structural variation in EEG (Yi et al., 2023; Xiang et al., 2025; Sukhbaatar et al., 2026), highlighting the importance of structural awareness in EEG representation learning.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

These developments suggest that structural differences should not simply be treated as discrepancies to be eliminated before joint pretraining. Variations in electrode montages and sampling rates reflect differences in spatial coverage, temporal resolution, and neurophysiological characteristics (Knyazev, 2012; Babiloni et al., 2021; Vempati & Sharma, 2023; Liu et al., 2025), potentially providing complementary information across datasets. This raises a fundamental question: *how can structurally diverse EEG datasets be jointly leveraged without discarding the information encoded in their native structures?*

To address this question, we propose **NeuroCodeX**, a structure-aware self-supervised framework for EEG joint pretraining. Rather than eliminating structural differences through aggressive homogenization, NeuroCodeX explicitly preserves and models the native spatial and spectral characteristics of diverse EEG datasets during representation learning. Specifically, the proposed *Hierarchical Topographic Encoder* (HTE) captures anatomy-aware spatial dependencies through hierarchical region-level graph aggregation, enabling effective modeling of EEG datasets with diverse electrode configurations. Complementarily, the proposed *Multi-Scale Quantizer* (MSQ) provides rhythm-aware discrete supervision across canonical EEG frequency bands, explicitly accounting for differences in temporal resolution induced by varying sampling rates. Our contributions are threefold:

- We identify structural heterogeneity as a key challenge in EEG joint pretraining and formulate the perspective that effective joint pretraining should explicitly model, rather than aggressively homogenize, structural differences across datasets.
- We propose NeuroCodeX, a structure-aware self-supervised framework that explicitly models spatial and spectral heterogeneity through the Hierarchical Topographic Encoder and Multi-Scale Quantizer, respectively.
- Extensive experiments on four EEG datasets with diverse acquisition settings demonstrate the effectiveness of NeuroCodeX across heterogeneous joint pretraining settings, with particularly strong benefits on data-scarce downstream tasks.

2 METHOD

Problem Setup. We consider a collection of M heterogeneous EEG datasets $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_M\}$, each acquired under distinct recording protocols. Dataset $\mathcal{D}_m$ contains N_m multichannel EEG recordings $\{\mathbf{R}_m^{(i)} \in \mathbb{R}^{C_m \times T_m^{(i)}}\}_{i=1}^{N_m}$, where C_m denotes the number of electrodes and $T_m^{(i)} = f_m \cdot \text{duration}_m^{(i)}$ depends on the dataset-specific sampling rate f_m. For downstream classification tasks, each recording is segmented into non-overlapping windows of L_m seconds (e.g., $L_m = 10$ for seizure detection or $L_m = 5$ for emotion recognition), yielding windowed EEG samples $\mathbf{S}_m \in \mathbb{R}^{C_m \times (L_m f_m)}$. As a result, both the spatial dimensionality C_m and the temporal resolution $L_m f_m$ vary substantially across datasets. We consider a data-imbalanced setting in which some datasets provide large-scale recordings, while others contain only limited data. Our goal is to learn an encoder $f_{encoder} : \mathbf{S}_m \mapsto \mathbf{u}_m$ from heterogeneous datasets under this imbalance, and leverage large-scale datasets to improve downstream classification performance on data-scarce datasets.

2.1 METHOD OVERVIEW

To enable effective joint pretraining across structurally diverse EEG datasets, NeuroCodeX explicitly addresses spatial and spectral mismatches arising from heterogeneous electrode configurations and sampling rates. These mismatches make it difficult to construct representations with comparable anatomical semantics across different montages and to jointly model signals observed at different temporal and spectral resolutions. NeuroCodeX addresses these challenges within a random vector quantization-based self-supervised learning framework through two complementary components. The *Hierarchical Topographic Encoder* (HTE) introduces spatial inductive biases by aggregating channel-wise signals into anatomy-aware regional representations through hierarchical graph aggregation, providing a shared anatomical abstraction for diverse electrode configurations. Building on these regional representations, the *Multi-Scale Quantizer* (MSQ) incorporates spectral inductive biases into the learning objective by performing random vector quantization over canonical EEG frequency bands, providing rhythm-aware discrete supervision for signals with different sampling rates.

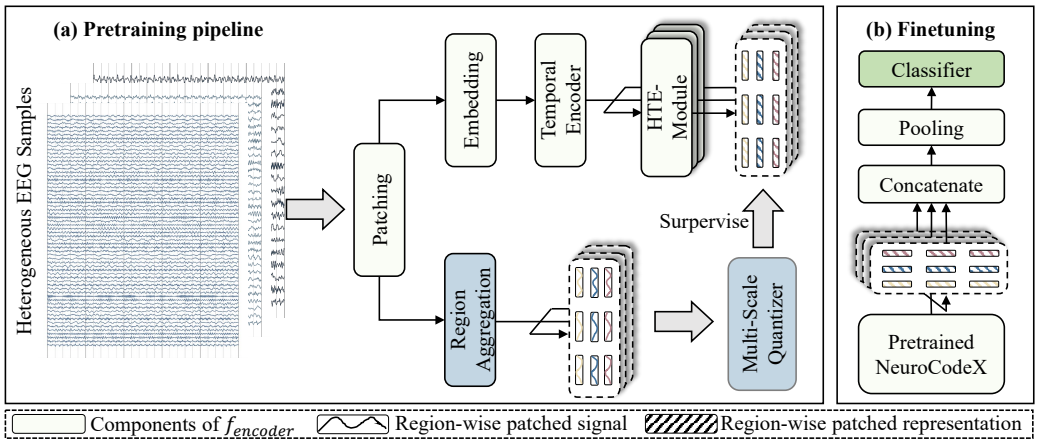

Figure 1: Architecture of NeuroCodeX.

Pretraining Pipeline. As illustrated in Figure 1, NeuroCodeX processes each windowed EEG sample $\mathbf{S} \in \mathbb{R}^{C\times Lf}$ (dataset index omitted) by first partitioning it into $P = \frac{Lf}{N}$ non-overlapping temporal patches of length of N (padding or truncation ensures $Lf \bmod N = 0$). This yields a set of channel-wise temporal patches

$$\mathbf{X} = [\mathbf{x}_{c,p} \mid c = 1, \cdots, C\ , p = 1, \ldots, P] \in \mathbb{R}^{C\times P\times N},$$

where $\mathbf{x}_{c,p} \in \mathbb{R}^N$ denotes the p-th patch from channel c. Then, NeuroCodeX employs two parallel processing flows operating on $\mathbf{X}$, a *representation flow* and a a *pseudo-label flow*. The former produces topology-aware region-wise embeddings while the latter generates frequency-consistent multi-scale discrete supervision signals. Both flows share common cortical region hierarchies, ensuring spatially consistent supervision.

Cortical Region Hierarchies. We abstract the scalp topology using a predefined cortical region hierarchy $\mathcal{R} = \{\mathcal{C}_1, \ldots, \mathcal{C}_R\}$, where each region $\mathcal{C}_r \subseteq \{1, \ldots, C\}$ indexes electrodes belonging to the same anatomical area (e.g., FP1 and FPZ assigned to the frontal region). Following Ye et al. (2022), we adopt three granularities $g \in \mathrm{C}, \mathrm{M}, \mathrm{F}$ (coarse, medium, fine) with $R^{\mathrm{C}} = 6$, $R^{\mathrm{M}} = 10$, and $R^{\mathrm{F}} = 17$. For simplicity, the description below focuses on a single granularity. The detailed definitions of region partitions at different granularities are provided in Section J

Representation Flow. Temporal dynamics play a central role in EEG analysis (Shen et al., 2025). We first encode each patch $\mathbf{x}_{c,p}$ into a D-dimensional vector embedding, followed by a vanilla Transformer (Vaswani et al., 2017) to model per-channel temporal dependencies across patches:

$$\begin{aligned} \mathbf{e}_{c,p} &= \mathrm{LayerNorm}(\mathrm{GeLU}(\mathbf{W}\mathbf{x}_{c,p})) \in \mathbb{R}^D \\ [\mathbf{e}^{\mathrm{tem}}_{c,p}]_{p=1}^{P} &= \mathrm{Transformer}\big([\mathbf{e}_{c,p}]_{p=1}^{P}\big) \in \mathbb{R}^{P\times D}, \end{aligned} \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{D\times N}$ is shared weights across all patches. For a given patch p, we collect the temporal embeddings across channels as $\mathbf{E}_p = [\mathbf{e}^{\mathrm{tem}}_{c,p}]_{c=1}^{C} \in \mathbb{R}^{C\times D}$. Since channel numbers and layouts vary across datasets, directly modeling $\mathbf{E}_p$ is not spatially consistent. Therefore, it is transformed by the Hierarchical Topographic Encoder (HTE), which captures the cortical topology of a specific cortex partition granularity g, yielding region-wise representations as follows:

$$\mathbf{H}_p = [\mathbf{h}_{r,p}]_{r=1}^{R} = \mathrm{HTE}(\mathbf{E}_p) \in \mathbb{R}^{R\times D}. \tag{2}$$

Each region-wise patch embedding $\mathbf{h}_{r,p}$ is quantized by the multi-scale pseudo-labels during pre-training.

