BEYOND PATCH-LEVEL EEG RECONSTRUCTION: CONTRASTIVE PRETRAINING AND MULTISCALE CONVOLUTIONAL TRANSFORMER FOR EEG DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised pretraining has shown early promise for improving non-invasive electroencephalogram (EEG) decoding. Recently, many large-scale EEG foundation models have converged on the recipe of raw signal tokenization followed by masked reconstruction pretraining. However, this approach may be suboptimal for data with low signal-to-noise ratio and sample-level dependencies at multiple timescales. To test this hypothesis, we develop a novel contrastive-pretrained EEG model with multiscale sample-level convolutions and Transformer encoder blocks (CoCoT-EEG). CoCoT achieves state-of-the-art performance on extensive benchmark decoding tasks with heterogeneous experimental configurations. Moreover, we systematically dissect the impact of architecture and pretraining objective through ablation experiments and mechanistic analyses. Consistent with our hypothesis, models benefit from contrastive pretraining, as well as sample-level feature extraction across patch boundaries and longer convolution kernels. Thus, we demonstrate the viability of contrastive learning for EEG modeling while suggesting key design considerations that complement existing architectural advances, prompting further investigations in alternative large-scale pretraining strategies.

1 INTRODUCTION

Accurate prediction of physiological and cognitive states from electroencephalography (EEG) enables many applications like brain-computer interface and clinical diagnostics, owing to its moderate costs, non-invasive nature, and high temporal resolution. Classically, decoding models are trained using supervised objectives to extract task-specific features from labeled data (Lawhern et al., 2016). But inspired by the success of large-scale, self-supervised language and vision models, i.e., *foundation models* (FM), efforts towards EEG decoding have quickly adopted similar strategies.

EEG FMs are *pretrained* on tens of thousands of hours of unlabeled data to learn general-purpose features, which subsequently transfer to a broad range of downstream decoding tasks. The overwhelming majority of self-supervised models have converged on a common recipe employing *reconstruction pretraining objectives* (49/55 models surveyed in Liu et al. (2026) and 9/10 in Kuruppu et al. (2026); Fig. 1A,B), such as masked autoencoding (MAE). By keeping to the established paradigm, recent works readily incorporated advances from masked vision and language modeling, such as discrete codebooks (Jiang et al., 2024), dedicated spatial and temporal attention (Wang et al., 2024; Zhou et al., 2025), alongside EEG-specific innovations such as electrode position encoding strategies (Ouahidi et al., 2025), and state space models (SSMs) for efficient cross-token processing (Ma et al., 2025).

However, while the focus on architectural innovations and scaling has resulted in performance gains, a large, complementary part of the design space *beyond patch-level reconstruction* remains unexplored despite strong theoretical and practical arguments to do so. Consider its signal properties: EEG has sample-level dependencies within and across timescales, i.e., is scale-free (He, 2014). Meanwhile, EEG has high temporal resolution but low signal-to-noise ratio, and is riddled with physiological and environmental noise (Cohen, 2014; Lai et al., 2018). As such, it's unclear if the common recipe using reconstruction pretraining with patch-level tokenization is also *the most effective one*:

First, reconstruction-based objectives may be suboptimal for pretraining EEG tokenizers and encoders. Recent works have shown in general settings, both theoretically and empirically, that reconstruction

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

biases models to focus on low-level features that are irrelevant for tasks relying on perceptual semantics (Balestriero & LeCun, 2024), and are less suited to high-noise regimes (Van Assel et al., 2025). In EEG specifically, MAE objectives result in embeddings that are biased to capture the overall "1/f noise" or aperiodic component (Donoghue et al., 2020) while discarding information about weaker oscillations, especially those at high frequencies (Kommineni et al., 2026).

Furthermore, adherence to mask reconstruction naturally leads to the inheritance of *patch-level tokenization*, which avoids leakage across masks. Almost all EEG FMs first divide the raw signal into short patch-tokens. While this is sensible for sequences with clear boundaries (e.g., text), for EEG, early-patching imposes arbitrary boundaries that remove context between otherwise adjacent samples, artificially creating two scales that the tokenizer and encoder separately see. Classical supervised models, on the other hand, very effectively capture dependencies in EEG samples with multiscale temporal convolution kernels (Lawhern et al., 2016; Ingolfsson et al., 2020).

Finally, as a result of these frequently coupled design choices, we lack a thorough understanding of the EEG FM design space in general. In particular, while ablation experiments usually isolate the effect of a new tokenizer or encoder architecture, evidence for how reconstruction objectives impact task performance and model representations is sparse: even with the few models that are not reconstruction-pretrained, meta-comparisons are confounded by vastly different architectures (Liu et al., 2026; Kuruppu et al., 2026), while no single study has systematically reported the effect of pretraining objectives for a fixed architecture. As such, we lack critical evidence that can inform principled design of EEG FMs that potentially combine performant components and objectives.

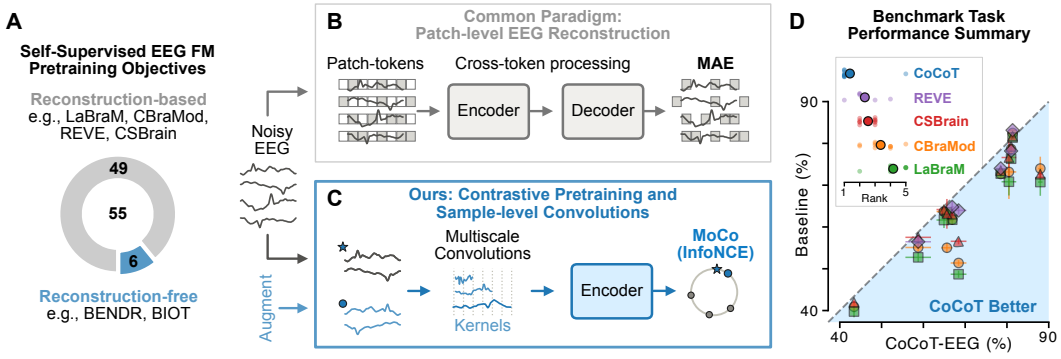

Figure 1: **Motivation and summary.** **(A)** The vast majority of current EEG FMs use reconstruction-based pretraining objectives; adapted from (Liu et al., 2026). **(B)** The common design pattern follows patch-level tokenization and reconstruction, resulting in hard boundaries between adjacent samples and model capacity dedicated to reconstructing noise; *MAE: masked autoencoding.* **(C)** We here adopt a novel design employing multiscale convolutions and contrastive-pretraining for self-supervised EEG modeling; *MoCo: Momentum Contrast.* **(D)** Our model—CoCoT-EEG—achieves state of the art performance on diverse downstream decoding tasks; inset: model rankings across 11 tasks.

Main contributions. In this study, we look beyond the prevailing paradigm of patch-token reconstruction and investigate the effectiveness of contrastive pretraining and (sample-level) multiscale convolutions for large-scale EEG FMs:

1. We develop a novel **Co**ntrastive-pretrained EEG model with multiscale 1D **Co**nvolutions and **T**ransformer encoders (CoCoT-EEG, Fig. 1C). The pretrained model achieves state-of-the-art performance on diverse downstream tasks compared to recent EEG FMs (Fig. 1D), while the architecture itself is effective when trained from scratch, outperforming many existing single-task models.

2. Through extensive ablation experiments and interpretability analyses, we find that the convolution-first tokenizer consistently outperforms patch-first designs, even with matched kernel lengths. These findings confirm the importance of sample-level feature extraction that crosses arbitrary patch boundaries, while longer kernels further impart selectivity to biologically-relevant EEG frequencies.

3. Lastly, we isolate the contribution of pretraining and its augmentations. Critically, we develop a sparse convolution-based tokenizer compatible with MAE pretraining (Woo et al., 2023). While

this matched architecture is already performant under MAE, contrastive pretraining provides further benefits, which we trace to better preservation of signal content and task-informative representations.

Overall, our results demonstrate *and dissect* the benefits of contrastive pretraining and multiscale convolution-based feature extraction for EEG FMs. More broadly, we highlight the importance of decoupling architectural considerations from objectives, offering concrete insights for designing self-supervised pretraining strategies that readily complement existing architectural innovations.

2 Related works

Self-Supervised EEG Foundation Models. Large-scale EEG FMs have predominantly been pretrained with reconstruction objectives: One of the first to massively scale data and model capacity, LaBraM (Jiang et al., 2024) introduced vector-quantized neural spectrum prediction for masked EEG modeling, pretraining on ∼2,500 hours across 20 datasets. CBraMod (Wang et al., 2024) proposed a criss-cross Transformer that separately models spatiotemporal dependencies through parallel spatial and temporal attention. REVE (Ouahidi et al., 2025) further scaled up to 60k hours of EEG recordings while leveraging 4D positional encoding to align heterogeneous datasets. Lastly, CSBrain (Zhou et al., 2025) and CodeBrain (Ma et al., 2025) propose cross-scale learning under MAE pretraining, but similarly separate sample-level feature extraction within patches from token-level processing, with sparse attention and state space models for the latter, respectively.

Theoretical Foundations of Contrastive Learning. Our work builds on theoretical insights from the joint embedding literature. Contrastive losses asymptotically optimize two key properties (Wang & Isola, 2020): alignment, which ensures that semantically similar samples are mapped to nearby representations, and uniformity, which encourages features to be evenly distributed to preserve maximal information. Recent theoretical and empirical work suggests that joint-embedding methods are more robust than reconstruction-based objectives in regimes dominated by high-magnitude nuisance variability (Van Assel et al., 2025), a setting particularly relevant for EEG signals.

Contrastive Learning for EEG modeling. Contrastive learning has remained a viable strategy for pretraining language and vision models (Radford et al., 2021). For EEG, early contrastive self-supervised models include BENDR (Kostas et al., 2021), which adapts wav2vec2.0's framework (Baevski et al., 2020) to learn transferable representations. BIOT (Yang et al., 2023a)—the single model out of 55 surveyed in (Liu et al., 2026) that is pretrained using a contrastive objective with augmentations only—breaks the raw signal into small, fixed-size patches, before tokenizing and embedding their Fourier spectra. More recently, hybrid objectives combining masked reconstruction with contrastive learning have been explored: CoMET (Li et al., 2025), for instance, combines masked EEG modeling with contrastive objectives.

Tokenization in EEG models: patch-first vs. convolution-first. Convolutional feature extraction on the continuous raw signal has long been the default in single-task EEG decoding (Lawhern et al., 2016; Song et al., 2023), yet EEG foundation models have converged on the opposite order: the signal is first segmented into short patches—non-overlapping in LaBraM, CBraMod, and CSBrain (Jiang et al., 2024; Wang et al., 2024; Zhou et al., 2025), and lightly overlapping in BIOT and REVE (Yang et al., 2023a; Ouahidi et al., 2025). CSBrain applies convolutions *after patchification and only within short windows*, while the concurrent NeuroRVQ tokenizer (Barmpas et al., 2025), whose multi-scale kernel bank is close to ours, also applies it on non-overlapping patches. BENDR (Kostas et al., 2021) is the exception that tokenizes by convolution, but with a single-scale.

Thus, despite recent progress, EEG FMs mainly use a paradigm centered around patchification-first and masked reconstruction losses. *CoCoT-EEG is so far the only EEG FM employing multi-scale convolutions at the sample-level across patches, and is further improved with contrastive pretraining*.

