EEG-X: Toward Device-Agnostic and Noise-Robust Foundation Models for EEG

Navid Mohammadi Foumani
Emotiv Research
Melbourne, Australia
navid@emotiv.com

Soheila Ghane*
Google
Melbourne, Australia
soheila@google.com

Nam Nguyen
Emotiv Research
Sydney, Australia
namnguyen@emotiv.com

Mahsa Salehi
Monash University
Melbourne, Australia
mahsa.salehi@monash.edu

Geoffrey I. Webb
Monash University
Melbourne, Australia
geoff.webb@monash.edu

Geoffrey Mackellar
Emotiv Research
Sydney, Australia
geoff@emotiv.com

Abstract

Foundation models for EEG analysis are still in their infancy, limited by two key challenges: (1) variability across datasets caused by differences in recording devices and electrode configurations, and (2) the low signal-to-noise ratio (SNR) of EEG, where neural activity is often buried under artifacts and non-brain sources. To address these challenges, we present *EEG-X*, a device-agnostic and noise-robust foundation model for EEG representation learning. EEG-X introduces a location-based channel embedding that enables robust transfer across heterogeneous EEG devices and electrode layouts. To improve robustness against noise, EEG-X employs a novel noise-aware masking-reconstruction strategy that reconstructs artifact-cleaned signals rather than raw noisy EEG and introduces a novel Dictionary-based Convolution Transformation (DiCT) that improves reconstruction-based pretraining by comparing signals in a structured feature space instead of directly in the raw signal space. Experiments across datasets collected from diverse EEG devices show that EEG-X consistently outperforms state-of-the-art methods across multiple downstream tasks and demonstrates strong cross-domain generalization when pretraining and downstream datasets differ in electrode layouts and recording configurations paving the way towards foundation models. The models and code are available at github.com/Emotiv/EEG-X.

1 Introduction

Electroencephalography (EEG) is a non-invasive method for recording brain activity by placing electrodes on the scalp, offering high temporal resolution and relatively low cost [Teplan, 2002, Müller-Putz, 2020]. It is widely used in both neuroscience and clinical practice, where it supports the diagnosis of neurological disorders such as epilepsy and sleep disorders [Lotte et al., 2018, Boonyakitanont et al., 2020], as well as in applications like sleep stage classification, affective state analysis, motor imagery decoding, stress detection, and seizure classification [Craik et al., 2019, Hosseini et al., 2021, Raković et al., 2024]. Despite these benefits, effectively modeling EEG data often requires large amounts of annotated data, expert knowledge, and techniques specially designed to the unique characteristics of each task [Weng et al., 2024].

Recently, self-supervised pretraining has achieved remarkable success in natural language processing (NLP), computer vision (CV), and speech/audio processing [Bommasani et al., 2021, Yu et al., 2022,

*Work done while at Emotiv Research.

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

Chen et al., 2023]. Models pretrained on large unlabeled datasets can be efficiently adapted to a wide range of downstream tasks, often requiring little or no labeled data. A major advantage of these models is their ability to handle multiple tasks within a single framework, unlike traditional approaches that typically require large labeled datasets and task-specific algorithms. This representation learning approach simplifies development by reducing the need for task-specific models and extensive labeling, while also improving generalization by leveraging shared patterns from diverse data. This makes the model more adaptable and efficient across various tasks, forming the basis of what is known as a foundation model.

While foundation models have shown great success in vision, language, and speech, their application to EEG data remains limited and has yet to achieve comparable breakthroughs [Weng et al., 2024, Mohammadi Foumani et al., 2024b]. Two key challenges hinder the broader adoption of EEG foundation models: (1) substantial variability across EEG datasets due to differences in recording equipment, which introduce variations in sampling frequencies, electrode placements, and the number of channels, and (2) the inherently low signal-to-noise ratio of EEG, where neural activity is often buried under artifacts and non-brain sources [Jiang et al., 2019, Pion-Tonachini et al., 2019, Blum et al., 2019]. To address these challenges, we propose EEG-X, an EEG representation learning model that learns device-agnostic and noise-robust representations through a location-based channel embedding to handle variability across devices, a noise-aware masking–reconstruction strategy for robust pretraining, and a dictionary-based convolution transformation layer that improves reconstruction loss by enabling a balanced frequency- and shape-aware comparisons.

The first component, location-based channel embedding, is designed to handle device variability. Earlier models typically rely on learnable channel embeddings [Yang et al., 2023, Jiang et al., 2024, Wang et al., 2024, 2025, Yang et al., 2026] to indicate signal source, which can work within a single dataset but transfer poorly across devices, as the same channels may not exist across different EEG devices. In addition, such embeddings do not account for electrode positions or preserve the similarity between neighboring regions on the scalp, both of which are important for modeling brain activity. EEG-X instead encodes electrode positions together with their local spatial neighborhoods, enabling nearby electrodes to share similar representations. Furthermore, EEG-X introduces a universal scalp mapping strategy based on a dense scalp mesh spanning the standard 10-05, 10-10, and 10-20 EEG systems, allowing electrodes from heterogeneous devices to be mapped into a shared spatial representation.

The second component addresses the low Signal-to-Noise Ratio (low-SNR) nature of EEG signals. Unlike previous reconstruction-based methods that reconstruct the original noisy EEG signals [Kostas et al., 2021, Chien et al., 2022, Wang et al., 2024, 2025, Ouahidi et al., 2025, Yang et al., 2026], EEG-X reconstructs artifact-cleaned signals obtained through an artifact removal process. This encourages the model to focus on neural activity rather than noise and improves the robustness and quality of the learned representations.

Finally, EEG-X introduces a **Di**ctionary-based **C**onvolution **T**ransformation (DiCT) layer to improve the reconstruction objective in MAE-based architectures. While Mean Squared Error (MSE) is widely used in self-supervised prediction tasks [Baevski et al., 2023, 2022], it may be suboptimal for EEG signals due to its point-wise nature and tendency to overemphasize large-amplitude components [Thompson, 1990, Foumani et al., 2024a]. Consequently, MSE may inadequately capture frequency-specific and structural similarities in EEG signals. To address these limitations, DiCT projects both reconstructed and target EEG signals into a structured feature space using random convolutional kernels with multiple dilation scales before computing reconstruction loss. By computing MSE on projected temporal features rather than directly in the raw signal space, DiCT improves sensitivity to multi-scale temporal patterns while increasing robustness to amplitude variation and temporal misalignment. Importantly, DiCT has no learnable parameters and can be seamlessly integrated into existing reconstruction-based methods. Here is a summary of our contributions:

- **Location-Based Channel Embedding**: Introduces a spatially-aware channel embedding that encodes electrode locations and local scalp neighborhoods, enabling robust transfer across heterogeneous EEG devices and channel layouts.
- **Noise-Aware Reconstruction**: Proposes a reconstruction strategy based on artifact-cleaned EEG signals, encouraging representations to focus on neural activity while improving robustness to low-SNR EEG recordings.

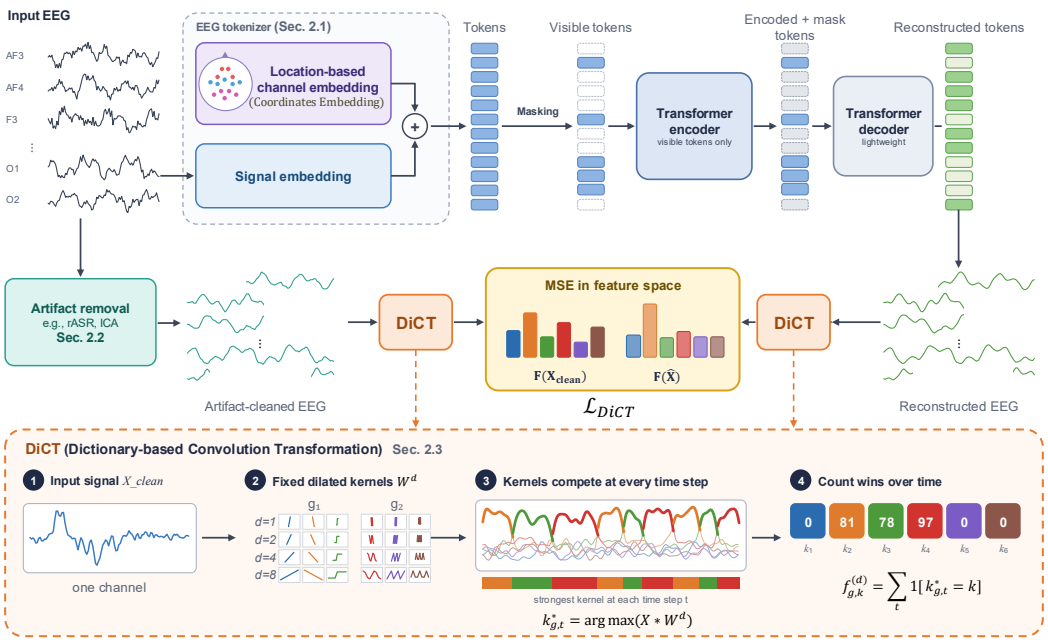


Figure 1: Architecture of EEG-X. EEG-X is built upon a masked autoencoding (MAE) framework and consists of three main components: (1) a signal embedding module augmented with location-based channel embeddings for handling heterogeneous EEG devices and channel layouts, (2) an artifact removal module that produces artifact-cleaned reconstruction targets from raw EEG signals, and (3) the proposed Dictionary-based Convolution Transformation (DiCT) objective for structured reconstruction learning.

- **Dictionary-based Convolution Transformation (DiCT)**: Introduces a lightweight structured reconstruction objective that mitigates limitations of point-wise MSE by improving sensitivity to multi-scale frequency content and temporal structure.
- **Extensive Evaluation**: Demonstrates that EEG-X serves as a strong representation learning model for EEG, achieving state-of-the-art performance across diverse downstream tasks and datasets while exhibiting robust cross-domain generalization across heterogeneous EEG devices and recording configurations.

2 Methodology of EEG-X

Fig. 1 illustrates the overall architecture of EEG-X, which consists of three main components: (1) an input tokenizer with location-based channel embedding, (2) a noise-aware masked reconstruction framework, and (3) a Dictionary-based Convolution Transformation (DiCT) reconstruction objective. Given an EEG signal, EEG-X first divides the input into fixed-length temporal segments and projects them into token representations using a signal embedding module. To incorporate spatial information, each channel is augmented with a location-based channel embedding derived from electrode coordinates and local scalp neighborhoods, enabling robust representation learning across heterogeneous EEG devices and channel layouts.

The resulting tokens are masked and processed using a Masked Autoencoder (MAE) [He et al., 2022] framework, where the model reconstructs masked EEG tokens using contextual information from the remaining visible tokens. Similar to vision MAE frameworks, the encoder operates only on visible tokens to efficiently learn contextual representations, while the decoder reconstructs the missing regions from latent representations together with mask tokens. EEG-X adopts an asymmetric encoder–decoder design, where the decoder is intentionally smaller than the encoder to reduce computational cost and encourage the encoder to learn richer representations.

Unlike previous reconstruction-based methods that reconstruct the original noisy EEG, EEG-X reconstructs artifact-cleaned EEG obtained through an artifact removal process, encouraging the

model to focus on neural activity rather than noise. To further improve reconstruction learning, EEG-X introduces a Dictionary-based Convolution Transformation (DiCT) objective. Instead of computing reconstruction loss directly in the raw signal space, both the reconstructed and target EEG signals are projected into a structured feature space using random convolutional kernels with multiple dilation scales before computing MSE loss. By comparing the occurrence and distribution of temporal patterns rather than enforcing strict point-wise alignment, DiCT improves sensitivity to multi-scale temporal patterns while increasing robustness to amplitude variation and temporal misalignment. For downstream tasks, the tokenizer and encoder are frozen, and only a lightweight classifier is trained on top of the learned EEG representations.