Pseudo-label Flow. To ensure spatially aligned supervision, pseudo-labels are generated at the same region granularity as the representation flow. For the p-th patch, the raw signals belonging to region $\mathcal{C}_r$ are aggregated via channel-wise average pooling:

$$\mathbf{z}_{r,p} = \frac{1}{\mathcal{C}_r} \sum_{c\in\mathcal{C}_r} \mathbf{x}_{c,p} \in \mathbb{R}^N, \tag{3}$$

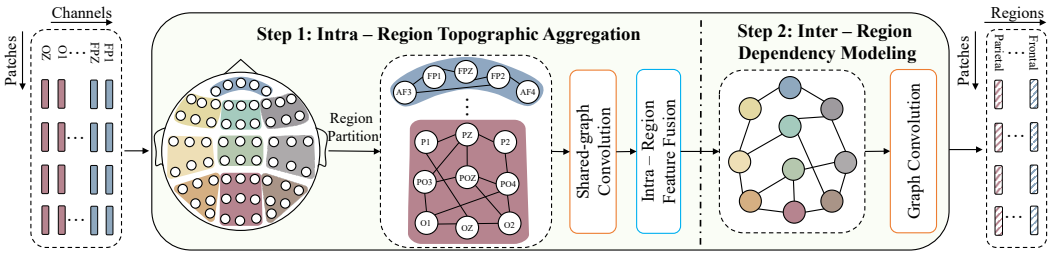

Figure 2: Illustration of HTE module in the coarse cortex region granularity.

which is then transformed into multi-scale discrete pseudo-labels in two steps:

$$\mathbf{y}_{r,p} = \mathrm{MSQ}(\mathrm{RaD}(\mathbf{z}_{r,p})) \in \mathbb{R}^{|\mathcal{B}|}, \tag{4}$$

where $\mathrm{RaD}(\cdot)$ denotes the Rhythm-aware Decomposition and $\mathrm{MSQ}(\cdot)$ denotes the Multi-Scale Quantizer that generates one discrete index per canonical EEG band[1] $\mathcal{B}$. Precise formulas for RaD and MSQ are provided in Section 2.3.

2.2 Hierarchical Topographic Encoder

Different EEG acquisition systems (e.g., 10–10 vs. 10–20 montages (Jurcak et al., 2007)) exhibit substantial variations in channel coverage and spatial density, resulting in diverse electrode configurations across datasets. Under such structural variation, channel-level representations are not sufficient as a shared semantic unit for joint pretraining, since the same cortical regions may be represented by different sets of electrodes across datasets. We therefore define the semantic unit of spatial representation at the level of anatomically grounded cortical regions rather than individual electrodes. We introduce the *Hierarchical Topographic Encoder* (HTE), which maps channel-wise EEG signals into anatomy-aware regional representations while preserving the native electrode configurations. As shown in Figure 2, each HTE module operates at a spatial granularity level g and consists of two coordinated steps: (i) *intra-region topographic aggregation*, which captures functional connectivity among electrodes within the same anatomical region; and (ii) *inter-region dependency modeling*, which models long-range spatial interactions across cortical regions.

(1) Intra-Region Topographic Aggregation. For a temporal patch p, we group channel-wise temporal embeddings according to the predefined cortical partition:

$$\mathbf{E}_{\mathcal{C}_{r,p}} = [\mathbf{e}^{\mathrm{tem}}_{c,p}]_{c\in\mathcal{C}_r} \in \mathbb{R}^{|\mathcal{C}_r|\times D}, \tag{5}$$

where $\mathcal{C}_r$ denotes the set of channels assigned to region r. To explicitly model functional dependencies among channels within each region, we construct a region-specific adjacency matrix $\mathbf{A}_{\mathcal{C}r,p}$ using $\mathrm{Adj}(\cdot)$[2] by computing pairwise connectivity between channel embeddings in $\mathbf{E}_{\mathcal{C}r,p}$.The resulting adjacency matrix defines the functional graph structure, which is used by a shared graph convolutional network $\mathrm{IntraGCN}$ to propagate information among channels, enabling each channel representation to incorporate contextual information from other channels within the same anatomical region:

$$\begin{aligned}\mathbf{A}_{\mathcal{C}_r,p} &= \mathrm{Adj}(\mathbf{E}_{\mathcal{C}_r,p}) \in \mathbb{R}^{|\mathcal{C}_r|\times|\mathcal{C}_r|}\\ \mathbf{G}_{\mathcal{C}_r,p} &= \mathrm{IntraGCN}(\mathbf{A}_{\mathcal{C}_r,p};\mathbf{E}_{\mathcal{C}_r,p}) \in \mathbb{R}^{|\mathcal{C}_r|\times D}.\end{aligned} \tag{6}$$

Finally, channel-wise responses are summarized into a region-level representation via average pooling:

$$\mathbf{g}_{r,p} = \mathrm{AvgPool}(\mathbf{G}_{\mathcal{C}_r,p}) \in \mathbb{R}^{D}. \tag{7}$$

(2) Inter-Region Dependency Modeling. To capture long-range spatial dependencies across cortical regions and encode global topographic structure, we collect the region-level embeddings $\mathbf{G}_p =$

[1] Concretely, we adopt five canonical EEG bands (Yuan et al., 2023) as $\mathcal{B} = \{\delta, \theta, \alpha, \beta, \gamma\}$ with typical frequency ranges (Hz): δ(0.5–4), θ(4–8), α(8–13), β(13–30), γ(>30).

[2] The adjacency matrix is constructed based on magnitude-squared coherence (MSC) (Miskovic & Keil, 2015; Cui et al., 2018), which measures frequency-domain functional connectivity between channel pairs. Details are provided in Section E.

$[\mathbf{g}_{r,p}]_{r=1}^{R} \in \mathbb{R}^{R\times D}$ for patch p. Analogously to the intra-region stage, we construct an inter-region adjacency matrix $\mathbf{A}_p$[2] among region-wise embeddings and apply a graph convolutional network over the region graph to yield their connectivity:

$$\begin{aligned}\mathbf{A}_p &= \mathrm{Adj}(\mathbf{G}_p) \in \mathbb{R}^{R\times R}\\ \mathbf{H}_p &= \mathrm{InterGCN}(\mathbf{A}_p; \mathbf{G}_p) \in \mathbb{R}^{R\times D}.\end{aligned} \tag{8}$$

The final output $\mathbf{H}_p$ constitutes anatomy-grounded, topology-aware region representations that are invariant to montage-specific channel configurations.

2.3 Multi-Scale Quantizer

While HTE addresses spatial heterogeneity by providing anatomy-aware regional representations, NeuroCodeX further considers the spectral heterogeneity arising from different sampling rates. Varying sampling rates result in different temporal and spectral resolutions across EEG datasets, which complicates the use of a shared discretization scheme for joint pretraining. As illustrated in Fig. 3, temporal segments with the same number of sampling points correspond to different physical durations under different sampling rates and may therefore capture different amounts of rhythmic information. Applying a single global quantization scheme (Haokun et al., 2024) to such representations may consequently associate a shared discrete code with rhythmically different signal content, leading to inconsistent quantization semantics across datasets.

We therefore introduce the *Multi-Scale Quantizer* (MSQ), which incorporates canonical EEG rhythms into the discretization process by factorizing the quantization space across frequency bands to capture physiologically meaningful rhythmic patterns (Sterman, 1996; Salansky et al., 1998). (Sterman, 1996; Salansky et al., 1998). Specifically, MSQ consists of two components: (i) *Rhythm-aware Decomposition*, which extracts band-limited signal components while aligning their frequency supports across sampling rates; and (ii) *Multi-Scale Quantization*, which assigns independent discrete codes to region representations at each rhythmic scale, promoting consistent quantization semantics across datasets with different temporal resolutions.

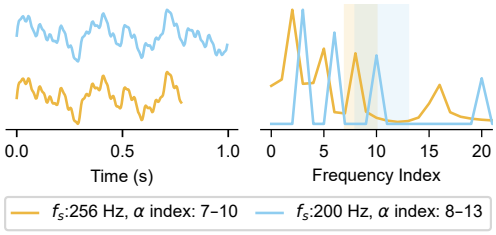

Figure 3: Example of sampling rate induced time-frequency mismatch. **Left**: Temporal segments with identical length (200 samples) span different physical durations when recorded at different sampling rates (200 Hz vs. 256 Hz). **Right**: The same canonical EEG α rhythm corresponds to different discrete Fourier index ranges due to sampling-rate-dependent spectral resolution.

Rhythm-aware Decomposition. Given a region-wise temporal patch $\mathbf{z} \in \mathbb{R}^N$ (region and patch indices omitted for clarity), we first compute its real-valued Fourier transform:

$$\hat{\mathbf{z}} = \mathrm{rFFT}(\mathbf{z}) \in \mathbb{C}^K, \quad K = \left\lfloor \frac{N}{2} \right\rfloor + 1. \tag{9}$$

To obtain rhythm-aligned spectral representations under heterogeneous sampling rates, we explicitly define canonical EEG bands in the physical frequency domain. For each rhythm $b \in \mathcal{B}$ with frequency range $[f^b_{\min}, f^b_{\max}]$ (in Hz), we map it to the corresponding Fourier-bin interval determined by the dataset-specific sampling rate f_m:

$$[k^b_{\min}, k^b_{\max}] = \left[\left\lfloor N \cdot \frac{f^b_{\min}}{f_m} \right\rfloor, \left\lceil N \cdot \frac{f^b_{\max}}{f_m} \right\rceil \right]. \tag{10}$$

The rhythm-specific spectrum is then obtained by masking coefficients outside the corresponding interval:

$$\hat{z}^b_k = \begin{cases} \hat{z}_k, & k^b_{\min} \le k < k^b_{\max}, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

This operation yields a set of band-limited spectral representations, each corresponding to a canonical EEG rhythm with aligned physical frequency support across datasets.

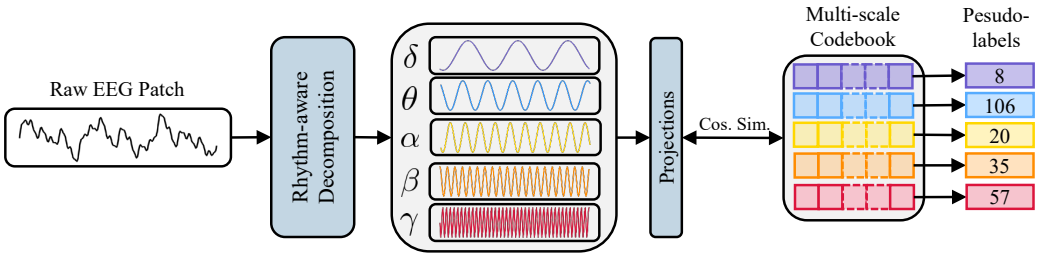

Figure 4: Illustration of MSQ. Separated rhythm-specific projections are visually presented as a single block for clarity.