3 Model architecture and training

3.1 Overview

We propose **CoCoT-EEG**, a novel convolution-Transformer hybrid architecture for EEG representation learning and decoding. Given an EEG sample $X \in \mathbb{R}^{C \times T}$ (channels C, timepoints T), this model 1) extracts multi-scale temporal features *on the full sequence in raw signal space* with parallel

1D convolutional branches, 2) *patchifies after embedding* through L2Pooling to produce tokens $Z \in \mathbb{R}^{C \times S \times D}$ with sequence length S patches and embedding dimension D, for each channel C, and 3) *encodes tokens* with a Transformer stack to produce per-token features. Here, given patch size P, the sequence length is expressed as $S = \lfloor T/P \rfloor$.

A central design choice is to *embed before patchification*: unlike conventional tokenizers that segment the raw signal in patches prior to feature extraction, our convolution-first approach selectively filters out frequency-dependent noise *and* preserves temporal continuity during early processing, especially long-timescale signals, effectively presenting later Transformer blocks with more information-dense features. Detailed descriptions of the architecture are provided in Sec. 3.2-3.3, and Fig. 4 in Appendix.

The model can be pretrained in a self-supervised manner using contrastive objectives (Sec. 3.4) and further finetuned with supervision labels, or similarly be trained from scratch end-to-end in a fully supervised setting. We ablate key model components and pretraining objective in Sec. 4.3.

3.2 Multiscale convolutional tokenizer

Multi-scale temporal branches. We use d parallel temporal convolution branches with kernel sizes $\{k_i\}_{i=1}^{d}$, each operating on the full sequence independently for each EEG channel, using D_i filters with weights shared across channels: $h_i = \mathrm{Conv1D}_{k_i}(X), \quad \text{with} \quad h_i \in \mathbb{R}^{C \times D_i \times T}$, followed by group normalization and GELU nonlinearity. Branch dimensions satisfy $\sum_{i=1}^{d} D_i = D$, and their outputs are concatenated along the feature axis to yield $H = [h_1, \ldots, h_d] \in \mathbb{R}^{C \times D \times T}$.

Patchification in feature space. Patchification in feature space is performed by applying L_2 pooling with both kernel size and stride P to produce $Z \in \mathbb{R}^{C \times D \times S}$ whose entries, for patch index $s \in \{0, \ldots, S-1\}$, are

$$Z[c, d, s] = \left(\sum_{t=0}^{P-1} |H[c, d, sP + t]|^2 \right)^{\frac{1}{2}}, \quad c \in \{1, \ldots, C\},\ d \in \{1, \ldots, D\}. \tag{1}$$

We then permute Z to shape $C \times S \times D$. Layer normalization is applied after adding temporal positional encoding.

Temporal Fourier Positional Encoding. Positional information is injected using temporal Fourier positional encodings only. Let $s \in \{0, \ldots, S-1\}$ denote the patch index along time. We build deterministic Fourier features $\phi(s) \in \mathbb{R}^D$: $\phi_{2m}(s) = \sin\left(\frac{s}{\tau_m}\right), \quad \phi_{2m+1}(s) = \cos\left(\frac{s}{\tau_m}\right)$, where $\{\tau_m\}$ are linearly spaced timescales. We then add the same temporal encoding to all channels: this provides temporal order information while keeping the model free of electrode geometry assumptions. See additional experiments with electrode position encoding schemes in Sec. 4.3.1.

3.3 Transformer Encoder & Prediction head

Backbone. The backbone encoder is a stack of L transformer layers operating on $Z \in \mathbb{R}^{C \times S \times D}$. We apply multi-head self-attention (MHSA) over all channel–time tokens. Given token features Z, we flatten them into a sequence $\tilde{Z}$ of length $N = C \cdot S$. Each transformer layer applies MHSA followed by a position-wise feed-forward network (FFN). We use a pre-layer normalization design with residual connections, stochastic depth (DropPath), and learnable LayerScale parameters initialized to 10^{-4}. The FFN employs a gated SwiGLU activation.

Residual MLP head for supervised finetuning. For downstream classification or regression, we flatten the final-layer token features before applying an MLP head composed of L_h residual GLU blocks. The flattened features are first projected to a hidden dimension d_{proj} with GELU and dropout; residual GLU blocks with scaled residual factor α are then applied, followed by a skip connection from the input and a final linear prediction layer after layer normalization.

3.4 Self-Supervised Pretraining

Contrastive pretraining (MoCo). We pretrain the tokenizer and encoder in a self-supervised manner using Momentum Contrast (MoCo) (He et al., 2019). Given an unlabeled EEG segment X, we

sample two stochastic augmentations to produce two views $X^q = t_1(X)$ and $X^k = t_2(X)$. The query encoder f_θ and key encoder $f_{\bar{\theta}}$, updated through an exponential moving average of the query encoder, share the same architecture. We pool the token features into a single vector representation through attention pooling before passing it through a SimCLR-style two-layer MLP projector (Chen et al., 2020). The projected query and key vectors are then ℓ_2-normalized. We maintain a queue of negative keys $\{k_i^-\}_{i=1}^K$ and optimize the InfoNCE loss:

$$\mathcal{L}_{\text{MoCo}} = -\log \frac{\exp(q^\top k^+/\tau)}{\exp(q^\top k^+/\tau) + \sum_{i=1}^K \exp(q^\top k_i^-/\tau)}, \tag{2}$$

where τ is the temperature and k^+ is the positive key corresponding to q. Details in Appendix B-C.

Augmentations. We construct the two views of each EEG sample by applying random compositions of EEG-specific augmentations, including time masking, local time reversal, channel dropout, phase jitter, spectral corruption, and Fourier-phase surrogates. These augmentations mimic common acquisition artifacts and noise observed in EEG recordings. They encourage invariance to corrupted or missing electrodes, sensor noise, temporal occlusions, and moderate spectral perturbations. The augmentation probabilities, parameters, and specific configurations are provided in Appendix D.

Masked reconstruction (MAE). For the pretraining objective ablation, we additionally train the same encoder architecture using masked autoencoding, under matched pretraining corpus and preprocessing. Since convolutional receptive fields cross patch boundaries, masking tokens after convolution would expose information from hidden samples. We therefore randomly mask 50% of channel–time patches before convolution and use sparse, visibility-aware tokenization (Woo et al., 2023), retaining the full kernel bank while preventing this leakage. Only visible tokens enter the Transformer encoder. A four-layer Transformer decoder reconstructs the masked waveform samples using an L1 loss. Implementation details and settings are provided in Appendix B.

4 Experiments

We evaluate CoCoT-EEG on a suite of downstream EEG tasks spanning multiple configurations and application domains (see Appendix E), while comparing to recent state-of-the-art (SOTA) models. In this section, we first describe the overall experimental setup (Sec 4.1). We then show that our model is competitive in both pretrained and "from-scratch" settings compared to foundation and single-task models, respectively (Sec. 4.2). Finally, we perform ablation experiments to assess the contribution of key design elements that distinguish our architecture and pretraining strategy (Sec. 4.3).

4.1 Experimental setup

Pretraining corpora. We pretrain on the Temple University EEG Corpus (TUEG; (Obeid & Picone, 2016)), a clinical EEG dataset comprising over 16,000 sessions from more than 10,800 subjects and totaling around 27,000 hours of recordings, downsampled to 100 Hz. For evaluation on TUAB, which is a subset of TUEG, we use a checkpoint pretrained with recordings from TUAB excluded.

Downstream benchmarks. We evaluate CoCoT-EEG on eleven EEG datasets from the benchmark suites used by CBraMod and CSbrain (Wang et al., 2024; Zhou et al., 2025). These cover motor imagery classification, emotion recognition, seizure and abnormality detection, sleep stage classification and vigilance estimation. Dataset descriptions and preprocessing procedures are detailed in Appendix E and in Table 4. To ensure fair comparisons, datasets were split into train, validation and test following the same process implemented in Wang et al. (2024) and downsampled to 100 Hz.

Evaluation protocol. We evaluate CoCoT-EEG both through end-to-end finetuning of pretrained checkpoints and through supervised training from scratch. The encoder and task-specific prediction head are optimized jointly, with checkpoints selected using validation performance. Our primary metric is balanced accuracy for classification and Pearson correlation for vigilance estimation on SEED-VIG; both reported on a scale of 0–100. We report means and 95% confidence intervals across training four seeds. Task-specific hyperparameters are detailed in the appendix App. F.

Baselines. In the pretrained setting, we compare CoCoT-EEG against recent SOTA EEG foundation models, including LaBraM (Jiang et al., 2024), CBraMod (Wang et al., 2024), REVE (Ouahidi et al., 2025), and CSBrain (Zhou et al., 2025). We also compare CoCoT-EEG trained from scratch against

competitive single-task model performances reported in (Zhou et al., 2025), including ContraWR (Yang et al., 2023b), CNN-Transformer (Peh et al., 2022), and Spatiotemporal-Transformer (Song et al., 2021), which achieved best-of-class performance on one or more tasks. Published baseline results are taken from the cited studies under their respective evaluation settings.

4.2 Model performance on benchmark task suite

Pretrained. CoCoT-EEG pretrained under a contrastive objective (MoCo) and finetuned per task achieves SOTA performance on 10 out of 11 tasks, beating or matching EEG FMs pretrained on TUEG (CBraMod and CSBrain) or even larger EEG data sources (REVE). For BCI-IV, to match the protocol evaluation of (Ouahidi et al., 2025), we finetune our model using Euclidean Alignment (EA; He & Wu (2020)). Without EA, CoCoT-EEG's balanced-accuracy on BCI-IV (59.25 ± 2.12) ranks above all EEG FMs evaluated without adaptation. When pretrained on a different dataset (subset of HBN, Shirazi et al. (2024)) comprising around 1500 hours of EEG (18 times fewer hours), CoCoT still remains competitive and yields higher mean performance than training from scratch on all tasks included in the corpus comparison (Appendix L, Table 10). These results suggest that CoCoT-EEG benefits from pretraining across distinct populations and recording configurations.

From scratch. We also evaluate CoCoT-EEG trained from scratch on each downstream task independently. While pretraining improves performance on all downstream tasks, the base CoCoT-EEG architecture still outperforms single-task models on most tasks (Table 9), including those with similar design patterns such as CNN-Transformer and Spatiotemporal-Transformer, and remains competitive with several pretrained EEG foundation models (Table 1, SCRATCH).

Table 1: Performance of CoCoT-EEG trained from scratch, or first pretrained (PT) with MoCo on TUEG. Baseline performances are taken from published values in (Ouahidi et al., 2025) and (Zhou et al., 2025). mean ± std is reported for the baselines, and mean ±95% CI over 4 seeds is reported for CoCoT. **Bold** denotes SOTA mean. ⋆ denotes runs using domain-adaptation, specifically, Euclidean Alignment (He & Wu, 2020), to be consistent with REVE (Ouahidi et al., 2025).

