

---

# REVE: A Foundation Model for EEG Adapting to Any Setup with Large-Scale Pretraining on 25,000 Subjects

---

Yassine El Ouahidi<sup>1\*</sup>, Jonathan Lys<sup>1</sup>, Philipp Thölke<sup>2</sup>, Nicolas Farrugia<sup>1</sup>,  
Bastien Pasdeloup<sup>1</sup>, Vincent Gripon<sup>1</sup>, Karim Jerbi<sup>2,3,4</sup>, Giulia Lioi<sup>1\*</sup>

<sup>1</sup> IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238 Brest, France

<sup>2</sup> Psychology Department, Université de Montréal, Montreal, QC, Canada

<sup>3</sup> Mila (Quebec AI research institute), Montreal, QC, Canada

<sup>4</sup> UNIQUE (Quebec Neuro-AI research center), QC, Canada

## Abstract

Foundation models have transformed AI by reducing reliance on task-specific data through large-scale pretraining. While successful in language and vision, their adoption in EEG has lagged due to the heterogeneity of public datasets, which are collected under varying protocols, devices, and electrode configurations. Existing EEG foundation models struggle to generalize across these variations, often restricting pretraining to a single setup, resulting in suboptimal performance, in particular under linear probing. We present REVE (Representation for EEG with Versatile Embeddings), a pretrained model explicitly designed to generalize across diverse EEG signals. REVE introduces a novel 4D positional encoding scheme that enables it to process signals of arbitrary length and electrode arrangement. Using a masked autoencoding objective, we pretrain REVE on over 60,000 hours of EEG data from 92 datasets spanning 25,000 subjects, representing the largest EEG pretraining effort to date. REVE achieves state-of-the-art results on 10 downstream EEG tasks, including motor imagery classification, seizure detection, sleep staging, cognitive load estimation, and emotion recognition. With little to no fine-tuning, it demonstrates strong generalization, and nuanced spatio-temporal modeling. We release code, pretrained weights, and tutorials<sup>2</sup> to support standardized EEG research and accelerate progress in clinical neuroscience.

## 1 Introduction

Electroencephalography (EEG) is a non-invasive technique widely used to study brain activity, with applications spanning brain-computer interfaces (BCIs), clinical diagnostics, and neuroscience research. Despite its potential, the adoption of EEG-based technologies remains limited (Lotte et al., 2018). A key challenge is developing models that generalize effectively to new subjects. EEG data varies widely in electrode configurations, recording conditions, and subject-specific factors, complicating model transferability. This heterogeneity has led to a fragmented ecosystem of datasets and task-specific models, many of which struggle to generalize across settings.

Foundation models have transformed natural language processing (Achiam et al., 2023; Dubey et al., 2024; Warner et al., 2024) and computer vision (Radford et al., 2021; Caron et al., 2021; Kirillov et al., 2023) by leveraging large-scale pretraining to enable transfer with minimal supervision. Their

\*Corresponding authors: [yassine.elouahidi@mistral.ai](mailto:yassine.elouahidi@mistral.ai), [giulia.lioi@imt-atlantique.fr](mailto:giulia.lioi@imt-atlantique.fr)

<sup>2</sup>Project page: <https://brain-bzh.github.io/reve/>

ability to produce general-purpose representations has sparked growing interest in building similar models for EEG (Yang et al., 2024; Wang et al., 2024b; Jiang et al., 2024; Cui et al., 2024; Yuan et al., 2024b; Wang et al., 2024a). Yet, EEG poses unique challenges including data heterogeneity, low signal-to-noise ratio, and the lack of standardized positional encoding to accommodate varying electrode configurations.

Recent EEG foundation models such as BIOT (Yang et al., 2024), Labram (Jiang et al., 2024), CBraMod (Wang et al., 2024b), and NeuroGPT (Cui et al., 2024) adopt self-supervised learning (SSL) techniques for pretraining. While promising, many of these models rely solely on the TUH database (Obeid and Picone, 2016) which uses a fixed 19 or 21-channel montage. As a result, they often fail to generalize to datasets with different electrode layouts or recording setups. Furthermore, existing positional encoding schemes, whether absolute (Yang et al., 2024; Jiang et al., 2024) or convolutional (Wang et al., 2024b), lack the flexibility to accommodate spatial diversity, often necessitating full fine-tuning for transfer.

To address the limitations in current EEG foundation models, we consider three core contributions that enable scalable, generalizable representation learning across diverse, large-scale EEG datasets.

First, we propose a novel 4D positional encoding scheme that enables flexible modeling of EEG signals with varying temporal lengths and electrode configurations. Unlike existing absolute or convolutional encodings, our formulation naturally supports spatial and temporal variability, eliminating the need for fixed montages or fine-tuning of positional priors.

Thanks to this flexible positional encoding method, we are able to train with a wider range of EEG configurations, allowing to scale to larger and more heterogeneous datasets. To this end, we curate the largest and most diverse EEG corpus to date, comprising over 60,000 hours of data from 92 datasets and 25,000 subjects. This diverse collection spans clinical, BCI, and research domains, providing the scale and diversity necessary for robust pretraining.

Combining architectural flexibility with large-scale data results in REVE (Representation for EEG with Versatile Embeddings), a spatio-temporal transformer model trained with a modified masked autoencoder (MAE) (He et al., 2022) objective that promotes learning better representations in the model. REVE learns general-purpose EEG representations that transfer effectively across a wide range of downstream tasks.

REVE achieves state-of-the-art performance across numerous benchmarks, including BCI and clinical datasets, outperforming prior EEG foundation models. Our scaling studies further show improved generalization with larger model sizes, reinforcing the benefits of large-scale pretraining. To support adoption, we release open-source code, pretrained models of multiple sizes, and detailed tutorials for applying REVE to various EEG tasks. By addressing the unique challenges of EEG with scalable architectures and flexible spatial encoding, REVE establishes a unified foundation for EEG analysis and paves the way for new advances in neuroscience and clinical applications.

## 2 Methods

We pretrain our encoder using a masked autoencoder objective. The REVE encoder consists of a patch embedding module, a 4D position encoding module, and a transformer backbone. During pretraining, we apply spatio-temporal contiguous masking to the patch embeddings and jointly train the encoder and decoder to reconstruct the missing segments of EEG, enabling the encoder to learn robust feature representations. Subsequent hyperparameter values are listed in Table 5 in the Appendix.

### 2.1 EEG Representation and Block Masking strategy

We represent multi-channel EEG data as  $\mathbf{X} \in \mathbb{R}^{C \times T}$ , where  $C$  is the number of electrodes and  $T$  the number of time samples, electrode positions are given by  $\mathbf{P} \in \mathbb{R}^{C \times 3}$ , corresponding to their 3D coordinates. To process the data, we segment each channel into patches of size  $w$  with overlap  $o$ , following BIOT (Yang et al., 2024). This yields  $p = \left\lceil \frac{T-w}{w-o} \right\rceil + \mathbb{1}[(T-w) \bmod (w-o) \neq 0]$  non-overlapping patches (discarding any incomplete ones), and reshapes  $\mathbf{X}$  into  $\mathbf{X}_p \in \mathbb{R}^{C \times p \times w}$ . Each patch is linearly embedded, resulting in  $\mathbf{E} \in \mathbb{R}^{C \times p \times D_E}$ , where  $D_E$  is the embedding dimension.

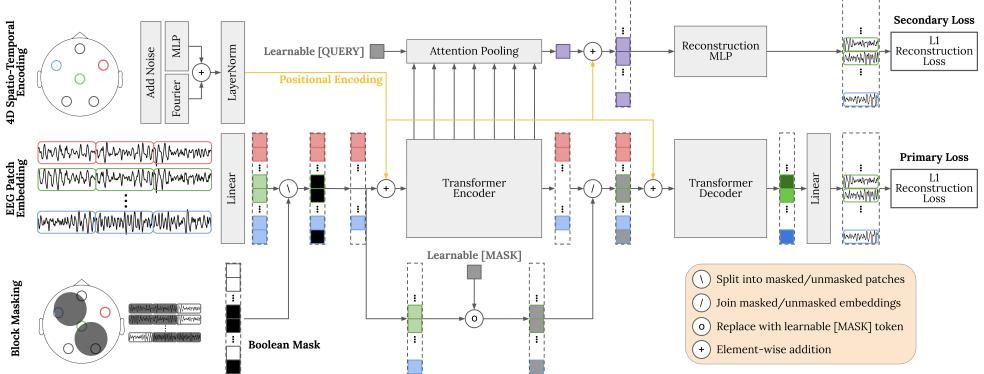


Figure 1: Overview of the **REVE** pretraining framework. The model processes multi-channel EEG data through a linear **Patch Embedding** where signals are divided into overlapping temporal patches for each channel and embedded with a linear layer. **4D Spatio-Temporal Position Encoding** combines spatial coordinates of electrodes with temporal patch indices, augmented with noise for robust generalization. A **Block Masking Strategy** masks contiguous regions across spatial and temporal dimensions to simulate realistic disruptions. The transformer encoder processes unmasked embeddings. Updated embeddings are joined with learnable placeholders for the masked tokens, from which raw EEG is reconstructed using the decoder. The **Primary Task** predicts raw EEG signals directly, while the **Secondary Task** trains a single global token via attention pooling to summarize the input. Both tasks minimize an  $L_1$  reconstruction loss.

To enhance learning during pretraining, we apply a joint spatio-temporal block masking strategy that masks structured regions across both spatial and temporal dimensions. Random masking, proposed for EEG by Chien et al. (2022), was later improved through spatial masking (Mohammadi Foumani et al., 2024; Guetschel et al., 2024). In this work, we extend the masking strategy to the temporal domain. This builds on insights from image modeling, where structured masking outperforms random masking (Xie et al., 2022), a trend also supported by our ablation results (Table 18, Appendix). As neighboring segments of EEG, in both spatial and temporal domain, are typically similar, naive random masking could leave redundant information exposed, reducing the difficulty of reconstruction. In contrast, block masking better disrupts these patterns, encouraging more effective learning.

Our block masking strategy is governed by the following parameters: The masking Ratio  $M_r$  controls the overall proportion of masked tokens. The spatial Masking Radius  $R_s$  and Temporal Masking Radius  $R_t$  respectively define the spatial extent (around a selected channel) and the time window (around a selected token) to be masked. Similarly, the Dropout Ratio  $D_r$  sets the fraction of masked tokens for which the entire time series of the corresponding channel is dropped, while the dropout Radius  $R_d$  determines the spatial neighborhood affected by dropout. For tokens not dropped, temporal masking is applied within radius  $R_t$ . This process yields a binary mask  $\mathbf{B} \in \mathbb{R}^{C \times p}$ , containing  $N_m = \lfloor (1 - M_r) \cdot C \cdot p \rfloor$  masked entries (zeros) and  $N_{\bar{m}} = C \cdot p - N_m$  unmasked entries (ones).

## 2.2 4D Position Encoding Strategy

Unlike prior works that rely on learned embedding tables for spatial encoding vectors (Jiang et al., 2024; Wang et al., 2024b), we directly generate position encodings from the spatio-temporal coordinates of the tokens, allowing the processing of signals of any length or EEG layout and enabling better generalization to unseen setups. More specifically, our method uses a transformation applicable to each position, utilizing the actual 3D coordinates and timestep of each EEG patch, enabling the model to handle arbitrary electrode configurations and sequence lengths without relying on learned embeddings.

**4D Positional Encoding and Spatial Augmentation.** We start with the spatial positions of the EEG electrodes  $\mathbf{P} \in \mathbb{R}^{C \times 3}$ , where each row of  $\mathbf{P}$  contains the  $(x, y, z)$  coordinates of a channel, to which we add Gaussian noise with standard deviation  $\sigma_{\text{noise}}$ . This improves generalization to diverse electrode positions and ensures robustness to variability in head size or electrode placement. We extend  $\mathbf{P}$  with a temporal component, resulting in  $\mathbf{P}_{\text{ext}} \in \mathbb{R}^{C \times p \times 4}$ , where  $p$  is the number of

patches obtained from segmenting EEG signal, as defined in Section 2.1. The temporal dimension is represented as discrete values from 1 to  $p$ , scaled by a factor  $s_t$  to ensure a scale similar to the spatial dimensions.

**4D Fourier-Based Position Encoding.** Building on the 2D approach proposed by Défossez et al. (2023), we extend the Fourier positional encoding method to 4D in our encoding strategy, as follows. Each positional component  $(x, y, z, t)$  of  $\mathbf{P}_{\text{ext}}$  is projected into a multi-frequency space, using  $n_{\text{freq}}$  frequencies per dimension. The frequency assignment follows a Cartesian product structure, *i.e.*, all combinations of frequencies across the four dimensions contribute to the encoding, resulting in a flattened vector of dimension  $n_{\text{freq}}^4$ . A hierarchical periodicity emerges: the period of  $x$  is  $n_{\text{freq}}^1$ , of  $y$  is  $n_{\text{freq}}^2$ , of  $z$  is  $n_{\text{freq}}^3$ , and of  $t$  is  $n_{\text{freq}}^4$ . Then, applying sine and cosine transformations doubles the embedding size, producing a positional vector of dimension  $2 \cdot n_{\text{freq}}^4$ . We ensure that the embedding dimension matches the hidden size required by the 4DPE module, with  $n_{\text{freq}} \in \{3, 4, 5\}$  resulting in the final embedding  $\mathbf{F}_{\text{pe}} \in \mathbb{R}^{C \times p \times D_E}$ . The 4D encoding adds minimal compute overhead, with sinusoidal computations and a small linear layer. Computational cost scales linearly with the number of input tokens (channels  $\times$  temporal patches) and is negligible relative to the transformer backbone.