To encourage discretization invariance within each rhythm, we further construct rhythm-preserving augmented views via band-wise phase rotation, following Haokun et al. (2024). Let $|\hat{z}_k|$ and ϕ_k denote the amplitude and phase of coefficient k. For each rhythm b, phases are shifted relative to the first non-DC component:

$$\begin{aligned} \phi_k^{b_{\text{aug}}} &= \phi_k^b - k \cdot \phi_1, \\ \hat{z}_k^{b_{\text{aug}}} &= |\hat{z}_k^b| \, e^{i\phi_k^{b_{\text{aug}}}}, \\ \mathbf{z}^{b_{\text{aug}}} &= \text{irFFT}(\hat{\mathbf{z}}^{b_{\text{aug}}}). \end{aligned} \tag{12}$$

The resulting $\mathbf{z}^{b_{\text{aug}}}$ preserves rhythm-specific spectral structure while providing phase-shifted views of the same underlying dynamics, encouraging the quantizer to assign consistent discrete codes to semantically equivalent rhythm-level representations.

Multi-Scale Quantization. For each frequency band b, MSQ assigns a discrete pseudo-label that reflects its band-specific spectral signature. To accomplish this, we equip each band with an independent random-projection quantizer. Let $\mathcal{V}$ denote the discrete code vocabulary, we denote the neural codebook by $\mathbf{C}^b = [\mathbf{c}_i^b \in \mathbb{R}^D]_{i=1}^{|\mathcal{V}|}$ and projection matrix $\mathbf{W}^b \in \mathbb{R}^{D \times N}$ for band b, respectively. We first map the corresponding band-limited signal $\mathbf{z}^{b_{\text{aug}}}$ into the quantization space and then assign a discrete pseudo-label by nearest-neighbor matching in terms of cosine similarity as follows:

$$\begin{aligned} \mathbf{v}^b &= \mathbf{W}^b \mathbf{z}^{b_{\text{aug}}}, \\ y^b &= \arg\max_{i \in [|\mathcal{V}|]} \left\langle \frac{\mathbf{c}_i^b}{\|\mathbf{c}_i^b\|_2}, \frac{\mathbf{v}^b}{\|\mathbf{v}^b\|_2} \right\rangle. \end{aligned} \tag{13}$$

Both neural embedding $\mathbf{c}_i^b$ and $\mathbf{W}^b$ are randomly initialized and non-learnable throughout pretraining. Collecting the band-wise discrete labels forms a multi-scale pseudo-label:

$$\mathbf{y} = [y^b]_{b \in \mathcal{B}}, \tag{14}$$

which is used to supervise the corresponding region-wise embedding during pretraining. By factorizing the discretization space according to canonical EEG rhythms (e.g., δ, θ, α, β, γ), MSQ avoids enforcing a single global codebook over mismatched spectral content, yielding rhythm-consistent discrete supervision across heterogeneous sampling rates.

2.4 Pretraining and Finetuning

Pretraining Objective. Pretraining is formulated as a multi-band classification task, where each region-wise patch representation $\mathbf{h}_{r,p}$ is trained to predict the discrete pseudo-label assigned by MSQ for each canonical EEG frequency band. Formally, the pretraining objective is the sum of cross-entropy losses across all regions, patches, and bands:

$$\mathcal{L}_{\text{pre}} = \sum_{r=1}^{R} \sum_{p=1}^{P} \sum_{b \in \mathcal{B}} \text{CE}\left(\mathbf{W}_{\text{cls}}^b \mathbf{h}_{r,p}, y_{r,p}^b\right), \tag{15}$$

where $\mathbf{W}^b_{\text{cls}}$ is the learnable band-specific classifier and $y^b_{r,p}$ denotes the pseudo-label. Since each band is quantized independently, the objective explicitly encourages the encoder to learn frequency-aware and sampling-invariant cortical representations.

Finetuning on Downstream Tasks. During downstream finetuning, the entire representation branch—including the temporal encoder and all HTEs—is updated end-to-end. Given the topology-aware outputs from all granularity levels $g \in \{\mathrm{C}, \mathrm{M}, \mathrm{F}\}$, We obtain the unified multi-granularity representation by concatenating each $\mathbf{H}^g$ along the region dimension:

$$\mathbf{H} = \text{Concat}(\mathbf{H}^g) \in \mathbb{R}^{\sum_g R^g \times P \times D}. \tag{16}$$

Temporal information is then summarized via average pooling over patches, followed by flattening across regions:

$$\mathbf{u} = \text{Flatten}\left(\text{AvgPool}(\mathbf{H})\right) \in \mathbb{R}^{\sum_g R^g D}. \tag{17}$$

The resulting feature $\mathbf{u}$ is passed to a task-specific classifier (e.g., seizure detection, sleep staging), depending on the downstream applications.

3 Experiments

3.1 Experimental Setup

Datasets. We evaluate NeuroCodeX on four representative EEG datasets spanning diverse recording protocols, sampling rates, electrode montages, task definitions, and data scales: TUAB (López et al., 2017) for normal/abnormal classification, TUEV (Harati et al., 2015) for EEG event classification, CHB-MIT (Shoeb, 2009) for seizure detection with bipolar montages, and SEED-IV (Zheng et al., 2018) for emotion recognition in a non-clinical setting. A detailed summary is provided in Section F.

Baselines. We compare NeuroCodeX with representative EEG representation learning methods, including (i) supervised models: SPaRCNet (Jing et al., 2023), CNN-Transformer (Peh et al., 2022), FFCL (Li et al., 2022), and ST-Transformer (Song et al., 2021); (ii) self-supervised EEG models: ContraWR (Yang et al., 2021), VQ-MTM (Haokun et al., 2024), and BIOT (Yang et al., 2023); and (iii) EEG foundation models: LaBraM (Jiang et al., 2024), EEGPT (Wang et al., 2024), and CBraMod (Wang et al., 2025). Detailed descriptions and implementation settings for all baselines are provided in Section G.

Evaluation Protocol and Metrics. We consider two pretraining configurations: (i) **Single-dataset pretraining**, where the model is pretrained using only the unlabeled data from the same dataset used for downstream evaluation; and (ii) **Joint pretraining**, where the model is pretrained using unlabeled data jointly collected from multiple EEG datasets, including the dataset used for downstream evaluation. For each pretrained model, we fine-tune it on the labeled training split of the corresponding dataset and evaluate it on the held-out test split. Following prior works (Haokun et al., 2024; Wang et al., 2025), we report **Balanced Accuracy**, **AUC-PR**, and **AUROC** for binary classification tasks, and **Balanced Accuracy**, **Weighted F1-score**, and **Cohen's Kappa** for multi-class classification tasks. All results are averaged over three independent runs with different random seeds.

3.2 Main Results

Our experiments are designed to answer three questions concerning heterogeneous EEG joint pretraining. First, *how does NeuroCodeX perform under single-dataset pretraining?* Second, *can explicitly modeling structural heterogeneity improve joint pretraining across heterogeneous EEG datasets?* Third, *does incorporating more diverse pretraining data consistently improve joint pretraining?* Table 1 reports the main results on TUEV, while the complete results of other datasets are provided in Tables 5 to 7. Several observations are noteworthy.

First, *NeuroCodeX effectively learns informative EEG representations under single-dataset pretraining.* Across the four benchmarks, NeuroCodeX achieves performance comparable to or better than representative self-supervised and foundation models. This demonstrates the effectiveness of the proposed structure-aware design for EEG representation learning and establishes a foundation for examining the additional benefits of heterogeneous joint pretraining.

Table 1: Evaluation results on TUEV. Best results are in bold, and the second best are underscored

Models	TUEV		
	Balanced Acc.	Cohen's Kappa	Weighted F1
Supervised Models			
SPaRCNet (2023)	0.3671 ± 0.0047	0.2169 ± 0.0057	0.5773 ± 0.0039
CNN-Transformer (2022)	0.2943 ± 0.0052	0.2853 ± 0.0053	0.6526 ± 0.0027
FFCL (2022)	0.3640 ± 0.0019	0.3007 ± 0.0004	0.6535 ± 0.0014
ST-Transformer (2021)	0.3429 ± 0.0036	0.2545 ± 0.0003	0.6028 ± 0.0022
Self-Supervised Models (pretrained from scratch)			
ContraWR (2021)	0.3020 ± 0.0039	0.2602 ± 0.0009	0.6390 ± 0.0026
VQ-MTM (2024)	0.3225 ± 0.0017	0.3344 ± 0.0004	0.6542 ± 0.0013
BIOT (2023)	0.3567 ± 0.0029	0.3593 ± 0.0021	0.6208 ± 0.0019
EEG Foundation Models			
$\text{EEGPT}_{\text{Large}}$ (2024)	0.3956 ± 0.0045	<u>0.3847 ± 0.0041</u>	0.6409 ± 0.0033
$\text{LaBraM}_{\text{Base}}$ (2024)	0.3858 ± 0.0035	0.3742 ± 0.0051	0.6508 ± 0.0020
CBraMod (2025)	0.4068 ± 0.0016	0.3233 ± 0.0066	0.6601 ± 0.0013
NeuroCodeX (Ours)			
Single-dataset pretraining	0.3900 ± 0.0016	0.3708 ± 0.0004	0.6396 ± 0.0005
Joint pretraining			
+ TUAB	0.4104 ± 0.0037	**0.3911 ± 0.0027**	**0.6717 ± 0.0028**
+ CHB-MIT	**0.4286 ± 0.0028**	0.3762 ± 0.0006	<u>0.6662 ± 0.0021</u>
+ All	<u>0.4266 ± 0.0026</u>	0.3550 ± 0.0007	0.6567 ± 0.0018

Table 2: Ablation results on TUEV under joint pretraining with TUAB.

Model Variant	TUEV		
	Balanced Acc.	Cohen's Kappa	Weighted F1
Full model (Ours)	0.4104 ± 0.0037	0.3911 ± 0.0027	0.6717 ± 0.0028
w/o HTE	0.3796 ± 0.0038	0.3526 ± 0.0026	0.6175 ± 0.0017
w/o MSQ	0.3691 ± 0.0089	0.3426 ± 0.0057	0.6077 ± 0.0094

Second, *heterogeneous joint pretraining can improve downstream performance*. Joint pretraining with CHB-MIT improves NeuroCodeX on both TUEV and SEED-IV, while joint pretraining with TUAB yields consistent gains on TUEV. These results are consistent with our hypothesis that structurally heterogeneous EEG datasets can provide complementary information for representation learning, while the varying magnitude of improvement further suggests that the benefits depend on the degree of complementarity among the pretraining datasets.