2.1 EEG Tokenizer

Our goal is to model heterogeneous EEG signals—recorded using different devices with varying channel configurations, combinations, and recording lengths—within a unified framework.

Signal Embedding: Let $\mathbf{X} = \{x_1, \ldots, x_C\}$ denote an EEG recording with C channels, where each channel $x_c \in \mathbb{R}^L$ has L samples. To handle recordings of different lengths, we split each channel into fixed-length tokens of w samples using a sliding window with stride $s = w - o$, where $o < w$ is the overlap (typically $o = w/4$). The i-th token of channel c is

$$t_{c,i} = x_c\big[(i-1)s + 1 : (i-1)s + w\big], \qquad i = 1, \ldots, N, \tag{1}$$

where $N = \lfloor (L - w)/s \rfloor + 1$ is the number of tokens per channel. To capture both temporal and spectral characteristics of EEG signals, we combine raw time-domain and frequency-domain representations for each token. Specifically, we apply the Short-Time Fourier Transform (STFT) to each token and extract the magnitude of its frequency representation [Yang et al., 2023]. The raw token and its corresponding STFT magnitude are then independently projected into embedding spaces using learned linear transformations. The resulting representations are concatenated to obtain the final token embedding $e_{c,i}$:

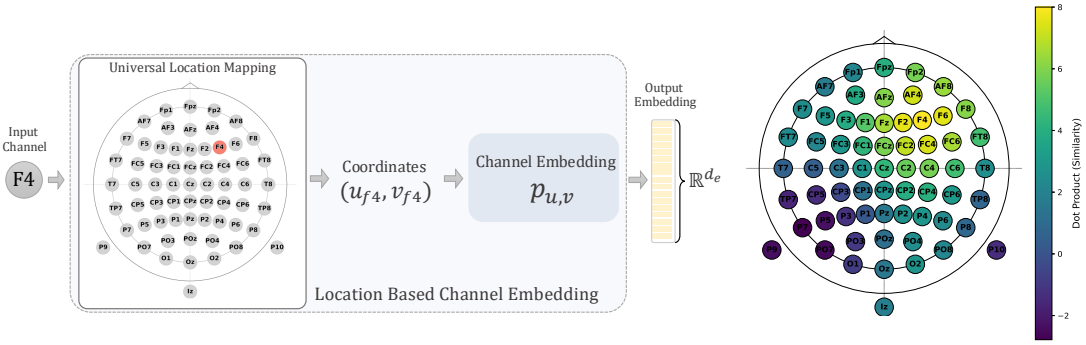
$$e_{c,i} = \left[W_{\text{raw}} t_{c,i} \; ; \; W_{\text{freq}} \left|\text{STFT}(t_{c,i})\right|\right], \quad e_{c,i} \in \mathbb{R}^{d_e}. \tag{2}$$

Here, $\mathbb{R}^{d_e}$ denotes the embedding dimension, and the transformations are parameterized by learned weight matrices W_{raw} and W_{freq}. The concatenation combines information from both the original temporal signal and its spectral representation, allowing the model to learn complementary features from the two domains. This tokenization process is similar to natural language processing, where each EEG sample is treated as a sequence of tokens, similar to words in a sentence.

While this signal embedding addresses variations in recording length, handling channel variability remains a challenge. Previous approaches train an embedding parameter for all channels in the data and add the corresponding channel embedding to the token [Yang et al., 2023, Jiang et al., 2024, Wang et al., 2024, 2025]; however, they present several limitations: (1) The same channels may not exist across different EEG devices, making the learned embeddings less transferable across diverse recording setups. (2) Channel embeddings may not account for the spatial locations of electrodes on the scalp, which are crucial for accurately modeling the brain's spatial activity patterns. (3) They do not preserve the similarity between neighboring regions on the scalp, which can lead to the loss of meaningful spatial relationships. These limitations highlight the need for a new embedding strategy that incorporates both the spatial location of the channels and their relationships across different devices, which our method aims to address.

To address these limitations, EEG-X incorporates electrode location information directly into the token representations. Each electrode is mapped to Cartesian scalp coordinates using a universal scalp mapping (Fig. 2a). For visualization purposes, Fig. 2a illustrates a 64-channel configuration. In practice, EEG-X uses a substantially denser universal scalp mesh spanning standard 10-05, 10-10, and 10-20 EEG systems, enabling heterogeneous EEG devices to be represented within a shared spatial coordinate space (Appendix B.6). Recent work such as REVE [Ouahidi et al., 2025] also combines spatial location encoding with learnable channel embeddings. However, REVE employs a joint multidimensional Fourier encoding, whereas EEG-X adopts a simpler separable sinusoidal formulation together with the proposed universal scalp mapping. In practice, we find that the separable formulation is more robust and remains stable even at relatively small embedding dimensions.

Given an electrode location (u, v), where u and v denote the horizontal and vertical coordinates (i.e., x and y), we generate spatial embeddings using a separable sinusoidal encoding adapted from transformer positional embeddings [Vaswani et al., 2017, Foumani et al., 2024b]:



(a) (b)

Figure 2: (a) Location-based channel embedding using a universal scalp coordinate mapping to assign each channel Cartesian coordinates. (b) Embedding similarity of electrode F4 with all other channels, showing higher similarity to nearby electrodes (e.g., F2, F6) than distant ones (e.g., P7).

$$\begin{aligned} p_{u,v}(4k) &= \sin u\omega_k, \quad p_{u,v}(4k+1) = \cos u\omega_k, \\ p_{u,v}(4k+2) &= \sin v\omega_k, \quad p_{u,v}(4k+3) = \cos v\omega_k \end{aligned} \tag{3}$$

where

$$\omega_k = 1000^{-4k/d_e},$$

$k \in \{0, \ldots, d_e/4 - 1\}$, and d_e is the embedding dimension. The frequency term ω_k assigns multiple spatial frequencies across embedding dimensions, enabling the model to represent both coarse and fine-grained spatial relationships between electrodes.

This approach allows signals from nearby electrode positions to receive similar embeddings. Specifically:

$$\text{Dist}((u_i, v_i), (u_j, v_j)) < \text{Dist}((u_i, v_i), (u_k, v_k)) \quad \Longrightarrow \quad p_{u_i,v_i} \cdot p_{u_j,v_j} > p_{u_i,v_i} \cdot p_{u_k,v_k}$$

Where Dist() is the distance function in Cartesian coordinates, and $p_{u_i,v_i}, p_{u_j,v_j} \in \mathbb{R}^{d_e}$ represent the channel embeddings of the electrodes at positions (u_i, v_i) and (u_j, v_j), respectively. The equation demonstrates that electrodes closer in physical space (i.e., neighboring electrodes) will have higher similarity in their embeddings, as indicated by the dot product.

Fig. 2b shows the embedding similarity between electrods $F4$ with all other electrodes. For example, as shown in Fig. 2b, signals from the $F4$ electrode will have embeddings more similar to those of $F2$ and $F6$ than to other electrodes (such as $P7$). As a result, when any of these electrodes are present in a dataset, the model can effectively utilize them. If an electrode is missing or too noisy to be useful, the model can reconstruct the signal primarily using data from adjacent electrodes. This approach mitigates the issue of missing channels across devices, as it maintains consistent channel relationships based on location, even if the channel labeling differs. Additionally, this encoding provides the model with spatial awareness, enabling it to recognize regional brain activity patterns, where neighboring areas exhibit similar behaviors while differing from other regions.

2.2 Noise-Aware Reconstruction

EEG signals exhibit inherently low signal-to-noise ratio (SNR), where neural activity is often contaminated by artifacts and non-brain sources. Previous MAE-based self-supervised methods primarily reconstruct the original raw EEG signal [Chien et al., 2022, Yang et al., 2023, Wang et al., 2024, 2025, Ouahidi et al., 2025, Yang et al., 2026]; however, this is particularly challenging for EEG, as artifacts can dominate the underlying neural activity. Consequently, reconstruction objectives may encourage models to preserve artifact-related patterns together with meaningful brain signals, reducing the quality and robustness of the learned representations.

To address this limitation, EEG-X uses artifact-cleaned EEG signals as reconstruction targets instead of the original noisy input. Our framework is agnostic to the choice of denoising method; in this work, we adopt Riemannian Artifact Subspace Reconstruction (rASR) [Blum et al., 2019] due to its robustness, computational efficiency, and suitability for large-scale EEG processing (see Appendix B.7 for a broader discussion of alternative denoising methods). Given an input EEG signal $\mathbf{X}$, rASR produces an artifact-cleaned signal $\mathbf{X}_{\text{clean}}$, which serves as the reconstruction target during self-supervised pretraining. Detailed descriptions of the rASR calibration process, parameter selection,

and implementation details are provided in Appendix B.7. By reconstructing artifact-cleaned EEG rather than noisy raw signals, EEG-X encourages the learned representations to focus on neural activity while improving robustness to artifacts and recording variability.

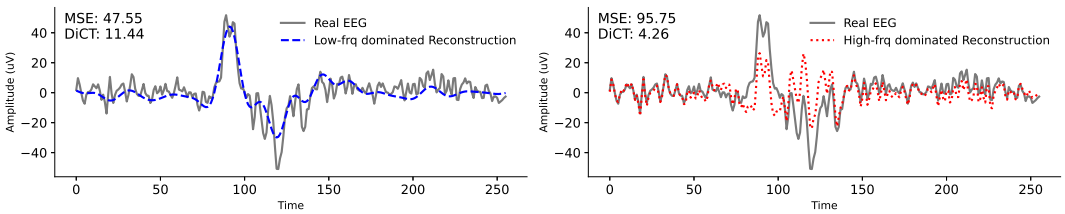


Figure 3: Real EEG signal and reconstructed variants. **Left:** low-frequency dominant reconstruction preserving mainly Delta and Theta components. **Right:** mid/high-frequency dominant reconstruction preserving Theta, Alpha, Beta, and Gamma components. These examples illustrate the amplitude bias of direct time-domain MSE, where low-frequency high-amplitude components can dominate the reconstruction objective despite missing informative higher-frequency temporal patterns.

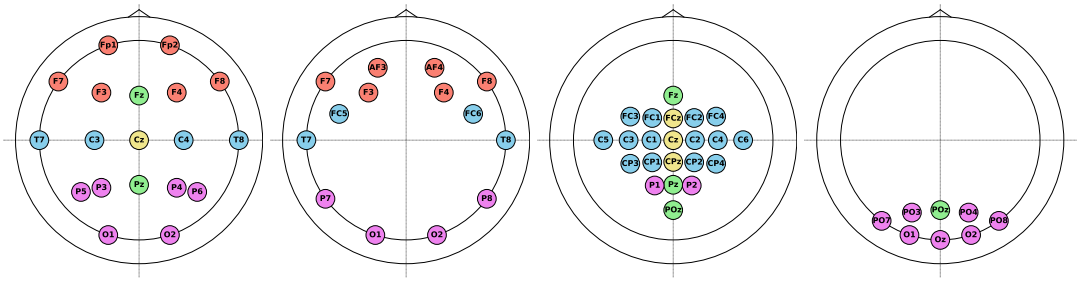
2.3 Dictionary-based Convolution Transformation (DiCT)

As illustrated in Fig. 1, EEG-X is built upon a masked autoencoding (MAE) framework [He et al., 2022], where portions of the input EEG tokens are masked and the model is trained to reconstruct the missing regions from the remaining visible tokens. A straightforward approach is to compute reconstruction loss directly in the raw signal space using Mean Squared Error (MSE), as commonly adopted in previous self-supervised EEG methods [Chien et al., 2022, Yang et al., 2023, Wang et al., 2024, 2025, Ouahidi et al., 2025, Yang et al., 2026]. However, EEG signals are inherently multi-scale, where meaningful neural activity is distributed across multiple temporal resolutions and frequency bands with substantially different amplitudes and dynamics.