		Baselines				CoCoT-EEG (ours)	
Dataset	Metric	LaBraM	CBraMod	REVE	CSBrain	Scratch	TUEG PT
BCI-IV	B-Acc	48.69 ± 0.85	51.38 ± 0.66	63.96⋆ ± 0.95	56.57 ± 0.71	54.40 ± 2.99	**68.42⋆** ± 1.89
	F1-W	47.58 ± 1.03	49.84 ± 0.85	63.39 ± 1.10	56.37 ± 0.87	53.25 ± 0.36	67.99 ± 2.10
CHB-MIT	B-Acc	70.75 ± 3.58	73.98 ± 2.84	—	72.62 ± 1.15	60.55 ± 0.79	**87.98** ± 0.53
	AUROC	86.79 ± 1.99	88.92 ± 1.54		89.15 ± 3.21	80.32 ± 1.67	91.60 ± 0.53
FACED	B-Acc	52.73 ± 1.07	55.09 ± 0.89	56.46 ± 1.64	57.52 ± 0.42	55.52 ± 1.00	**58.76** ± 3.06
	F1-W	52.88 ± 1.02	56.18 ± 0.93	56.59 ± 1.72	57.96 ± 0.31	55.19 ± 1.15	58.54 ± 2.96
HMC	B-Acc	72.77 ± 1.01	72.69 ± 0.41	74.01 ± 0.75	73.45 ± 0.47	73.69 ± 0.21	**78.51** ± 0.26
	F1-W	74.54 ± 0.27	73.95 ± 0.89	76.38 ± 0.74	75.06 ± 0.42	74.27 ± 0.32	78.58 ± 0.37
ISRUC	B-Acc	76.33 ± 1.02	78.65 ± 1.10	78.19 ± 0.78	79.25 ± 0.30	74.63 ± 0.28	**80.99** ± 0.26
	F1-W	78.10 ± 1.33	80.11 ± 0.99	80.05 ± 1.35	79.90 ± 0.91	76.70 ± 0.37	81.66 ± 0.12
PHYSIO	B-Acc	61.73 ± 1.22	64.17 ± 0.91	64.80 ± 1.40	63.04 ± 0.90	62.44 ± 1.31	**66.90** ± 0.37
	F1-W	61.77 ± 1.41	64.27 ± 0.10	64.84 ± 1.70	63.08 ± 0.95	62.46 ± 1.28	67.04 ± 0.36
SEED-V	B-Acc	39.76 ± 1.38	40.91 ± 0.97	—	41.97 ± 0.33	41.41 ± 0.44	**43.40** ± 0.33
	F1-W	39.74 ± 1.11	41.01 ± 1.08		42.80 ± 0.23	41.95 ± 0.61	43.83 ± 0.28
SHU	B-Acc	61.66 ± 1.92	63.70 ± 1.51	—	64.17 ± 0.37	62.25 ± 1.65	**64.93** ± 1.79
	AUROC	66.04 ± 0.91	69.88 ± 0.68		72.00 ± 0.58	68.44 ± 1.38	73.00 ± 0.91
TUAB	B-Acc	81.40 ± 0.19	82.49 ± 0.25	**83.15** ± 0.14	81.72 ± 0.43	79.17 ± 0.14	81.35 ± 0.91
	AUROC	90.22 ± 0.09	86.06 ± 0.57	92.45 ± 0.13	89.57 ± 0.46	85.98 ± 0.55	88.96 ± 0.89
SIENA	B-Acc	70.82 ± 3.29	73.17 ± 6.47	—	76.62 ± 4.71	75.52 ± 1.32	**80.47** ± 1.76
	AUROC	88.14 ± 3.28	90.38 ± 2.18		90.76 ± 1.19	89.21 ± 0.22	91.81 ± 0.52
SEEDVIG	Corr.	63.47 ± 1.35	55.02 ± 1.15	—	63.14 ± 3.56	60.37 ± 2.04	**65.63** ± 1.32
	R^2	18.08 ± 9.58	7.37 ± 1.67		23.63 ± 5.19	16.73 ± 0.73	29.25 ± 4.54

4.3 Ablation experiments and analysis

To assess the contribution of our main design choices to downstream performance, we systematically ablate CoCoT-EEG's architecture, pretraining, and data augmentation strategy.

Table 2: Architecture and pretraining objective ablations: Convolution-first vs. Patch-first model variants pretrained with MoCo (with truncated kernels, maximum kernel length restricted to the patch size), and CoCoT-EEG's backbone with the full multi-scale kernel bank pretrained with MoCo vs. MAE. Pearson correlation for SEED-VIG, balanced accuracy (%) otherwise; mean ± 95% CI over 4 seeds. **Bold**: best per dataset. Underlined: MAE values would rank first among baselines in Table 1.

	BCI-IV	CHB-MIT	FACED	HMC	ISRUC	PHYSIO	SEED-V	SHU	TUAB	SIENA	SEED-VIG
Conv-1st	58.9 ±2.8	87.6 ±0.8	58.4 ±1.8	78.0 ±0.7	80.5 ±0.6	66.1 ±1.1	43.0 ±0.9	64.9 ±1.0	81.1 ±0.4	**81.4** ±1.6	63.7 ±1.4
Patch-1st	48.9 ±3.8	86.8 ±1.2	54.0 ±6.1	76.6 ±0.3	80.6 ±0.3	64.1 ±1.0	42.9 ±0.9	64.4 ±1.8	80.5 ±0.5	81.1 ±1.3	61.5 ±6.2
MoCo	**59.2** ±2.1	**88.0** ±0.5	**58.8** ±3.1	**78.5** ±0.3	**81.0** ±0.3	**66.9** ±0.4	**43.4** ±0.3	**64.9** ±1.8	81.4 ±0.9	80.5 ±1.8	**65.6** ±1.3
MAE	54.8 ±2.3	84.8 ±2.0	56.8 ±1.4	76.7 ±0.4	79.4 ±0.1	64.3 ±1.1	42.3 ±0.7	64.5 ±0.3	**82.0** ±0.5	77.8 ±1.5	64.0 ±2.4

4.3.1 Model architecture ablation

Convolution-first. In this work, we argue that convolutional feature extraction should precede patchification. To examine this, we evaluate two tokenization variants—a convolution-first tokenizer (Conv-first) and a non-overlapping patch tokenizer (Patch-first)—both using a truncated convolutional kernel bank where the maximum kernel length is below the patch size to control for receptive field size, under matched pretraining conditions (Table 2). Conv-first achieves higher performance on ten of the eleven evaluated tasks, including gains of 10.07 points on BCI-IV and 4.34 on FACED.

To characterize how tokenization order impacts sensitivity to patch boundaries, we translated oscillatory bursts across patch edges on held-out EEG backgrounds and decoded their frequency using linear probes fitted to frozen tokenizer features (Fig. 2A). We found that frequency-decoding error increases near patch boundaries for Patch-first, whereas Conv-first maintains lower and uniform error across event positions. This suggests that extracting features before patchification makes spectral information more consistently accessible, possibly explaining the downstream advantage of Conv-first tokenization for tasks relying on frequency-specific EEG patterns, such as motor imagery.

Multiscale kernels. Conv-first tokenization also permits convolutional kernels to extend beyond individual patches, allowing feature extraction over longer temporal contexts. Comparing the full Conv-first kernel bank, spanning 3–121 samples, with its truncated 3–45-sample counterpart under matched MoCo pretraining, longer temporal context generally results in higher mean downstream performances (Table 2). To characterize contributions from different kernels in more detail, we examine the learned filters spectra and measure the full nonlinear tokenizer's responses to sinusoidal perturbations added to held-out EEG backgrounds (Fig. 2B). Relative to the 50-sample patch size, long kernels (61–121 samples) favor lower frequencies and short kernels (3–45 samples) favor higher frequencies, emphasizing the benefit of sample-level feature extraction across diverse timescales.

Spatial positional encoding ablation. CoCoT-EEG intentionally uses only temporal positional encoding, allowing its backbone to accommodate heterogeneous and non-standard EEG montages without requiring electrode-coordinate metadata. Since EEG signals exhibit meaningful spatial organization, we evaluate whether explicitly encoding electrode geometry improves downstream performance by augmenting CoCoT-EEG with a 3D spatial positional encoding (3DPE) inspired from (Ouahidi et al., 2025). We observe that explicit spatial encoding yields inconsistent benefits, supporting temporal-only encoding as a practical default in the evaluated setting (Appendix I).

4.3.2 Pretraining objective ablation

MoCo vs. MAE. A key motivation for our study is to better understand the impact of pretraining objective under architecture-matched settings. Therefore, we compare contrastive (MoCo) and masked reconstruction (MAE) pretraining using the same full-kernel Conv-first encoder architecture. Our MAE implementation reconstructs masked inputs with an L1 loss and a four-layer Transformer decoder (details in Appendix B). We find that MoCo-pretraining achieves higher performance than MAE on ten of the eleven tasks (Table 2), including gains of 4.41 percentage points on BCI-IV, 3.20

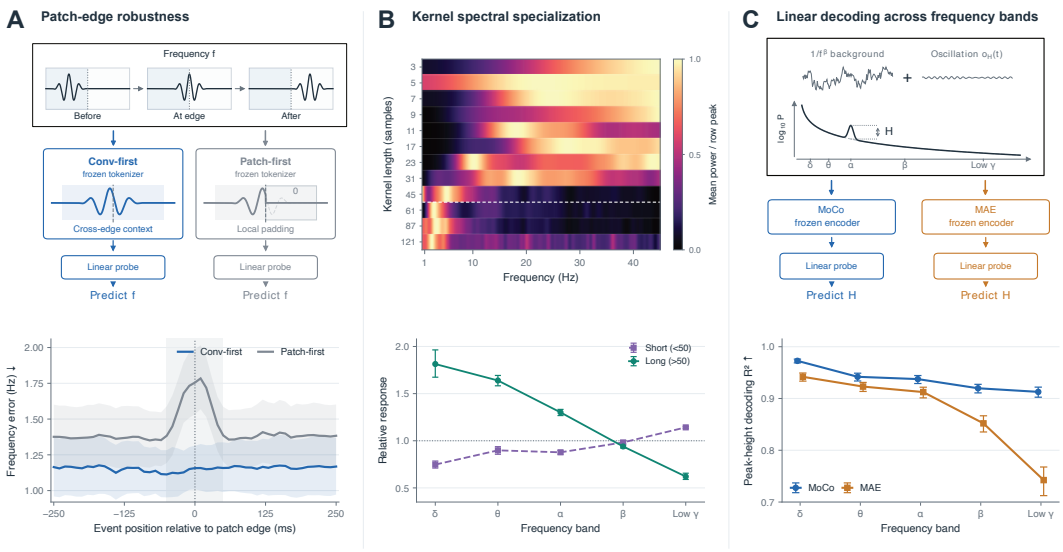

Figure 2: **Architectural robustness and spectral information accessibility in CoCoT-EEG. (A)** Linear probes trained to decode burst frequencies across event positions on frozen tokenizer features reveal better and more uniform frequency decoding across patch boundaries for Conv-first than Patch-first. **(B)** Top: learned filter-power spectra, averaged by kernel size and normalized to each row's maximum. Bottom: background-subtracted tokenizer responses to sinusoidal perturbations, normalized by each branch's frequency-sweep mean. Long kernels (> 50 samples) preferentially respond to slower activity, complementing shorter kernels. **(C)** Linear decoding of spectral peak height from frozen embeddings, using synthetic signals with an aperiodic background and matched total RMS. MoCo exhibits a smaller decline in decoding performance toward higher frequencies. Ribbons and error bars indicate pointwise 95% bootstrap confidence intervals over 20 subjects (A–B) or 512 independent synthetic draws (C). Full experimental protocols are provided in the Appendix M.

on CHB-MIT and 2.56 on PHYSIO. Nevertheless, our MAE-pretrained model remains competitive with state-of-the-art EEG foundation models, exceeding the strongest reported baseline on the primary evaluation metric for seven of the eleven tasks (Table 6).