**Final Adjusted Position Encoding.** To complement the fixed Fourier features, we also process  $\mathbf{P}_{\text{ext}}$  through a linear layer followed by GELU (Hendrycks and Gimpel, 2016) and LayerNorm (Lei Ba et al., 2016), producing a learnable representation  $\mathbf{F}_{\text{lin}} \in \mathbb{R}^{C \times p \times D_E}$ . This component adapts the positional encoding to the specific dataset and task, and can compensate for any truncation in the Fourier basis. The final positional encoding is given by  $\mathbf{P}_{\text{enc}} = \text{LayerNorm}(\mathbf{F}_{\text{pe}} + \mathbf{F}_{\text{lin}})$ , combining the structured inductive bias of Fourier features with the flexibility of learned adaptation. This vector is added to the non-masked patch embeddings before being passed to the transformer encoder similarly to MAE (He et al., 2022), and is consistent with standard absolute positional encoding practices (Vaswani, 2017). The ablation study in Table 19 confirms that this method outperforms both fixed learnable and purely MLP-based positional encoding schemes.

### 2.3 Transformer

Our model extends the standard Transformer architecture (Vaswani, 2017) with enhancements that improve efficiency and stability. We use RMSNorm (Zhang and Sennrich, 2019) in lieu of LayerNorm as a **normalization layer** for better training stability, and choose GEGLU (Shazeer, 2020) as the **activation** function in the feed-forward network (FFN) layers as it outperforms standard GELU through more expressive gating mechanisms (Geiping and Goldstein, 2023). This choice is further supported by the ablation results in Table 20. Our **FFN layers** follow a two-layer structure with an expansion ratio of  $\frac{8}{3}$ , consistent with designs from LLaMA (Touvron et al., 2023), Qwen (Bai et al., 2023) or Mistral (Jiang et al., 2023). Following Dayma et al. (2021), we **remove bias terms** from all linear layers except the final decoder layer. This reallocates the parameter budget to linear transformations, improving efficiency. We use **Flash Attention v2** (Dao, 2024) for memory and computational efficiency in the attention as it reduces the softmax overhead and ensures scalability to long sequences, while maintaining the core transformer formulation.

### 2.4 Masked EEG Reconstruction Methodology

During pretraining, our model reconstructs EEG signal of masked patches using information from the visible, unmasked patches. The overall pretraining framework is illustrated in Figure 1.

Let  $\mathbf{P}_m \in \mathbb{R}^{N_m \times w}$ , and  $\mathbf{P}_{\bar{m}} \in \mathbb{R}^{N_{\bar{m}} \times w}$  denote the masked and visible patches, respectively, with  $N_m$  and  $N_{\bar{m}}$  as defined in Section 2.1. The associated patch embeddings are denoted as  $\mathbf{E}_m$  for masked patches and  $\mathbf{E}_{\bar{m}}$  for visible patches.

We adopt the MAE structure from He et al. (2022), with a larger encoder and a lighter decoder each following the architecture described in Section 2.3. Only the embeddings of the visible patches  $\mathbf{E}_{\bar{m}}$ , enriched with their positional encodings are passed through the encoder, to produce latent representations  $\mathbf{F}_{\bar{m}}$ . Masked patches are represented using a learned embedding, repeated  $N_m$  times and also augmented with positional encodings. Before entering the decoder, positional encodings are re-added to both visible and masked latent patches. Together, they form the decoder input from which the raw EEG signal of the masked patches is reconstructed.

Unlike the original MAE, which uses a separate set of fixed positional encodings for the decoder, we reuse the same encoding for both the encoder and decoder. This design ensures flexibility for processing EEG signals with varying temporal lengths and electrode configurations.

The output of the decoder transformer, is passed through a linear projection layer that maps latent patches back into the signal space, reconstructing the raw EEG signal of the masked patches. Reconstructed patches minimize the  $L_1$  loss relative to the original raw EEG patches:

$$\mathcal{L} = \frac{1}{|\mathbf{P}_m|} \sum_{i \in \mathbf{P}_m} \left\| \hat{\mathbf{P}}_m^{(i)} - \mathbf{P}_m^{(i)} \right\|_1 \quad (1)$$

where  $\hat{\mathbf{P}}_m^{(i)}$  represents the reconstructed signal for patch  $i$ , and  $\mathbf{P}_m^{(i)}$  is the original signal. We chose  $L_1$  loss over  $L_2$  due to the inherently noisy nature of EEG signals. While  $L_2$  amplifies the influence of noise,  $L_1$  loss offers greater robustness by reducing the impact of outliers.

In addition to the main reconstruction loss, we introduce a secondary task that reconstructs masked patches from a compact global representation. We apply attention pooling over the outputs of all Multi-Head Attention (MHA) layers in the encoder: the output tokens (after FFN) from each MHA block are concatenated and attended by a learned query token. This pooled token is then repeated, enriched with positional encodings, and passed through a 2-layer FFN to reconstruct the masked patches. As with the primary loss, we use  $L_1$  loss for reconstruction. The total loss is a weighted sum: Loss = Primary Loss +  $\lambda \cdot$  Secondary Loss

This secondary loss encourages the encoder to distribute useful information across all layers, mitigating over-specialization in the final layer and yielding more generalizable representations.

The secondary loss mitigates a limitation of the MAE framework: the final encoder layer can overfit to the reconstruction task, especially with a shallow decoder (He et al., 2022). By pooling features across all transformer layers (Alkin et al., 2024), the learned token captures a compact, global EEG representation, encouraging more balanced use of the encoder depth. This leads to stronger, more generalizable features for downstream tasks like linear probing, few-shot learning, and transfer without fine-tuning.

After the pretraining phase, the decoder is discarded, and only the encoder is used. In this case, no embeddings are masked, *i.e.*,  $\mathbf{P}_m = \mathbf{E}_m = \emptyset$ . All patches are processed as usual by retaining their associated positional encoding.

To avoid confusion regarding terminology, we clarify that the terms “encoder” and “decoder” are used here in the context of masked auto-encoders (MAE), not in the autoregressive Transformer sense. All Transformer blocks in REVE are non-causal and operate within a standard encoder-style attention pattern; no autoregressive training is involved. The “decoder” refers solely to the lightweight reconstruction head used to recover masked EEG segments during self-supervised pretraining.

### 3 Experiments

#### 3.1 Pretraining

This section outlines the data sources and preprocessing steps used for pretraining, followed by our strategy for scalable and effective representation learning across diverse datasets.

##### 3.1.1 Dataset Collection & Preprocessing

To enable large-scale pretraining, we assembled a massive and diverse collection of EEG recordings from open-source or request-accessible datasets. It comprises 19 TB of raw data, spanning 24,274 subjects, 150,833 unique sessions, and 61,415 hours of recordings drawn from 92 different sources, including OpenNeuro (Markiewicz et al., 2021), MOABB (Aristimunha et al., 2023), and TUH (Obeid and Picone, 2016). To our knowledge, this is the largest and most diverse EEG dataset assembled for training a foundation model. The most extensive prior effort, by Yuan et al. (2024a), comprised approximately 40,000 hours of recordings, but primarily relied on intracranial EEG (iEEG) rather than non-invasive EEG. A summary of the dataset composition and a full list of included sources are provided in Appendix B. While the majority of the data consists of clinical EEG recordings, we also include a substantial subset of cognitive and BCI-related data which, although smaller in proportion,

tend to be cleaner and more diverse. We also collected electrode positional information for each recording. When 3D coordinates were available, they were used directly; otherwise, positions were inferred from standard labels. Channels without identifiable names or positional data were excluded. The dataset spans a wide range of EEG systems and formats—including BrainVision, BioSemi, EDF, GDF, and EEGLAB, with most recordings adhering to the 10-5 system (Oostenveld and Praamstra, 2001). In total, the dataset includes 396 unique electrode names.

Our preprocessing pipeline is designed to preserve signal diversity and prioritize robustness when scaling. We only removed recordings shorter than 10 seconds, and those used in downstream tasks. Remaining signals were resampled to 200 Hz, band-pass filtered (0.5–99.5 Hz), and converted to float32, resulting in a 6 TB dataset. To address amplitude variations across recordings, we applied Z-score normalization with statistics computed across the recording sessions to ensure robust statistics. After normalization, values exceeding 15 standard deviations were clipped, as in Défossez et al. (2023). Unlike CBraMod (Wang et al., 2024b), which excluded signals above 100  $\mu$ V, our approach retains them, resulting in about 60,000 hours of EEG, compared to 9,000 in CBraMod and 2,534 in LaBraM.

### 3.1.2 Pretraining Strategy & Scaling

## 3.2 Training and Scaling Strategy

We present the training procedure used for pretraining the Small model and detail how it scales to larger architectures under constrained resources. Our training framework builds upon recent advances in state-of-the-art NLP methodologies (Warner et al., 2024). We use the StableAdamW (Wortsman et al., 2023) optimizer, designed for low precision frameworks and improved stability, thanks to the Adafactor-style gradient clipping (Shazeer and Stern, 2018). Table 5 of the Appendix lists the optimizer hyperparameters.

The learning rate follows a Warmup Stable Decay (trapezoidal) schedule (Hu et al., 2024), known for its robustness to learning rate variations (Hägle et al., 2024). We use a linear warmup over 10% of the first epoch, followed by 80% at peak LR, and a linear decay to 1% of the maximum. Unlike one-cycle schedules that reset every epoch, our cyclic trapezoidal variant allows multiple cooldown phases across epochs, particularly beneficial for EEG training where masked token sampling introduces variability. We apply Megatron-style initialization (Shoeybi et al., 2019) with a standard deviation of 0.02 for all transformer layers and the mask token, ensuring stable dynamics. Other parameters use PyTorch’s default initializations.

A key factor for the success of foundation models is the simultaneous scaling of both training datasets and model architectures (Touvron et al., 2023). We describe our scaling methodology to maximize computational efficiency and accommodate larger models within constrained resources. To scale model capacity, we adjust depth, width, and number of attention heads while maintaining a fixed FFN ratio. Table 6 of the Appendix summarizes these configurations. This scaling strategy enables efficient capacity expansion while preserving architectural consistency across model sizes.

Recent advances in NLP provide strong theoretical and empirical evidence for the existence of scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022), which govern the relationship between model size, training dynamics, optimization and initialization hyperparameters. We follow the power law  $\eta \propto D^{\alpha_D}$ , with  $\alpha_D = -0.90$  and  $D$  the model dimension, for the learning rate, as derived in Everett et al. (2024). The optimal LR is first swept on the small model and then scaled accordingly.

To efficiently train models, we use data parallelism, maintaining a constant batch size by reducing per-GPU loads for large models. A load-aware data-shuffling strategy groups samples by electrode count, shuffles within and across buckets, and balances batches across GPUs to avoid bottlenecks, for constant optimization steps and maximized throughput.

Although scaling laws exist for adjusting AdamW momentum terms (Malladi et al., 2022), our use of a constant effective batch size across models allows us to fix  $\beta_1$  and  $\beta_2$ . Regarding initialization, while Hägle et al. (2024) suggests scaling  $\sigma_{\text{init}} \propto D^{-0.5}$ , our width increase (from 200 to 1,216) leads us to keep  $\sigma_{\text{init}} = 0.02$  fixed across scales.

## 3.3 Downstream tasks

**Downstream task datasets** To evaluate the performance and generalizability of our EEG foundation model, we perform extensive assessments across 10 diverse downstream tasks, selected to ensure

comparability with existing models in the field. These tasks span a variety of EEG-based applications, including sleep staging, emotion and event classification, detection of stress and mental disorder , across the following datasets: PhysioNet-MI (Goldberger et al., 2000), BCIC-IV-2a (Tangermann et al., 2012), TUEV (Obeid and Picone, 2016), TUAB (Obeid and Picone, 2016), HMC (Alvarez-Estevez and Rijsman, 2021), ISRUC (Khalighi et al., 2016), FACED (Chen et al., 2023), Mumtaz (Mumtaz, 2016), Mental Arithmetic (MAT) (Zyma et al., 2019), and BCI2020-IV-3 (Jeong et al., 2022). A summary of these datasets is provided in Table 1, with a more detailed description available in the Appendix.

Table 1: Overview of downstream tasks and datasets.