EEG recordings are also highly susceptible to artifacts and non-brain sources. For example, eye blinks and movement artifacts often dominate low-frequency regions, while muscle activity and line noise introduce strong high-frequency components. As a result, both low- and high-frequency regions can be dominated by non-neural, high-amplitude noise, causing point-wise losses such as MSE to be disproportionately influenced by large-amplitude components. Fig. 3 illustrates this limitation using a real EEG signal together with two reconstructed variants. The left reconstruction preserves only low-frequency components, specifically Delta (1–4 Hz) and Theta (4–8 Hz) bands, while the right reconstruction preserves richer temporal structure including Theta (4–8 Hz), Alpha (8–14 Hz), Beta (14–30 Hz), and Gamma (>30 Hz) components (details of signal construction are provided in Appendix B). Despite the right reconstruction containing more informative temporal dynamics, direct time-domain MSE favors the low-frequency dominant reconstruction because its large-amplitude components align more closely with the original signal. Specifically, the low-frequency dominant reconstruction achieves a lower MSE (47.55) compared to the richer mid/high-frequency reconstruction (95.75).

To address these limitations, EEG-X introduces the **Di**ctionary-based **C**onvolution **T**ransformation (DiCT), which projects both the reconstructed and the target EEG signals into a structured feature space before computing the reconstruction loss. DiCT is inspired by the success of random convolutional kernels [Dempster et al., 2020] and competing kernel groups [Dempster et al., 2023] in time series classification. Instead of enforcing exact point-wise alignment, it compares the occurrence and distribution of temporal patterns across multiple temporal scales.

Let $\mathbf{X}_{\text{clean}} \in \mathbb{R}^{C \times L}$ denote the artifact-cleaned target EEG signal (shown as $\mathbf{X}$ for simplicity) and $\hat{\mathbf{X}}$ its reconstruction. We apply random convolutional kernels organized into G groups, each containing K kernels. To better capture EEG temporal dynamics, DiCT employs multiple kernel sizes and dilation scales designed to model neural activity across different temporal resolutions and frequency ranges. As shown in Fig. 1, kernels within each group compete to represent local temporal patterns. The kernels operate at multiple dilation scales d, where d controls the temporal spacing between convolutional elements. The dilation ranges and kernel configurations are selected according to EEG temporal characteristics and sampling frequencies (Details of kernel design and parameter selection are provided in Appendix B).



(a) TUH Cap 19-channel (b) Emotiv Epoc 14-channel (c) BCICIV Cap 22-channel (d) SSVEP Cap 8-channel

Figure 4: Comparison of four different EEG headset configurations used for data collection

The convolutional responses are computed as:

$$\mathbf{R}^{(d)} = \mathbf{X} *_d \mathbf{W}^{(d)}, \quad \mathbf{R}^{(d)} \in \mathbb{R}^{G \times K \times T}, \tag{4}$$

where $*_d$ denotes dilated convolution. The convolutional kernels $\mathbf{W}^{(d)}$ are shared across EEG channels and applied independently to each channel. For each group g and time step t, the kernel with the strongest response is selected:

$$k^*_{g,t} = \arg\max_k \mathbf{R}^{(d)}_{g,k,t}. \tag{5}$$

The responses are then aggregated across time:

$$f^{(d)}_{g,k} = \sum_t \mathbf{1}[k^*_{g,t} = k]. \tag{6}$$

For simplicity, we illustrate the kernel selection using hard assignments in Equations (5) and (6). However, hard argmax assignments are non-differentiable and do not allow gradients to flow through the assignment operation. Therefore, during training, we use a differentiable softmax relaxation of the assignments (detailed in Appendix B.4). The final representation is obtained by concatenating features across groups, dilation scales, and channels. Reconstruction loss is then computed in this structured feature space:

$$\mathcal{L}_{\text{DiCT}} = MSE(\mathcal{F}(\mathbf{X}) - \mathcal{F}(\hat{\mathbf{X}})). \tag{7}$$

By comparing temporal patterns instead of enforcing strict point-wise alignment, DiCT improves robustness to amplitude imbalance, artifacts, and temporal misalignment while preserving multi-scale temporal structure. As shown in Fig. 3, DiCT assigns a substantially lower reconstruction error to the richer mid/high-frequency reconstruction (4.26) compared to the low-frequency dominant reconstruction (11.44), indicating improved sensitivity to temporal and frequency-aware structure (additional quantitative comparisons are provided in Appendix B).

3 Experimental Results

We report the main experimental results of EEG-X, with additional setup details and supplementary analyses provided in Appendix C.

3.1 Datasets and Pre-processing

We evaluate EEG-X across eight diverse datasets from five EEG headsets, ranging from portable consumer-grade devices to clinical-grade systems, in both real-world and clinical settings. We use larger datasets for in-domain evaluation and smaller ones, recorded with different headsets and channel configurations (Fig. 4), for cross-domain evaluation. A summary of the datasets is provided in Table 1, with additional preprocessing details in Appendix C.5.

3.2 Comparison with Baseline

We conducted comprehensive evaluations against eight state-of-the-art EEG foundation and representation learning models: MAEEG [Chien et al., 2022], BioT [Yang et al., 2023], LaBraM [Jiang et al., 2024], EEGPT [Wang et al., 2024], EEG2Rep [Mohammadi Foumani et al., 2024a], CBraMod [Wang et al., 2025], REVE [Ouahidi et al., 2025], and ST-EEGFormer [Yang et al., 2026], using publicly

Table 1: Datasets for pre-training and downstream evaluation (headset configurations in Fig. 4).

Category	Datasets	Channel	Length	#Samples	Task
In-Domain (Pre-training)	DREAMER	14	2s	170,246	Emotion Detection
	STEW	14	2s	56,928	Mental Workload Classification
	TUAB	19	10s	409,455	Abnormal EEG Classification
	TUEV_binary	19	2s	98,684	Epileptiform Classification
Cross-Domain (Downstream)	Crowdsourced	14	2s	12,296	Eyes Open/Close Detection
	BCICIV_2A	22	4s	5,184	Motor Imagery
	PhysioNetMI	64	2s	19,022	Motor Imagery
	SSVEP	8	2s	960	Steady-State Visual Evoked Potentials

available code and official implementations to ensure fair comparisons. All models use identical preprocessing pipelines, windowing strategies, and sampling rates. We report **Accuracy** (balanced), defined as the average recall across classes (*Acc* for binary and *B-Acc* for multi-class tasks). For binary classification, we also report **AUROC** (Area Under the ROC Curve) to summarize performance across thresholds. For multi-class tasks, we use the **Weighted F1** score, defined as the class-weighted average of F1 scores proportional to sample counts (see Appendix C.2 for detailed experimental settings).

Table 2: Average classification accuracy over three runs in the in-domain setting across four EEG tasks

Model	STEW		DREAMER		TUAB		TUEV_binary	
	Acc	AUROC	Acc	AUROC	Acc	AUROC	Acc	AUROC
MAEEG	75.98 ±2.87	81.73 ±2.95	52.36 ±3.12	54.45 ±2.98	77.56 ±1.45	86.56 ±1.67	85.75 ±2.03	91.36 ±1.89
BioT	75.79 ±1.12	86.48 ±1.45	52.85 ±1.34	54.19 ±1.23	79.21 ±1.56	87.42 ±1.78	86.28 ±1.43	93.06 ±1.67
LaBraM	75.26 ±1.35	84.37 ±1.67	55.85 ±1.45	55.34 ±1.23	80.61 ±1.54	88.77 ±1.76	85.43 ±1.48	90.10 ±1.59
EEGPT	77.65 ±2.95	85.98 ±3.01	54.11 ±3.12	54.91 ±2.87	78.91 ±1.49	87.16 ±1.85	84.05 ±2.03	90.26 ±1.78
EEG2Rep	76.71 ±1.23	85.89 ±1.45	57.09 ±1.32	56.35 ±1.21	80.21 ±1.54	88.43 ±1.67	89.83 ±1.38	95.57 ±1.47
CBraMod	75.89 ±1.43	84.89 ±1.37	56.09 ±1.09	55.56 ±1.21	80.05 ±1.40	88.63 ±1.70	87.35 ±1.42	94.71 ±1.25
REVE	79.95 ±1.28	86.72 ±1.41	56.84 ±1.17	56.14 ±1.24	80.48 ±1.45	88.82 ±1.62	90.21 ±1.36	95.83 ±1.31
ST-EEGFormer	80.02 ±1.36	88.11 ±1.42	57.21 ±1.25	56.58 ±1.20	79.88 ±1.39	88.02 ±1.55	88.05 ±1.41	94.63 ±1.28
EEG-X	**80.32** ±1.21	**88.37** ±1.43	**58.14** ±1.35	**57.64** ±1.29	**81.02** ±1.48	**88.97** ±1.62	**91.15** ±1.41	**96.18** ±1.39

In-domain evaluation: Table 2 reports average classification accuracy over three runs compared with other state-of-the-art methods. **Bold** indicates the highest performance for each dataset, while underlined values denote the second-best results, consistent across all tables in this paper. In the in-domain evaluation, each model is pre-trained and fine-tuned on the same dataset, ensuring evaluation within a consistent recording setup for each task.

Overall, EEG-X achieves the best average performance across all EEG tasks. Recent methods such as REVE and ST-EEGFormer also demonstrate strong performance across multiple datasets, frequently achieving second-best results. This highlights the effectiveness of spatially-aware channel representations and reconstruction-based pretraining for EEG foundation models. In particular, latent-space reconstruction methods such as EEG2Rep generally outperform direct raw-space reconstruction approaches like MAEEG and EEGPT on noisy datasets such as TUAB and TUEV_binary, suggesting the importance of noise-robust representation learning. EEG-X further improves upon these methods by reconstructing artifact-cleaned EEG signals together with the proposed DiCT-enhanced reconstruction objective, enabling the model to learn noise-robust representations while preserving informative temporal and frequency-aware structure from the input signals.

Cross-domain evaluation as a foundation model: Table 3 presents the average classification accuracy of EEG-X, highlighting its role as a *foundation model* for EEG representation learning. We compare EEG-X against recent transformer-based EEG representation learning models that employ channel-wise tokenization and are therefore capable of handling cross-domain scenarios with varying channel configurations between pre-training and downstream tasks (note that MAEEG and EEG2Rep are excluded as they rely on convolutional tokenizers). In this setup, all models are pre-trained solely

Table 3: Cross-domain performance comparison of models pre-trained on TUAB dataset.

Model	Crowdsourced		BCICIV_2A		SSVEP		PhysioNetMI	
	Acc	AUROC	B-Acc	Weighted F1	B-Acc	Weighted F1	B-Acc	Weighted F1
BioT	77.76 ±1.12	83.62 ±1.34	27.26 ±1.05	25.26 ±1.23	29.92 ±1.41	27.27 ±1.36	48.51 ±2.70	59.03 ±2.44
LaBraM	77.28 ±1.25	89.83 ±1.47	32.44 ±1.13	27.37 ±1.21	29.69 ±1.28	27.41 ±1.19	47.41 ±1.60	57.92 ±1.48
EEGPT	79.55 ±1.36	87.79 ±1.52	30.07 ±1.17	24.68 ±1.08	26.06 ±1.11	27.40 ±1.07	46.96 ±2.10	57.37 ±1.95
CBraMod	78.41 ±1.31	88.73 ±1.45	30.07 ±1.12	24.68 ±1.09	28.24 ±1.34	27.79 ±1.25	47.88 ±3.10	58.26 ±2.87
REVE	83.92 ±1.29	90.72 ±1.38	35.91 ±1.11	33.88 ±1.16	31.18 ±1.35	27.94 ±1.27	50.33 ±1.43	60.88 ±1.31
ST-EEGFormer	81.47 ±1.27	91.42 ±1.41	34.16 ±1.09	31.74 ±1.12	30.65 ±1.31	27.36 ±1.24	50.61 ±1.35	61.26 ±1.29
EEG-X	**86.89** ±1.28	**93.51** ±1.41	**37.58** ±1.19	**35.32** ±1.15	**32.97** ±1.42	**28.15** ±1.33	**51.89** ±1.20	**62.31** ±1.08

on the TUAB dataset, while the pre-trained tokenizer and encoder are frozen and only a lightweight linear classifier is trained on top of the learned representations.