These results demonstrate that our architecture supports effective transfer under both pretraining objectives, with contrastive pretraining providing an additional benefit. Grad-CAM analysis further shows that MoCo models exhibit focal, contralateral activations for the hand classes, consistent with sensorimotor organization, while MAE models are less lateralized (Appendix M, Fig. 5C).

Oscillatory information accessibility. To better understand this performance gap, we investigate whether MoCo makes oscillatory information more accessible in the learned representations. Reconstruction methods can prioritize high-variance components over features useful for discrimination (Balestriero & LeCun, 2024), whereas joint-embedding methods can be preferable when high-magnitude, task-irrelevant variation dominates the input (Van Assel et al., 2025). This distinction is relevant to EEG, whose aperiodic background concentrates power at lower frequencies, potentially disadvantaging weaker, higher-frequency oscillations that may carry discriminative information. Consistent with this concern, Kommineni et al. (2026) report a reduced decodability of higher-frequency oscillatory power in reconstruction-pretrained EEG foundation models, which we investigate below.

To test whether MoCo mitigates this bias, we fit linear probes to frozen MoCo and MAE embeddings to decode spectral peak height above a $1/f^{1.5}$ background, using synthetic inputs matched in total RMS (Fig. 2C). MoCo achieves higher decoding R^2 across all five evaluated bands, and a slower and weaker decline toward higher frequencies. MAE's decoding performance declines more sharply, with the greatest difference in the low-γ band, where MoCo achieves $R^2 = 0.913$ compared with 0.742 for MAE. This more consistent accessibility of oscillatory information offers a possible explanation for MoCo's downstream advantage on tasks relying on frequency-specific EEG patterns.

4.3.3 PRETRAINING AUGMENTATION ABLATION

To dissect which data augmentations contribute most to the learned representations, we pretrained CoCoT-EEG using an augmentation pipeline composed of 6 different augmentations (full description in Appendix D). To properly compare the impact of augmentations, we first pretrain CoCoT using only one single augmentation at a time, then two distinct combinations of three augmentations. We find that no single augmentation consistently recovers the benefit of the full pipeline (Fig. 3). Taken together, these findings suggest that the gains from augmentation are driven by diversity and complementary invariances rather than by any individual transformation alone, highlighting the importance of designing EEG-specific augmentations in contrastive-pretraining.

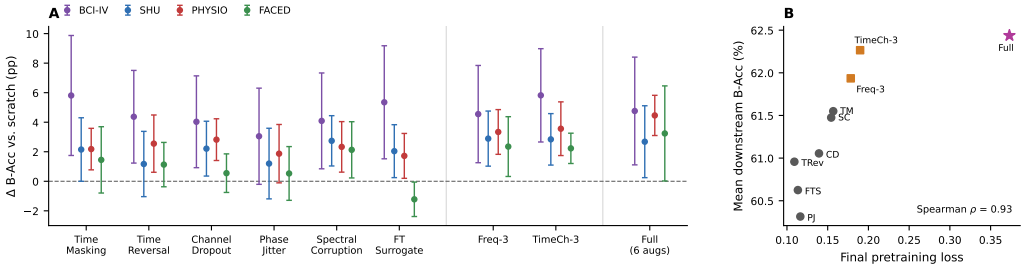

Figure 3: **Effect of pretraining augmentations on downstream performance. (A)** Gap to training from scratch in balanced accuracy (Δ = pretrained $-$ scratch, in % points): single augmentations, two combinations of three distinct augmentations — FREQ-3 (Phase Jitter, Spectral Corruption, FT Surrogate) and TIMECH-3 (Time Masking, Time Reverse, Channel Dropout) — and the full pipeline. Error bars: 95% CIs over 4 seeds. **(B)** Final pretraining loss against downstream balanced accuracy averaged over the four datasets: harder contrastive tasks seem to transfer better (Spearman $\rho = 0.98$).

5 DISCUSSION

In this study, we challenge the prevailing reconstruction-based pretraining paradigm for EEG foundation models. To do so, we develop CoCoT-EEG, a contrastive-pretrained, convolutional Transformer architecture that embeds the raw signal through parallel multi-scale 1D convolutions and tokenizes through L2 pooling, preserving temporal continuity and filtering noise before Transformer encoding. We report three main findings: first, our model achieves state-of-the-art performance on diverse downstream tasks compared to recent EEG FMs, while the architecture itself is effective when trained from scratch. Second, convolution-first tokenizer consistently outperforms patch-first designs, highlighting the importance of sample-level feature extraction that crosses arbitrary patch boundaries. Third, decoupling architectural considerations from objectives, contrastive learning provides consistent benefits over MAE pretraining, thus offering concrete insights for designing self-supervised pretraining strategies that readily complement existing architectural innovations.

One main limitation warrants discussion: our pretraining datasets (TUEG: 27,000 hours; HBN: 1,500 hours) are smaller than those used by some recent EEG foundation models, leaving open questions about how contrastive objectives scale with data. EEG FMs hold substantial promise for clinical and societal impact: by learning transferable representations from large unlabeled corpora, they can lower the data requirements for downstream applications, potentially expanding access to neurological screening in settings where expert-labeled data and trained clinicians are scarce. A natural next step is therefore scaling contrastive pretraining to substantially larger and more heterogeneous corpora that span clinical EEG (e.g., TUEG), developmental cohorts (e.g., HBN), sleep recordings (e.g., MASS, SHHS), and other BCI datasets. To do so robustly raises several concrete challenges: harmonizing across heterogeneous montages and channel counts (from 4-channel portable devices to 256-channel research caps), reconciling sampling rates and reference schemes, accounting for site- and device-specific noise distributions, and designing sampling strategies that balance data sources and discourage reliance on site-specific shortcuts.

AI USE STATEMENT

We used generative AI tools to assist with coding and implementation, provide feedback on experimental design, and aid with writing and polishing the manuscript. The authors take full responsibility for the methods, results, and final content.

ETHICS STATEMENT

This work studies EEG representation learning using existing research datasets and does not involve new participant recruitment or data collection. Improved EEG decoding may support brain–computer interfaces and clinical research, but EEG recordings can contain sensitive information about health and cognitive states. Their use requires appropriate consent, access controls, and respect for the original datasets' conditions of use. CoCoT-EEG has not been validated for autonomous clinical decision-making, and any deployment would require application-specific validation and safeguards against unauthorized inference of sensitive information.

REPRODUCIBILITY STATEMENT

Section 3 describes the model architecture and pretraining objectives, and Section 4.1 summarizes the evaluation protocol. Appendix B details the MoCo and MAE implementations and preprocessing, while Appendix D specifies the augmentation procedures. Downstream dataset descriptions are provided in Appendix E, and task-specific training settings are reported in Appendix F. Appendix M documents the controlled diagnostics. Results are reported with the uncertainty estimates specified in the corresponding tables and figure captions, and published baseline results are attributed to their original sources.

REFERENCES

Diego Alvarez-Estevez and Roselyne M Rijsman. Inter-database validation of a deep learning approach for automatic sleep scoring. *PloS One*, 16(8):e0256111, 16 August 2021. doi: 10.1371/journal.pone.0256111.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. Wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv [cs.CL]*, June 2020.

Randall Balestriero and Yann LeCun. Learning by reconstruction produces uninformative features for perception. *arXiv [cs.CV]*, February 2024.

Konstantinos Barmpas, Na Lee, Alexandros Koliousis, Yannis Panagakis, Dimitrios Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. Neurorvq: Multi-scale eeg tokenization for generative large brainwave models. 2025.

Jingjing Chen, Xiaobin Wang, Chen Huang, Xin Hu, Xinke Shen, and Dan Zhang. A large finer-grained affective computing eeg dataset. *Scientific Data*, 10(1):740, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv [cs.LG]*, February 2020.

Mike X Cohen. *Analyzing neural time series data: Theory and practice*. The MIT Press, January 2014.

Paolo Detti, Garazi Zuriarrain Martin de Lara, Marcos Santos, and Olivia Bruni. Eeg synchronization analysis for seizure prediction: A study on patients with drug-resistant epilepsy. *Processes*, 8(7), 2020. ISSN 2227-9717. doi: 10.3390/pr8070846. URL `https://mdpi.com`.

Thomas Donoghue, Matar Haller, Erik J Peterson, Paroma Varma, Priyadarshini Sebastian, Richard Gao, Torben Noto, Antonio H Lara, Joni D Wallis, Robert T Knight, Avgusta Shestyuk, and Bradley Voytek. Parameterizing neural power spectra into periodic and aperiodic components. *Nature neuroscience*, 23(12): 1655–1665, December 2020. doi: 10.1038/s41593-020-00744-x.

Yassine El Ouahidi, Vincent Gripon, Bastien Pasdeloup, Ghaith Bouallegue, Nicolas Farrugia, and Giulia Lioi. A strong and simple deep learning baseline for BCI motor imagery decoding. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 32:3338–3347, September 2024.

A L Goldberger, L A Amaral, L Glass, J M Hausdorff, P C Ivanov, R G Mark, J E Mietus, G B Moody, C K Peng, and H E Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):E215–20, June 2000.

Biyu J He. Scale-free brain activity: past, present, and future. *Trends in cognitive sciences*, 18(9):480–487, September 2014. doi: 10.1016/j.tics.2014.04.003.

He He and Dongrui Wu. Transfer learning for brain–computer interfaces: A euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*, 67(2):399–410, 2020. doi: 10.1109/TBME.2019.2913914.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv [cs.CV]*, November 2019.

Thorir Mar Ingolfsson, Michael Hersche, Xiaying Wang, Nobuaki Kobayashi, Lukas Cavigelli, and Luca Benini. EEG-TCNet: An accurate temporal convolutional network for embedded motor-imagery brain-machine interfaces. *arXiv [eess.SP]*, May 2020.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. *arXiv [cs.LG]*, May 2024.

Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. ISRUC-sleep: A comprehensive public dataset for sleep researchers. *Computer Methods and Programs in Biomedicine*, 124:180–192, 1 February 2016. doi: 10.1016/j.cmpb.2015.10.013.

Aditya Kommineni, Emily Zhou, Kleanthis Avramidis, Simon Bock Segaard, Jeppe Roden Münster, Andreas Peter Juhl Hansen, Takfarinas Medani, Tiantian Feng, Richard Leahy, and Shrikanth Narayanan. Aperiodic and low-frequency spectral bias in reconstruction based eeg foundation models. *arXiv preprint arXiv:2605.26434*, 2026.

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. *Front. Hum. Neurosci.*, 15:653659, June 2021.

Gayal Kuruppu, Neeraj Wagh, Vaclav Kremen, and Yogatheesan Varatharajah. EEG foundation models: a critical review of current progress and future directions. *Journal of Neural Engineering*, 23(2):021001, 10 March 2026. doi: 10.1088/1741-2552/ae4455.