Task	Dataset	# Channels	Duration	# Samples	Rate	# Classes
Motor Imagery	PhysioNet-MI	64	4s	9,837	160Hz	4
	BCIC-IV-2a	22	4s	5,184	250Hz	4
Event Type	TUEV	16	5s	112,491	256Hz	6
Abnormal detection	TUAB	16	10s	409,455	256Hz	2
Sleep staging	HMC	4	30s	137,243	256 Hz	5
	ISRUC	6	30s	89,240	200Hz	5
Emotion recognition	FACED	32	10s	10,332	250Hz	9
Mental disorder	Mumtaz	19	5s	7,143	256Hz	2
Mental stress	MAT	20	5s	1,707	500Hz	2
Imagined speech	BCIC2020-3	64	3s	6,000	256Hz	5

Our evaluation process maintains strict consistency with prior works by adhering to the same train/val/test splits used in earlier studies, ensuring that our results are directly comparable to baseline models. Specifically, we follow the protocols from CBraMod (Wang et al., 2024b), LaBraM (Jiang et al., 2024), and BIOT (Yang et al., 2024), guaranteeing fair comparisons across tasks. For fairness in preprocessing, we adopt the same pipeline as the baselines. A notable correction was made for the ISRUC dataset, where we identified and removed a bug in the baseline code involving the inclusion of a chin electrode instead of an EEG electrode. Our results for REVE exclude the chin electrode, aligning with proper electrode placement.

**Finetuning** Fine-tuning EEG-based models presents unique challenges due to the small size of available datasets and the high noise levels in EEG recordings. Unlike large-scale vision datasets, EEG datasets are often limited in size, subject-dependent, and prone to distribution shifts across different recording setups. Effective fine-tuning must therefore maximize generalization while mitigating the risk of overfitting. To address this, we adopt a two-step fine-tuning strategy, incorporating techniques specifically designed to enhance stability and adaptability. This includes the use of parameter-efficient fine-tuning techniques (Suzumura et al., 2024) tailored to this domain.

For downstream classification tasks, the two-step strategy, inspired by Kumar et al. (2022), goes as follow: We first train a linear probe while keeping the encoder frozen, aligning the classifier with the pretrained feature space. Next, we unfreeze the encoder and fine-tune the entire network for task-specific adaptation, preserving the robustness of the pretrained model. Importantly, this two-step strategy is implemented as a single continuous training run, where the backbone is initially frozen (i.e., only the head is trained) and later unfrozen. This approach is well-suited for EEG data, where distributions can shift significantly across datasets. We employ dropout and Mixup (Zhang et al., 2018) as data augmentation for improved robustness. To further mitigate catastrophic forgetting and improve efficiency, we integrate Low-Rank Adaptation (LoRA) into the attention blocks, within the query, key, value, and output (QKVO) projection layers. Instead of fine-tuning the entire model, LoRA introduces trainable low-rank matrices that enable effective adaptation while preserving the integrity of the pretrained model’s knowledge (Hu et al., 2022).

Each training step includes a warmup phase (Kalra and Barkeshli, 2024) followed by a cooldown phase. The cooldown phase employs a Reduce-on-Plateau learning rate scheduler, which dynamically lowers the learning rate when training convergence slows to preventing overfitting.

To further enhance robustness, we explore model souping (Wortsman et al., 2022), which averages the weights of multiple fine-tuning runs to improve accuracy. Given the stochasticity and noise inherent in EEG datasets, souping smooths gradients and reduces variance across different fine-tuning

trajectories. Our experiments confirm that this approach enhances generalization and produces more stable performance across diverse EEG tasks.

By integrating structured fine-tuning with data augmentation, LoRA and model souping, our approach effectively addresses the small-scale and noisy nature of EEG datasets. These techniques effectively ensure robust and generalized adaptation to downstream tasks.

## 4 Results and Discussion

We evaluate REVE against non-foundation and foundation model baselines on the previously discussed datasets.

**Non-Foundation Models:** We compare to EEGNet (Lawhern et al., 2018), EEGConformer (Song et al., 2022), SPaRCNet (Jing et al., 2023), ContraWR (Yang et al., 2021), CNN-Transformer (Peh et al., 2022), FFCL (Li et al., 2022), and ST-Transformer (Song et al., 2021).

**Foundation Models:** We compare to BIOT (Yang et al., 2024), LaBraM (Jiang et al., 2024) and CBraMod (Wang et al., 2024b). We report results displayed in existing studies.

We report the balanced accuracy for each dataset and provide additional evaluation metrics in the appendix.

Table 2: Balanced accuracy ( $\pm$  std) of different methods across 9 EEG classification task

Methods	TUAB	TUEV	PhysioNetMI	BCI-IV-2a	FACED
EEGNet	0.7642 $\pm$ 0.0036	0.3876 $\pm$ 0.0143	0.5814 $\pm$ 0.0125	0.4482 $\pm$ 0.0094	0.4090 $\pm$ 0.0122
EEGConformer	0.7758 $\pm$ 0.0049	0.4074 $\pm$ 0.0164	0.6049 $\pm$ 0.0104	0.4696 $\pm$ 0.0106	0.4559 $\pm$ 0.0125
SPaRCNet	0.7896 $\pm$ 0.0018	0.4161 $\pm$ 0.0262	0.5932 $\pm$ 0.0152	0.4635 $\pm$ 0.0117	0.4673 $\pm$ 0.0155
ContraWR	0.7746 $\pm$ 0.0041	0.4384 $\pm$ 0.0349	0.5892 $\pm$ 0.0133	0.4678 $\pm$ 0.0125	0.4887 $\pm$ 0.0078
CNN-Transformer	0.7777 $\pm$ 0.0022	0.4087 $\pm$ 0.0161	0.6053 $\pm$ 0.0118	0.4600 $\pm$ 0.0108	0.4697 $\pm$ 0.0132
FFCL	0.7848 $\pm$ 0.0038	0.3979 $\pm$ 0.0104	0.5726 $\pm$ 0.0092	0.4470 $\pm$ 0.0143	0.4673 $\pm$ 0.0158
ST-Transformer	0.7966 $\pm$ 0.0023	0.3984 $\pm$ 0.0228	0.6035 $\pm$ 0.0081	0.4575 $\pm$ 0.0145	0.4810 $\pm$ 0.0079
BIOT	0.7959 $\pm$ 0.0057	0.5281 $\pm$ 0.0225	0.6153 $\pm$ 0.0154	0.4748 $\pm$ 0.0093	0.5118 $\pm$ 0.0118
LaBraM-Base	0.8140 $\pm$ 0.0019	0.6409 $\pm$ 0.0065	0.6173 $\pm$ 0.0122	0.4869 $\pm$ 0.0085	0.5273 $\pm$ 0.0107
CBraMod	0.8289 $\pm$ 0.0022	0.6671 $\pm$ 0.0107	0.6417 $\pm$ 0.0091	0.5138 $\pm$ 0.0066	0.5509 $\pm$ 0.0089
REVE-Base	<b>0.8315</b> $\pm$ 0.0014	<b>0.6759</b> $\pm$ 0.0229	<b>0.6480</b> $\pm$ 0.0140	<b>0.6396</b> $\pm$ 0.0095	<b>0.5646</b> $\pm$ 0.0164
	ISRUC	Mumtaz	MAT	BCI-2020-3	Average
EEGNet	0.7154 $\pm$ 0.0121	0.9232 $\pm$ 0.0104	0.6770 $\pm$ 0.0116	0.4413 $\pm$ 0.0096	0.5941 $\pm$ 0.0037
EEGConformer	0.7400 $\pm$ 0.0133	0.9308 $\pm$ 0.0117	0.6805 $\pm$ 0.0123	0.4506 $\pm$ 0.0133	0.6128 $\pm$ 0.0044
SPaRCNet	0.7487 $\pm$ 0.0075	0.9316 $\pm$ 0.0095	0.6879 $\pm$ 0.0107	0.4426 $\pm$ 0.0156	0.6156 $\pm$ 0.0047
ContraWR	0.7402 $\pm$ 0.0126	0.9195 $\pm$ 0.0115	0.6631 $\pm$ 0.0097	0.4257 $\pm$ 0.0162	0.6119 $\pm$ 0.0053
CNN-Transformer	0.7363 $\pm$ 0.0087	0.9305 $\pm$ 0.0068	0.6779 $\pm$ 0.0268	0.4533 $\pm$ 0.0092	0.6133 $\pm$ 0.0045
FFCL	0.7277 $\pm$ 0.0182	0.9314 $\pm$ 0.0038	0.6798 $\pm$ 0.0142	0.4678 $\pm$ 0.0197	0.6085 $\pm$ 0.0044
ST-Transformer	0.7381 $\pm$ 0.0205	0.9135 $\pm$ 0.0103	0.6631 $\pm$ 0.0173	0.4126 $\pm$ 0.0122	0.6071 $\pm$ 0.0048
BIOT	0.7527 $\pm$ 0.0121	0.9358 $\pm$ 0.0052	0.6875 $\pm$ 0.0186	0.4920 $\pm$ 0.0086	0.6438 $\pm$ 0.0044
LaBraM-Base	0.7633 $\pm$ 0.0102	0.9409 $\pm$ 0.0079	0.6909 $\pm$ 0.0125	0.5060 $\pm$ 0.0155	0.6653 $\pm$ 0.0031
CBraMod	<b>0.7865</b> $\pm$ 0.0110	0.9560 $\pm$ 0.0056	0.7256 $\pm$ 0.0132	0.5373 $\pm$ 0.0108	0.6898 $\pm$ 0.0031
REVE-Base	0.7819 $\pm$ 0.0078 <sup>3</sup>	<b>0.9644</b> $\pm$ 0.0097	<b>0.7660</b> $\pm$ 0.0355	<b>0.5635</b> $\pm$ 0.0123	<b>0.7150</b> $\pm$ 0.0057

Table 2 shows that REVE achieves state-of-the-art performance on the downstream tasks in this study, with an average gain of 2.5%, compared to CBraMod the highest performing baseline. The results on ISRUC and HMC (Appendix C.6) show that the model effectively generalizes beyond the 10-second segments it was pretrained on, performing well on tasks with 30-second inputs, which highlights the strength of our positional encoding method. The results on TUEV highlight the model’s ability to generalize to unseen electrode configurations, including bipolar setups never encountered during training.

In addition to the detailed evaluation metrics provided in Appendix C, we report the performance of the Large model across our downstream tasks in Table 4. We observe that the Large model consistently produces richer embeddings, leading to improved linear probing performance compared to the Base model. Model souping consistently improved performance, averaging a 1.5% gain when

<sup>3</sup>NB: our preprocessing pipeline is different from the baseline and fixes a potential bug

Table 3: Impact of pretraining (PT) and weight freezing on REVE and baselines for PhysioNet-MI

Settings	PhysioNet-MI, 4-class		
	Balanced Accuracy	Cohen's Kappa	Weighted F1
CBraMod (w/ PT)	0.6417 ± 0.0091	0.5222 ± 0.0169	0.6427 ± 0.0100
BIOT (w/ PT)	0.6153 ± 0.0154	0.4875 ± 0.0272	0.6158 ± 0.0197
LaBraM-Base (w/ PT)	0.6173 ± 0.0122	0.4912 ± 0.0192	0.6177 ± 0.0141
REVE-Base (w/ PT)	<b>0.6480</b> ± 0.0140	<b>0.5306</b> ± 0.0187	<b>0.6484</b> ± 0.0170
CBraMod (w/o PT)	<u>0.6196</u> ± 0.0143	<u>0.4994</u> ± 0.0289	<u>0.6289</u> ± 0.0179
REVE-Base (w/o PT)	0.5409 ± 0.0094	0.3879 ± 0.0125	0.5421 ± 0.0101
Cbramod (Frozen)	0.3845 ± 0.0345	0.2983 ± 0.0498	0.3946 ± 0.0378
BIOT (Frozen)	0.3698 ± 0.0318	0.2703 ± 0.0472	0.3723 ± 0.0364
LaBraM (Frozen)	0.3715 ± 0.0458	0.2814 ± 0.0586	0.3796 ± 0.0472
REVE-Base (Frozen)	<u>0.5371</u> ± 0.0052	<u>0.3827</u> ± 0.0070	<u>0.5376</u> ± 0.0033

Table 4: **Linear probing** results on downstream tasks for REVE and CBraMod models with (Pool) and without pooling across multiple EEG downstream tasks. Best results are highlighted in bold. To ensure a fair comparison, we reproduced CBraMod (Wang et al., 2024b) using their official code and pretrained checkpoint, carefully following their classification pipeline (notably, no pooling) and matched architectural details to avoid any bias.