As shown in Table 3, EEG-X consistently outperforms existing methods across all downstream tasks. Interestingly, recent methods such as REVE, which also incorporate spatially-aware channel embeddings, demonstrate stronger cross-domain performance compared to earlier approaches based purely on learnable channel embeddings. This observation further highlights the importance of incorporating electrode location information when training foundation models on heterogeneous EEG datasets. The TUAB pre-training dataset contains 19 channel configurations, while downstream tasks involve 14-, 22-, 64-, and 8-channel EEG setups with limited channel overlap. By leveraging location-based channel embeddings, EEG-X preserves spatial relationships among neighboring electrodes and enables robust generalization across different devices, channel counts, and recording configurations.

3.3 Ablation study

Channel Embedding: We evaluate the impact of channel embeddings on model performance in both in-domain and cross-domain scenarios by comparing three configurations: without channel embedding, with learnable channel embedding, and our proposed location-based channel embedding in EEG-X. As shown in Table 4, incorporating channel embeddings improves the average performance compared with removing them in both settings. In the in-domain setting, learnable channel embedding improves the average score from 76.50 to 77.07, while the full EEG-X model achieves 77.66. The improvement is more pronounced in the cross-domain setting, where the average increases from 47.46 without channel embedding to 47.53 with learnable channel embedding and 52.33 with the full EEG-X model. These results highlight the importance of incorporating spatial information into channel representations. In particular, our location-based channel embedding provides additional benefits in cross-domain scenarios, potentially helping the model preserve spatial relationships among neighboring electrodes and generalize across datasets with different channel configurations.

DiCT: We further investigate the contribution of the Dictionary Convolution Transformation (DiCT) by comparing the full EEG-X model against a variant without DiCT. Removing DiCT reduces the average performance from 77.66 to 74.54 in the in-domain setting and from 52.33 to 49.34 in the cross-domain setting. The performance degradation is observed across all four datasets in both settings. These results demonstrate the importance of DiCT in learning effective EEG representations. By incorporating dictionary-based convolutional transformations into the reconstruction objective, DiCT helps the model capture meaningful temporal and spectral patterns, contributing to more discriminative representations and improved generalization across recording domains.

Artifact Removal Raw Space: We examine the impact of artifact-aware reconstruction by comparing reconstruction of the original raw EEG signals against reconstruction of the rASR-cleaned targets used in EEG-X. Specifically, we compare the full EEG-X model, which reconstructs artifact-cleaned signals, with a variant that uses the same architecture and DiCT objective but reconstructs the original raw signals ($\mathcal{L}_{\text{rec (Raw)}}$). As shown in Table 4, raw-space reconstruction consistently underperforms the full EEG-X model across all four in-domain and cross-domain datasets. The average performance decreases from 77.66 to 75.24 in the in-domain setting and from 52.33 to 50.20 in the cross-domain setting. These results suggest that using artifact-cleaned EEG signals as reconstruction targets helps the model learn more robust and meaningful representations by reducing the influence of artifacts and non-neural noise during self-supervised training.

Table 4: Ablation study of EEG-X components. The average column reports the mean score across datasets, with the difference from full EEG-X in parentheses.

In-Domain					
Component	**STEW**	**DREAMER**	**TUAB**	**TUEV_binary**	**Average**
Without channel embedding	79.44	56.17	79.94	90.46	76.50 (↓ 1.16)
Learnable channel embedding	80.07	57.12	80.94	90.14	77.07 (↓ 0.59)
Without DiCT	76.91	54.36	78.88	88.01	74.54 (↓ 3.12)
Raw reconstruction	77.90	55.41	79.41	88.22	75.24 (↓ 2.42)
Full EEG-X	**80.32**	**58.14**	**81.02**	**91.15**	**77.66**
Cross-Domain					
Component	**Crowdsourced**	**BCICIV_2A**	**SSVEP**	**PhysioNetMI**	**Average**
Without channel embedding	81.64	33.55	26.82	47.82	47.46 (↓ 4.88)
Learnable channel embedding	81.65	33.25	27.00	48.21	47.53 (↓ 4.81)
Without DiCT	84.12	34.57	29.35	49.33	49.34 (↓ 2.99)
Raw reconstruction	85.34	35.31	30.02	50.12	50.20 (↓ 2.14)
Full EEG-X	**86.89**	**37.58**	**32.97**	**51.89**	**52.33**

rASR Preprocessing vs. Noise-Aware Reconstruction: A key question is: Why not use artifact removal (rASR) directly as a preprocessing step rather than incorporating it into the self-supervised learning process? To investigate this, we compare three settings, as shown in Table 5: supervised EEG-X trained for 100 epochs using raw inputs, supervised EEG-X trained for 100 epochs using rASR-preprocessed inputs, and the self-supervised EEG-X model trained using raw inputs with rASR-cleaned reconstruction targets. The results show that applying rASR as a preprocessing step does not consistently improve downstream performance: it improves results on DREAMER and TUAB, but reduces performance on STEW and TUEV. In contrast, the self-supervised setting achieves the best performance across all four datasets. These findings suggest that directly applying artifact removal as a preprocessing step may not always preserve the signal information most useful for downstream tasks. Incorporating rASR into the self-supervised reconstruction objective instead allows the model to learn from raw inputs while using artifact-cleaned signals as reconstruction targets.

Table 5: Performance of EEG-X under different data and training settings. Supervised models are evaluated using raw or rASR-preprocessed inputs, while the self-supervised model is pretrained using raw inputs with rASR-cleaned reconstruction targets.

Setting	**STEW**	**DREAMER**	**TUAB**	**TUEV_binary**
Supervised (Raw input)	75.12 $\pm$2.64	55.69 $\pm$1.34	78.12 $\pm$0.98	88.24 $\pm$2.09
Supervised (rASR input)	73.68 $\pm$1.74	56.41 $\pm$1.04	78.32 $\pm$1.11	83.36 $\pm$1.68
Self-Supervised (Raw input with rASR target)	**80.32** $\pm$**1.21**	**58.14** $\pm$**1.35**	**81.02** $\pm$**1.48**	**91.15** $\pm$**1.41**

4 Conclusion

In this paper, we introduced *EEG-X*, a device-agnostic and noise-robust model for EEG representation learning. EEG-X addresses two major challenges in EEG analysis: variability across recording devices and the inherently low signal-to-noise ratio of EEG signals. To tackle these challenges, EEG-X combines location-based channel embeddings for robust cross-device generalization with a noise-aware reconstruction strategy that reconstructs artifact-cleaned EEG signals instead of noisy raw inputs. We further introduced the Dictionary-based Convolution Transformation (DiCT), a lightweight structured reconstruction objective that improves robustness to amplitude variation and temporal misalignment while capturing multi-scale temporal patterns. Extensive experiments across diverse EEG datasets demonstrate that EEG-X consistently outperforms state-of-the-art methods in both in-domain and cross-domain settings, establishing a strong foundation model for EEG representation learning.

References

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *Proceedings of the 39th International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022.

Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In *Proceedings of the 40th International Conference on Machine Learning*, pages 1416–1429. PMLR, 2023.

Sarah Blum, Nadine S. J. Jacobsen, Martin G. Bleichner, and Stefan Debener. A riemannian modification of artifact subspace reconstruction for EEG artifact handling. *Frontiers in Human Neuroscience*, 13:141, 2019.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Poomipat Boonyakitanont, Apiwat Lek-Uthai, Krisnachai Chomtho, and Jitkomut Songsiri. A review of feature extraction and performance evaluation in epileptic seizure detection using EEG. *Biomedical Signal Processing and Control*, 57:101702, 2020.

Clemens Brunner, Robert Leeb, Gernot Müller-Putz, Alois Schlögl, and Gert Pfurtscheller. BCI competition 2008–graz data set a. Institute for Knowledge Discovery, Graz University of Technology, 2008.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. VAST: A vision-audio-subtitle-text omni-modality foundation model and dataset. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Hsiang-Yun Sherry Chien, Hanlin Goh, Christopher M. Sandino, and Joseph Y. Cheng. MAEEG: Masked auto-encoder for EEG representation learning. In *NeurIPS Workshop*, 2022.

Alexander Craik, Yongtian He, and Jose L. Contreras-Vidal. Deep learning for electroencephalogram (EEG) classification tasks: A review. *Journal of Neural Engineering*, 16(3):031001, 2019.

Angus Dempster, François Petitjean, and Geoffrey I. Webb. ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34(5):1454–1495, 2020.

Angus Dempster, Daniel F. Schmidt, and Geoffrey I. Webb. Hydra: Competing convolutional kernels for fast and accurate time series classification. *Data Mining and Knowledge Discovery*, 37(5): 1779–1805, 2023.

Angus Dempster, Navid Mohammadi Foumani, Chang Wei Tan, Lynn Miller, Amish Mishra, Mahsa Salehi, Charlotte Pelletier, Daniel F. Schmidt, and Geoffrey I. Webb. MONSTER: Monash scalable time series evaluation repository. *Journal of Data-centric Machine Learning Research*, 2025. URL `https://openreview.net/attachment?id=XauSqSfZfc&name=pdf`.

Navid Mohammadi Foumani, Chang Wei Tan, Geoffrey I. Webb, Hamid Rezatofighi, and Mahsa Salehi. Series2Vec: Similarity-based self-supervised representation learning for time series classification. *Data Mining and Knowledge Discovery*, 38(4):2520–2544, 2024a.

Navid Mohammadi Foumani, Chang Wei Tan, Geoffrey I. Webb, and Mahsa Salehi. Improving position encoding of transformers for multivariate time series classification. *Data Mining and Knowledge Discovery*, 38(1):22–48, 2024b.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

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

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

Mohammad-Parsa Hosseini, Amin Hosseini, and Kiarash Ahi. A review on machine learning for EEG signal processing in bioengineering. *IEEE Reviews in Biomedical Engineering*, 14:204–218, 2021.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. *arXiv preprint arXiv:2405.18765*, 2024.

Xiao Jiang, Gui-Bin Bian, and Zean Tian. Removal of artifacts from EEG signals: A review. *Sensors*, 19(5):987, 2019.

Emmanuel K. Kalunga, Sylvain Chevallier, Quentin Barthélemy, Karim Djouani, Eric Monacelli, and Yskandar Hamam. Online SSVEP-based BCI using riemannian geometry. *Neurocomputing*, 191: 55–68, 2016.

Stamos Katsigiannis and Naeem Ramzan. DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE Journal of Biomedical and Health Informatics*, 22(1):98–107, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Demetres Kostas, Stéphane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15, 2021.

W. L. Lim, O. Sourina, and Lipo P. Wang. STEW: Simultaneous task EEG workload data set. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(11):2106–2114, 2018.

S. Lopez, G. Suarez, D. Jungreis, I. Obeid, and J. Picone. Automated identification of abnormal adult EEGs. In *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–5. IEEE, 2015.

Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update. *Journal of Neural Engineering*, 15(3):031005, 2018. doi: 10.1088/1741-2552/aab2f2.

Navid Mohammadi Foumani, Geoffrey Mackellar, Soheila Ghane, Saad Irtza, Nam Nguyen, and Mahsa Salehi. EEG2Rep: Enhancing self-supervised EEG representation through informative masked inputs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5544–5555, 2024a.

Navid Mohammadi Foumani, Lynn Miller, Chang Wei Tan, Geoffrey I. Webb, Germain Forestier, and Mahsa Salehi. Deep learning for time series classification and extrinsic regression: A current survey. *ACM Computing Surveys*, 56(9):1–45, 2024b.