Chi Qin Lai, Haidi Ibrahim, Mohd Zaid Abdullah, Jafri Malin Abdullah, Shahrel Azmin Suandi, and Azlinda Azman. Artifacts and noise removal for electroencephalogram (EEG): A literature review. In *2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, pp. 326–332. IEEE, April 2018.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. EEGNet: A compact convolutional network for EEG-based brain-computer interfaces. *arXiv [cs.LG]*, November 2016.

Ang Li, Zikai Wang, Liuyin Yang, Zhenyu Wang, Tianheng Xu, Honglin Hu, and Marc M. Van Hulle. Comet: A contrastive-masked brain foundation model for universal eeg representation, 2025. URL https://arxiv.org/abs/2509.00314.

Dingkun Liu, Yuheng Chen, Zhu Chen, Zhenyao Cui, Yaozhi Wen, Jiayu An, Jingwei Luo, and Dongrui Wu. EEG-FM-compass: Progress, benchmarking, and future directions for EEG foundation models. *arXiv [cs.LG]*, 25 January 2026. doi: 10.48550/arXiv.2601.17883.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 2021.

Jingying Ma, Feng Wu, Qika Lin, Yucheng Xing, Chenyu Liu, Ziyu Jia, and Mengling Feng. CodeBrain: Bridging decoupled tokenizer and multi-scale architecture for EEG foundation model. In *The Fourteenth International Conference on Learning Representations*, 8 October 2025.

Jun Ma, Banghua Yang, Wenzheng Qiu, Yunzhe Li, Shouwei Gao, and Xinxing Xia. A large EEG dataset for studying cross-session variability in motor imagery brain-computer interface. *Sci. Data*, 9(1):531, September 2022.

Iyad Obeid and Joseph Picone. The temple university hospital EEG data corpus. *Front. Neurosci.*, 10:196, May 2016.

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 [cs.LG]*, October 2025.

Wei Yan Peh, Yuanyuan Yao, and Justin Dauwels. Transformer convolutional neural networks for automated artifact detection in scalp EEG. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2022: 3599–3602, July 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv [cs.CV]*, February 2021.

Seyed Yahya Shirazi, Alexandre Franco, Maurício Scopel Hoffmann, Nathalia B Esper, Dung Truong, Arnaud Delorme, Michael P Milham, and Scott Makeig. HBN-EEG: The FAIR implementation of the healthy brain network (HBN) electroencephalography dataset. *bioRxiv*, pp. 2024.10.03.615261, October 2024.

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

Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie. Transformer-based spatial-temporal feature learning for EEG decoding. *arXiv [eess.SP]*, June 2021.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J Miller, Gernot R Müller-Putz, et al. Review of the bci competition iv. *Frontiers in neuroscience*, 6:55, 2012.

Hugues Van Assel, Mark Ibrahim, Tommaso Biancalani, Aviv Regev, and Randall Balestriero. Joint embedding vs reconstruction: Provable benefits of latent space prediction for self supervised learning. *arXiv [cs.LG]*, October 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv [cs.LG]*, July 2018.

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. *arXiv [eess.SP]*, December 2024.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *arXiv [cs.LG]*, May 2020.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. *arXiv [cs.CV]*, 2 January 2023. doi: 10.48550/arXiv.2301.00808.

Chaoqi Yang, M Brandon Westover, and Jimeng Sun. BIOT: Cross-data biosignal learning in the wild. *arXiv [eess.SP]*, May 2023a.

Chaoqi Yang, Cao Xiao, M Brandon Westover, and Jimeng Sun. Self-supervised electroencephalogram representation learning for automatic sleep staging: Model development and evaluation study. *JMIR AI*, 2(1): e46769, January 2023b.

Wei-Long Zheng and Bao-Liang Lu. A multimodal approach to estimating vigilance using EEG and forehead EOG. *J. Neural Eng.*, 14(2):026017, April 2017.

Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. CSBrain: A cross-scale spatiotemporal brain foundation model for EEG decoding. *arXiv [cs.HC]*, June 2025.

A CoCoT-EEG Architecture and Pretraining Schematic

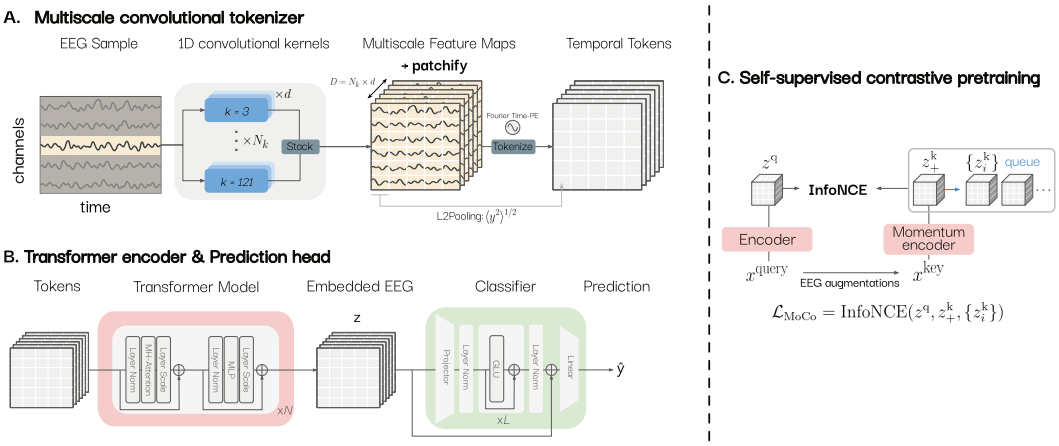

Figure 4: **CoCoT-EEG. A.** Raw EEG signals are processed through a bank of 1D convolutions with varying kernel sizes ($k = 3$ to 121). The resulting multiscale feature maps are patchified into temporal tokens via L2 pooling, followed by the addition of a Fourier positional encoding. **B.** Self-attention is applied to the tokens to produce EEG embeddings (z). A residual MLP head with two GLU blocks is used for the downstream classification or regression. **C.** For pretraining, a MoCo-style contrastive learning framework He et al. (2019) is employed. The encoder and momentum encoder produce embeddings (z^q and z^k_+), which are trained with InfoNCE van den Oord et al. (2018) against a queue of negative samples z^k_i. GLU: gated linear unit; MH: multi-head, MoCo: momentum contrast; N: number of transformer blocks; PE: positional encoding.

B MoCo and MAE Pretraining

Shared preprocessing. Both pretraining methods operate on the same preprocessed TUEG corpus (Obeid & Picone, 2016), using 19 EEG channels sampled at 100 Hz and segmented into non-overlapping 5 s windows. We apply channel-wise robust normalization:

$$\widetilde{X}_{c,t} = \frac{X_{c,t} - a_c}{\alpha\,\mathrm{MAD}_c + \varepsilon}, \qquad \varepsilon = 10^{-6}, \tag{3}$$

where a_c is the median for channel c, MAD_c is the median absolute deviation from a_c, and $\alpha = 1/\Phi^{-1}(0.75) \approx 1.4826$, with Φ denoting the standard normal cumulative distribution function. Under Gaussianity, $\alpha\,\mathrm{MAD}_c$ consistently estimates the standard deviation, making this normalization asymptotically equivalent to standard z-scoring, apart from the numerical stabilizer. Statistics are estimated once from training windows, with a configured maximum of 50,000 windows, and held fixed thereafter. Normalization precedes augmentation for MoCo and masking for MAE.

Shared training setup. For the pretraining objective ablation, both methods use the full-kernel Conv-first encoder, with 12 Transformer layers, embedding dimension 512, 8 attention heads, and temporal positional encoding. During pretraining, a learned token-wise linear projection of dimension $512 \to 512$ is applied to the Transformer outputs. Both methods are trained for 150,000 optimizer updates with a global batch size of 1,336. We use AdamW with learning rate 10^{-4}, weight decay 10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and optimizer stabilizer 10^{-8}. Biases, normalization parameters, and other one-dimensional parameters are excluded from weight decay. The learning rate increases linearly from 10^{-6} to 10^{-4} during the first 3,000 updates and remains constant thereafter. Downstream finetuning uses the momentum backbone for MoCo and the trained backbone for MAE, after discarding the pretraining heads and the encoder's pretraining output projection. A summary of the pretraining and architecture parameters can be found in Table 3.

Contrastive pretraining (MoCo). MoCo uses two independently augmented views of each normalized EEG window, following the configurations described in Appendix D. The query and momentum models each contain the encoder, an attention pooler, and a projection head. Given the projected token features $U = (u_1, \ldots, u_N)$, with $N = CS$ for C channels and S patches per channel, attention pooling produces

$$\alpha_i = \frac{\exp(w^\top u_i + b)}{\sum_{j=1}^{N} \exp(w^\top u_j + b)}, \qquad z = \sum_{i=1}^{N} \alpha_i u_i, \tag{4}$$

where w and b are learned parameters. A two-layer MLP maps the pooled representation from 512 to 2048 and then to 128 dimensions, with BatchNorm and ReLU between the two linear layers. The projected query and key vectors are ℓ_2-normalized when computing the InfoNCE objective described in Section 3.4. We use a temperature of $\tau = 0.07$ and a queue containing $K = 65{,}536$ negative keys. The queue is initialized with normalized random vectors and updated with the current momentum keys, gathered across workers, replacing the oldest entries. Gradients update the query model only and after each optimizer step, the momentum model parameters, including the pooling and projection heads, are updated by exponential moving average with coefficient $m = 0.999$.

Masked reconstruction (MAE). Applying masked autoencoding to a Conv-first tokenizer requires preventing hidden waveform samples from entering visible token representations through overlapping convolutional receptive fields. Masking tokens only after convolution would therefore permit information leakage. We address this using sparse, visibility-aware convolutional tokenization, which preserves the full multiscale kernel bank while restricting feature extraction to visible waveform samples. This allows us to compare MoCo and MAE using the same underlying encoder architecture without restricting convolutions to individual patches.

MAE uses a single normalized waveform without the input augmentations used by MoCo. Each channel is divided into patches of $P = 50$ samples, corresponding to $0.5\,\mathrm{s}$. A $5\,\mathrm{s}$ input therefore contains $S = 10$ patches per channel and $CS = 190$ channel–time patches. For each input and forward pass, we uniformly sample 95 patches without replacement to mask, leaving 95 visible tokens. Sampling occurs across the entire channel–time grid. The patch mask is expanded to the sample level and applied before convolution. Hidden samples are set to zero, convolutional outputs at hidden positions are kept inactive, and GroupNorm statistics are computed using visible temporal positions within each sample, electrode, and feature group. The tokenizer then applies GELU and the usual non-overlapping L2 pooling. Temporal positional encodings retain the original patch positions, and only visible tokens enter the Transformer encoder. Sparsity is implemented through explicit input and feature masks using dense convolution operations.

The encoded visible tokens are restored to their original positions in the complete channel–time grid, while hidden positions receive a shared learned mask token. We add learned electrode-identity embeddings and fixed temporal Fourier positional encodings to all decoder positions. The electrode embeddings identify the 19 TUEG channels and belong exclusively to the decoder; the retained encoder uses temporal positional encoding only. The decoder contains four Transformer blocks of width 512 with eight attention heads. Following the decoder design used in REVE (Ouahidi et al., 2025), each block uses pre-RMSNorm, full self-attention, and a GEGLU feed-forward residual branch. The feed-forward hidden dimension is 1360, and the RMSNorm numerical stabilizer is 10^{-6}. The decoder uses no dropout. A final linear layer maps each token to its 50 waveform samples.