Dataset	REVE-B (Pool)	REVE-B	REVE-L (Pool)	REVE-L	CBraMod (Pool)	CBraMod
Mumtaz	0.962 ± 0.003	0.931 ± 0.021	<b>0.985 ± 0.006</b>	0.980 ± 0.009	0.859 ± 0.009	0.907 ± 0.027
M. Arithmetic	0.725 ± 0.010	<b>0.740 ± 0.073</b>	0.712 ± 0.008	0.665 ± 0.103	0.500 ± 0.000	0.605 ± 0.020
TUAB	0.810 ± 0.007	0.809 ± 0.004	<b>0.821 ± 0.004</b>	0.809 ± 0.004	0.500 ± 0.000	0.500 ± 0.000
PhysioNetMI	0.537 ± 0.005	0.510 ± 0.012	0.551 ± 0.001	<b>0.617 ± 0.000</b>	0.256 ± 0.002	0.531 ± 0.015
BCIC-IV-2a	0.432 ± 0.004	0.517 ± 0.015	0.534 ± 0.001	<b>0.603 ± 0.011</b>	0.287 ± 0.023	0.376 ± 0.006
ISRIC	0.697 ± 0.011	0.662 ± 0.030	0.743 ± 0.004	<b>0.758 ± 0.001</b>	0.407 ± 0.049	0.430 ± 0.043
HMC	0.647 ± 0.008	0.604 ± 0.008	0.703 ± 0.003	<b>0.710 ± 0.007</b>	0.368 ± 0.001	0.538 ± 0.009
BCIC2020-3	0.234 ± 0.009	<b>0.390 ± 0.017</b>	0.274 ± 0.001	0.378 ± 0.021	0.214 ± 0.003	0.374 ± 0.007
TUEV	0.592 ± 0.008	0.508 ± 0.073	<b>0.630 ± 0.003</b>	0.550 ± 0.014	0.219 ± 0.009	0.482 ± 0.037
Faced	0.240 ± 0.010	0.422 ± 0.028	0.283 ± 0.003	<b>0.469 ± 0.007</b>	0.117 ± 0.005	0.261 ± 0.013
Avg.	0.586	0.609	0.623	<b>0.654</b>	0.373	0.501

combining at least 5 Base or Large models. For example, REVE-Base achieved 69.6% balanced accuracy on TUEV using the 10 models from Table 2. However, souping showed limited benefits for the small models and sometimes led to negative outcomes.

Table 3 highlights the importance of REVE’s pretraining phase. Without pretraining, CBraMod outperforms REVE by at least 8%. However, pretraining improves REVE-Base by 11%, while CBraMod gains only 2%, a trend also observed in the LaBraM paper. This suggests that REVE benefits more significantly from pretraining, whereas other models derive most of their performance from architectural design rather than pretraining learned representations. A key advantage of REVE is its ability to produce high-quality latent spaces without heavy fine-tuning, as evidenced by linear probing results in Table 4: REVE consistently outperforms CBraMod across all downstream tasks and model sizes, with REVE-Large achieving nearly 17% higher performance. These results also highlight REVE’s ability to scale effectively with model size, yielding richer and more generalizable embeddings as capacity increases. Providing rich, ready-to-use embeddings is crucial for enabling zero-shot analysis, faster BCI calibration, and improved performance in low-data or sparsely annotated settings. REVE also benefits from its spatial encoding strategy, which enables transfer across diverse EEG configurations. In Appendix D, we further demonstrate the contribution of our secondary loss function, a novel component of our framework, which proves particularly effective in frozen-feature scenarios. The secondary objective reconstructs masked tokens using a compressed, global representation from attention pooling. This pooling acts as an information bottleneck, forcing the model to distill key information from the entire input sequence into a single vector. As shown by Table 17, the secondary loss mainly improves the quality of the frozen embeddings of the model.

## 5 Limitations and Future Work

The model has some limitations, requiring signals to be at least one second and multiples of one second. A way to address this could be to leverage padding with causal masking.

While the focus has been on collecting large EEG datasets for pretraining, an important next step could be to curate this data more selectively. This includes removing low-quality recordings, balancing distributions, and identifying representative subsets, especially given the inherently noisy nature of EEG signals. Our current pretraining corpus aggregates 92 publicly available EEG datasets spanning over 25,000 subjects, which helps reduce overfitting to any single source. However, most public EEG data originates from North America and Europe, resulting in limited demographic diversity—a key limitation that calls for broader, more equitable data collection efforts. To partially mitigate such imbalances, we leverage self-supervised learning (MAE), which has been shown to be robust to long-tailed and heterogeneous data distributions (Xu et al., 2023). Targeted selection strategies, combined with robust SSL objectives, could help focus on the most informative and complementary data for building stronger, fairer, and more efficient foundation models. Thanks to its flexibility in handling any EEG configuration, REVE could itself guide this curation process.

We also plan to extend our study to diverse tasks, including zero-/few-shot regimes. This first iteration uses a simple MAE approach and a standard transformer, but future improvements could leverage more advanced SSL techniques and architectures. We release the model’s code, weights and guidelines for adapting it to mainstream EEG tasks. In parallel, our findings point toward the presence of scaling effects in EEG foundation models. Identifying precise scaling laws that capture how model size, data volume, and downstream performance interact would be valuable for future work.

## 6 Conclusion

EEG research has lacked a foundation model that transfers robustly across devices, montages, and tasks—especially under linear probing. REVE contributes to bridging this gap. Trained on 60,000 hours from 92 datasets and 25,000 subjects, REVE combines a 4D Fourier positional encoding that natively supports arbitrary electrode layouts and sequence lengths with masked autoencoding enhanced by spatio-temporal block masking and a global-token secondary loss. Across 10 benchmarks, it sets a new state of the art (average +2.5% balanced accuracy over prior foundation models), delivers up to 17% gains in linear probing, and generalizes to unseen/bipolar montages and longer inputs than used in pretraining. These properties enable faster BCI calibration, more reliable cross-site clinical deployment, and standardized embeddings for downstream analytics. We release code, weights, loaders for arbitrary 3D coordinates, and training/eval recipes. We invite the community to extend REVE to broader populations and modalities (MEG/iEEG/OPM-MEG), and to co-build a cross-montage benchmark for fair, scalable EEG evaluation.

## 7 Acknowledgments

This research was supported by the French National Research Agency (ANR) through its AI@IMT program and grant ANR-24-CE23-7365, as well as by a grant from the Brittany region. Further support was provided by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC), by funding from the Canada Research Chairs program and the Fonds de recherche du Québec – Nature et technologies (FRQ-NT). This work was granted access to the HPC resources of IDRIS under the allocation 2024-AD011015237R1 made by GENCI, as well as HPC provided by Digital Alliance Canada.

## References

- Bernd Accou, Lies Bollens, Marlies Gillis, Wendy Verheijen, Hugo Van hamme, and Tom Francart. SparrKULee: A Speech-Evoked Auditory Response Repository of the KU Leuven, Containing EEG of 85 Participants. *BioRxiv*, pages 2023–07, 2023.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 Technical Report. *ArXiv Preprint ArXiv:2303.08774*, 2023.
- Blanca Aguado-Lopez, Ana F. Palenciano, Jose M. G. Penalver, Paloma Diaz-Gutierrez, David Lopez-Garcia, Chiara Avancini, Luis F. Ciria, and Maria Ruz. "Proactive Selective Attention Across Competition Contexts", 2024.
- Lindsay M Alexander, Jasmine Escalera, Lei Ai, Charissa Andreotti, Karina Febre, Alexander Mangone, Natan Vega-Potler, Nicolas Langer, Alexis Alexander, Meagan Kovacs, et al. An Open Resource for Transdiagnostic Research in Pediatric Mental Health and Learning Disorders. *Scientific Data*, 4(1):1–26, 2017.
- Benedikt Alkin, Lukas Miklautz, Sepp Hochreiter, and Johannes Brandstetter. MIM-Refiner: A Contrastive Learning Boost from Intermediate Pre-Trained Representations. *ArXiv Preprint ArXiv:2402.10093*, 2024.
- Diego Alvarez-Estevez and Roselyne M Rijsman. Inter-Database Validation of a Deep Learning Approach for Automatic Sleep Scoring. *PLOS One*, 16(8):e0256111, 2021.
- Edilberto Amorim, Wei-Long Zheng, Jong Woo Lee, Susan Herman, Mohammad Ghassemi, Adithya Sivaraju, Nicolas Gaspard, Jeannette Hofmeijer, Michel JAM van Putten, Matthew Reyna, et al. I-CARE: International Cardiac Arrest Research Consortium Database. <https://physionet.org/content/i-care/2.0>, 2023.
- Carlos Valle Araya, Carolina Mendez-Orellana, and Maria Rodriguez-Fernandez. "Large Spanish EEG", 2023.
- Pietro Aricò, F Aloise, Francesca Schettini, Serenella Salinari, D Mattia, and Febo Cincotti. Influence of P300 Latency Jitter on Event Related Potential-Based Brain–Computer Interface Performance. *Journal of Neural Engineering*, 11(3):035008, 2014.
- Bruno Aristimunha, Igor Carrara, Pierre Guetschel, Sara Sedlar, Pedro Rodrigues, Jan Sosulski, Divyesh Narayanan, Erik Bjareholt, Barthelemy Quentin, Robin Tibor Schirrmeister, Emmanuel Kalunga, Ludovic Darmet, Cattan Gregoire, Ali Abdul Hussain, Ramiro Gatti, Vladislav Goncharenko, Jordy Thielen, Thomas Moreau, Yannick Roy, Vinay Jayaram, Alexandre Barachant, and Sylvain Chevallier. Mother of All BCI Benchmarks, 2023. URL <https://moabb.neurotechx.com/docs/index.html>.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen Technical Report. *ArXiv Preprint ArXiv:2309.16609*, 2023.
- Imad J. Bajwa1, Andre S. Nilsen1, 3 René Skukies1, Arnfinn Aamodt1, Gernot Ernst2, Johan F. Storm1, and 2 Bjørn E. Juel1. "A Repeated Awakening Study Exploring the Capacity of Complexity Measures to Capture Dreaming During Propofol Sedation", 2024.
- Alexandre Barachant. *Commande Robuste d'un Effecteur par une Interface Cerveau Machine EEG Asynchrone*. PhD thesis, Université de Grenoble, 2012.
- Cemre Baykan and Alexander C. Schütz. "Electroencephalographic Responses to the Number of Objects in Partially Occluded and Uncovered Scenes", 2024.
- Ole Bialas, Emily Teoh, Andrew Anderson, and Edmund Lalor. "Invariant Encoding of Phonemes in Neural Responses to Continuous Speech", 2023.
- Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

- James F Cavanagh and Trevor C J Jackson. "Mood Manipulation and PST, Experiment 1", 2022.
- Luis Alberto Barradas Chacón and Selina C. Wriessnegger. "Toonfaces", 2023.
- Jingjing Chen, Xiaobin Wang, Chen Huang, Xin Hu, Xinken Shen, and Dan Zhang. A Large Finer-Grained Affective Computing EEG Dataset. *Scientific Data*, 10(1):740, 2023.
- Hsiang-Yun Sherry Chien, Hanlin Goh, Christopher Michael Sandino, and Joseph Yitan Cheng. MaEEG: Masked Auto-Encoder for EEG Representation Learning. In *NeurIPS 2022 Workshop on Learning from Time Series for Health*, 2022.
- Hohyun Cho, Minkyu Ahn, Sangtae Ahn, Moonyoung Kwon, and Sung Chan Jun. EEG Datasets for Motor Imagery Brain–Computer Interface. *Gigascience*, 6(7):gix034, 2017.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.
- Jose Cordoba-Silva, Rafael Maya, Mario Valderrama, Luis Felipe Giraldo, William Betancourt-Zapata, Andrés Salgado-Vascob, Juliana Marín-Sánchez, Viviana Gómez-Ortega, and Mark Ettenberger. "Dataset of Electrophysiological Signals (EEG, ECG, EMG) During Music Therapy with Adult Burn Patients in the Intensive Care Unit", 2023.
- Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A Joshi, and Richard M Leahy. Neuro-GPT: Towards a Foundation Model for EEG. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.
- Ian Daly, Nicoletta Nicolaou, Duncan Williams, Faustina Hwang, Alexis Kirke, Eduardo Miranda, and Slawomir J. Nasuto. "An EEG Dataset Recorded During Affective Music Listening", 2020.
- Tri Dao. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.
- Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phúc Lê Khac, Luke Melas, and Ritobrata Ghosh. Dall·E Mini, 7 2021. URL <https://github.com/borisdayma/dalle-mini>.
- Alexandre Défossez, Charlotte Caucheteux, Jérémie Rapin, Ori Kabeli, and Jean-Rémi King. Decoding Speech Perception from Non-Invasive Brain Recordings. *Nature Machine Intelligence*, 5(10):1097–1107, 2023.
- Arnaud Delorme and Claire Braboszcz. "Meditation Vs Thinking Task", 2021.
- Arnaud Delorme and Tracy Brandmeyer. "EEG Meditation Study", 2024.
- Arnaud Delorme and Michele Fabre-Thorpe. "Go-Nogo Categorization and Detection Task", 2020.
- Paolo Detti. Siena Scalp EEG Database. *Physionet. Doi*, 10:493, 2020.
- Paolo Detti, Giampaolo Vatti, and Garazi Zabalo Manrique de Lara. EEG Synchronization Analysis for Seizure Prediction: A Study on Data of Noninvasive Recordings. *Processes*, 8(7):846, 2020.
- Pauline Dreyer, Aline Roc, Léa Pillette, Sébastien Rimbert, and Fabien Lotte. A Large EEG Database with Users' Profile Information for Motor Imagery Brain-Computer Interface Research. *Scientific Data*, 10(1):580, 2023.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 Herd of Models. *ArXiv Preprint ArXiv:2407.21783*, 2024.
- 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 Transactions on Neural Systems and Rehabilitation Engineering*, 2024.

Pablo Rodríguez-San Esteban, Ana B. Chica, and José A. González-López. "Neural Representation of Consciously Seen and Unseen Information", 2024.