Third, *the benefit of joint pretraining does not monotonically increase with the amount of pretraining data*. Joint pretraining on all available datasets does not consistently outperform alternative dataset combinations. Moreover, existing EEG foundation models pretrained on large-scale EEG collections do not consistently achieve better downstream performance across the four benchmarks. These observations suggest that simply increasing the scale or diversity of pretraining data is not sufficient to ensure improved downstream performance, further supporting our structure-aware approach to leveraging heterogeneous EEG datasets rather than indiscriminately aggregating them.

3.3 Model Analysis

Component Ablation. We first examine the contributions of the two core components of NeuroCodeX under heterogeneous joint pretraining. As shown in Table 2, removing either HTE or MSQ consistently degrades performance on TUEV, with removing MSQ causing a larger drop in all three evaluation metrics. Results on SEED-IV (provided in appendix Table 8) show consistent performance degradation after removing either component, although the relative impact varies across

Table 3: Comparison of different spatial aggregation strategies in HTE on TUEV.

Spatial Aggregation	TUEV		
	Balanced Acc.	Cohen's Kappa	Weighted F1
Graph-based Aggregation	0.3900 ± 0.0011	0.3708 ± 0.0004	0.6396 ± 0.0012
Mean Aggregation	0.3819 ± 0.0009	0.3267 ± 0.0016	0.6383 ± 0.0010
Channel Padding	0.3889 ± 0.0017	0.3395 ± 0.0002	0.6324 ± 0.0005

Table 4: Band-level ablation study on SEED-IV

Model Variant	SEED-IV		
	Balanced Acc.	Cohen's Kappa	Weighted F1
Full model (Ours)	0.3127 ± 0.0007	0.0839 ± 0.0003	0.3140 ± 0.0012
w/o δ	0.2910 ± 0.0005	0.0864 ± 0.0002	0.2756 ± 0.0013
w/o θ	0.3036 ± 0.0004	0.0717 ± 0.0003	0.2952 ± 0.0009
w/o α	0.3079 ± 0.0008	0.0800 ± 0.0005	0.2794 ± 0.0004
w/o β	0.3031 ± 0.0009	0.0942 ± 0.0007	0.2712 ± 0.0010
w/o γ	0.2958 ± 0.0015	0.0837 ± 0.0009	0.2639 ± 0.0011

evaluation metrics. These results indicate that HTE and MSQ make complementary contributions to heterogeneous EEG representation learning by addressing distinct aspects of structural heterogeneity.

Analysis of Spatial Aggregation. To further investigate the role of spatial structure modeling in HTE, we compare its graph-based aggregation with topology-agnostic alternatives, including mean aggregation and channel padding. As shown in Table 3, graph-based aggregation achieves the best overall performance, with a notable improvement in Cohen's Kappa over both alternatives. These results suggest that the benefit of HTE extends beyond mapping variable-sized channel sets to region-level representations: explicitly modeling spatial relationships among electrodes provides additional benefits for heterogeneous EEG pretraining.

Effect of Frequency Bands. We further examine the contribution of individual frequency bands on SEED-IV. As shown in Table 4, removing any frequency band leads to a noticeable degradation in Weighted F1, with more pronounced drops observed for β and γ. These results suggest that different rhythmic components provide complementary information for EEG representation learning. Although Cohen's Kappa slightly improves after removing δ or β, the consistent degradation in Weighted F1 indicates that retaining multiple frequency bands is beneficial for learning robust EEG representations.

4 Conclusion

In this work, we proposed NeuroCodeX, a vector-quantization–based pretraining framework for learning EEG representations from structurally heterogeneous datasets without explicitly homogenizing their acquisition structures. Rather than removing dataset-specific variations, NeuroCodeX preserves native sampling rates and electrode configurations while modeling their underlying spatial and spectral structures. Extensive experiments show that these structure-aware designs enable heterogeneous joint pretraining to effectively exploit complementary information across datasets, improving the generalization of pretrained representations to data-scarce target datasets under substantial structural mismatch. More broadly, our results suggest that effective EEG foundation pretraining does not necessarily require making heterogeneous datasets identical; instead, explicitly modeling their intrinsic structural differences offers a principled alternative to homogenizing heterogeneous EEG data for representation learning.

AI USE STATEMENT

In this work, we used generative AI tools for editing the research paper to improve readability, including English proofreading, grammar correction, and rewriting of sentences and paragraphs. We have not used generative AI tools for any of the required disclosure tasks, including generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating or proving mathematical claims, proposing or refining hypotheses, designing or evaluating research methodology or experiments, implementing methods, translating content, cleaning or reformatting datasets, performing qualitative or thematic analysis, or interpreting research results.

All AI-assisted text was carefully reviewed and revised by the authors. The authors independently verified the accuracy, intended meaning, and consistency of the AI-assisted text with the scientific content of the paper. Generative AI tools were not used to generate scientific claims, experimental findings, or conclusions. The authors take full responsibility for the final content of this work, including all text, claims, and scientific conclusions.

REPRODUCIBILITY STATEMENT

We have made efforts to facilitate the reproducibility of our work by providing detailed descriptions of the proposed NeuroCodeX framework. The experimental setup, datasets, baselines, evaluation protocols, and metrics are described in the main paper, while additional experimental results and analyses are provided in the appendix. Further implementation details, including preprocessing procedures, model configurations, and training settings, are provided in the supplementary materials. All datasets used in this work are publicly available, and their usage and preprocessing procedures are documented in the appendix and supplementary materials.

REFERENCES

Claudio Babiloni, Xianghong Arakaki, Hamed Azami, Karim Bennys, Katarzyna Blinowska, Laura Bonanni, Ana Bujan, Maria C Carrillo, Andrzej Cichocki, Jaisalmer de Frutos-Lucas, et al. Measures of resting state eeg rhythms for clinical trials in alzheimer's disease: recommendations of an expert panel. *Alzheimer's & Dementia*, 2021.

Junru Chen, Yang Yang, Tao Yu, Yingying Fan, Xiaolong Mo, and Carl Yang. Brainnet: Epileptic wave detection from seeg with hierarchical graph diffusion learning. In *ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD)*, 2022.

Dong Cui, Shunai Qi, Guanghua Gu, Xiaoli Li, Zhaohui Li, Lei Wang, and Shimin Yin. Magnitude squared coherence method based on weighted canonical correlation analysis for eeg synchronization analysis in amnesic mild cognitive impairment of diabetes mellitus. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2018.

Ahmet Cengizhan Dirican and Mehmet Göktürk. Psychophysiological measures of human cognitive states applied in human computer interaction. *Procedia Computer Science*, 2011.

GUI Haokun, Xiucheng Li, and Xinyang Chen. Vector quantization pretraining for eeg time series with random projection and phase alignment. In *International Conference on Machine Learning (ICML)*, 2024.

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

Thi Kieu Khanh Ho and Narges Armanfard. Self-supervised learning for anomalous channel detection in eeg graphs: application to seizure analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

Weibang Jiang, Liming Zhao, and Bao-liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. In *International Conference on Learning Representations (ICLR)*, 2024.

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*, 2023.

Valer Jurcak, Daisuke Tsuzuki, and Ippeita Dan. 10/20, 10/10, and 10/5 systems revisited: their validity as relative head-surface-based positioning systems. *Neuroimage*, 2007.

Pieter-Jan Kindermans, David Hübner, Thibault Verhoeven, Klaus-Robert Müller, and Michael Tangermann. Unsupervised learning for brain–computer interfaces based on event-related potentials. *Signal Processing and Machine Learning for Brain–Machine Interfaces*, 2018.

Gennady G Knyazev. Eeg delta oscillations as a correlate of basic homeostatic and motivational processes. *Neuroscience & Biobehavioral Reviews*, 2012.

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*, 2022.

Yang Li, Wenming Zheng, Lei Wang, Yuan Zong, and Zhen Cui. From regional to global brain: A novel hierarchical spatial-temporal neural network model for eeg emotion recognition. *IEEE Transactions on Affective Computing*, 2019.

Hao Liu, Zihui Qi, Yihang Wang, Zhengyi Yang, Lingzhong Fan, Nianming Zuo, and Tianzi Jiang. A novel real-time phase prediction network in eeg rhythm. *Neuroscience Bulletin*, 2025.

Silvia López, Iyad Obeid, and Joseph Picone. Automated interpretation of abnormal adult electroencephalograms. *MS Thesis, Temple University*, 2017.

Dennis J McFarland and Jonathan R Wolpaw. Eeg-based brain–computer interfaces. *Current Opinion in Biomedical Engineering*, 2017.

Vladimir Miskovic and Andreas Keil. Reliability of event-related eeg functional connectivity during visual entrainment: Magnitude squared coherence and phase synchrony estimates. *Psychophysiology*, 2015.

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)*, 2022.

N Salansky, A Fedotchev, and A Bondar. Responses of the nervous system to low frequency stimulation and eeg rhythms: clinical implications. *Neuroscience & Biobehavioral Reviews*, 1998.

Xinke Shen, Runmin Gan, Kaixuan Wang, Shuyi Yang, Qingzhu Zhang, Quanying Liu, Dan Zhang, and Sen Song. Dynamic-attention-based eeg state transition modeling for emotion recognition. *IEEE Transactions on Affective Computing*, 2025.

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

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

M Barry Sterman. Physiological origins and functional correlates of eeg rhythmic activities: implications for self-regulation. *Biofeedback and Self-regulation*, 1996.

Jamiyan Sukhbaatar, Satoshi Imamura, and Toshihisa Tanaka. Bandvq: Band-wise vector-quantized eeg foundation model. *arXiv preprint arXiv:2605.24921*, 2026.