Gernot R. Müller-Putz. Electroencephalography. In *Handbook of Clinical Neurology*, volume 168, pages 249–262. Elsevier, 2020.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG–adapting to any setup with large-scale pretraining on 25,000 subjects. *arXiv preprint arXiv:2510.21585*, 2025.

Luca Pion-Tonachini, Ken Kreutz-Delgado, and Scott Makeig. ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, 198:181–197, 2019.

Mladen Raković, Yuheng Li, Navid Mohammadi Foumani, Mahsa Salehi, Levin Kuhlmann, Geoffrey Mackellar, Roberto Martinez-Maldonado, Gholamreza Haffari, Zachari Swiecki, Xinyu Li, et al. Measuring affective and motivational states as conditions for cognitive and metacognitive processing in self-regulated learning. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, pages 701–712, 2024.

Amarpal Sahota, Navid Mohammadi Foumani, Raul Santos-Rodriguez, and Zahraa S. Abdallah. Knoweeg: Explainable knowledge driven eeg classification. *Neurocomputing*, 710:135337, 2027. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2026.135337. URL `https://www.sciencedirect.com/science/article/pii/S0925231226027359`.

Michal Teplan. Fundamentals of EEG measurement. *Measurement Science Review*, 2(2):1–11, 2002.

Patrick A. Thompson. An MSE statistic for comparing forecast accuracy across series. *International Journal of Forecasting*, 6(2):219–227, 1990.

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*, volume 30, 2017.

Zhijiang Wan, Manyu Li, Shichang Liu, Jiajin Huang, Hai Tan, and Wenfeng Duan. EEGFormer: A transformer-based brain activity classification method using EEG signal. *Frontiers in Neuroscience*, 17:1148855, 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*, volume 37, 2024. URL `https://openreview.net/forum?id=lvS2b8CjG5`.

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

Weining Weng, Yang Gu, Shuai Guo, Yuan Ma, Zhaohua Yang, Yuchen Liu, and Yiqiang Chen. Self-supervised learning for electroencephalogram: A systematic survey. *arXiv preprint arXiv:2401.05446*, 2024.

Chaoqi Yang, M. Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36, 2023.

Liuyin Yang, Qiang Sun, Ang Li, and Marc M. Van Hulle. Are EEG foundation models worth it? comparative evaluation with traditional decoders in diverse BCI tasks. In *The Fourteenth International Conference on Learning Representations*, 2026.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

A Related Work

A line of work tackles the variability in EEG data caused by differences in recording devices and configurations, particularly variations in channel layouts, electrode positions, and recording lengths. BIOT [Yang et al., 2023] addresses this challenge by tokenizing biosignals into a "sentence-like" structure, segmenting EEG time series in each channel into fixed-length tokens and adding temporal and channel embeddings. Similarly, EEGFormer [Wan et al., 2023], LaBraM [Jiang et al., 2024], and EEGPT [Wang et al., 2024] tokenize EEG signals on a channel-wise basis and process them using transformer-based architectures. LaBraM uses learnable embeddings for both channel and temporal positions, while EEGPT incorporates rotary position encodings for temporal representation alongside learnable channel embeddings. More recently, REVE [Ouahidi et al., 2025] introduces a 4D Fourier-based positional encoding together with learnable channel embeddings to handle varying electrode layouts and temporal lengths through coupled spatial-temporal frequency interactions, making it the closest work to our location-based channel embedding approach. Despite these advances, most existing methods either rely on learnable channel embeddings tied to specific electrode identities or require fixed spatial encoding strategies that may not generalize efficiently across heterogeneous EEG devices. In particular, learnable channel embeddings often lack explicit spatial awareness, as they do not directly reflect electrode placement on the scalp—critical for modeling meaningful EEG patterns. Moreover, the same electrode may not exist across different devices, making such embeddings impractical for handling channel mismatches and mixed-device training.

A parallel line of work has explored self-prediction approaches, such as Masked Autoencoders [He et al., 2022], to adapt pre-trained models for EEG [Chien et al., 2022, Kostas et al., 2021, Wang et al., 2024, Mohammadi Foumani et al., 2024a, Wang et al., 2025, Ouahidi et al., 2025, Yang et al., 2026]. BENDR [Kostas et al., 2021], MAEEG [Chien et al., 2022], CBraMod [Wang et al., 2025], REVE [Ouahidi et al., 2025], and ST-EEGFormer [Yang et al., 2026] use masking and reconstruction as pretext tasks for self-supervised learning, directly reconstructing masked segments in the raw signal space. However, raw-space reconstruction is highly sensitive to noise, often leading models to preserve artifact-related patterns together with neural activity. Alternatively, LaBraM [Jiang et al., 2024] uses a neural tokenizer that transforms raw EEG signals into vocabulary-based tokens, predicting tokens instead of reconstructing raw signals. However, since these approaches still rely on raw EEG representations, they remain affected by the inherently low signal-to-noise ratio of EEG, where neural activity is often buried under external artifacts and non-brain sources. EEG2Rep [Mohammadi Foumani et al., 2024a] improves robustness against noise by introducing latent-space reconstruction instead of raw-space reconstruction, reducing sensitivity to noise and encouraging semantically rich representations through high-level feature prediction. However, it remains limited in handling datasets with varying channel configurations and recording lengths.

B Dictionary Convolution Transformation (DiCT)

B.1 Motivation

Reconstruction losses such as Mean Squared Error (MSE) are widely used in self-supervised learning and perform well in high signal-to-noise ratio (SNR) domains such as vision, language, and speech. However, they are not well aligned with the characteristics of EEG time series.

EEG signals are inherently *multi-scale*, meaning that meaningful information is distributed across different temporal resolutions or frequency bands. These bands exhibit varying amplitudes and dynamics, and their relative importance depends on the underlying neural processes and recording conditions.

EEG recordings are also highly susceptible to artifacts, which often have large amplitudes and can occur across a wide range of frequencies. For example, eye blinks and movement artifacts typically dominate the low-frequency range, while muscle activity and line noise introduce strong high-frequency components. As a result, both low- and high-frequency regions of the signal can be dominated by non-neural, high-amplitude noise.

Because MSE measures point-wise differences, it is disproportionately influenced by large-amplitude components—many of which in EEG arise from artifacts—causing the model to prioritize fitting noise over meaningful neural structure.

This leads to three key limitations:

- **Amplitude bias:** MSE penalizes errors proportionally to their magnitude, causing high-amplitude components—often dominated by artifacts—to drive the loss. As a result, models are encouraged to reconstruct noise or slow dominant fluctuations while ignoring informative neural activity.
- **Lack of frequency awareness:** MSE treats all point-wise deviations equally, regardless of their frequency origin. For example, failing to reconstruct a high-frequency neural oscillation may contribute little to the loss compared to a small error in a large low-frequency artifact.
- **Lack of shape awareness:** MSE compares signals point by point, making it highly sensitive to temporal misalignment. Even small shifts or phase differences can result in large penalties, despite the signals having similar structure and frequency content.

These limitations are particularly problematic for EEG, where meaningful neural activity is distributed across frequency bands, and artifacts can dominate the signal across both low- and high-frequency ranges. To address these limitations, we introduce the Dictionary Convolution Transformation (DiCT).

B.2 Demonstrating the Limitations of Time-Domain MSE

To illustrate the limitations of direct time-domain MSE and motivate DiCT, we construct a controlled example using real EEG frequency bands.

Setup. Given a signal $\mathbf{X}$, we decompose it into standard EEG frequency bands using band-pass filters:

$$\mathbf{X} = \sum_{b \in \mathcal{B}} \mathbf{X}^{(b)}, \tag{8}$$

where $\mathcal{B} = \{\text{Delta}, \text{Theta}, \text{Alpha}, \text{Beta}, \text{Gamma}\}$ and $\mathbf{X}^{(b)}$ denotes the component of $\mathbf{X}$ in band b.

We define the bands as: Delta (1–4 Hz), Theta (4–8 Hz), Alpha (8–14 Hz), Beta (14–30 Hz), and Gamma (30–47 Hz).

Reconstructed signals. We construct three target signals:

(1) Low-frequency dominant (Delta + Theta):

$$\hat{\mathbf{X}}_{\text{low}} = \mathbf{X}^{(\text{Delta})} + \mathbf{X}^{(\text{Theta})}. \tag{9}$$

(2) Mid/High-frequency dominant (Theta + Alpha + Beta + Gamma):

$$\hat{\mathbf{X}}_{\text{high}} = \mathbf{X}^{(\text{Theta})} + \mathbf{X}^{(\text{Alpha})} + \mathbf{X}^{(\text{Beta})} + \mathbf{X}^{(\text{Gamma})}. \tag{10}$$

(3) Phase-shifted signal:

$$\hat{\mathbf{X}}_{\text{shift}}(t) = \mathbf{X}(t - \Delta t), \tag{11}$$

where Δt is a small temporal shift.

Note: Theta is included in both reconstructions to ensure that differences arise from the relative dominance of low- versus higher-frequency components, rather than complete band removal.

Fig. 5 visually illustrates the original signal and the three reconstructed variants.

Results. Table 6 reports reconstruction errors before and after applying DiCT.

Target	MSE (Time)	MSE (DiCT)
Low-frequency dominant	47.55	11.44
Mid/High-frequency dominant	95.75	4.26
Phase-shifted signal	126.79	0.02

Table 6: Comparison of reconstruction error before and after DiCT transformation.

Analysis. These results reveal a fundamental mismatch between time-domain MSE and the structural properties of EEG signals.

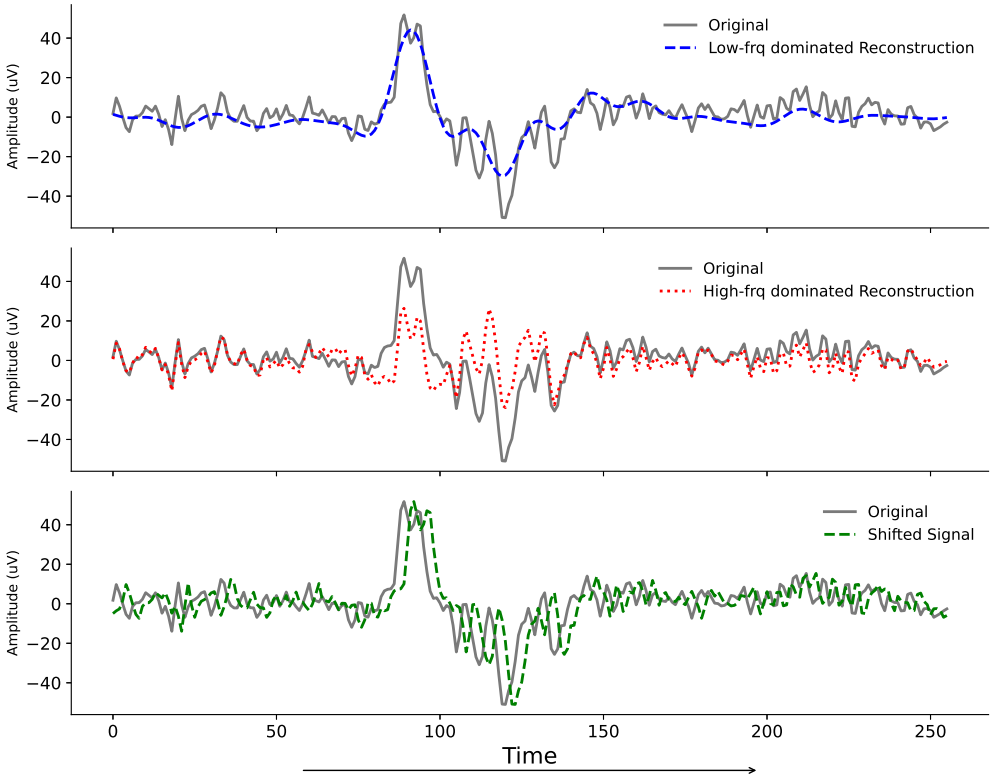


Figure 5: Original EEG signal (top) and reconstructed variants: (1) low-frequency dominant (Delta + Theta), (2) mid/high-frequency dominant (Theta + Alpha + Beta + Gamma), and (3) phase-shifted signal. These examples highlight amplitude bias and phase sensitivity in time-domain MSE.