Katie E Everett, Lechao Xiao, Mitchell Wortsman, Alexander A Alemi, Roman Novak, Peter J Liu, Izzeddin Gur, Jascha Sohl-Dickstein, Leslie Pack Kaelbling, Jaehoon Lee, and Jeffrey Pennington. Scaling Exponents Across Parameterizations and Optimizers. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41St International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12666–12700. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/everett24a.html>.

Josef Faller, Carmen Vidaurre, Teodoro Solis-Escalante, Christa Neuper, and Reinhold Scherer. Autocalibration and Recurrent Adaptation: Towards a Plug and Play Online ERD-BCI. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 20(3):313–319, 2012.

Lukas Gehrke, Sezen Akman, Albert Chen, Pedro Lopes, and Klaus Gramann. "Prediction Error", 2024.

Jonas Geiping and Tom Goldstein. Cramming: Training a Language Model on a Single GPU in One Day. In *International Conference on Machine Learning*, pages 11117–11143. PMLR, 2023.

Alessandro T Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M Cichy. A Large and Rich EEG Dataset for Modeling Human Visual Object Recognition. *NeuroImage*, 264:119754, 2022.

Ary L Goldberger, Luis AN 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.

Tijl Grootswagers, Ivy Zhou, Amanda Robinson, Martin Hebart, and Thomas Carlson. "Human Electroencephalography Recordings from 50 Subjects for 22,248 Images from 1,854 Object Concepts", 2022.

Tijl Grootswagers, Amanda Robinson, Sofia Shatek, and Thomas Carlson. "EEG-attention-rsvp-exp1", 2023a.

Tijl Grootswagers, Amanda Robinson, Sofia Shatek, and Thomas Carlson. "EEG-attention-rsvp-exp2", 2023b.

Tijl Grootswagers, Amanda Robinson, Sofia Shatek, and Thomas Carlson. "Features-EEG", 2024.

Pierre Guetschel, Thomas Moreau, and Michael Tangermann. S-JEPA: Towards Seamless Cross-Dataset Transfer Through Dynamic Spatial Attention. *ArXiv Preprint ArXiv:2403.11772*, 2024.

Christoph Guger, Shahab Daban, Eric Sellers, Clemens Holzner, Gunther Krausz, Roberta Carabalona, Furio Gramatica, and Guenter Edlinger. How Many People Are Able To Control a P300-Based Brain-Computer Interface (BCI)? *Neuroscience Letters*, 462(1):94–98, 2009.

Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna Ben Allal, Leandro Von Werra, and Martin Jaggi. Scaling Laws and Compute-Optimal Training Beyond Fixed Training Durations. *ArXiv Preprint ArXiv:2405.18392*, 2024.

Cameron D. Hassall, Yan Yan, and Laurence T. Hunt. "Drum Trainer", 2022a.

Cameron D. Hassall, Yan Yan, and Laurence T. Hunt. "Steer the Ship", 2022b.

Cameron D. Hassall, Laurence T. Hunt, and Clay B. Holroyd. "Average Task Value", 2024.

Christoffer Hatlestad-Hall, Trine Waage Rygvold, and Stein Andersson. "SRM Resting-State EEG", 2022.

Marleen Haupt, Monika Graumann, Santani Teng, Carina Kaltenbach, and Radoslaw M. Cichy. "Braille Letters - EEG", 2024.

- 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, 2019.
- 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.
- Jason Helbing, Dejan Draschkow, and Melissa L.-H. Võ. "Search Superiority Recollection Familiarity", 2024.
- Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs). *ArXiv Preprint ArXiv:1606.08415*, 2016.
- Jordan Hoffmann, Sébastien Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training Compute-Optimal Large Language Models. *ArXiv Preprint ArXiv:2203.15556*, 2022.
- Ulrich Hoffmann, Jean-Marc Vesin, Touradj Ebrahimi, and Karin Diserens. An Efficient P300-Based Brain–Computer Interface for Disabled Subjects. *Journal of Neuroscience Methods*, 167(1): 115–125, 2008.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=nZeVKeeFYf9>.
- Shengding Hu, Yuge Tu, Xu Han, Ganqu Cui, Chaoqun He, Weilin Zhao, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Xinrong Zhang, Zhen Leng Thai, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, da-hai li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the Potential of Small Language Models with Scalable Training Strategies. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=3X2L2TFr0f>.
- Ji-Hoon Jeong, Jeong-Hyun Cho, Young-Eun Lee, Seo-Hyun Lee, Gi-Hwan Shin, Young-Seok Kweon, José del R Millán, Klaus-Robert Müller, and Seong-Whan Lee. 2020 International Brain–Computer Interface Competition: A Review. *Frontiers in Human Neuroscience*, 16:898300, 2022.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *ArXiv Preprint ArXiv:2310.06825*, 2023.
- Weibang Jiang, Liming Zhao, and Bao liang Lu. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024.
- Jin Jing, Wendong Ge, Shenda Hong, Marta Bento Fernandes, Zhen Lin, Chaoqi Yang, Sungtae An, Aaron F Struck, Aline Herlopian, Ioannis Karakis, et al. Development of Expert-Level Classification of Seizures and Rhythmic and Periodic Patterns During EEG Interpretation. *Neurology*, 100(17):e1750–e1762, 2023.
- Michael J. Kahana, Joseph H. Rudoler, Lynn J. Lohnas, Karl Healey, Ada Aka, Adam Broitman, Elizabeth Crutchley, Patrick Crutchley, Kylie H. Alm, Brandon S. Katerman, Nicole E. Miller, Joel R. Kuhn, Yuxuan Li, Nicole M. Long, Jonathan Miller, Madison D. Paron, Jesse K. Pazdera, Isaac Pedisich, and Christoph T. Weidemann. "Penn Electrophysiology of Encoding and Retrieval Study (Peers)", 2023.
- Dayal Singh Kalra and Maissam Barkeshli. Why Warmup the Learning Rate? Underlying Mechanisms and Improvements. *ArXiv Preprint ArXiv:2406.09405*, 2024.
- Emmanuel K Kalunga, Sylvain Chevallier, and Quentin Barthélémy. Using Riemannian Geometry for SSVEP-Based Brain Computer Interface. *ArXiv Preprint ArXiv:1501.03227*, 2015.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models. *ArXiv Preprint ArXiv:2001.08361*, 2020.

Zoltan Kekecs and Yeganeh Farahzadi. "OTKA PLB-HYP Study1", 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, 2016.

Hassan Aqeel Khan, Rahat Ul Ain, Awais Mehmood Kamboh, Hammad Tanveer Butt, Saima Shafait, Wasim Alamgir, Didier Stricker, and Faisal Shafait. The NMT Scalp EEG Dataset: An Open-Source Annotated Dataset of Healthy and Pathological EEG Recordings for Predictive Modeling. *Frontiers in Neuroscience*, 15:755817, 2022.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

Louis Korczowski, Martine Cederhout, Anton Andreev, Grégoire Cattan, Pedro Luiz Coelho Rodrigues, Violette Gautheret, and Marco Congedo. *Brain Invaders Calibration-Less P300-Based BCI With Modulation of Flash Duration Dataset (Bi2015A)*. PhD thesis, GIPSA-lab, 2019a.

Louis Korczowski, Ekaterina Ostaschenko, Anton Andreev, Grégoire Cattan, Pedro Luiz Coelho Rodrigues, Violette Gautheret, and Marco Congedo. *Brain Invaders Calibration-Less P300-Based BCI Using Dry EEG Electrodes Dataset (Bi2014A)*. PhD thesis, GIPSA-lab, 2019b.

Louis Korczowski, Ekaterina Ostaschenko, Anton Andreev, Grégoire Cattan, Pedro Luiz Coelho Rodrigues, Violette Gautheret, and Marco Congedo. *Brain Invaders Solo Versus Collaboration: Multi-User P300-Based Brain-Computer Interface Dataset (Bi2014B)*. PhD thesis, GIPSA-lab, 2019c.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-Tuning Can Distort Pretrained Features and Underperform Out-Of-Distribution. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=UYneFzXSJWh>.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. EEGnet: A Compact Convolutional Neural Network for EEG-Based Brain-Computer Interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018.

Min-Ho Lee, O-Yeon Kwon, Yong-Jeong Kim, Hong-Kyung Kim, Young-Eun Lee, John Williamson, Siamac Fazli, and Seong-Whan Lee. EEG Dataset and OpenBMI Toolbox for Three BCI Paradigms: An Investigation Into BCI Illiteracy. *Gigascience*, 8(5):giz002, 2019.

Robert Leeb, Felix Lee, Claudia Keinrath, Reinhold Scherer, Horst Bischof, and Gert Pfurtscheller. Brain-Computer Communication: Motivation, Aim, and Impact of Exploring a Virtual Apartment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 15(4):473–482, 2007.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer Normalization. *ArXiv e-Prints*, pages ArXiv–1607, 2016.

Hongli Li, Man Ding, Ronghua Zhang, and Chunbo Xiu. Motor Imagery EEG Classification Algorithm Based on CNN-LSTM Feature Fusion Network. *Biomedical Signal Processing and Control*, 72:103342, 2022.

Weilong Li and Jiaxin Zhao. "PerceiveImagine", 2024.

Haijie Liu, Penghu Wei, Haochong Wang, Xiaodong Lv, Wei Duan, Meijie Li, Yan Zhao, Qingmei Wang, Xinyuan Chen, Gaige Shi, et al. An EEG Motor Imagery Dataset for Brain Computer Interface in Acute Stroke Patients. *Scientific Data*, 11(1):131, 2024.

- 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.
- Benjamin Lowe, Jonathan Robinson, Naohide Yamamoto, Hinze Hogendoorn, and Patrick Johnston. "Visual Attribute-Specific Contextual Trajectory Paradigm", 2023.
- Dominique Makowski, An-Shu Te, Stephanie Kirk, and Zi Liang Ngoi. "FakeFaceEmo Data", 2023.
- Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the SDEs and Scaling Rules for Adaptive Gradient Algorithms. *Advances in Neural Information Processing Systems*, 35: 7697–7711, 2022.
- Christopher J Markiewicz, Krzysztof J Gorgolewski, Franklin Feingold, Ross Blair, Yaroslav O Halchenko, Eric Miller, Nell Hardcastle, Joe Wexler, Oscar Esteban, Mathias Goncalves, et al. The OpenNeuro Resource for Sharing of Neuroscience Data. *eLife*, 10:e71774, 2021.
- Donia Metwalli, Eslam Ahmed, Antony Emil, Yousef A. Radwan, Mariam Barakat, and Anas Ahmed. "ArEEG: Arabic Inner Speech EEG Dataset", 2024.
- Denise Moerel, Tijl Grootswagers, Amanda Robinson, Sophia Shatek, Alexandra Woolgar, Thomas Carlson, and Anina Rich. "The Time-Course of Feature-Based Attention Effects Dissociated from Temporal Expectation and Target-Related Processes", 2022.
- Navid Mohammadi Foumani, Geoffrey Mackellar, Soheila Ghane, Saad Irtza, Nam Nguyen, and Mahsa Salehi. EEG2Rep: Enhancing Self-Supervised EEG Representation Through Informative Masked Inputs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5544–5555, 2024.
- Wajid Mumtaz. MDD Patients and Healthy Controls EEG Data (New). *Figshare, Dataset*, 2016.
- Masaki Nakanishi, Yijun Wang, Yu-Te Wang, and Tzyy-Ping Jung. A Comparison Study of Canonical Correlation Analysis Based Methods for Detecting Steady-State Visual Evoked Potentials. *PLOS One*, 10(10):e0140703, 2015.
- Iyad Obeid and Joseph Picone. The Temple University Hospital EEG Data Corpus. *Frontiers in Neuroscience*, 10:196, 2016.
- Patrick Ofner, Andreas Schwarz, Joana Pereira, and Gernot R Müller-Putz. Upper Limb Movements Can Be Decoded From the Time-Domain of Low-Frequency EEG. *PLOS One*, 12(8):e0182578, 2017.
- Julie Onton and Scott Makeig. "Imagined Emotion Study", 2022.
- Robert Oostenveld and Peter Praamstra. The Five Percent Electrode System for High-Resolution EEG and ERP Measurements. *Clinical Neurophysiology*, 112(4):713–719, 2001.
- Tasos Papastylianou, Rodrigo Ramele, Luca Citi, Caterina Cinel, and Riccardo Poli. "PES - Pandemic Emergency Scenario", 2023.
- Wei Yan Peh, Yuanyuan Yao, and Justin Dauwels. Transformer Convolutional Neural Networks for Automated Artifact Detection in Scalp EEG. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3599–3602. IEEE, 2022.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models from Natural Language Supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.
- Shanker Ram, Sambhu Ganesan, and Yajat Nagaraj Kiran. Harmful Brain Activity Classification of Spectrograms with Transfer Deep Learning. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 499–502. IEEE, 2024.
- Maria J. Ribeiro and Miguel Castelo-Branco. "EEG, ECG and Pupil Data from Young and Older Adults: Rest and Auditory Cued Reaction Time Tasks", 2021.