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 (ICLR)*, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Raveendrababu Vempati and Lakhan Dev Sharma. Eeg rhythm based emotion recognition using multivariate decomposition and ensemble machine learning classifier. *Journal of Neuroscience Methods*, 2023.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. In *International Conference on Learning Representations (ICLR)*, 2025.

Tian-Yu Xiang, Wen-Kai Wang, Xiao-Hu Zhou, Hong-Jun Yang, Mei-Jiang Gui, Xiao-Liang Xie, Shi-Qi Liu, Jia-Xing Wang, Jia-Mou Liu, and Zeng-Guang Hou. Cross-dataset montage alignment via electroencephalogram source imaging for enhanced motor decoding. *IEEE Transactions on Cognitive and Developmental Systems*, 2025.

Chaoqi Yang, Danica Xiao, M Brandon Westover, and Jimeng Sun. Self-supervised eeg representation learning for automatic sleep staging. *arXiv preprint arXiv:2110.15278*, 2021.

Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Mengqing Ye, CL Philip Chen, and Tong Zhang. Hierarchical dynamic graph convolutional network with interpretability for eeg-based emotion recognition. *IEEE transactions on neural networks and learning systems*, 2022.

Ke Yi, Yansen Wang, Kan Ren, and Dongsheng Li. Learning topology-agnostic eeg representations with geometry-aware modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zhizhang Yuan, Daoze Zhang, Yang Yang, Junru Chen, and Yafeng Li. Ppi: Pretraining brain signal model for patient-independent seizure detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics*, 2018.

A Related Work

Large-scale EEG Pretraining. Recent studies have explored large-scale and cross-dataset EEG pretraining to exploit the growing availability of EEG corpora for representation learning. BIOT (Yang et al., 2023) introduces a unified "biosignal sentence" representation to enable pretraining across multiple physiological signals. EEG foundation models (Wang et al., 2024; Jiang et al., 2024; Wang et al., 2025) further scale transformer-based architectures to large collections of EEG segments and demonstrate the potential of large-scale pretraining for diverse downstream applications. These studies establish the value of leveraging diverse EEG data for learning generalizable representations. However, heterogeneous EEG datasets differ substantially in electrode configurations and temporal resolutions, making direct joint pretraining non-trivial. Existing approaches often rely on a shared input structure or preprocessing-based alignment, which can obscure dataset-specific spatial or spectral characteristics.

Topography-Aware EEG Modeling. EEG recordings exhibit intrinsic spatial organization arising from functional interactions among brain regions, motivating the incorporation of topographic priors into EEG representation learning. Region-based graph models, such as R2G-STNN (Li et al., 2019), HD-GCN (Ye et al., 2022), and BrainNet (Chen et al., 2022), explicitly model regional interactions to capture spatial dependencies. PPi (Yuan et al., 2023) further leverages predefined scalp regions to improve robustness across subjects. More recently, MMM (Yi et al., 2023) and E-MAP (Xiang et al., 2025) address montage variations by learning topology-aware representations or mapping heterogeneous electrode configurations into shared spaces. These studies highlight the importance of preserving spatial structure in EEG modeling. Nevertheless, their primary focus is on spatial topology or montage adaptation, without explicitly addressing the concurrent spectral heterogeneity introduced by different sampling rates in cross-dataset pretraining.

Neural Tokenizers for EEG. Inspired by discrete representation learning in computer vision and speech processing, recent studies have explored neural tokenization for EEG representation learning. LaBraM (Jiang et al., 2024) introduces a vector-quantized spectrum prediction framework for channel-wise EEG patches, while VQ-MTM (Haokun et al., 2024) employs random vector quantization with phase alignment to improve token consistency. These approaches demonstrate the potential of discrete representations for capturing temporal–spectral patterns in EEG signals. More recently, BandVQ (Sukhbaatar et al., 2026) incorporates canonical EEG frequency bands by learning band-specific VQ-VAE tokenizers for masked token modeling, explicitly introducing frequency-specific structure into discrete EEG representations. In contrast, NeuroCodeX is designed for heterogeneous joint pretraining and employs multi-scale quantization to provide rhythm-aware discrete supervision across datasets with different temporal resolutions.

Motivation. Despite these advances, existing approaches largely address spatial topology and spectral structure as separate aspects of EEG representation learning. Spatially oriented methods primarily focus on handling heterogeneous electrode configurations, whereas neural tokenizers mainly investigate how to construct informative discrete representations from EEG signals. In heterogeneous EEG joint pretraining, however, these sources of structural heterogeneity can arise simultaneously: different datasets may employ incompatible electrode montages while also being recorded at different sampling rates, leading to mismatched spatial organization and temporal–spectral characteristics. A joint pretraining framework therefore needs to preserve meaningful spatial structure across heterogeneous montages while maintaining consistent discretization semantics across different temporal resolutions. NeuroCodeX is developed from this perspective, explicitly modeling these two forms of structural heterogeneity through hierarchical topographic representation learning and multi-scale, rhythm-aware quantization.

B More Results and Analysis

Analysis of Results on CHB-MIT. One notable observation is the substantial performance gap between NeuroCodeX and existing methods on CHB-MIT, particularly in AUC-PR. CHB-MIT adopts bipolar montages, where each channel represents the potential difference between a pair of electrodes, making the spatial relationship among channels an intrinsic part of the signal representation. Many existing EEG models do not explicitly model such spatial relationships under heterogeneous channel configurations, which may partly explain their relatively limited performance on CHB-MIT. In

Table 5: Evaluation results on TUAB. Best results are in bold, and the second best are underscored.

Models	TUAB		
	Balanced Acc.	AUC-PR	AUROC
Supervised Models			
SPaRCNet (2023)	0.7731 ± 0.0110	0.8440 ± 0.0079	0.8534 ± 0.0128
CNN-Transformer (2022)	<u>0.7809 ± 0.0095</u>	0.8243 ± 0.0334	0.8531 ± 0.0146
FFCL (2022)	0.7775 ± 0.0035	0.8461 ± 0.0103	0.8599 ± 0.0039
ST-Transformer (2021)	0.7790 ± 0.0053	0.8488 ± 0.0113	0.8738 ± 0.0046
Self-Supervised Models (pretrained from scratch)			
ContraWR (2021)	0.7756 ± 0.0097	0.8424 ± 0.0247	0.8551 ± 0.0114
VQ-MTM (2024)	0.7673 ± 0.0025	0.8391 ± 0.0211	0.8460 ± 0.0125
BIOT (2023)	0.7550 ± 0.0125	0.8256 ± 0.0198	0.8352 ± 0.0169
EEG Foundation Models			
$\text{EEGPT}_{\text{Large}}$ (2024)	0.7551 ± 0.0111	0.8252 ± 0.0072	0.8339 ± 0.0105
$\text{LaBraM}_{\text{Base}}$ (2024)	0.6176 ± 0.1143	0.6179 ± 0.1594	0.6562 ± 0.1473
CBraMod (2025)	0.7409 ± 0.0052	0.7901 ± 0.0246	0.8168 ± 0.0163
NeuroCodeX (Ours)			
Single-dataset pretraining	0.7723 ± 0.0143	<u>0.8526 ± 0.0082</u>	0.8595 ± 0.0100
Joint pretraining			
+ SEED-IV	0.7762 ± 0.0069	**0.8549 ± 0.0030**	<u>0.8598 ± 0.0054</u>
+ TUEV	0.7803 ± 0.0047	0.8456 ± 0.0057	**0.8606 ± 0.0046**
+ All	**0.7840** ± 0.0054	0.8442 ± 0.0092	0.8564 ± 0.0072

comparison, NeuroCodeX explicitly models electrode topology through HTE and maps heterogeneous channel configurations into anatomy-aligned regional representations, enabling it to better preserve structural information under such a montage. Moreover, NeuroCodeX achieves an AUC-PR of 0.5963 and an AUROC of 0.9412 under single-dataset pretraining, substantially exceeding the strongest foundation-model baseline, CBraMod (0.3689 and 0.8892, respectively). These results highlight the importance of explicitly modeling EEG spatial structure when learning representations across heterogeneous channel configurations.

Effect of Sampling-Rate Normalization. We further investigate whether explicitly normalizing sampling rates benefits heterogeneous EEG pretraining. As shown in Table 9, upsampling SEED-IV from 200 Hz to 250 Hz does not consistently improve performance for either conventional baselines or NeuroCodeX compared with their native-sampling results in Table 7, suggesting that interpolation alone is insufficient to resolve cross-dataset temporal-resolution heterogeneity. In contrast, Table 10 shows that forcing all signals to a common 200 Hz, a commonly used preprocessing strategy, consistently degrades NeuroCodeX on both TUEV and SEED-IV. These results suggest that sampling-rate heterogeneity should be modeled rather than normalized away, further motivating MSQ as a mechanism for learning spectrally consistent representations while preserving the native temporal resolution of each dataset.

C Discussion

This work revisits heterogeneous EEG pretraining from a structural perspective and shows that effective joint pretraining is not governed by data scale alone. Empirically, simply aggregating datasets with mismatched sampling rates and electrode montages does not reliably improve performance and can even degrade representations. Notably, joint pretraining on all available datasets does not consistently outperform partial combinations, further indicating that naive data scaling may be detrimental when structural mismatches are ignored. These findings challenge the common assumption that more data necessarily leads to better representations, and instead highlight that heterogeneity introduces non-trivial conflicts across datasets. Overall, our results suggest that

Table 6: Evaluation results on CHB-MIT. Best results are in bold, and the second best are underscored.