(i) Amplitude bias. Low-frequency components dominate the signal energy in EEG. Consequently, MSE is largely driven by these high-amplitude components, allowing reconstructions that discard higher-frequency structure to achieve relatively low error.

(ii) Frequency insensitivity. The mid/high-frequency dominant reconstruction preserves rich temporal dynamics but differs in amplitude distribution. MSE penalizes it heavily, indicating that it fails to reflect the presence of informative spectral content.

(iii) Phase sensitivity. A small temporal shift results in a large MSE (126.79), despite the signal being structurally identical. This highlights that MSE is sensitive to alignment rather than underlying pattern similarity.

Importantly, MSE can assign lower error to structurally incomplete reconstructions than to those preserving richer signal content.

Why DiCT works better. In contrast, DiCT significantly reduces these discrepancies. By projecting signals into a convolutional feature space, it compares the occurrence of temporal patterns rather than enforcing point-wise alignment. As a result, it (i) mitigates amplitude bias, (ii) captures multi-scale frequency content, and (iii) remains robust to temporal shifts.

Notably, the near-zero error for the shifted signal (0.02) demonstrates that DiCT preserves structural similarity under temporal misalignment—an essential property for EEG, where phase variability across trials and subjects is common.

B.3 Parameter Selection

The key hyperparameters in DiCT are the number of groups (G) and the number of kernels per group (K), which together define the structure of the convolutional dictionary. These parameters control the trade-off between pattern diversity and the strength of competition among patterns.

Given a fixed computational budget, increasing G increases the number of independent groups, encouraging more localized and shape-preserving pattern matching. In contrast, increasing K increases the number of candidate patterns within each group, improving coverage of temporal and frequency variations, but reducing the relative competition between patterns.

Empirical evaluation. We evaluate several configurations of (G, K) on the STEW and DREAMER datasets, as summarized in Table 7.

Groups (G)	**Kernels (K)**	**STEW Acc (%)**	**DREAMER Acc (%)**
4	16	80.15	58.00
8	32	**80.32**	**58.14**
64	16	80.28	58.10
8	64	80.30	58.12

Table 7: Effect of the number of groups and kernels per group on downstream performance.

Analysis. The results highlight a clear trade-off between G and K:

- **Increasing G (more groups):** promotes more structured representations by enforcing stronger competition within smaller groups, leading to improved shape sensitivity.
- **Increasing K (more kernels per group):** increases pattern diversity and frequency coverage, but reduces the relative distinctiveness of patterns and increases computational cost.

We observe that extreme configurations (very large G or K) do not yield significant improvements, indicating diminishing returns once sufficient pattern diversity is achieved. Instead, performance is maximized when balancing these two effects.

As shown in Table 7, the performance is relatively stable across different configurations of (G) and (K). In particular, configurations such as $(G = 8, K = 32)$, $(G = 64, K = 16)$, and $(G = 8, K = 64)$ achieve very similar results on STEW and DREAMER. We therefore select $(G = 8, K = 32)$ as a practical trade-off between model configuration complexity and downstream performance, rather than as a definitive optimum.

Kernel and dilation selection. To better capture EEG temporal dynamics across multiple frequency bands, DiCT employs convolutional kernels with sizes 9 and 15 together with dilation scales of $\{1, 2, 4, 8, 16\}$. These configurations are selected to model temporal patterns spanning approximately the 1–50 Hz EEG frequency range, covering Delta, Theta, Alpha, Beta, and Gamma bands under the sampling rates used in our experiments. Smaller kernels and dilations focus on high-frequency local oscillations, while larger receptive fields capture slower temporal dynamics and long-range structures.

B.4 Differentiable Soft Assignment for DiCT

In Section 2.3, we introduce DiCT using hard winner assignments, where the kernel with the strongest response is selected at each time step. While this formulation provides an intuitive explanation of the dictionary-based transformation, the argmax operation is non-differentiable and prevents gradients from flowing through the assignment operation. To enable end-to-end optimization, we use a differentiable softmax relaxation during training.

Let $R^{(d)}_{g,k,t}$ denote the convolutional response for group g, kernel k, time step t, and dilation scale d. The soft assignment probability is defined as:

$$q^{(d)}_{g,k,t} = \frac{\exp\left(R^{(d)}_{g,k,t}/\tau\right)}{\sum_{j=1}^{K} \exp\left(R^{(d)}_{g,j,t}/\tau\right)}, \tag{12}$$

where τ is the temperature parameter controlling the sharpness of the assignments. A lower temperature produces more concentrated assignments, whereas a higher temperature distributes probability more evenly across competing kernels. In our experiments, we use $\tau = 1.0$.

The soft feature count for each kernel is then obtained by summing its assignment probabilities over time:

$$f_{g,k}^{(d)} = \sum_{t=1}^{T} q_{g,k,t}^{(d)}. \tag{13}$$

Unlike hard assignments, which assign each time step exclusively to a single kernel, soft assignments distribute the contribution across competing kernels according to their relative responses. The resulting feature representation is constructed by concatenating the soft counts across EEG channels, kernel groups, kernel lengths, and dilation scales.

We apply the same fixed convolutional dictionary independently to the reconstructed EEG signal and its corresponding target signal. The dictionary kernels are shared across channels and remain fixed throughout training. Let $\mathbf{F}(\cdot)$ denote the resulting soft feature transformation. The target features are computed without gradient propagation, while gradients are retained through the reconstructed signal. The DiCT reconstruction loss is defined as:

$$\mathcal{L}_{\mathrm{DiCT}} = \mathrm{MSE}\left(\mathbf{F}(\hat{\mathbf{X}}), \mathrm{stopgrad}\left(\mathbf{F}(\mathbf{X}_{\mathrm{clean}})\right)\right). \tag{14}$$

Here, $\hat{\mathbf{X}}$ denotes the reconstructed EEG signal and $\mathbf{X}_{\mathrm{clean}}$ denotes the corresponding processed target signal. The mean squared error is computed over the batch and all concatenated channel and dictionary features. The temporal feature counts are summed without normalization by the window length.

B.5 Worked example of DiCT

Fig. 6 walks through DiCT on a 1-second EEG segment (128 time steps) and two generated targets: one dominated by high+mid frequencies and one dominated by low frequencies. Fig. 6 (a) shows the three signals. The low-frequency target follows the large slow deflections of the input but lacks its fast fluctuations, whereas the high-frequency target reproduces the fast activity but with a distorted large-scale shape. Fig. 6 (b) shows DiCT applied to the input EEG signal. Two groups of three fixed kernels are used: g_1 contains edge-like kernels (rising ramp, falling ramp, step), and g_2 contains oscillatory kernels (a medium wave and a fast sine/cosine pair). Each group is applied at dilations $d \in \{1, 2, 4\}$, so the same kernel shape covers an increasingly long time span. At every time step the kernel with the strongest response wins, and the number of time steps won by each kernel forms the count $f_{g,k}^{(d)}$. Fig. 6 (c) concatenates these counts into the 18-value representation $\mathcal{F}(\cdot)$ of the input and of each target; each signal is standardized to zero mean and unit variance before the transform. Fig. 6 (d) compares the two error measures. The raw signal MSE (point-wise) prefers the low-frequency target (62.9 versus 185.6) because it matches the large-amplitude deflections that dominate the MSE. The DiCT representation MSE, the mean squared difference between the two count vectors and that of the input, prefers the high+mid frequency target (26.6 versus 163.6), because it preserves the pattern content that decides which kernels win.

For clarity, this example uses hand-picked kernels and hard winner counts; the model uses random fixed kernels with the softmax relaxation (temperature τ). The kernels were chosen to illustrate the behavior on this segment and are not evidence of the same ordering on all data.

B.6 Universal Location Mapping

Fig. 7 shows a dense scalp mesh including 348 channels, covering virtually all electrode positions from the 10-05, 10-10, and 10-20 standard EEG systems. This mesh enables universal location-based channel embeddings: even if electrode names differ across devices, each sensor can be mapped to the nearest mesh point. Since typical EEG sensors are larger than 5 mm in diameter, any sensor smaller than 10 mm placed anywhere on the scalp will intersect with at least one mesh point, ensuring consistent spatial representation across devices and subjects.

B.7 Artifact Removal Details

We adopt Riemannian Artifact Subspace Reconstruction (rASR) [Blum et al., 2019] for artifact removal due to its robustness and suitability for large-scale EEG processing. Instead of explicitly

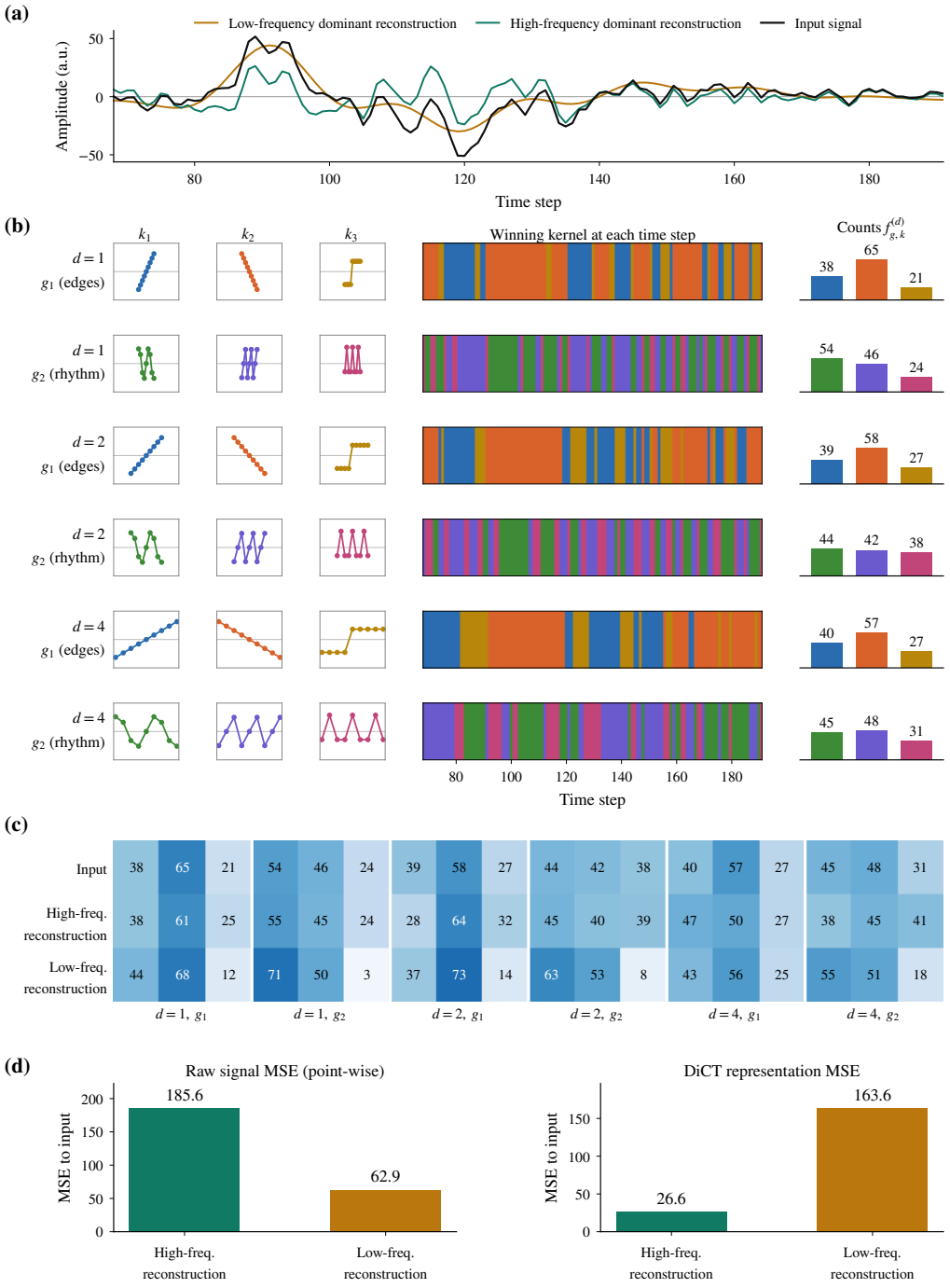


Figure 6: Real example of DiCT on an EEG segment. **(a)** Real EEG signal and two reconstructions; **(b)** DiCT on the input EEG. **(c)** The representation of the input and each reconstruction; **(d)** The raw signal MSE (point-wise) is lower for the low-frequency reconstruction, whereas the DiCT representation MSE is lower for the high-frequency one, which preserves the input's fast pattern content.

separating sources, rASR models clean EEG activity as a low-variance subspace and removes deviations from this subspace.