Angela Riccio, Luca Simione, Francesca Schettini, Alessia Pizzimenti, Maurizio Inghilleri, Marta Olivetti Belardinelli, Donatella Mattia, and Febo Cincotti. Attention and P300-Based BCI Performance in People with Amyotrophic Lateral Sclerosis. *Frontiers in Human Neuroscience*, 7:732, 2013.

Alexander P. Rockhill, Nicko Jackson, Jobi George, Adam Aron, and Nicole C. Swann. "UC San Diego Resting State EEG Data from Patients with Parkinson's Disease", 2020.

Joseph H. Rudoler, Matthew R. Dougherty, Brandon S. Katerman, James P. Bruska, Woohyeuk Chang, David J. Halpern, Nicholas B. Diamond, and Michael J. Kahana. "Spatial Memory and Non-Invasive Closed-Loop Stimulus Timing", 2023.

Reinhold Scherer, Josef Faller, Elisabeth VC Friedrich, Eloy Opisso, Ursula Costa, Andrea Kübler, and Gernot R Müller-Putz. Individually Adapted Imagery Improves Brain-Computer Interface Performance in End-Users with Disability. *PLOS One*, 10(5):e0123727, 2015.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep Learning with Convolutional Neural Networks for EEG Decoding and Visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017.

Tong Shan, Madeline S. Cappelloni, and Ross K. Maddox. "Music and Speech Elicit Similar Subcortical Responses in Human Listeners", 2022.

Sophia M. Shatek, Amanda K. Robinson, Tijl Grootswagers, and Thomas A. Carlson. "Capacity for Movement Is a Major Organisational Principle in Object Representations: EEG Data from Experiment 2", 2021.

Sophia M. Shatek, Amanda K. Robinson, Tijl Grootswagers, and Thomas A. Carlson. "Capacity for Movement Is an Organisational Principle in Object Representations: EEG Data from Experiment 2", 2023.

Noam Shazeer. GLU Variants Improve Transformer. *ArXiv Preprint ArXiv:2002.05202*, 2020.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive Learning Rates with Sublinear Memory Cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR, 2018.

Jaeyoung Shin, Alexander von Lühmann, Benjamin Blankertz, Do-Won Kim, Jichai Jeong, Han-Jeong Hwang, and Klaus-Robert Müller. Open Access Dataset for EEG+ Nirs Single-Trial Classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1735–1745, 2016.

Seyed Yahya Shirazi, Alexandre Franco, Maurício Scopel Hoffmann, Nathalia B. Esper, Dung Truong, Arnaud Delorme, Michael Milham, and Scott Makeig. "Healthy Brain Network (HBN) EEG - Release 4", 2024a.

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*, pages 2024–10, 2024b.

Seyed Yahya Shirazi, Alexandre Franco, Maurício Scopel Hoffmann, Nathalia B. Esper, Dung Truong, Arnaud Delorme, Michael Milham, and Scott Makeig. "Healthy Brain Network (HBN) EEG - Release 5", 2025.

Mohammad Shoeybi, Mostafa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. *ArXiv Preprint ArXiv:1909.08053*, 2019.

Elizabeth M. Siefert, Sindhuja Uppuluri, Jianing Mu, Marlie C. Tandoc, James W. Antony, and Anna C. Schapiro. "Siefert2024", 2024.

Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie. Transformer-Based Spatial-Temporal Feature Learning for EEG Decoding. *ArXiv Preprint ArXiv:2106.11170*, 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, 2022.
- Jan Sosulski, David Hübner, Aaron Klein, and Michael Tangermann. Online Optimization of Stimulation Speed in an Auditory Brain-Computer Interface Under Time Constraints. *ArXiv Preprint ArXiv:2109.06011*, 2021.
- Toyotaro Suzumura, Hiroki Kanezashi, and Shotaro Akahori. Graph Adapter of EEG Foundation Models for Parameter Efficient Fine Tuning. *ArXiv Preprint ArXiv:2411.16155*, 2024.
- 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.
- Jack E. Taylor, Rasmus Sinn, Cosimo Iaia, and Christian J. Fiebach. "Alphabetic Decision Task (Arial Light Font)", 2024.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and Efficient Foundation Language Models. *ArXiv Preprint ArXiv:2302.13971*, 2023.
- Hanneke Van Dijk, Guido Van Wingen, Damiaan Denys, Sebastian Olbrich, Rosalinde Van Ruth, and Martijn Arns. The Two Decades Brainclinics Research Archive for Insights in Neurophysiology (TDBRAIN) Database. *Scientific Data*, 9(1):333, 2022.
- A Vaswani. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 2017.
- J. Veillette, S. Heald, B. Wittenbrink, and H. Nusbaum. "EEG-Neuroforecasting", 2022.
- John Veillette, Pedro Lopes, and Howard Nusbaum. "Illusion of Agency Over Electrically-Actuated Movements", 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 *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*, 2024a.
- 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 Preprint ArXiv:2412.07236*, 2024b.
- Yulin Wang, Wei Duan, Debo Dong, Lihong Ding, and Xu Lei. "A Test-Retest Resting and Cognitive State EEG Dataset", 2022.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. *ArXiv Preprint ArXiv:2412.13663*, 2024.
- Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model Soups: Averaging Weights of Multiple Fine-Tuned Models Improves Accuracy Without Increasing Inference Time. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR, 17–23 Jul 2022. URL <https://proceedings.mlr.press/v162/wortsman22a.html>.
- Mitchell Wortsman, Tim Dettmers, Luke Zettlemoyer, Ari Morcos, Ali Farhadi, and Ludwig Schmidt. Stable and Low-Precision Training for Large-Scale Vision-Language Models. *Advances in Neural Information Processing Systems*, 36:10271–10298, 2023.
- Chuqin Xiang, Xinrui Fan, Duo Bai, Ke Lv, and Xu Lei. "A Resting-State EEG Dataset for Sleep Deprivation", 2024.

- Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A Simple Framework for Masked Image Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022.
- Zhengzhuo Xu, Ruiyang Liu, Shuo Yang, Zenghao Chai, and Chun Yuan. Learning Imbalanced Data with Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15793–15803, 2023.
- Chaoqi Yang, Danica Xiao, M Brandon Westover, and Jimeng Sun. Self-Supervised EEG Representation Learning for Automatic Sleep Staging. *ArXiv Preprint ArXiv:2110.15278*, 2021.
- Chaoqi Yang, M Westover, and Jimeng Sun. BIOT: Biosignal Transformer for Cross-Data Learning in the Wild. *Advances in Neural Information Processing Systems*, 36, 2024.
- Weibo Yi, Shuang Qiu, Kun Wang, Hongzhi Qi, Lixin Zhang, Peng Zhou, Feng He, and Dong Ming. Evaluation of EEG Oscillatory Patterns and Cognitive Process During Simple and Compound Limb Motor Imagery. *PLOS One*, 9(12):e114853, 2014.
- Zhizhang Yuan, Fanqi Shen, Meng Li, Yuguo Yu, Chenhao Tan, and Yang Yang. BrainWave: A Brain Signal Foundation Model for Clinical Applications, 2024a. URL <https://arxiv.org/abs/2402.10251>.
- Zhizhang Yuan, Daoze Zhang, Junru Chen, Geifei Gu, and Yang Yang. Brant-2: Foundation Model for Brain Signals. *ArXiv Preprint ArXiv:2402.10251*, 2024b.
- Biao Zhang and Rico Sennrich. Root Mean Square Layer Normalization. *Advances in Neural Information Processing Systems*, 32, 2019.
- Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=r1Ddp1-Rb>.
- Bangyan Zhou, Xiaopei Wu, Zhao Lv, Lei Zhang, and Xiaojin Guo. A Fully Automated Trial Selection Method for Optimization of Motor Imagery Based Brain-Computer Interface. *PLOS One*, 11(9):e0162657, 2016.
- Natalia Zhozhikashvili, Maria Protopova, Tatiana Shkurenko, Marie Arsalidou, Ilya Zakharov, Boris Kotchoubey, Sergey Malykh, and Yuri Pavlov. "Sternberg Difficult", 2024.
- Igor Zyma, Sergii Tukaev, Ivan Seleznev, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms During Mental Arithmetic Task Performance. *Data*, 4(1):14, 2019.

## Appendix

### A Configurations

We report the hyperparameters used to train the REVE suite of models, including data preprocessing steps, self-supervised masking configurations, and optimizer settings governing the training dynamics. Notations are consistent with those in the main text.

Table 5: Exhaustive list of all hyperparameter values

Variable	Meaning	Value
<b>Data preprocessing</b>		
$w$	Window size	1s
$o$	Overlap	0.1s
$\sigma_{\text{noise}}$	Position noise std	0.25cm
<b>Masking parameters</b>		
$M_r$	Total masking ratio	55%
$R_s$	Spatial masking radius	3 cm
$R_t$	Temporal masking radius	3 seconds
$D_r$	Dropout ratio	10%
$R_d$	Dropout spatial radius	4 cm
<b>Training dynamics</b>		
Optimizer	StableAdamW	
Scheduler	Warmup Stable Decay	
$\eta$	Peak learning rate	$\eta = 2.4 \cdot 10^{-4}$
$\beta_1, \beta_2$	Momentum constants	0.9, 0.95
$\varepsilon$	Numerical stability bias	$10^{-9}$
$\sigma_{\text{init}}$	Initialization std	0.02
	Batch size	4,096
$\lambda$	Secondary loss multiplier	0.1

We report how the scaled number of parameters is allocated across our models. We also indicate the number of Fourier frequencies encoded (see Section 2.2). Note that no frequency truncation was required, as we closely matched the hidden dimension of our models to the number of components generated by the 4D PE module.

Table 6: Summary of encoder configurations for different sizes

Size	depth	n_heads	dim	params (M)	$n_{\text{freq}}$
Small	4	8	512	12	4
Base	22	8	512	69	4
Large	22	19	1250	408	5

### B Pretraining dataset

We include a summarized description of the pretraining dataset composition, grouped by category, platform of origin and number of channels. The final dataset spans 61,415 hours of recordings from 92 datasets encompassing 24,274 subjects.

Table 7: Detailed overview of the pretraining datasets.

Group	Subjects	Duration (hours)	Datasets
<b>Category</b>			
BCI	791	457	28
Cognition	4,193	10,376	56
Clinic	19,290	50,581	8
<b>Platform</b>			
TUH	14,987	26,847	1
Physionet	607	22,707	2
OpenNeuro	4153	10,194	56
MOABB	711	384	27
Other	3,802	1,250	6
<b>Channels</b>			
[3 – 30[	19,871	50,870	31
[30 – 80[	1,781	1,516	48
[80 – 129]	2,622	9,027	13
<b>Total</b>	<b>24,274</b>	<b>61,415</b>	<b>92</b>

We provide and exhaustive list of the datasets in the pretraining set, along with their respective licenses.

**MOABB (Aristimunha et al., 2023):** AlexMI (Barachant, 2012), BNCI2014004 (Leeb et al., 2007), BNCI2015001 (Faller et al., 2012), BNCI2015004 (Scherer et al., 2015), Cho2017 (Cho et al., 2017), Lee2019MI (Lee et al., 2019), Liu2024 (Liu et al., 2024), Ofner2017 (Ofner et al., 2017), Shin2017A (Shin et al., 2016), Weibo2014 (Yi et al., 2014), Zhou2016 (Zhou et al., 2016), Schirrmeister2017 (Schirrmeister et al., 2017), Kalunga2016 (Kalunga et al., 2015), Lee2019SSVEP (Lee et al., 2019), Nakanishi2015 (Nakanishi et al., 2015), BI2014a (Korczowski et al., 2019b), BI2014b (Korczowski et al., 2019c), BNCI2014008 (Riccio et al., 2013), BNCI2014009 (Aricò et al., 2014), BNCI2015003 (Guger et al., 2009), EPFLP300 (Hoffmann et al., 2008), BI2015a (Korczowski et al., 2019a), BI2015b (Korczowski et al., 2019c), Sosulski2019 (Sosulski et al., 2021), Lee2019ERP (Lee et al., 2019)

MOABB is under a BSD 3-Clause License.