The reconstruction target is the normalized input waveform, reshaped into patches without additional per-patch normalization. Let $\mathcal{M}_b$ denote the set of masked patches for input b, with $|\mathcal{M}_b| = M = 95$. We minimize the mean absolute reconstruction error over masked waveform samples:

$$\mathcal{L}_{\mathrm{MAE}} = \frac{1}{BMP} \sum_{b=1}^{B} \sum_{j \in \mathcal{M}_b} \sum_{p=1}^{P} \left| \widehat{x}_{b,j,p} - x_{b,j,p} \right|, \tag{5}$$

where B is the batch size, $x_{b,j,p}$ is a target sample, and $\widehat{x}_{b,j,p}$ is its reconstruction.

Table 3: Pretraining hyperparameters.

HYPERPARAMETER	VALUES
Optimization	
LEARNING RATE	1×10^{-4}
BATCH SIZE	1336
WEIGHT DECAY	1×10^{-4}
WARMUP STEPS	3000
MAX STEPS	150K
MoCo Loss	
TEMPERATURE	0.07
QUEUE SIZE	65536
FEATURE DIMENSION	128
CoCoT-EEG	
PATCH SIZE	50
MODEL DIMENSION	512
FFN DIMENSION	2048
LAYERS	12
ATTENTION HEADS	8
DROP PATH RATE	0.1
Projector	
HIDDEN DIMENSION	2048
OUTPUT DIMENSION	128
LAYERS	2
Data	
SAMPLING RATE	100 HZ
WINDOW SIZE	5 SECONDS
WINDOW STRIDE	5 SECONDS

C PRETRAINING AND MODEL HYPERPARAMETERS

D EEG-SPECIFIC AUGMENTATIONS

To construct positive pairs for contrastive pretraining, we apply stochastic compositions of augmentations designed for EEG signals. These transformations either mimic common acquisition artifacts and noise sources encountered in EEG recordings, or encourage invariance to nuisance variability while preserving task-relevant neural activity.

Composition. Each view is produced by offering every transform in the pipeline independently with probability $p = 0.40$, in the order listed below.

Time masking. A contiguous temporal span is masked by setting the signal to zero, with length sampled uniformly between 10% and 25% of the sequence. This encourages robustness to transient artifacts and partial temporal occlusions.

Time reversal. A random temporal window, covering between 10% and 50% of the sequence, is reversed in time. This perturbs local temporal ordering while preserving the amplitude distribution and spectral content of the affected span.

Channel dropout. Channels are zeroed independently with probability $p_{\text{drop}} = 0.28$, simulating corrupted or disconnected electrodes and encouraging robustness to missing sensors.

Phase jitter. Random perturbations are applied to the STFT phase while preserving spectral magnitudes, with phase offsets sampled uniformly between 5° and 110°. This introduces controlled temporal distortions while approximately preserving the signal power spectrum.

Table 4: Overview of downstream task datasets.

Dataset	Task	#Ch.	Dur.	#Samp.	#Cls.
BCI-IV	Motor Im.	22	4s	5K	4
PHYSIONET	Motor Im.	64	4s	10K	4
SHU	Motor Im.	32	4s	12K	2
FACED	Emotion	32	10s	10K	9
SEED-V	Emotion	62	1s	118K	5
CHB-MIT	Seizure	16	10s	327K	2
SIENA	Seizure	29	10s	51k	2
TUAB	Abnormal	16	10s	409K	2
SEED-VIG	Vigilance	17	8s	20K	Regr.
HMC	Sleep Staging	4	30s	137K	4
ISRUC	Sleep Staging	6	30s	90K	5

Spectral corruption. Frequency components are masked in the Fourier domain, with the masked fraction of the spectrum fixed to 0.45. This encourages robustness to frequency-specific corruption and reduces dependence on isolated spectral components.

Fourier-transform surrogate. Fourier phases are randomized by adding uniform noise of magnitude 0.13 (as a fraction of 2π) to all frequency components, producing a surrogate signal with identical power spectrum but scrambled phase relationships.

E Downstream task datasets

We evaluate on eleven datasets spanning motor imagery classification, emotion recognition, seizure detection, abnormality detection, sleep staging, and vigilance estimation. Table 4 summarizes the tasks, input channel counts, window durations, sample counts, and prediction targets. All signals are resampled to 100 Hz.

Motor Imagery Classification. We evaluate on three motor imagery datasets: `BCI-IV` (BCI Competition IV-2a) Tangermann et al. (2012), `PHYSIO` (PhysioNet Motor Movement/Imagery) Goldberger et al. (2000), and `SHU-MI` Ma et al. (2022). BCI Competition IV-2a contains 22-channel EEG from 9 subjects performing four motor imagery tasks (left hand, right hand, feet, tongue). PhysioNet-MI comprises 64-channel recordings from 109 subjects across four classes (left fist, right fist, both fists, both feet). SHU-MI provides 32-channel binary classification data (left/right hand) from 25 subjects.

Emotion Classification. We use `FACED` Chen et al. (2023) and `SEED-V` Liu et al. (2021). FACED is a large-scale affective computing dataset with 32-channel EEG from 123 subjects, covering nine fine-grained emotion categories (amusement, inspiration, joy, tenderness, anger, fear, disgust, sadness, neutral). SEED-V contains 62-channel EEG from 16 subjects across three sessions, with five emotion categories (happy, sad, neutral, disgust, fear).

Seizure Detection. We evaluate binary seizure detection on `CHB-MIT` (CHB-MIT Scalp EEG Database) Shoeb (2009) and `SIENA` Detti et al. (2020). CHB-MIT contains long-term scalp EEG recordings from pediatric patients with intractable seizures, together with seizure onset and offset annotations. Our benchmark representation uses 16 bipolar EEG derivations and ten-second windows. SIENA contains clinical scalp EEG recordings from 14 patients with epilepsy, acquired at the University of Siena.

Abnormality Detection. `TUAB` is part of the Temple University database Obeid & Picone (2016), where 16-channel EEG segments are labeled as pathological or non-pathological, similar to seizure detection but in a more general medical context.

Sleep staging. We use `HMC` Alvarez-Estevez & Rijsman (2021) and `ISRUC` Khalighi et al. (2016) for sleep stage classification evaluation. HMC contains 151 whole-night polysomnographic (PSG)

recordings collected from a heterogeneous population referred for clinical sleep assessment. For ISRUC, we use subgroup I, comprising recordings from 100 participants with sleep disorders, with six EEG channels. For both datasets, inputs are 30-second epochs.

Vigilance Estimation. The `SEED-VIG` dataset Zheng & Lu (2017) provides continuous vigilance labels derived from PERCLOS (percentage of eye closure) during simulated driving. This 17-channel dataset enables regression-based evaluation of drowsiness estimation, collected from 23 subjects over approximately 2-hour driving sessions.

Evaluation Partitions. To compare with existing EEG foundation models, we follow dataset-specific benchmark protocols established in prior work, including CBraMod Wang et al. (2024), CSBrain Zhou et al. (2025), and REVE Ouahidi et al. (2025). For each dataset, we use fixed training, validation, and test partitions that remain unchanged across training seeds. Published baseline results are taken from the cited studies under their respective evaluation settings.

F FINETUNING HYPERPARAMETERS

Table 5: Finetuning hyperparameters.

HYPERPARAMETER	VALUE
Optimization	
LEARNING RATE	1×10^{-4}
–	EXCEPT CHB-MIT = 1×10^{-7}, SIENA AND ISRUC = 1×10^{-5})
BATCH SIZE	16 (EXCEPT CHB-MIT AND SEEDV = 64)
WEIGHT DECAY	5×10^{-2}
EARLY STOPPING PATIENCE	6 EPOCHS
Classification head	
HIDDEN DIMENSION	1024
LAYERS	2
Data	
SAMPLING RATE	100 HZ
SEQUENCE LENGTH	DEPENDS ON THE DATASET

G COMPUTING RESOURCES

Pretraining on either TUEG or HBN was performed on a single H100-SXM-8-80G GPU machine, for 150,000 steps. Both main pretraining runs accounted for a total of 20 hours. Including exploration and ablation experiments, around 30 pretrainings were run in total. Finetuning on all the downstream datasets was done using a H100-SXM-8-80G GPU machine and would take around 4 hours per checkpoint. Accounting for all the ablations ran in the paper, around 100 hours of H100-SXM-8-80G was used for finetuning.

H COMPLETE DOWNSTREAM RESULTS

Table 6 extends the main comparison with additional task-specific metrics and results for the MAE-pretrained CoCoT-EEG variant across the eleven downstream tasks. Baseline results are taken from their respective sources, with missing entries indicating unavailable results. To summarize relative performance, we report mean ranks based on the primary metric for the six datasets with results available for all seven models. Rank 1 denotes the best performance.

Table 6: Downstream performance across eleven EEG benchmarks. All values are percentage points (SEED-VIG RMSE scaled by 100); **bold** marks the best model in each row, _underlined marks second best score. Baselines are reported as mean ± standard deviation over 5 runs (REVE-Base, 10 hyperparameter search runs), as published. CoCoT is mean ± 95% confidence-interval half-width over 4 seeds. "–" indicates a value the source paper does not report. ⋆ denotes runs using generalization domain-adaptation techniques such as Euclidean Alignment (He & Wu, 2020).