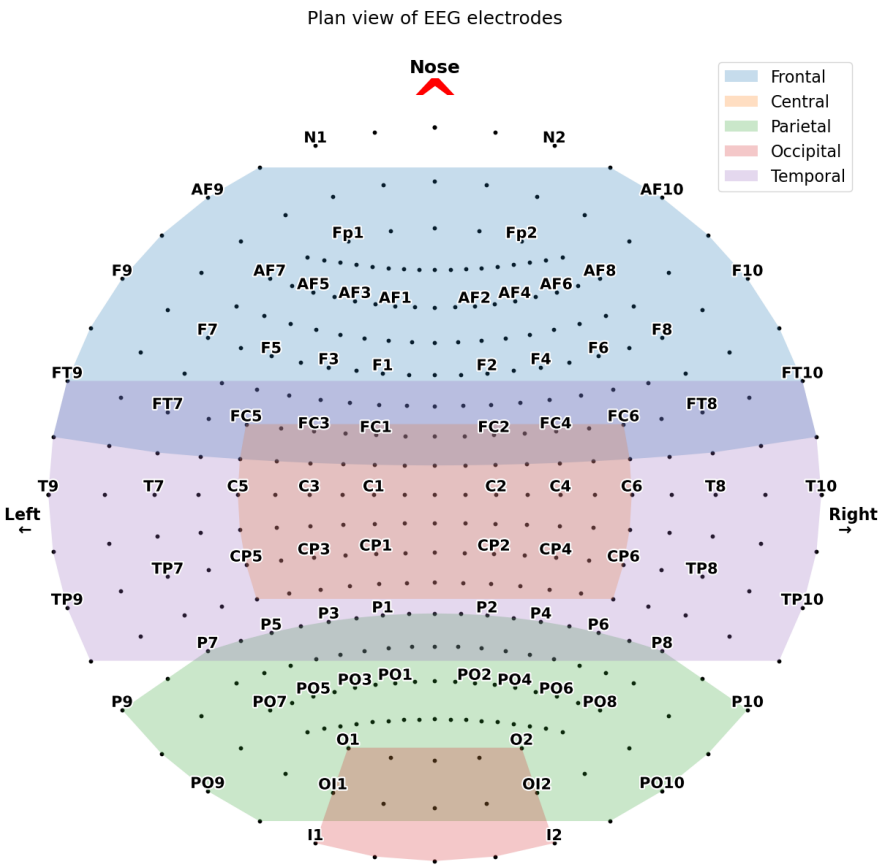


Figure 7: Dense scalp electrode mesh combining positions from the 10-05, 10-10, and 10-20 systems, used as a universal location mapping.

Calibration. rASR requires a short segment of clean EEG data. We obtain this calibration data by selecting low-variance segments using a simple signal quality criterion (e.g., RMS thresholding). Given calibration data $\mathbf{X}_{\text{cal}}$, we estimate a reference covariance matrix:

$$\mathbf{\Sigma}_{\text{ref}} = \frac{1}{L}\mathbf{X}_{\text{cal}}\mathbf{X}_{\text{cal}}^{\top}. \tag{15}$$

Artifact suppression. For each incoming signal window $\mathbf{X}$, we compute its covariance and compare it to the reference covariance. Deviations are measured via whitening, where directions with abnormally large variance are identified as artifacts. These directions are suppressed, and the cleaned signal is reconstructed by projecting back to the original space.

Parameter selection. We tune rASR hyperparameters by injecting synthetic ERP signals into contaminated data and evaluating their recovery after denoising, ensuring a balance between artifact removal and signal preservation. We search over combinations of window length, window overlap, and rejection cutoff, and obtain the best performance with:

$$[2,\ 0.5,\ 7.5] \quad \text{and} \quad [2,\ 0.75,\ 9],$$

where each tuple corresponds to (window length in seconds, overlap ratio, rejection cutoff in standard deviations). Unless otherwise stated, we use the second configuration as default due to its more stable performance across datasets. For scalability, we provide a pre-trained rASR model for data cleaning, avoiding per-dataset calibration in most cases.

Why rASR? Alternative artifact removal approaches include ICA-based decomposition, wavelet denoising, regression-based filtering, and spatial filtering methods such as CSP. ICA-based approaches often require explicit component decomposition and manual or semi-automatic component labeling, making them computationally expensive and less scalable for large heterogeneous EEG corpora.

Table 8: Holm-adjusted p-values for pairwise comparisons between EEG-X and competing methods.

Setting	REVE	ST-EEGFormer	EEG2Rep	CBraMod	LaBraM	EEGPT	BioT	MAEEG
In-domain	0.5324	0.1700	**0.0264**	**0.0056**	**0.0036**	**0.0036**	**0.0036**	**0.0036**

Wavelet and regression-based methods can effectively suppress specific artifact types but typically rely on hand-crafted assumptions or auxiliary channels such as EOG/EMG, which are not consistently available across datasets. Spatial filtering approaches are often task-specific and may not generalize well across diverse EEG paradigms. In contrast, rASR provides a lightweight and scalable artifact removal framework that does not require explicit component labeling while remaining robust across different recording conditions and EEG devices, making it particularly suitable for large-scale foundation model training.

C Additional Experimental Details and Results

C.1 Architecture Details

EEG-X employs an asymmetric masked autoencoder architecture consisting of an 8-layer Transformer encoder and a lightweight 4-layer Transformer decoder. Similar to the standard Transformer architecture [Vaswani et al., 2017], we use eight attention heads in both encoder and decoder blocks to capture diverse temporal patterns from EEG signals. The embedding dimension is set to $d_e = 64$, and the feed-forward network (FFN) expands the hidden dimension by a factor of 4 before projecting it back to the original size. We follow the masking strategy of EEG2Rep [Mohammadi Foumani et al., 2024a] (semantic subsequence-preserving masking), with a masking ratio of 75%. A new mask is drawn at every training step, and the encoder processes only the visible tokens.

For reconstruction learning, EEG-X uses the proposed DiCT objective, which is randomly initialized and contains no learnable parameters. EEG signals are segmented into patches of length 32 (0.25 second) with an overlap of 8 time steps, balancing temporal resolution and computational efficiency.

C.2 Experimental Setup

In our experiments, EEG-X was trained using the Adam optimizer [Kingma and Ba, 2014] with a batch size of 64. The learning rate was initialized at 1×10^{-3} and updated using a cosine learning rate scheduler [Baevski et al., 2022]. The model was pretrained for 100 epochs with early stopping based on validation loss to reduce overfitting. During in-domain fine-tuning, the tokenizer, encoder, and classification head were optimized jointly for 10 epochs using cross-entropy loss. In the cross-domain setting, the pretrained tokenizer and encoder were frozen, and lightweight linear classifier was applied on top of the learned EEG representations.

C.3 Statistical Significance Analysis

We performed pairwise two-sided Wilcoxon signed-rank tests on all matched results, using three runs per dataset and method, followed by Holm correction to account for multiple comparisons.

For the in-domain results in Table 2, EEG-X achieved the highest average rank among all evaluated methods. However, its improvements over REVE and ST-EEGFormer were not statistically significant after Holm correction, with adjusted p-values of 0.5324 and 0.1700, respectively. In contrast, EEG-X significantly outperformed all other baselines, including EEG2Rep, CBraMod, LaBraM, EEGPT, BioT, and MAEEG, with adjusted p-values below 0.05. The detailed adjusted p-values are reported in Table 8.

For the cross-domain results in Table 3, EEG-X achieved the highest average rank and significantly outperformed all competing methods after Holm correction. The adjusted p-value was 0.0021 for all pairwise comparisons. These results demonstrate the effectiveness of EEG-X in transferring learned representations across datasets with different recording configurations and channel layouts.

C.4 Computational Efficiency and Scalability

To study how EEG-X scales, we increase the encoder embedding width from 64 to 512 while keeping the depth (8 layers), architecture and training recipe fixed. Encoder parameters grow from 0.40M to 25.2M. We report compute as nominal *dense-operation FLOP estimates*, which are hardware-independent. They count multiply-adds as 2 FLOPs and include the matrix products in the tokenizer, attention, feed-forward layers, reconstruction head and SoftDiCT convolutions. FFT, bias additions, normalization, activations, softmax, other elementwise operations, optimizer updates and data movement are excluded. Inference FLOPs cover the tokenizer, full encoder and classification head for one 2 s EEG window (14 channels, 256 samples). Pretraining FLOPs cover one epoch (forward and backward) of the masked encoder, Predictor, decoder and reconstruction head, including both SoftDiCT branches. The pretraining batch size is 64 for all configurations. These are FLOP counts, not measured wall-clock time or hardware throughput.

Emb. dim	Enc. params	Total params	Inference GFLOPs / window	Pretrain TFLOPs / epoch	STEW	DREAMER
64	0.40M	0.65M	0.114	44.97	80.32	58.14
128	1.59M	2.59M	0.405	72.70	80.70	58.68
256	6.33M	10.28M	1.515	179.87	80.75	59.12
512	25.2M	41.01M	5.848	601.05	80.78	60.38

Table 9: Scalability of EEG-X across encoder widths. Parameter counts and dense FLOP estimates are measured from our implementation; FLOPs are per 2 s window (inference) or per pass over the STEW training set (39,138 windows; pretraining). Downstream columns report accuracy on STEW and DREAMER.

C.5 Dataset Overview and Processing

For the Temple University Hospital (TUH) datasets, we use the default train and test splits. For the remaining datasets, we randomly select 20% of the data subject-wise, ensuring that the same subject is not present in both the training and test sets.

C.5.1 Emotiv

We applied a bandpass filter to all Emotiv datasets and then segmented them into windows of 256 time steps, with each window corresponding to 2 seconds of recording, following EEG2Rep [Mohammadi Foumani et al., 2024a] and KnowEEG [Sahota et al., 2027].

DREAMER DREAMER is a multimodal dataset containing electroencephalogram (EEG) and electrocardiogram (ECG) signals recorded during affect elicitation using audio-visual stimuli [Katsigiannis and Ramzan, 2018], captured with a 14-channel Emotiv EPOC headset. The dataset includes data from 23 participants, along with their self-assessments of affective states (valence, arousal, and dominance) after each stimulus. For our classification task, we focus specifically on the valence labels. The DREAMER dataset can be accessed here[2]. We use the raw EEG data, applying low-pass and high-pass filters[3]. It's important to note that for our analysis, we exclude the ECG signals and only focus on the EEG data.

Crowdsourced Crowdsourced EEG data was collected while participants were engaged in a resting state task, involving periods with eyes open and eyes closed, each lasting 2 minutes. Out of 60 participants, only 13 successfully completed both conditions using 14-channel EPOC+, EPOC X, and EPOC devices. The data was initially recorded at 2048 Hz and later downsampled to 128 Hz. Raw EEG data for these 13 participants, along with preprocessing, analysis, and visualization scripts, are openly available on the Open Science Framework (OSF)[4].