**Physionet (Goldberger et al., 2000):** Siena (Detti, 2020; Detti et al., 2020), under the Creative Commons Attribution 4.0 International Public License, ICARE (Amorim et al., 2023) under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Public License,

**OpenNeuro:** ds004706 (Rudoler et al., 2023), ds004582 (Makowski et al., 2023), ds004356 (Shan et al., 2022), ds004817 (Grootswagers et al., 2023b), ds005189 (Helbing et al., 2024), ds003887 (Shatek et al., 2023), ds004043 (Moerel et al., 2022), ds003885 (Shatek et al., 2021), ds004357 (Grootswagers et al., 2024), ds003825 (Grootswagers et al., 2022), ds004816 (Grootswagers et al., 2023a), ds004840 (Cordoba-Silva et al., 2023), ds005262 (Metwalli et al., 2024), ds004477 (Papastylianou et al., 2023), ds005273 (Esteban et al., 2024), ds004561 (Veillette et al., 2023), ds004951 (Haupt et al., 2024), ds004324 (Chacón and Wriessnegger, 2023), ds005095 (Zhozhikashvili et al., 2024), ds005509 (Shirazi et al., 2025), ds005505, ds005506, ds005507, ds005510, ds005511, ds005512, ds005514 (Shirazi et al., 2024b; Alexander et al., 2017) ds001787 (Delorme and Brandmeyer, 2024), ds003690 (Ribeiro and Castelo-Branco, 2021), ds004603 (Lowe et al., 2023), ds003969 (Delorme and Braboszcz, 2021), ds004147 (Hassall et al., 2024), ds003004 (Onton and Makeig, 2022), ds002721 (Daly et al., 2020), ds004152 (Hassall et al., 2022a), ds005089 (Aguado-Lopez et al., 2024), ds004264 (Hassall et al., 2022b), ds004315 (Cavanagh and Jackson, 2022), ds004408 (Bialas et al., 2023), ds005121 (Siefert et al., 2024), ds003775 (Hatlestad-Hall et al., 2022), ds004572 (Kekecs and Farahzadi, 2024), ds002778 (Rockhill et al., 2020), ds003846 (Gehrke et al., 2024), ds004279 (Araya et al., 2023), ds004148 (Wang et al., 2022), ds004902 (Xiang et al., 2024), ds002680 (Delorme and Fabre-Thorpe, 2020), ds004284 (Veillette et al., 2022), ds004395 (Kahana et al., 2023),

ds005508 (Shirazi et al., 2024a), ds005697 (Li and Zhao, 2024), ds005620 (Bajwa1 et al., 2024), ds005594 (Taylor et al., 2024), ds005586 (Baykan and Schütz, 2024). OpenNeuro is under the Creative Commons CC0 license.

**Other sources:** NMT (Khan et al., 2022) under the Creative Commons Attribution License (CC BY), HMS (Ram et al., 2024) under the Attribution-NonCommercial 4.0 International (CC-BY-NC-4.0), SparrKULee (Accou et al., 2023) under the Attribution-Non Commercial 4.0 International (CC-BY-NC-4.0), Inria Large (Dreyer et al., 2023) the data on Zenodo being under the Creative Commons Attribution 4.0 International, THINGS2 (Gifford et al., 2022), under the CC-By Attribution 4.0 International license, TDBRAIN (Van Dijk et al., 2022), under the GPL-3.0 license, TUH (Obeid and Picone, 2016), freely available with registration required.

## C Detailed results

This section presents detailed results on downstream tasks along with concise descriptions of the datasets.

### C.1 Emotion Recognition

**FACED** (Chen et al., 2023) We evaluate on the FACED dataset, which contains 32-channel EEG recordings (originally at 250 Hz, resampled to 200 Hz) from 123 subjects across nine emotion classes. The data is segmented into 10,332 samples of 10 seconds each. We follow the standard split: subjects 1–80 for training, 81–100 for validation, and 101–123 for testing.

Table 8: The results of different methods on emotion recognition (FACED, 9-class).

Methods	Balanced Accuracy	Cohen’s Kappa	Weighted F1
EEGNet	$0.4090 \pm 0.0122$	$0.3342 \pm 0.0251$	$0.4124 \pm 0.0141$
EEGConformer	$0.4559 \pm 0.0125$	$0.3858 \pm 0.0186$	$0.4514 \pm 0.0107$
SPaRCNet	$0.4673 \pm 0.0155$	$0.3978 \pm 0.0289$	$0.4729 \pm 0.0133$
ContraWR	$0.4887 \pm 0.0078$	$0.4231 \pm 0.0151$	$0.4884 \pm 0.0074$
CNN-Transformer	$0.4697 \pm 0.0132$	$0.4017 \pm 0.0168$	$0.4720 \pm 0.0125$
FFCL	$0.4673 \pm 0.0158$	$0.3987 \pm 0.0383$	$0.4699 \pm 0.0145$
ST-Transformer	$0.4810 \pm 0.0079$	$0.4137 \pm 0.0133$	$0.4795 \pm 0.0096$
BIOT	$0.5118 \pm 0.0118$	$0.4476 \pm 0.0254$	$0.5136 \pm 0.0112$
LaBraM-Base	$0.5273 \pm 0.0107$	$0.4698 \pm 0.0188$	$0.5288 \pm 0.0102$
CBraMod	$0.5509 \pm 0.0089$	$0.5041 \pm 0.0122$	$0.5618 \pm 0.0093$
REVE-Base (ours)	<b><math>0.5646 \pm 0.0164</math></b>	<b><math>0.5080 \pm 0.0191</math></b>	<b><math>0.5659 \pm 0.0172</math></b>

### C.2 Mental Disorder Diagnosis

**Mumtaz** (Mumtaz, 2016) We use the Mumtaz2016 dataset, which includes EEG recordings from 34 individuals with major depressive disorder (MDD) and 30 healthy controls, acquired from 19 electrodes (10–20 system) at 256 Hz. Only the eyes-open and eyes-closed sessions are used. Signals are band-pass filtered (0.3–75 Hz), notch filtered at 50 Hz, resampled to 200 Hz, and segmented into 7,143 samples of 5 seconds each. The split includes 24 MDD and 19 control subjects for training, 5 MDD and 4 controls for validation, and 5 MDD and 5 controls for testing. The dataset is under CC BY 4.0.

Table 9: The results of different methods on mental disorder diagnosis (Mumtaz2016, 2-class).

Methods	Balanced Accuracy	AUC-PR	AUROC
EEGNet	$0.9232 \pm 0.0104$	$0.9626 \pm 0.0095$	$0.9639 \pm 0.0093$
EEGConformer	$0.9308 \pm 0.0117$	$0.9684 \pm 0.0105$	$0.9702 \pm 0.0101$
SPaRCNet	$0.9316 \pm 0.0095$	$0.9754 \pm 0.0065$	$0.9781 \pm 0.0083$
ContraWR	$0.9195 \pm 0.0115$	$0.9589 \pm 0.0102$	$0.9621 \pm 0.0092$
CNN-Transformer	$0.9305 \pm 0.0068$	$0.9757 \pm 0.0074$	$0.9742 \pm 0.0059$
FFCL	$0.9314 \pm 0.0038$	$0.9717 \pm 0.0021$	$0.9753 \pm 0.0033$
ST-Transformer	$0.9135 \pm 0.0103$	$0.9578 \pm 0.0086$	$0.9594 \pm 0.0059$
BIOT	$0.9358 \pm 0.0052$	$0.9736 \pm 0.0034$	$0.9758 \pm 0.0042$
LaBraM-Base	$0.9409 \pm 0.0079$	$0.9798 \pm 0.0093$	$0.9782 \pm 0.0057$
CBraMod	$0.9560 \pm 0.0056$	$0.9923 \pm 0.0032$	$0.9921 \pm 0.0025$
REVE-Base (ours)	<b><math>0.9644 \pm 0.0097</math></b>	<b><math>0.9961 \pm 0.0013</math></b>	<b><math>0.9957 \pm 0.0015</math></b>

### C.3 Mental Stress Detection

**MAT** (Zyma et al., 2019) The MentalArithmetic dataset contains EEG recordings from 36 subjects, labeled as “with” or “without” mental stress depending on whether a mental arithmetic task was being performed. Signals were recorded from 20 electrodes (10–20 system) at 500 Hz, band-pass filtered (0.5–45 Hz), resampled to 200 Hz, and segmented into 1,707 samples of 5 seconds. Subjects 1–28 are used for training, 29–32 for validation, and 33–36 for testing. The MentalArithmetic dataset is under the Open Data Commons Attribution License v1.0.

Table 10: The results of different methods on mental stress detection (MAT, 2-class).

Methods	Balanced Accuracy	AUC-PR	AUROC
EEGNet	$0.6770 \pm 0.0116$	$0.5763 \pm 0.0102$	$0.7321 \pm 0.0108$
EEGConformer	$0.6805 \pm 0.0123$	$0.5829 \pm 0.0134$	$0.7424 \pm 0.0128$
SPaRCNet	$0.6879 \pm 0.0107$	$0.5825 \pm 0.0193$	$0.7418 \pm 0.0132$
ContraWR	$0.6631 \pm 0.0097$	$0.5787 \pm 0.0164$	$0.7332 \pm 0.0082$
CNN-Transformer	$0.6779 \pm 0.0268$	$0.5777 \pm 0.0285$	$0.7258 \pm 0.0336$
FFCL	$0.6798 \pm 0.0142$	$0.5786 \pm 0.0266$	$0.7330 \pm 0.0198$
ST-Transformer	$0.6631 \pm 0.0173$	$0.5672 \pm 0.0259$	$0.7132 \pm 0.0174$
BIOT	$0.6875 \pm 0.0186$	$0.6004 \pm 0.0195$	$0.7536 \pm 0.0144$
LaBraM-Base	$0.6909 \pm 0.0125$	$0.5999 \pm 0.0155$	$0.7721 \pm 0.0093$
CBraMod	$0.7256 \pm 0.0132$	$0.6267 \pm 0.0099$	$0.7905 \pm 0.0073$
REVE-Base (ours)	<b><math>0.7660 \pm 0.0355</math></b>	<b><math>0.7470 \pm 0.0807</math></b>	<b><math>0.8450 \pm 0.0514</math></b>

### C.4 Imagined Speech

**BCIC2020-3** (Jeong et al., 2022) BCIC2020-3 is an imagined speech EEG dataset from 15 subjects, recorded with 64 channels at 256 Hz while subjects silently imagined five phrases (“hello”, “help me”, “stop”, “thank you”, “yes”) without any articulation. Each phrase has 80 trials per subject, totaling 6,000 3-second samples. The data is resampled to 200 Hz. The official split includes 60 trials per class for training, 10 for validation, and 10 for testing. BCIC2020-3 is under the Creative Commons Attribution No Derivatives license (CC BY-ND 4.0).

Table 11: The results of different methods on imagined speech classification (BCIC2020-3, 5-class).

Methods	Balanced Accuracy	Cohen's Kappa	Weighted F1
EEGNet	$0.4413 \pm 0.0096$	$0.3016 \pm 0.0123$	$0.4413 \pm 0.0102$
EEGConformer	$0.4506 \pm 0.0133$	$0.3133 \pm 0.0183$	$0.4488 \pm 0.0154$
SPaRCNet	$0.4426 \pm 0.0156$	$0.3033 \pm 0.0233$	$0.4420 \pm 0.0108$
ContraWR	$0.4257 \pm 0.0162$	$0.3078 \pm 0.0218$	$0.4407 \pm 0.0182$
CNN-Transformer	$0.4533 \pm 0.0092$	$0.3166 \pm 0.0118$	$0.4506 \pm 0.0127$
FFCL	$0.4678 \pm 0.0197$	$0.3301 \pm 0.0359$	$0.4689 \pm 0.0205$
ST-Transformer	$0.4126 \pm 0.0122$	$0.2941 \pm 0.0159$	$0.4247 \pm 0.0138$
BIOT	$0.4920 \pm 0.0086$	$0.3650 \pm 0.0176$	$0.4917 \pm 0.0079$
LaBraM-Base	$0.5060 \pm 0.0155$	$0.3800 \pm 0.0242$	$0.5054 \pm 0.0205$
CBraMod	$0.5373 \pm 0.0108$	$0.4216 \pm 0.0163$	$0.5383 \pm 0.0096$
REVE-Base (ours)	<b><math>0.5635 \pm 0.0123</math></b>	<b><math>0.4543 \pm 0.0154</math></b>	<b><math>0.5633 \pm 0.0124</math></b>

## C.5 Motor Imagery Classification

**PhysioNet-MI** (Goldberger et al., 2000) is used for motor imagery classification. It contains recordings with 64 channels at a 160 Hz sampling rate and includes 4 classes: left fist, right fist, both fists, and feet. As in CBraMod, we select 4-second samples of the signals, resulting in 9,837 samples. Following CBraMod’s protocol, subjects 1–70 are used for training, 71–89 for validation, and 90–109 for testing. We retain all subjects and use full 4-second windows to stay consistent with CBraMod. To handle lower sampling rates in some recordings, we load all data at 128 Hz (using a 64 Hz low-pass filter) before resampling to 200 Hz. Physionet-MI is under the Open Data Commons Attribution License v1.0.

**BCIC-IV-2a** (Tangermann et al., 2012) is also used for motor imagery classification. It contains EEG recordings from 9 subjects performing 4 motor imagery tasks: left hand, right hand, both feet, and tongue. Data were collected over 2 sessions with 22 electrodes at 250 Hz. Each session includes 288 trials (72 per task). We use the [2,6] second window from each trial, apply a 0.5–99.5 Hz band-pass filter, resample to 200 Hz, and apply Euclidean Alignment (He and Wu, 2019), proven to be effective on this task (El Ouahidi et al., 2024), resulting in 5184 4-second samples.

Table 12: The results of different methods on Motor Imagery classification.