Dataset	Metric	LaBraM-B.	CBraMod	REVE-B.	CSBrain	CoCoT-MAE	**CoCoT**
BCI-IV	Balanced acc.	48.69 ± 0.85	51.38 ± 0.66	$63.96 \pm 0.95^{\star}$	56.57 ± 0.71	54.84 ± 2.29	$\mathbf{68.42 \pm 1.89}^{\star}$
	Cohen's κ	31.59 ± 1.54	35.18 ± 0.94	$51.94 \pm 1.26^{\star}$	42.09 ± 0.93	39.79 ± 3.05	$\mathbf{57.89 \pm 2.58}^{\star}$
	Weighted F1	47.58 ± 1.03	49.84 ± 0.85	$63.39 \pm 1.10^{\star}$	56.37 ± 0.87	53.06 ± 3.29	$\mathbf{67.99 \pm 2.10}^{\star}$
FACED	Balanced acc.	52.73 ± 1.07	55.09 ± 0.89	56.46 ± 1.64	57.52 ± 0.42	56.81 ± 1.41	$\mathbf{58.76 \pm 3.06}$
	Cohen's κ	46.98 ± 1.88	50.41 ± 1.22	50.80 ± 1.91	52.04 ± 0.36	50.85 ± 1.68	$\mathbf{53.13 \pm 3.30}$
	Weighted F1	52.88 ± 1.02	56.18 ± 0.93	56.59 ± 1.72	57.96 ± 0.31	56.41 ± 1.57	$\mathbf{58.54 \pm 2.96}$
HMC	Balanced acc.	72.77 ± 1.01	72.69 ± 0.41	74.01 ± 0.75	73.45 ± 0.47	76.68 ± 0.37	$\mathbf{78.51 \pm 0.26}$
	Cohen's κ	68.13 ± 0.53	66.85 ± 1.04	69.82 ± 0.78	68.18 ± 0.46	70.28 ± 0.37	$\mathbf{72.08 \pm 0.49}$
	Weighted F1	74.54 ± 0.27	73.95 ± 0.89	76.38 ± 0.74	75.06 ± 0.42	77.15 ± 0.29	$\mathbf{78.58 \pm 0.37}$
ISRUC	Balanced acc.	76.33 ± 1.02	78.65 ± 1.10	78.19 ± 0.78	79.25 ± 0.30	79.45 ± 0.05	$\mathbf{80.99 \pm 0.26}$
	Cohen's κ	72.31 ± 1.82	74.42 ± 1.52	75.00 ± 1.56	74.06 ± 1.02	74.80 ± 0.20	$\mathbf{76.63 \pm 0.08}$
	Weighted F1	78.10 ± 1.33	80.11 ± 0.99	80.05 ± 1.35	79.90 ± 0.91	80.77 ± 0.09	$\mathbf{81.66 \pm 0.12}$
PHYSIO	Balanced acc.	61.73 ± 1.22	64.17 ± 0.91	64.80 ± 1.40	63.04 ± 0.90	64.34 ± 1.08	$\mathbf{66.90 \pm 0.37}$
	Cohen's κ	49.12 ± 1.92	52.22 ± 1.69	53.06 ± 1.87	50.71 ± 1.20	52.45 ± 1.43	$\mathbf{55.87 \pm 0.49}$
	Weighted F1	61.77 ± 1.41	64.27 ± 1.00	64.84 ± 1.70	63.08 ± 0.95	64.44 ± 1.21	$\mathbf{67.04 \pm 0.36}$
SEED-V	Balanced acc.	39.76 ± 1.38	40.91 ± 0.97	–	41.97 ± 0.33	42.29 ± 0.73	$\mathbf{43.40 \pm 0.33}$
	Cohen's κ	23.86 ± 2.09	25.69 ± 1.43	–	27.85 ± 0.34	27.80 ± 0.85	$\mathbf{29.22 \pm 0.38}$
	Weighted F1	39.74 ± 1.11	41.01 ± 1.08	–	42.80 ± 0.23	42.68 ± 0.58	$\mathbf{43.83 \pm 0.28}$
CHB-MIT	Balanced acc.	70.75 ± 3.58	73.98 ± 2.84	–	72.62 ± 1.15	84.78 ± 2.03	$\mathbf{87.98 \pm 0.53}$
	AUC-PR	32.87 ± 4.02	36.89 ± 3.82	–	51.64 ± 4.49	44.88 ± 1.23	$\mathbf{53.13 \pm 1.79}$
	AUROC	86.79 ± 1.99	88.92 ± 1.54	–	89.15 ± 3.21	90.53 ± 0.06	$\mathbf{91.60 \pm 0.53}$
SIENA	Balanced acc.	70.82 ± 3.29	73.17 ± 6.47	–	76.62 ± 4.71	77.77 ± 1.46	$\mathbf{80.47 \pm 1.76}$
	AUC-PR	31.22 ± 9.76	41.07 ± 7.20	–	$\mathbf{48.71 \pm 3.43}$	37.80 ± 1.21	43.31 ± 3.05
	AUROC	88.14 ± 3.28	90.38 ± 2.18	–	90.76 ± 1.19	91.63 ± 0.49	$\mathbf{91.81 \pm 0.52}$
SHU	Balanced acc.	61.66 ± 1.92	63.70 ± 1.51	–	64.17 ± 0.37	64.45 ± 0.31	$\mathbf{64.93 \pm 1.79}$
	AUC-PR	67.61 ± 0.83	71.39 ± 0.88	–	72.30 ± 0.79	73.10 ± 0.41	$\mathbf{74.20 \pm 1.36}$
	AUROC	66.04 ± 0.91	69.88 ± 0.68	–	72.00 ± 0.58	71.54 ± 0.47	$\mathbf{73.00 \pm 0.91}$
TUAB	Balanced acc.	81.40 ± 0.19	82.49 ± 0.25	$\mathbf{83.15 \pm 0.14}$	81.72 ± 0.43	82.02 ± 0.45	81.35 ± 0.91
	AUC-PR	89.65 ± 0.16	92.21 ± 0.15	$\mathbf{92.81 \pm 0.09}$	90.05 ± 0.66	89.96 ± 0.43	89.70 ± 0.93
	AUROC	90.22 ± 0.09	91.56 ± 0.17	$\mathbf{92.45 \pm 0.13}$	89.57 ± 0.46	89.16 ± 0.40	88.96 ± 0.89
SEED-VIG	Pearson r	63.47 ± 1.35	55.02 ± 1.15	–	63.14 ± 3.56	64.02 ± 2.42	$\mathbf{65.63 \pm 1.32}$
	R^2	18.08 ± 9.58	7.37 ± 1.67	–	23.63 ± 5.19	25.43 ± 5.57	$\mathbf{29.25 \pm 4.54}$
	RMSE ↓	28.71 ± 1.66	30.57 ± 0.27	–	27.74 ± 0.94	27.42 ± 1.02	$\mathbf{26.71 \pm 0.86}$
Mean primary-metric ranks on six shared datasets							
Mean rank ↓		4.67	3.50	2.33	2.83	2.83	**1.83**

I 3D Spatial Positional Encoding

To evaluate whether explicit electrode topology improves representation learning, we augment CoCoT-EEG with a 3D spatial positional encoding (3DPE) inspired by REVE Ouahidi et al. (2025). Unlike REVE's joint 4D spatiotemporal encoding, we encode only spatial coordinates; temporal patch positions are handled separately using a temporal Fourier positional encoding.

Each EEG channel c is associated with a standardized 3D electrode coordinate

$$\mathbf{r}_c = (x_c, y_c, z_c) \in \mathbb{R}^3.$$

For every coordinate, we compute a Fourier embedding over the Cartesian product of spatial frequencies:

$$\gamma(\mathbf{r}_c) = [\cos(\theta), \sin(\theta)]\,,$$

where the phase θ is constructed from frequency combinations over (x_c, y_c, z_c), following the spatial component of REVE. We use $F = 4$ frequencies per dimension, resulting in $2F^3 = 128$ Fourier features per channel. When this dimensionality differs from the transformer dimension, the Fourier embedding is projected with a bias-free linear layer.

Following REVE, the Fourier representation is combined with a learnable projection of the raw coordinates:

$$\mathbf{p}_c^{\text{spatial}} = \text{LayerNorm}\left(W_\gamma \gamma(\mathbf{r}_c) + \text{MLP}(\mathbf{r}_c)\right),$$

where MLP is implemented as a bias-free linear layer followed by GELU and LayerNorm. During training, we also add small Gaussian noise to the coordinates before encoding.

After multiscale convolution and temporal patchification, the spatial embedding is added to each token together with the temporal Fourier positional encoding:

$$\tilde{\mathbf{h}}_{b,c,s} = \mathbf{h}_{b,c,s} + \mathbf{p}_c^{\text{spatial}} + \mathbf{p}_s^{\text{temporal}}.$$

To test whether the model truly exploits meaningful electrode topology rather than simply relying on additional channel-identifying embeddings, we evaluate a shuffled-position control setting. In this condition, electrode coordinates are randomly permuted once at finetuning initialization while preserving the original EEG channel order. This destroys the anatomical correspondence between electrodes and spatial coordinates while retaining individual channel identifiers.

Table 7: Spatial positional encoding ablation. We compare the default topology-agnostic CoCoT-EEG pretrained on TUEG (**No-3DPE**) against variants using a 3D spatial positional encoding (**3DPE**) and a control condition where electrode positions are randomly shuffled once before finetuning (**3DPE + Shuffled Pos.**). Results are reported as balanced accuracy (%) except for SEED-VIG (Pearson correlation $\times$ 100); mean $\pm$ 95% CI over 4 seeds. **Bold** marks the best mean per row.

Dataset	No-3DPE	3DPE + Shuffled Pos.	3DPE
BCI-IV	$\mathbf{59.25 \pm 2.12}$	57.75 ± 3.27	57.90 ± 1.54
FACED	58.76 ± 3.06	58.86 ± 3.95	$\mathbf{60.08 \pm 1.10}$
PHYSIO	66.90 ± 0.37	66.71 ± 0.92	$\mathbf{68.19 \pm 0.98}$
SEED-V	$\mathbf{43.40 \pm 0.33}$	42.38 ± 0.51	43.01 ± 1.10
SHU	64.93 ± 1.79	65.92 ± 0.85	$\mathbf{66.42 \pm 1.08}$
SEED-VIG	65.63 ± 1.32	$\mathbf{66.08 \pm 1.77}$	63.13 ± 2.66
HMC	78.51 ± 0.26	$\mathbf{78.84 \pm 0.22}$	78.83 ± 0.26
ISRUC	80.99 ± 0.26	80.91 ± 0.42	$\mathbf{81.15 \pm 0.15}$
CHB-MIT	87.98 ± 0.53	$\mathbf{88.31 \pm 1.12}$	87.00 ± 1.28
TUAB	81.35 ± 0.91	81.52 ± 0.32	$\mathbf{82.30 \pm 0.81}$

J From scratch Training Ablation of Truncated Architectures

Table 8: Scratch training comparison between truncated Conv-first and Patch-first architectures. Balanced accuracy (%) is reported for all tasks except SEED-VIG (Pearson correlation $\times 100$). Values are mean $\pm$ 95% CI. Bold denotes the best result per row.

Dataset	Patch-first Scratch	Conv-first Scratch
BCI-IV	46.83 $\pm$ 2.34	**48.09** $\pm$ 2.88
FACED	50.39 $\pm$ 0.95	**53.75** $\pm$ 0.76
PHYSIO	60.15 $\pm$ 0.48	**61.23** $\pm$ 0.71
SEED-V	39.92 $\pm$ 0.94	**40.45** $\pm$ 0.16
CHB-MIT	51.57 $\pm$ 4.99	**53.77** $\pm$ 5.62
SHU	**61.46** $\pm$ 2.27	60.84 $\pm$ 1.07
SEED-VIG	**63.75** $\pm$ 3.56	57.40 $\pm$ 5.23

K From scratch, single-task results

Table 9: Performance comparison including additional baselines from secondary sources. Balanced accuracy is displayed in %. mean ± std is reported for the baselines and mean ±95% CI is reported for CoCoT. Baseline performance is highlighted in light gray. $\star$ denotes SimpleConv runs that use additional finetuning generalization techniques such as Mixup and Euclidean Alignment (El Ouahidi et al., 2024).

	Baselines				CoCoT-EEG
Dataset	ContraWR	CNN-Trans	ST-Trans	SimpleConv	From Scratch
BCIIV	46.78	46.00	45.75	**62.38*** $\pm$ **2.38**	54.40 $\pm$ 2.99
CHB-MIT	63.44	**63.89**	59.15	59.78* $\pm$ 2.17	60.55 $\pm$ 0.79
FACED	48.87	46.97	48.10	39.51* $\pm$ 2.84	**55.52** $\pm$ **1.00**
PHYSIO	58.92	60.53	60.35	60.29* $\pm$ 0.62	**62.44** $\pm$ **1.31**
SEED-V	35.46	36.78	30.52	19.97* $\pm$ 0.04	**41.41** $\pm$ **0.44**
SHU	58.73	59.75	59.92	**62.92*** $\pm$ **1.33**	62.25 $\pm$ 1.65
TUAB	77.46	77.77	79.66	**81.74*** $\pm$ **0.39**	79.17 $\pm$ 0.14
SEED-VIG (Corr)	52.35	58.29	60.20	—	**60.37** $\pm$ **2.04**

L PRETRAINING DATASET ABLATION RESULTS

We compare pretraining on the full TUEG dataset (Obeid & Picone, 2016), on TUEG with TUAB recordings excluded (TUEG-Filt.), and a four tasks subset (contrast-change detection, surround suppression, resting state, and movie watching) of the HBN dataset (Shirazi et al., 2024).