Simultaneous Task EEG Workload (STEW) STEW dataset comprises raw EEG recordings from 48 participants using a 14-channel Emotiv EPOC headset involved in a multitasking workload experiment utilizing the SIMKAP multitasking test [Lim et al., 2018]. Additionally, the subjects'

[2] https://zenodo.org/records/546113

[3] https://torcheeg.readthedocs.io/en/v1.1.0/torcheeg.datasets.html

[4] https://osf.io/9bvgh/?view_only=70744f62157c46d5bd731480db1873df

baseline brain activity at rest was recorded before the test. The data was captured using the Emotiv Epoc device with a sampling frequency of 128Hz and 14 channels, resulting in 2.5 minutes of EEG recording for each case. Participants were instructed to assess their perceived mental workload after each stage using a rating scale ranging from 1 to 9, and these ratings are available in a separate file. Moreover, this dataset includes binary class labels, considering a workload rating of more than 4 as high and otherwise as low. We utilize these labels for our specific problem. STEW can be accessed upon request through the IEEE DataPort[5].

Alternatively, all the above Emotiv EEG datasets are also available on the Hugging Face platform under the `monster-monash` [Dempster et al., 2025] repository[6] and KnowEEG [Sahota et al., 2027].

C.5.2 Temple University Hospital (TUH) EEG Corpus

TUH Abnormal EEG Corpus (TUAB)
The TUH Abnormal EEG Corpus (TUAB) is a subset of the Temple University Hospital (TUH) EEG Corpus, which is one of the largest publicly available collections of clinical EEG data. The TUAB specifically focuses on EEG recordings labeled as abnormal, making it a valuable resource for studies on neurological disorders, brain function anomalies, and the development of diagnostic tools [Lopez et al., 2015].

TUH EEG Events (TUEV)
The TUH EEG Events Corpus (TUEV) contains annotations of EEG segments classified into six different categories: spike and sharp wave, generalized periodic epileptiform discharges, periodic lateralized epileptiform discharges, eye movement, artifact, and background [Harati et al., 2015]. For our binary classification task, we group the artifact and background categories into one class, while the remaining categories form the other class.

Acquisition and Preprocessing
The TUH Abnormal EEG Corpus (TUAB) [Lopez et al., 2015] and TUH EEG Events (TUEV) [Harati et al., 2015] can be accessed upon request through the Temple University Electroencephalography (EEG) Resources[7]. We processed both datasets to adhere to the 19-channel EEG montage following the 10-20 international system. The specific electrode configuration is as follows:

```
TUH_19_electrodes = ['fp1', 'fp2', 'f3', 'f4', 'c3', 'c4', 'p3', 'p4', 'o1',
'o2', 'f7', 'f8', 't7', 't8', 'p6', 'p5', 'fz', 'cz', 'pz']
```

C.5.3 BCICIV_2A: Motor Imagery Task:

The BCIC-IV-2a [Brunner et al., 2008] dataset was recorded using a 22-channel EEG headset (Fig.4c). The data were downsampled from 250 Hz to 128 Hz and consist of EEG recordings from 9 subjects performing four motor imagery tasks: left hand, right hand, feet, and tongue. Each subject completed two sessions on separate days, with 288 trials per session. This dataset was selected for its minimal electrode overlap with Emotiv and TUH devices (as shown in Fig. 4), providing a challenging cross-domain evaluation.

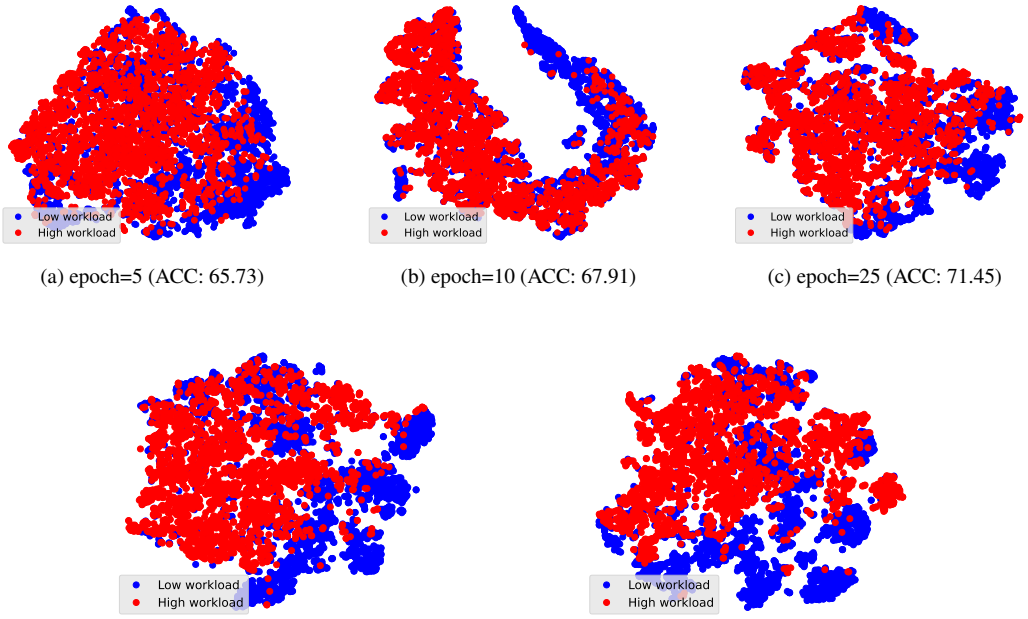
C.5.4 Online SSVEP-based BCI using Riemannian Geometry:

The dataset comprises 8-channel EEG recordings from 12 subjects (ages 20–28) performing SSVEP-based tasks [Kalunga et al., 2016]. Each trial lasts 5 seconds, initiated by an auditory cue indicating which LED to focus on, or a fixation point for the reject class. The data were resampled to 128 Hz and segmented into 2-second windows. The primary task involves classifying the subject's focus on one of three LED groups or the reject class, utilizing Riemannian geometry for classification. The dataset is publicly available for research purposes.

C.5.5 PhysioNet Motor Imagery (PhysioNetMI)

The PhysioNet Motor Imagery dataset [Goldberger et al., 2000] consists of 64-channel EEG recordings collected from 109 subjects performing different motor imagery and motor execution tasks.

[5] `https://ieee-dataport.org/open-access/stew-simultaneous-task-eeg-workload-dataset`
[6] `https://huggingface.co/monster-monash`
[7] `https://isip.piconepress.com/projects/tuh_eeg/html/downloads.shtml`



(a) epoch=5 (ACC: 65.73) (b) epoch=10 (ACC: 67.91) (c) epoch=25 (ACC: 71.45)

(d) epoch=50 (ACC: 77.02) (e) epoch=100 (ACC: 78.81)

Figure 8: 2D t-SNE visualization of the learned representation on the STEW dataset at different epochs during pre-training. At epoch 5, the classes are highly entangled, with separation increasing over time, becoming distinct by epoch 100. The accuracy improves from 65.73% at epoch 5 to 78.81% at epoch 100.

EEG signals were recorded using the international 10-10 electrode placement system at a sampling frequency of 160 Hz. Following previous EEG foundation model studies, we focus on the motor imagery task involving imagined left fist, right fist, both fists, and both feet movements. The data were resampled to 128 Hz and segmented into 2-second windows for training and evaluation. Due to its dense electrode configuration and large number of subjects, PhysioNetMI provides a challenging benchmark for evaluating cross-subject and cross-domain generalization in EEG representation learning.

C.6 Visualization

This section visualizes the evolution of learned representations during pre-training using t-SNE plots at various epoch. As shown in Fig. 8 In the early stages of training (e.g., epoch 5 and 10 Fig.8a and 8b, respectively), the embeddings from the two classes are highly entangled, with little distinction between them, showing that the model is just beginning to understand the data's structure. As training progresses, class separation improves, with noticeable distinction by epoch 25 (Fig.8c) and clear separation by epoch 100 (Fig.8e). Despite being unsupervised, the model leverages the STEW dataset's inherent structure to enhance class separation. The model's accuracy also increases from 65.73% at epoch 5 to 78.81% at epoch 100, highlighting the benefits of self-supervised training in learning meaningful representations.

C.7 Limitations

EEG-X has several limitations. First, it is evaluated on only eight datasets, which may not fully represent the diversity of EEG devices, populations, and recording protocols. Second, its noise-aware reconstruction strategy uses rASR to generate clean EEG targets. However, rASR may not be available or practical for all datasets and recording setups, requiring alternative artifact removal methods. Third, the DiCT objective introduces additional computational overhead, which increases with the input size. Finally, our evaluation mainly focuses on classification, and its effectiveness on other EEG tasks remains to be explored.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly describe the proposed EEG-X framework, including location-based channel embeddings, noise-aware reconstruction, and the DiCT reconstruction objective. The claims are supported by experimental results and aligned with the reported evaluations across multiple EEG datasets and cross-domain settings.

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper discusses limitations related to heterogeneous EEG datasets, artifact removal dependency, and computational considerations of large-scale EEG foundation models. Additional discussions on DiCT parameter sensitivity and denoising assumptions are included in the appendix.

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

   Justification: The paper does not present formal theoretical theorems or proofs. The work is primarily empirical and methodological.

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The paper describes the model architecture, training objectives, preprocessing pipeline, datasets, hyperparameters, and evaluation protocols in sufficient detail. Additional implementation details and parameter selections are provided in the appendix.

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

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

   Answer: [Yes]

   Justification: The EEG-X implementation code and pretrained models will be made publicly available for research and academic purposes only. Their use for commercial or business purposes is not permitted without prior authorization from the authors or the relevant rights holders. The accompanying materials provide instructions for preprocessing, model training, and reproducing the reported experiments. Public EEG benchmarks are appropriately referenced, while access to proprietary datasets is subject to their respective terms and conditions.

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The paper specifies dataset splits, preprocessing steps, optimizer settings, model configurations, masking strategies, and DiCT parameter selections. Additional experimental details and parameter configurations are provided in the appendix.

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification:The paper reports performance variability using mean and standard deviation across repeated runs. In addition, pairwise two-sided Wilcoxon signed-rank tests with Holm correction for multiple comparisons are conducted to assess the statistical significance of the performance differences between EEG-X and competing methods. The statistical analysis is reported alongside the corresponding experimental results.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The paper describes the hardware environment and computational resources used for pretraining and downstream evaluations, including the GPU and CPU specifications and available GPU memory. Additional computational details, including model complexity and FLOP estimates, are provided in the appendix.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics. The work focuses on methodological advances for EEG representation learning and uses appropriately licensed datasets and anonymized EEG recordings.

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The paper discusses potential positive impacts in healthcare, neuroscience, and brain-computer interfaces, as well as potential risks related to privacy, misuse of neurophysiological data, and biases arising from heterogeneous EEG datasets.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse?

    Answer: [N/A]

    Justification: The paper does not release generative models or high-risk systems. The work focuses on representation learning for EEG analysis and does not introduce assets with substantial misuse risk.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: Public datasets, preprocessing tools, and baseline methods are appropriately cited throughout the paper. The use of existing datasets and tools follows their respective licenses and terms of use. Access to proprietary datasets is subject to the respective owners' terms and conditions.

13. **New assets**

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

Answer: [Yes]

Justification: The paper introduces the EEG-X implementation and pretrained models, which will be released with documentation and reproduction instructions. These assets are intended for research and academic purposes only. Commercial or business use requires prior authorization from the authors or the relevant rights holders.

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or newly conducted human-subject studies.

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

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals were obtained?

    Answer: [N/A]

    Justification: The work relies on previously collected and anonymized EEG datasets and does not involve new human-subject experiments conducted by the authors.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

    Answer: [N/A]

    Justification: Large language models are not part of the proposed methodology or experimental framework.