Models	CHB-MIT		
	Balanced Acc.	AUC-PR	AUROC
Supervised Models			
SPaRCNet (2023)	0.5876 ± 0.0040	0.1274 ± 0.0080	0.8143 ± 0.0036
CNN-Transformer (2022)	0.6389 ± 0.0029	0.2479 ± 0.0026	0.8862 ± 0.0184
FFCL (2022)	0.6262 ± 0.0008	0.2049 ± 0.0041	0.8271 ± 0.0115
ST-Transformer (2021)	0.5915 ± 0.0052	0.1422 ± 0.0035	0.8237 ± 0.0029
Self-Supervised Models (pretrained from scratch)			
ContraWR (2021)	0.6344 ± 0.0010	0.2264 ± 0.0026	0.8097 ± 0.0079
VQ-MTM (2024)	0.5896 ± 0.0061	0.1240 ± 0.0014	0.8692 ± 0.0047
BIOT (2023)	0.7068 ± 0.0189	0.3277 ± 0.0017	0.8237 ± 0.0078
EEG Foundation Models			
$\text{EEGPT}_{\text{Large}}$ (2024)	0.5889 ± 0.0157	0.1234 ± 0.0024	0.6837 ± 0.0014
$\text{LaBraM}_{\text{Base}}$ (2024)	0.7075 ± 0.0016	0.3287 ± 0.0014	0.8679 ± 0.0045
CBraMod (2025)	0.7398 ± 0.0010	0.3689 ± 0.0086	0.8892 ± 0.0032
NeuroCodeX (Ours)			
Single-dataset pretraining	0.7599 ± 0.0038	0.5963 ± 0.0051	0.9412 ± 0.0027
Joint pretraining			
+ SEED-IV	0.7620 ± 0.0068	0.5894 ± 0.0048	0.9396 ± 0.0043
+ TUEV	**0.7654 ± 0.0035**	0.5576 ± 0.0031	0.9407 ± 0.0022
+ All	0.7610 ± 0.0023	**0.6080 ± 0.0017**	**0.9430 ± 0.0015**

Table 7: Evaluation results on SEED-IV. Best results are in bold, and the second best are underscored

Models	SEED-IV		
	Balanced Acc.	Cohen's Kappa	Weighted F1
Supervised Models			
SPaRCNet (2023)	0.2950 ± 0.0010	0.0604 ± 0.0015	0.2905 ± 0.1230
CNN-Transformer (2022)	0.3237 ± 0.0012	0.0948 ± 0.0017	0.3163 ± 0.0018
FFCL (2022)	0.2918 ± 0.0033	0.0272 ± 0.0041	0.2829 ± 0.0052
ST-Transformer (2021)	0.3015 ± 0.0003	0.0701 ± 0.0004	0.2954 ± 0.0006
Self-Supervised Models (pretrained from scratch)			
ContraWR (2021)	0.3116 ± 0.0016	0.0835 ± 0.0033	0.2973 ± 0.0055
VQ-MTM (2024)	0.2882 ± 0.0028	0.0516 ± 0.0033	0.2894 ± 0.0045
BIOT (2023)	0.2990 ± 0.0001	0.0301 ± 0.0005	0.2842 ± 0.0007
EEG Foundation Models			
$\text{EEGPT}_{\text{Large}}$ (2024)	0.2693 ± 0.0011	0.0307 ± 0.0009	0.2744 ± 0.0001
$\text{LaBraM}_{\text{Base}}$ (2024)	0.2639 ± 0.0020	0.0177 ± 0.0025	0.2631 ± 0.0023
CBraMod (2025)	0.2616 ± 0.0015	0.0168 ± 0.0015	0.1886 ± 0.0023
NeuroCodeX (Ours)			
Single-dataset pretraining	0.3127 ± 0.0007	0.0839 ± 0.0021	0.3140 ± 0.0033
Joint pretraining			
+ TUAB	0.3264 ± 0.0035	0.1088 ± 0.0035	0.3124 ± 0.0046
+ CHB-MIT	**0.3392 ± 0.0026**	**0.1185 ± 0.0033**	**0.3380 ± 0.0041**
+ All	0.3031 ± 0.0017	0.0914 ± 0.0008	0.3050 ± 0.0012

Table 8: Ablation results on SEED-IV under joint pretraining with TUAB.

Model Variant	SEED-IV		
	Balanced Acc.	Cohen's Kappa	Weighted F1
Full model (Ours)	0.3264 ± 0.0035	0.1088 ± 0.0035	0.3124 ± 0.0046
w/o HTE	0.3163 ± 0.0023	0.0742 ± 0.0102	0.2869 ± 0.0012
w/o MSQ	0.3039 ± 0.0100	0.0811 ± 0.0114	0.2956 ± 0.0173

Table 9: Evaluation results on SEED-IV after upsampling from 200 Hz to 250 Hz.

Models	SEED-IV (upsampling 250hz)		
	Balanced Acc.	Cohen's Kappa	Weighted F1
Supervised Models			
SPaRCNet (2023)	0.2831 ± 0.0001	0.0613 ± 0.0026	0.2955 ± 0.0003
CNN-Transformer (2022)	0.2919 ± 0.0024	**0.1101 ± 0.0037**	0.2933 ± 0.0008
FFCL (2022)	0.2896 ± 0.0013	0.0237 ± 0.0014	0.2690 ± 0.0005
ST-Transformer (2021)	0.2700 ± 0.0002	0.0841 ± 0.0003	0.2941 ± 0.0064
Self-Supervised Models (pretrained from scratch)			
ContraWR (2021)	0.3157 ± 0.0011	0.0800 ± 0.0013	0.2319 ± 0.0035
VQ-MTM (2024)	0.2636 ± 0.0017	0.0483 ± 0.0022	0.2232 ± 0.0019
BIOT (2023)	0.2612 ± 0.0021	0.0233 ± 0.0035	0.2324 ± 0.0007
EEG Foundation Models			
$\text{EEGPT}_{\text{Large}}$ (2024)	0.2667 ± 0.0036	0.0325 ± 0.0002	0.2551 ± 0.0003
$\text{LaBraM}_{\text{Base}}$ (2024)	0.2640 ± 0.0012	0.0200 ± 0.0005	0.2437 ± 0.0039
CBraMod (2025)	0.2839 ± 0.0015	0.0346 ± 0.0010	0.2147 ± 0.0016
NeuroCodeX (Ours)			
Single-dataset pretraining	0.3021 ± 0.0006	0.0793 ± 0.0017	**0.3050 ± 0.0003**
Joint pretraining			
+ TUAB	0.3146 ± 0.0013	0.0849 ± 0.0003	0.3046 ± 0.0022
+ CHB-MIT	**0.3235 ± 0.0036**	0.1040 ± 0.0006	0.2999 ± 0.0004
+ All	0.2942 ± 0.0007	0.0922 ± 0.0004	0.3028 ± 0.0018

Table 10: Effect of resampling all datasets to 200 Hz under heterogeneous joint pretraining.

Model Variant	TUEV			SEED-IV		
	Balanced Acc.	Cohen's Kappa	Weighted F1	Balanced Acc.	Cohen's Kappa	Weighted F1
Native Sampling	0.4104 ± 0.0037	0.3911 ± 0.0027	0.6717 ± 0.0028	0.3264 ± 0.0035	0.1088 ± 0.0035	0.3124 ± 0.0046
Resampling to 200HZ	0.3611 ± 0.0026	0.3364 ± 0.0033	0.6111 ± 0.0062	0.2938 ± 0.0109	0.0551 ± 0.0166	0.2729 ± 0.0064

Table 11: Dataset Statistics.

Dataset	Sampling Rate	Channels	Training / Evaluation / Testing Samples	Tasks	Sample Length
TUAB	250/256 Hz	19-23 (unipolar)	251359 / 58881 / 30762	Normal/Abnormal Classification	12s
TUEV	250 Hz	19-23 (unipolar)	65558 / 29421 / 19374	EEG Event Classification	5s
SEED-IV	200 Hz	62 (unipolar)	6220 / 1244 / 1244	Emotion Recognition	5s
CHB-MIT	256 Hz	16 (bipolar)	286962 / 23065 / 16962	Seizure Detection	10s

effective EEG pretraining requires heterogeneity-aware architectures that respect native structural and spectral properties, rather than relying solely on data scaling or uniform preprocessing.

D Limitation and Future Work

Despite its effectiveness, NeuroCodeX has several limitations that warrant further investigation. First, although NeuroCodeX explicitly models spatial and spectral heterogeneity, the current framework adopts a relatively uniform joint-pretraining strategy, leaving the optimal integration of heterogeneous EEG datasets largely unexplored. The varying benefits across dataset combinations suggest that dataset compatibility and information complementarity should be explicitly considered in future heterogeneous pretraining frameworks. Second, the cortical region partition is predefined and shared across all samples, without explicitly accounting for subject-specific variability in functional organization; adaptive or data-driven spatial partitioning may provide greater flexibility under unseen spatial configurations. Finally, the codebooks in the Multi-Scale Quantizer are randomly initialized at different temporal scales. More structured codebook construction or initialization strategies may further improve the efficiency and stability of multi-scale discrete representation learning.

E Magnitude Squared Coherence

We employ magnitude squared coherence (MSC) (Miskovic & Keil, 2015; Cui et al., 2018), a widely used frequency-domain measure of linear dependence, to estimate functional connectivity between EEG-derived signals. In our framework, pairwise MSC values are computed by aggregating cross-spectral information across frequencies, and are subsequently used to instantiate the adjacency function $\mathrm{Adj}(\cdot)$ for constructing functional connectivity graphs in the HTE. MSC is adopted as a practical and well-established choice rather than a strict requirement of our framework, and other functional connectivity measures can be readily incorporated (Tang et al., 2022; Ho & Armanfard, 2023). The formulation of MSC is given as follows.