Table 10: Effect of the pretraining corpus on CoCoT-EEG downstream performance. All models share the same architecture and finetuning protocol and differ only in the pretraining data: full TUEG, TUEG with TUAB recordings excluded, and HBN. Each cell reports the primary metric (top) and, in gray, the secondary metric (bottom), as indicated in the Metric column, in %. Mean ± 95% CI over seeds. **Bold** denotes the best mean per row.

DATASET	METRIC	TUEG PT	TUEG-FILT. PT	HBN PT
BCI-IV	B-ACC	59.25 ± 2.12	59.87 ± 3.63	**62.46** ± 0.70
	F1-W	57.80 ± 2.47	58.68 ± 4.14	61.78 ± 0.98
FACED	B-ACC	**58.76** ± 3.06	57.01 ± 4.89	57.04 ± 0.44
	F1-W	58.54 ± 2.96	56.85 ± 5.20	56.79 ± 0.82
HMC	B-ACC	78.51 ± 0.26	**78.54** ± 4.81	77.28 ± 0.74
	F1-W	78.58 ± 0.37	78.48 ± 7.17	77.28 ± 0.74
ISRUC	B-ACC	**80.99** ± 0.26	80.58 ± 1.65	79.49 ± 1.20
	F1-W	81.66 ± 0.12	81.03 ± 0.80	80.05 ± 0.91
PHYSIO	B-ACC	**66.90** ± 0.37	66.20 ± 0.73	65.32 ± 1.13
	F1-W	67.04 ± 0.36	66.29 ± 0.72	65.39 ± 1.25
SEED-V	B-ACC	43.40 ± 0.33	**43.55** ± 0.42	41.49 ± 0.60
	F1-W	43.83 ± 0.28	43.83 ± 0.51	41.53 ± 0.61
SHU	B-ACC	64.93 ± 1.79	65.29 ± 1.06	**65.42** ± 1.45
	AUROC	73.00 ± 0.91	73.25 ± 1.09	72.17 ± 1.30
TUAB	B-ACC	81.28 ± 0.31	81.35 ± 0.91	**81.41** ± 0.99
	AUROC	88.98 ± 0.33	88.96 ± 0.89	89.13 ± 0.45
SIENA	B-ACC	**80.47** ± 1.76	79.66 ± 3.73	79.65 ± 2.27
	AUROC	91.81 ± 0.52	92.21 ± 2.02	89.86 ± 1.56
SEEDVIG	CORR.	**65.63** ± 1.32	63.94 ± 3.39	64.12 ± 2.32
	R^2	29.25 ± 4.54	24.79 ± 6.36	28.16 ± 1.36
MACRO-AVG	B-ACC/CORR.	**68.01**	67.60	67.36

M INTERPRETABILITY ANALYSIS OF COCOT-EEG

M.1 TOKENIZER AND PRETRAINING LOSS ANALYSES

We detail here the three diagnostics in Fig. 2. All encoders are frozen at 150,000 pretraining steps and evaluated without stochastic training operations. Panels A and B use four-second PhysioNet-MI Cz recordings, resampled to 100 Hz, with 70/19/20 disjoint training/validation/test subjects (6,300/1,734/1,803 records). We apply our standard robust EEG normalization, subtracting the median and dividing by the scaled median absolute deviation, with statistics fitted on the training split. We use the median RMS of these normalized training windows, denoted R_{ref}, as a common amplitude reference for the injected signals.

A. Frequency decoding across patch boundaries. We compare independently pretrained Conv-first and Patch-first MoCo tokenizers with the same kernel bank $\{3, 5, 7, 9, 11, 17, 23, 31, 45\}$ and 50-sample patches (500 ms). A 160 ms cosine-tapered burst with random phase and carrier frequency sampled uniformly from 12–30 Hz is added to each EEG background. Bursts are zero-mean, with RMS over their 160 ms support sampled uniformly between $1.5R_{\mathrm{ref}}$ and $2.5R_{\mathrm{ref}}$. We examine the three internal boundaries of the four central patches, at 1.5, 2.0, and 2.5 s. For each boundary separately, the same burst is translated from -250 to $+250$ ms in 10 ms increments, while the encoder receives the complete four-second recording. Separate ridge probes predict frequency from the ordered concatenation of the two adjacent tokenizer tokens (1,024 features in total). Training samples two stratified offsets per boundary and record, yielding 37,800 examples. Before fitting the probes, embedding coordinates are z-scored and frequency targets are mean-centered using training statistics.

Regularization is selected from $\{10^{-4}, 10^{-3}, \ldots, 10^{4}\}$ by validation mean absolute error. Test errors are averaged over boundaries, records within runs, runs within subjects, and subjects equally.

B. Kernel spectral specialization. *Filter spectra.* For the full Conv-first MoCo model, we compute the Fourier power spectra of individually unit-L2-normalized filters, average powers within kernel size, and normalize each row by its maximum over 0.5–45 Hz.

Effective responses. We use one record from each of the 20 held-out PhysioNet-MI subjects. Four-second sinusoids spanning 0.5–45 Hz in 0.5 Hz steps are added at eight different phases, each scaled to RMS $0.25R_{\mathrm{ref}}$. For each phase, we measure the RMS token difference relative to the unmodified background, then average over phases. Features are extracted after convolution, GroupNorm, GELU, and pooling. We retain interior tokens 2–5 to avoid convolutional recording-edge padding. A branch comprises all filters of one kernel size. For each subject and branch, the response is divided by its mean across the frequency sweep. We then average branches equally within the short ($k < 50$) or long ($k > 50$) group, followed by averaging frequencies within bands and subjects. Bands are $\delta : [0.5, 4)$, $\theta : [4, 8)$, $\alpha : [8, 13)$, $\beta : [13, 30)$, and low $\gamma : [30, 45]$ Hz.

C. Oscillatory peak-height decoding. Inspired by Kommineni et al. (2026), we generate five-second signals at 100 Hz from

$$\log_{10} P(f) = 1 - 1.5 \log_{10} f + H \exp\left[-\frac{(f - f_0)^2}{2(0.5)^2}\right] + \epsilon(f), \tag{6}$$

where $H \sim \mathcal{U}(0.2, 0.8)$ and $\epsilon(f) \sim \mathcal{N}(0, 0.005^2)$. Spectra are evaluated on the native DFT grid within 0.5–45 Hz. We combine $\sqrt{P(f)}$ with random Fourier phases, apply the inverse FFT, and normalize each waveform to unit RMS. The finite-width peak produces a narrowband oscillatory component with a fluctuating envelope. The target H corresponds to a local spectral peak contrast of $10H =$ 2–8 dB above the aperiodic background. Peak centers are $\{2, 2.5, 3\}$, $\{5, 6, 7\}$, $\{9, 10.5, 12\}$, $\{15, 20, 25\}$, and $\{32, 37, 42\}$ Hz for the five bands, respectively. We generate 1,024/256/512 independent training/validation/test draws. Within each draw, height, spectral noise, and Fourier phases are reused across peak centers to enable paired frequency comparisons; both encoders receive identical inputs. For each pretrained encoder and known peak center, a separate ridge probe predicts H from the mean-pooled output of the final Transformer block (512 features). Probe inputs and targets are z-scored using training statistics. Regularization uses the same grid as panel A, selected by validation R^2. Each plotted band score averages its three center-specific test R^2 values.

Uncertainty Pointwise 95% confidence intervals use 5,000 paired bootstrap resamples of the 20 held-out PhysioNet-MI subjects (A–B) or the 512 independent synthetic test draws (C). Each resampled unit retains its corresponding conditions and model predictions; in C, all 15 frequency variants remain grouped within a draw.

M.2 Attention maps and Grad-CAM

We complement the quantitative diagnostics with qualitative visualizations of cross-channel attention and class-specific attribution, comparing CoCoT-EEG pretrained with MoCo and MAE (Fig. 5).

Cross-channel attention. We visualize attention at Transformer layers 1, 7, and 12, averaging over attention heads and temporal tokens to obtain channel-to-channel attention matrices (Fig. 5A). We additionally show the inter-channel Pearson correlation matrix as a reference for the statistical structure of the input signals (Fig. 5B). Both models exhibit prominent diagonal structure in the first layer, with MoCo's attention map qualitatively resembling the positive correlation structure of the input signals. At layer 7, MoCo displays more visible off-diagonal structure, with some patterns resembling the sign-reversed correlation matrix, suggesting a shift toward attending to negatively correlated channel pairs. MAE's final-layer attention, by comparison, appears more uniform. These qualitative observations illustrate distinct, layer-dependent patterns of inter-channel weighting in the two pretrained models.

Class-specific attribution. We compute Grad-CAM maps for the finetuned MoCo and MAE models on the four BCI-IV motor-imagery classes: left hand, right hand, feet, and tongue. The

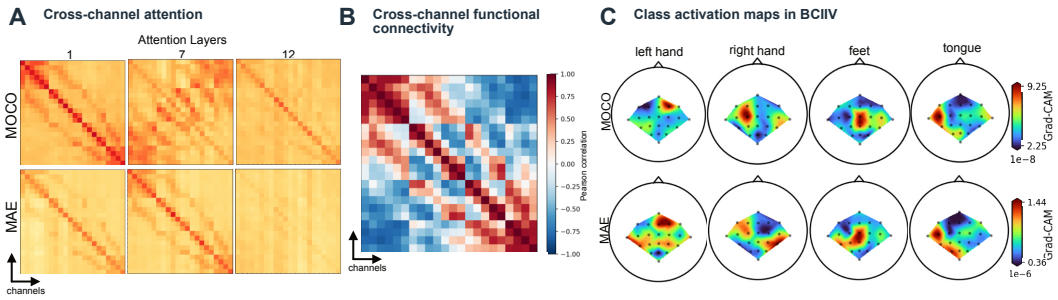

Figure 5: **Interpretability analyses: attention and class activation. A.** Cross-channel attention maps of CoCoT-EEG pretrained with MoCo (top) and MAE (bottom), averaged over heads and time tokens, at transformer layers 1, 7, and 12. Each matrix shows attention weights between pairs of EEG channels. **B.** Empirical cross-channel functional connectivity, computed as the Pearson correlation between channel time series. **C.** Grad-CAM class activation maps on BCI-IV for the four motor imagery classes, projected onto the scalp topography, for MoCo (top) and MAE (bottom) finetuned models.

resulting attributions are projected onto scalp topographies (Fig. 5C). While both models exhibit class-dependent spatial patterns, showing focal and contralateral activations for the hand classes consistent with sensorimotor organization, we note that MAE's maps are less lateralized than MoCo's.