Given two real-valued temporal signals $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$, we transform them into the frequency domain using the real-valued Fourier transform:

$$\hat{\mathbf{x}} = \mathrm{rFFT}(\mathbf{x}), \quad \hat{\mathbf{y}} = \mathrm{rFFT}(\mathbf{y}), \quad \hat{\mathbf{x}}, \hat{\mathbf{y}} \in \mathbb{C}^K,$$

where $K = \lfloor \frac{N}{2} \rfloor + 1$ denotes the number of retained frequencies. We define the aggregated cross-spectral magnitude as

$$S_{xy} = |\langle \hat{\mathbf{x}}, \hat{\mathbf{y}} \rangle| = \left| \sum_{k=1}^{K} \hat{x}_k \, \hat{y}_k^* \right|,$$

and the corresponding auto-spectral energies as

$$S_{xx} = \|\hat{\mathbf{x}}\|_2^2, \qquad S_{yy} = \|\hat{\mathbf{y}}\|_2^2,$$

where $(\cdot)^*$ denotes the complex conjugate. The magnitude squared coherence between $\mathbf{x}$ and $\mathbf{y}$ is then defined as

$$\mathrm{MSC}(\mathbf{x}, \mathbf{y}) = \frac{S_{xy}^2}{S_{xx} S_{yy}}.$$

F Dataset Statistics and Splits

We evaluate NeuroCodeX on three representative EEG datasets that exhibit substantial heterogeneity in recording protocols, sampling rates, electrode montages, task definitions, and data scale as shown

in Table 11. For the two TUH datasets, although recordings may contain between 19 and 23 channels, the additional channels beyond the standard 10–20 system are not consistently available across sessions. To ensure stable spatial correspondence within each dataset, we retain the 19 canonical 10–20 channels shared across all recordings. This choice establishes a consistent within-dataset configuration without enforcing cross-dataset homogenization, and all montage heterogeneity across datasets is preserved during joint pretraining. Notably, different from the other datasets used in this work that adopt unipolar (referential) recordings, CHB-MIT is represented in a bipolar montage, where each channel corresponds to the voltage difference between two electrodes. This leads to distinct spatial characteristics and channel semantics, introducing an additional form of structural heterogeneity across datasets.

TUAB. The TUH Abnormal EEG Corpus (TUAB) (López et al., 2017) is a large-scale clinical EEG dataset consisting of long-duration recordings collected under diverse clinical conditions. Recordings are sampled at 250 256Hz and annotated at the session level as either *normal* or *abnormal*. We adopt the official train/validation/test splits provided by the dataset to ensure fair and reproducible evaluation.

TUEV. The TUH EEG Events Corpus (TUEV) (Harati et al., 2015) is a curated subset of the TUH EEG Corpus with fine-grained event-level annotations. EEG segments are labeled into six categories, including spike and sharp waves (SPSW), generalized periodic epileptiform discharges (GPED), periodic lateralized epileptiform discharges (PLED), eye movements (EYEM), artifacts (ARTF), and background activity (BCKG). All recordings are sampled at 250 Hz. We follow the official data splits for all experiments.

SEED-IV. The SJTU Emotion EEG Dataset for Emotion Recognition with Four Emotions (SEED-IV) (Zheng et al., 2018) is a non-clinical EEG dataset designed for affective state recognition. It is recorded using a distinct acquisition system with a sampling rate of 200 Hz and a channel layout different from the TUH datasets. Following prior work, we randomly split the dataset into training, validation, and test sets with a ratio of 8:1:1 at the subject-independent level.

CHB-MIT. The CHB-MIT Scalp EEG Database (Shoeb, 2009) is a public EEG dataset collected at the Children's Hospital Boston, consisting of long-term recordings from 23 pediatric subjects with intractable seizures. The EEG signals are recorded under the international 10–20 system with a sampling rate of 256 Hz and annotated with binary labels (seizure vs. non-seizure). In our experiments, we follow a preprocessing protocol similar to prior work (Yang et al., 2023; Wang et al., 2025) and construct a 16-channel bipolar montage from the original recordings. We adopt a subject-wise split, where subjects 1–19 are used for training, 20–21 for validation, and 22–23 for testing. For each bipolar channel, we associate the signal with the anatomical regions corresponding to both constituent electrodes (e.g., "Fp1–F7" is associated with the prefrontal and left frontal regions), thereby preserving the cross-regional spatial information of bipolar recordings within the unified regional representation.

G Details of Baseline Models

This section provides brief descriptions of all baseline models used in our experiments. These baselines cover a broad spectrum of supervised EEG models, self-supervised representation learning approaches, and large-scale pretrained EEG models. Unlike these methods, NeuroCodeX explicitly aligns spatial topology and spectral structure during pretraining, enabling consistent representation learning across datasets with heterogeneous electrode montages and sampling rates

SPaRCNet (Jing et al., 2023) is a supervised neural network that consists of multiple convolutional blocks, where each block is composed of several dense layers with convolution and ELU activations, interleaved with transition blocks for feature compression and downsampling.

CNN-Transformer (Peh et al., 2022) is a supervised EEG artifact detection model that combines convolutional neural networks with a transformer module. The CNN backbone is responsible for extracting local temporal features from raw EEG signals, while the subsequent transformer captures long-range temporal dependencies that are difficult to model using convolution alone. The model is trained using a belief matching (BM) loss to enhance class-wise discrimination under imbalanced artifact distributions.

FFCL (Li et al., 2022) is a supervised motor imagery EEG classification model based on feature fusion between convolutional and recurrent neural networks. The architecture adopts a parallel design in which a CNN branch extracts spatial features from EEG signals, while an LSTM branch models temporal dynamics.

ST-Transformer (Song et al., 2021) is an attention-based EEG decoding model designed to capture global spatial and temporal dependencies beyond convolutional receptive fields. The model first applies preprocessing and spatial filtering to the input EEG signals, followed by attention transformation along the feature-channel dimension to enhance informative spatial representations.

ContraWR (Yang et al., 2021) is a self-supervised representation learning framework designed to learn robust vectorized EEG representations from large-scale unlabeled signals. The model adopts a contrastive learning paradigm, where EEG segments are encoded into latent representations and trained to discriminate individual samples against global statistics derived from the entire dataset.

VQ-MTM (Haokun et al., 2024) is a BERT-style self-supervised learning framework for EEG time-series representation learning based on masked modeling. It introduces a random-projection-based vector quantization module to discretize continuous EEG signals into a set of semantic units, analogous to tokens in natural language. VQ-MTM further incorporates a phase-aligning module guided by the time–phase shift equivariance property of the Fourier transform, which stabilizes the quantization process under temporal shifts. The model is trained to reconstruct masked segments using a single shared codebook, providing discrete supervision for self-supervised learning. VQ-MTM primarily focuses on single-dataset pretraining and does not explicitly model hierarchical spatial topology or cross-dataset structural heterogeneity.

BIOT (Yang et al., 2023) is a flexible Transformer-based encoder designed for learning representations from heterogeneous biosignals, including EEG, ECG, and wearable sensor data. To handle mismatched channel configurations, variable sequence lengths, and missing values across datasets, BIOT tokenizes each channel independently into fixed-length temporal segments, which are then rearranged into a unified token sequence. Each segment is treated as a token and augmented with channel embeddings and relative positional embeddings to preserve spatio-temporal information. This tokenization strategy enables joint pretraining across multiple datasets with different signal formats and supports downstream fine-tuning on diverse tasks. BIOT has demonstrated strong performance in cross-dataset biosignal learning scenarios, including seizure detection and multi-modal signal modeling. In our experiments, we re-train BIOT from scratch rather than using the official pretrained checkpoints. This is because the released BIOT models incorporate learnable channel embeddings tied to a fixed channel set and ordering, which is incompatible with our heterogeneous EEG datasets featuring varying electrode configurations.

LaBraM (Jiang et al., 2024) is a large-scale EEG model proposed to address the limited generalizability of task-specific EEG deep learning models. LaBraM aims to learn universal EEG representations via unsupervised pretraining across large-scale datasets. LaBraM segments EEG recordings into channel-wise patches and introduces a neural tokenizer based on vector-quantized neural spectrum prediction. LaBraM is pretrained on approximately 2,500 hours of EEG data collected from around 20 datasets and has demonstrated strong performance across a wide range of downstream tasks, including abnormal EEG detection, event type classification, and emotion recognition. In our experiments, we adopt the officially released pretrained base-version model provided by the authors, available at https://github.com/935963004/LaBraM.

EEGPT (Wang et al., 2024) is a pretrained Transformer-based EEG model designed for universal EEG representation learning. It adopts a mask-based dual self-supervised learning framework that emphasizes spatio-temporal representation alignment. EEGPT employs a hierarchical structure that decouples spatial and temporal modeling, enabling efficient extraction of EEG features while reducing computational complexity. By pretraining on a large-scale, mixed multi-task EEG corpus, EEGPT demonstrates strong transferability across diverse downstream EEG tasks under linear-probing settings. In our experiments, we adopt the officially released pretrained large-version model provided by the authors, available at https://github.com/BINE022/EEGPT.

CBraMod (Wang et al., 2025) is a pretrained EEG foundation model that adopts a criss-cross Transformer architecture to model spatial and temporal dependencies separately via parallel attention mechanisms. It further incorporates an asymmetric conditional positional encoding scheme to enhance adaptability to diverse EEG formats. In our experiments, we use the officially released pre-

trained model provided by the authors, available at https://github.com/wjq-learning/CBraMod.

H Implementation Details

All experiments are conducted on a Linux server equipped with four NVIDIA GeForce RTX 4090 GPUs (24GB memory each). The model is implemented in PyTorch 2.6.0 with CUDA 12.4 acceleration. NeuroCodeX is pretrained using the AdamW optimizer with a peak learning rate of 2×10^{-3} and a batch size of 512. A linear warm-up followed by cosine annealing is adopted for learning-rate scheduling, and pretraining is performed for 150 epochs. The embedding dimension is set to $D = 256$, and the temporal encoder consists of a single-layer Transformer with a feed-forward dimension of 1024. The random vector quantizers corresponding to the δ, θ, α, β, and γ bands are instantiated with vocabulary sizes of 128, 256, 512, 1024, and 1024, respectively. The model parameters are initialized using Xavier initialization, while the discrete embeddings used by random vector quantization are sampled from a standard normal distribution and kept fixed during training. During downstream finetuning, cosine annealing is applied with an initial learning rate of 1×10^{-4} for the classifier head, while a smaller learning rate of 1×10^{-5} is used for the pretrained encoder f_{encoder}. For fair comparison, all baseline methods are implemented using their official codebases with the recommended hyperparameter settings. For EEG foundation models, we use the officially released pretrained weights without additional pretraining. For self-supervised learning baselines, we perform pretraining on the same training datasets under the same preprocessing protocol as NeuroCodeX.

I Analysis of Codebook Utilization

We provide a detailed analysis of the utilization behavior of the band-specific codebooks employed in MSQ, aiming to understand how discrete representations are distributed across different EEG frequency bands and datasets. For each frequency band $b \in \mathcal{B}$, we collect the sequence of discrete pseudo-labels $\{y_i^b\}_{i=1}^{N_p}$ assigned according to the quantization rule in Eq. (13) over all patches in the evaluation set. Here, $N_p = \sum_g R^g P$ denotes the total number of quantized patches aggregated across all regions and granularities, and each $y_i^b \in \{1, \ldots, |\mathcal{V}|\}$ indexes a codeword in the band-specific discrete vocabulary $\mathcal{V}$. Based on these assignments, we compute the following utilization statistics.

Codebook Utilization Rate. We define the utilization rate as the fraction of codebook entries that are activated at least once:

$$\text{Usage}^b = \frac{\left|\{y_i^b \mid i = 1, \ldots, N_p\}\right|}{|\mathcal{V}|}.$$

Entropy. To quantify the diversity of code usage, we compute the normalized Shannon entropy of the empirical assignment distribution:

$$H^b = -\frac{1}{\log |\mathcal{V}|} \sum_{k=1}^{|\mathcal{V}|} p_k^b \log p_k^b,$$

where

$$p_k^b = \frac{1}{N_p} \sum_{i=1}^{N_p} \mathbb{I}\left(y_i^b = k\right)$$

denotes the empirical probability of selecting the k-th codeword in band b.

Gini Coefficient. To measure the concentration of code usage, we compute the Gini coefficient over the same empirical distribution:

$$G^b = 1 - 2 \sum_{k=1}^{|\mathcal{V}|} p_{(k)}^b \left(\frac{|\mathcal{V}| - k + 0.5}{|\mathcal{V}|} \right),$$

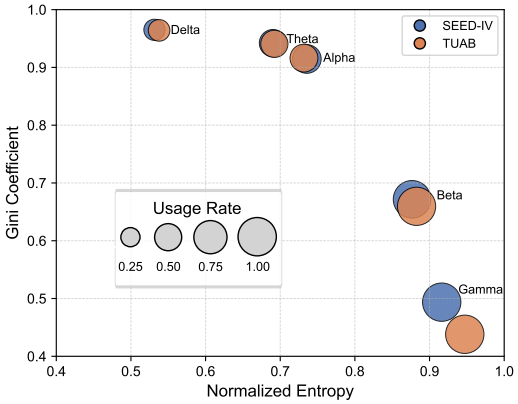

Figure 5: Codebook utilization statistics across frequency bands.

Table 12: Region partition ablations on TUEV and SEED-IV under joint pretraining with TUAB

Model Variant	TUEV			SEED-IV		
	Balanced Acc.	Cohen's Kappa	Weighted F1	Balanced Acc.	Cohen's Kappa	Weighted F1
Full model (Ours)	0.4104 ± 0.0037	0.3911 ± 0.0027	0.6717 ± 0.0028	0.3264 ± 0.0035	0.1088 ± 0.0035	0.3124 ± 0.0046
w/o Coarse	0.3886 ± 0.0018	0.3840 ± 0.0036	0.6505 ± 0.0044	0.3063 ± 0.0032	0.0956 ± 0.0021	0.2733 ± 0.0087
w/o Medium	0.3797 ± 0.0056	0.3706 ± 0.0059	0.6288 ± 0.0082	0.2805 ± 0.0103	0.0750 ± 0.0014	0.2539 ± 0.0093
w/o Fine	0.4011 ± 0.0033	0.3823 ± 0.0012	0.6146 ± 0.0016	0.2982 ± 0.0039	0.0888 ± 0.0055	0.2617 ± 0.0041

where $\{p^b_{(k)}\}$ denotes the probabilities $\{p^b_k\}$ sorted in descending order.

These metrics characterize complementary aspects of codebook behavior: utilization rate reflects effective capacity, entropy captures global diversity, and the Gini coefficient highlights concentration. High entropy combined with low Gini suggests uniformly distributed code usage, whereas low entropy and high Gini indicate that a small subset of codebook entries dominates the assignments. As shown in Figure 5, we observe a consistent frequency-dependent trend across datasets. Low-frequency rhythms (e.g., δ and θ) exhibit lower entropy and higher Gini coefficients, indicating that code usage is concentrated on a limited number of shared entries. This suggests that low-frequency EEG components exhibit relatively constrained structural variability and can be effectively represented with a compact shared codebook. In contrast, higher-frequency bands show increased utilization rates and entropy together with reduced concentration, reflecting inherently higher variability and less dominant code sharing. These results suggest that a fixed-size codebook induces frequency-dependent effects driven by the statistical properties of EEG rhythms, motivating future exploration of frequency-adaptive quantization strategies.

J Region Partitions

To model the hierarchical spatial organization of multichannel EEG signals under heterogeneous electrode configurations, we define region partitions at multiple spatial granularities. EEG channels are grouped into anatomically and topographically meaningful regions based on standard EEG montages and spatial proximity, enabling structured modeling of spatial dependencies across different scales. As illustrated in Fig. 6, we construct three levels of region partitions, corresponding to coarse-, medium-, and fine-grained granularities with 6, 10, and 17 regions, respectively. The coarse-grained partition aggregates channels into broad anatomical regions (e.g., frontal, temporal, central, parietal, and occipital) to capture global spatial context. The medium-grained partition further distinguishes hemispheric and midline structures, while the fine-grained partition introduces more localized subdivisions (e.g., frontocentral, temporo–parietal, and parieto–occipital regions) to model fine-scale spatial interactions. Detailed region definitions and channel assignments at each granularity are summarized in Table 13.

To assess the contribution of each spatial granularity, we conduct ablation studies by removing one level of region partition at a time while keeping the remaining components unchanged. The results on TUEV and SEED-IV under joint pretraining with TUAB are reported in Table 12. Overall, removing any level of the hierarchical partition leads to consistent performance degradation across most metrics, indicating that multi-granularity region modeling is beneficial. Among the three levels, the medium-grained partition exhibits the most pronounced impact, particularly on TUEV, suggesting the importance of explicitly modeling hemispheric and midline structures for event-related EEG signals. The coarse-grained partition provides complementary global spatial context, while the fine-grained partition mainly contributes to improvements in F1 score and AUROC, reflecting its role in capturing localized spatial interactions.

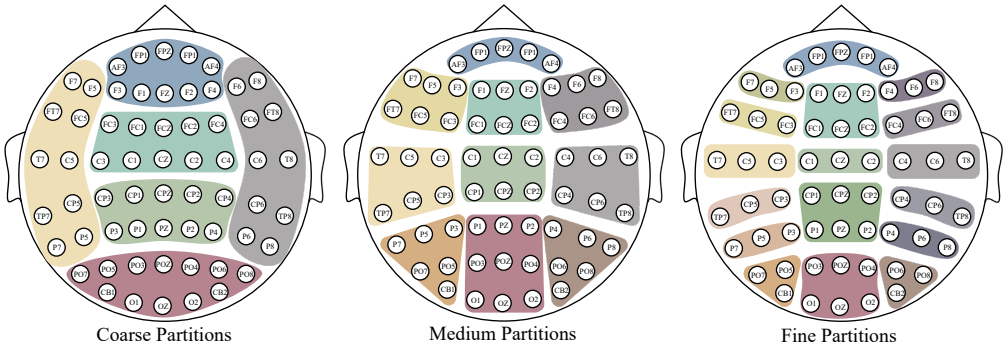

Figure 6: Example of region partitions with 62 channels.

Table 13: Detail region definitions at different spatial granularities.

Granularity	Region	Channels
Coarse	Frontal	FP1, FP2, FPZ, AF3, AF4, F7, F5, F3, F1, FZ, F2, F4, F6, F8, FC5, FC3, FC1, FCZ, FC2, FC4, FC6
	Temporal (Left)	T3, T5, FT7, T7, TP7, TP9
	Temporal (Right)	T4, T6, FT8, T8, TP8, TP10
	Central	C3, C4, CZ, C1, C2, C5, C6
	Parietal	P3, P4, PZ, CP5, CP3, CP1, CPZ, CP2, CP4, CP6, P7, P5, P1, P2, P6, P8, PO9, PO10
	Occipital	O1, O2, OZ, PO7, PO5, PO3, POZ, PO4, PO6, PO8, CB1, CB2
Medium	Prefrontal	FP1, FP2, FPZ, AF3, AF4
	Frontal (Left)	F7, F5, F3, FT7, FC5, FC3
	Frontal (Midline)	F1, FZ, F2, FC1, FCZ, FC2
	Frontal (Right)	F4, F6, F8, FT8, FC4, FC6
	Temporal (Left)	T3, T5, T7, TP7, TP9, C5, C3, CP3, CP5
	Temporal (Right)	T4, T6, T8, TP8, TP10, C6, C4, CP4, CP6
	Central	CZ, C1, CP1, C2, CPZ, CP2
	Parieto–Occipital (Midline)	P1, PZ, P2, PO3, POZ, PO4, O1, OZ, O2
	Parieto–Occipital (Left)	P3, P5, P7, PO5, PO7, PO9, CB1
	Parieto–Occipital (Right)	P4, P6, P8, PO6, PO8, PO10, CB2
Fine	Prefrontal	FP1, FP2, FPZ, AF3, AF4
	Frontal (Left)	F3, F5, F7
	Frontal (Right)	F4, F6, F8
	Frontocentral (Left)	FT7, FC5, FC3
	Frontocentral (Midline)	F1, FZ, F2, FC1, FCZ, FC2
	Frontocentral (Right)	FT8, FC6, FC4
	Temporal (Left)	T3, T5, T7, C5, C3
	Temporal (Right)	T4, T6, T8, C6, C4
	Temporo–Parietal (Left)	TP7, TP9, CP5, CP3
	Temporo–Parietal (Right)	TP8, TP10, CP6, CP4
	Central	C1, CZ, C2
	Centro–Parietal (Midline)	CP1, CPZ, CP2, P1, PZ, P2
	Parietal (Left)	P3, P5, P7
	Parietal (Right)	P4, P6, P8
	Occipital (Left)	PO5, PO7, PO9, CB1
	Occipital (Right)	PO6, PO8, PO10, CB2
	Occipital (Midline)	O1, OZ, O2, PO3, POZ, PO4