Methods	PhysioNet-MI, 4-class			BCIC-IV-2a, 4-class		
	Balanced Accuracy	Cohen's Kappa	Weighted F1	Balanced Accuracy	Cohen's Kappa	Weighted F1
EEGNet	$0.5814 \pm 0.0125$	$0.4468 \pm 0.0199$	$0.5796 \pm 0.0115$	$0.4482 \pm 0.0094$	$0.2693 \pm 0.0121$	$0.4226 \pm 0.0108$
EEGConformer	$0.6049 \pm 0.0104$	$0.4736 \pm 0.0171$	$0.6062 \pm 0.0095$	$0.4696 \pm 0.0106$	$0.2924 \pm 0.0141$	$0.4533 \pm 0.0128$
SPaRCNet	$0.5932 \pm 0.0152$	$0.4564 \pm 0.0234$	$0.5937 \pm 0.0147$	$0.4635 \pm 0.0117$	$0.2847 \pm 0.0147$	$0.4432 \pm 0.0126$
ContraWR	$0.5892 \pm 0.0133$	$0.4527 \pm 0.0248$	$0.5918 \pm 0.0116$	$0.4678 \pm 0.0125$	$0.2905 \pm 0.0160$	$0.4413 \pm 0.0142$
(CNN-Transformer)	$0.6053 \pm 0.0118$	$0.4725 \pm 0.0223$	$0.6041 \pm 0.0105$	$0.4600 \pm 0.0108$	$0.2800 \pm 0.0148$	$0.4460 \pm 0.0114$
FFCL	$0.5726 \pm 0.0092$	$0.4323 \pm 0.0182$	$0.5701 \pm 0.0079$	$0.4470 \pm 0.0143$	$0.2627 \pm 0.0176$	$0.4238 \pm 0.0139$
ST-Transformer	$0.6035 \pm 0.0081$	$0.4712 \pm 0.0199$	$0.6053 \pm 0.0075$	$0.4575 \pm 0.0145$	$0.2733 \pm 0.0198$	$0.4471 \pm 0.0142$
BIOT	$0.6153 \pm 0.0154$	$0.4875 \pm 0.0272$	$0.6158 \pm 0.0197$	$0.4748 \pm 0.0093$	$0.2997 \pm 0.0139$	$0.4607 \pm 0.0125$
LaBraM-Base	$0.6173 \pm 0.0122$	$0.4912 \pm 0.0192$	$0.6177 \pm 0.0141$	$0.4869 \pm 0.0085$	$0.3159 \pm 0.0154$	$0.4758 \pm 0.0103$
CBraMod	$0.6417 \pm 0.0091$	$0.5222 \pm 0.0169$	$0.6427 \pm 0.0100$	$0.5138 \pm 0.0066$	$0.3518 \pm 0.0094$	$0.4984 \pm 0.0085$
REVE-Base (ours)	<b><math>0.6480 \pm 0.0140</math></b>	<b><math>0.5306 \pm 0.0187</math></b>	<b><math>0.6484 \pm 0.0170</math></b>	<b><math>0.6396 \pm 0.0095</math></b>	<b><math>0.5194 \pm 0.0126</math></b>	<b><math>0.6339 \pm 0.0110</math></b>

## C.6 Sleep Staging

**ISRUUC** (Khalighi et al., 2016) We use the sleep staging task on the ISRUUC dataset (Subgroup 1), which contains PSG recordings from 100 subjects. Only EEG signals are used (6 channels, sampled at 200 Hz), segmented into 89,240 30-second epochs, each labeled with one of five sleep stages following AASM standards. Subjects 1–80 are used for training, 81–90 for validation, and 91–100 for testing. As in prior work, the task is framed as a sequence-to-sequence classification problem, using sequences of 20 consecutive epochs to model stage transitions. ISRUUC is freely accessible online.

Table 13: The results of different methods on sleep staging (ISRUC, 5-class). \* In the baseline code, a chin electrode might have been used instead of an EEG one; REVE results are reported without it.

Methods	Balanced Accuracy	Cohen’s Kappa	Weighted F1
EEGNet	$0.7154 \pm 0.0121$	$0.7040 \pm 0.0173$	$0.7513 \pm 0.0124$
EEGConformer	$0.7400 \pm 0.0133$	$0.7143 \pm 0.0162$	$0.7634 \pm 0.0151$
SPaRCNet	$0.7487 \pm 0.0075$	$0.7097 \pm 0.0132$	$0.7624 \pm 0.0092$
ContraWR	$0.7402 \pm 0.0126$	$0.7178 \pm 0.0156$	$0.7610 \pm 0.0137$
CNN-Transformer	$0.7363 \pm 0.0087$	$0.7129 \pm 0.0121$	$0.7719 \pm 0.0105$
FFCL	$0.7277 \pm 0.0182$	$0.7016 \pm 0.0291$	$0.7614 \pm 0.0197$
ST-Transformer	$0.7381 \pm 0.0205$	$0.7013 \pm 0.0352$	$0.7681 \pm 0.0175$
DeepSleepNet	$0.7419 \pm 0.0144$	$0.7036 \pm 0.0241$	$0.7643 \pm 0.0122$
USleep	$0.7586 \pm 0.0116$	$0.7209 \pm 0.0143$	$0.7805 \pm 0.0105$
BIOT	$0.7527 \pm 0.0121$	$0.7192 \pm 0.0231$	$0.7790 \pm 0.0146$
LaBraM-Base	$0.7633 \pm 0.0102$	$0.7231 \pm 0.0182$	$0.7810 \pm 0.0133$
CBraMod	<b><math>0.7865 \pm 0.0110</math></b>	$0.7442 \pm 0.0152$	<b><math>0.8011 \pm 0.0099</math></b>
REVE-Base*	$0.7819 \pm 0.0078$	<b><math>0.7500 \pm 0.0156</math></b>	$0.8005 \pm 0.0135$

**HMC** (Alvarez-Estevez and Rijsman, 2021). The Haaglanden Medisch Centrum (HMC) Sleep Staging Database is a sleep stage detection dataset, consisting of 151 full-night polysomnographic (PSG) recordings collected from patients referred for sleep studies. The data includes EEG, EOG, EMG, and ECG channels, with a sampling rate of 256 Hz, and annotations for five sleep stages (Wake, N1, N2, N3, REM) manually scored by trained sleep technicians. HMC is under the Creative Commons Attribution 4.0 International Public License.

Table 14: The results of different methods on sleep staging (HMC, 5-class).

Methods	Balanced Accuracy	Cohen’s Kappa	Weighted F1
SPaRCNet	$0.4756 \pm 0.1109$	$0.3147 \pm 0.1315$	$0.4108 \pm 0.1310$
ContraWR	$0.4242 \pm 0.0541$	$0.2340 \pm 0.0554$	$0.2987 \pm 0.0288$
CNN-Transformer	$0.6573 \pm 0.0141$	$0.5961 \pm 0.0105$	$0.6896 \pm 0.0065$
FFCL	$0.4427 \pm 0.0702$	$0.2542 \pm 0.0654$	$0.2902 \pm 0.0485$
ST-Transformer	$0.2559 \pm 0.0141$	$0.0503 \pm 0.0183$	$0.1428 \pm 0.0122$
BIOT	$0.6862 \pm 0.0041$	$0.6295 \pm 0.0113$	$0.7091 \pm 0.0147$
LaBraM-Base	$0.7286 \pm 0.0101$	$0.6812 \pm 0.0073$	$0.7554 \pm 0.0024$
REVE-Base	<b><math>0.7401 \pm 0.0075</math></b>	<b><math>0.6982 \pm 0.0078</math></b>	<b><math>0.7638 \pm 0.0074</math></b>

## C.7 Event Type Classification

**TUEV** (Obeid and Picone, 2016) is an EEG dataset with six annotated classes: spike and sharp wave, generalized periodic epileptiform discharges, periodic lateralized epileptiform discharges, eye movement, artifact, and background. The recordings use 23 channels at a 256 Hz sampling rate. For consistency with CBraMod, BIOT, and LaBraM, we used BIOT’s processing scripts which preprocess the dataset using 16 common bipolar montage channels in the 10-20 system, apply a 0.3–75 Hz band-pass filter, remove power line noise with a 60 Hz notch filter, and resample to 200 Hz. The dataset is split into 112,491 5-second samples. We follow the original training-test split and further divide the training set into 80% training and 20% validation, matching BIOT setting. To provide our model with the electrode positions, we used the average position of each bipolar montage. TUEV is part of the TUH dataset, which is freely available with registration required.

Table 15: The results of different methods on event type classification (TUEV, 6-class).

<b>Methods</b>	<b>Balanced Accuracy</b>	<b>Cohen’s Kappa</b>	<b>Weighted F1</b>
EEGNet	$0.3876 \pm 0.0143$	$0.3577 \pm 0.0155$	$0.6539 \pm 0.0120$
EEGConformer	$0.4074 \pm 0.0164$	$0.3967 \pm 0.0195$	$0.6983 \pm 0.0152$
SPaRCNet	$0.4161 \pm 0.0262$	$0.4233 \pm 0.0181$	$0.7024 \pm 0.0104$
ContraWR	$0.4384 \pm 0.0349$	$0.3912 \pm 0.0237$	$0.6893 \pm 0.0136$
CNN-Transformer	$0.4087 \pm 0.0161$	$0.3815 \pm 0.0134$	$0.6854 \pm 0.0293$
FFCL	$0.3979 \pm 0.0104$	$0.3732 \pm 0.0188$	$0.6783 \pm 0.0120$
ST-Transformer	$0.3984 \pm 0.0228$	$0.3765 \pm 0.0306$	$0.6823 \pm 0.0190$
BIOT	$0.5281 \pm 0.0225$	$0.5273 \pm 0.0249$	$0.7492 \pm 0.0082$
LaBraM-Base	$0.6409 \pm 0.0065$	$0.6637 \pm 0.0093$	$0.8312 \pm 0.0052$
LaBraM-Large	$0.6581 \pm 0.0156$	$0.6622 \pm 0.0136$	$0.8315 \pm 0.0040$
LaBraM-Huge	$0.6616 \pm 0.0170$	$0.6745 \pm 0.0195$	$0.8329 \pm 0.0086$
CBraMod	$0.6671 \pm 0.0107$	$0.6772 \pm 0.0096$	$0.8342 \pm 0.0064$
REVE-Base (ours)	<b><math>0.6759 \pm 0.0229</math></b>	<b><math>0.6783 \pm 0.0199</math></b>	<b><math>0.8451 \pm 0.0129</math></b>

### C.8 Abnormal Detection

**TUAB** (Obeid and Picone, 2016) is used for abnormal EEG detection, where recordings are labeled as normal or abnormal. It shares the same 23-channel, 256 Hz format as TUEV. The dataset is split into 409,455 10-second samples for binary classification. We follow the provided training-test split and apply an 80%-20% training-validation split, consistent with BIOT. We resampled at 200 Hz, band-pass at 0.5-99.5 Hz, and directly used all channels and their positions. TUAB is part of the TUH dataset, which is freely available with registration required.

Table 16: The results of different methods on abnormal detection (TUAB, 2-class).

<b>Methods</b>	<b>Balanced Accuracy</b>	<b>AUC-PR</b>	<b>AUROC</b>
EEGNet	$0.7642 \pm 0.0036$	$0.8299 \pm 0.0043$	$0.8412 \pm 0.0031$
EEGConformer	$0.7758 \pm 0.0049$	$0.8427 \pm 0.0054$	$0.8445 \pm 0.0038$
SPaRCNet	$0.7896 \pm 0.0018$	$0.8414 \pm 0.0018$	$0.8676 \pm 0.0012$
ContraWR	$0.7746 \pm 0.0041$	$0.8421 \pm 0.0104$	$0.8456 \pm 0.0074$
CNN-Transformer	$0.7777 \pm 0.0022$	$0.8433 \pm 0.0039$	$0.8461 \pm 0.0013$
FFCL	$0.7848 \pm 0.0038$	$0.8448 \pm 0.0065$	$0.8569 \pm 0.0051$
ST-Transformer	$0.7966 \pm 0.0023$	$0.8521 \pm 0.0026$	$0.8707 \pm 0.0019$
BIOT	$0.7959 \pm 0.0057$	$0.8792 \pm 0.0023$	$0.8815 \pm 0.0043$
LaBraM-Base	$0.8140 \pm 0.0019$	$0.8965 \pm 0.0016$	$0.9022 \pm 0.0009$
LaBraM-Large	$0.8226 \pm 0.0015$	$0.9130 \pm 0.0005$	$0.9127 \pm 0.0005$
LaBraM-Huge	$0.8258 \pm 0.0011$	$0.9204 \pm 0.0011$	$0.9162 \pm 0.0016$
CBraMod	$0.8289 \pm 0.0022$	$0.9258 \pm 0.0008$	$0.9227 \pm 0.0011$
REVE-Base (ours)	<b><math>0.8315 \pm 0.0014</math></b>	<b><math>0.9281 \pm 0.0009</math></b>	<b><math>0.9245 \pm 0.0013</math></b>

### D Ablation on the SSL Method

The final pretraining hyperparameters were selected based on a series of ablation studies, the results of which are presented in this section.

Table 17 reports the impact of the secondary pretraining loss on eight downstream tasks using REVE-Small, evaluated under frozen-backbone, linear probing (LP), and full fine-tuning (FT) settings. Results obtained with both losses are compared to those using only the primary loss. The secondary loss consistently improves performance across nearly all datasets, enhancing results in both LP and FT settings, while its removal leads to a substantial drop, underscoring its importance for the model to produce strong embeddings.

The results in Table 18 show that a block masking ratio of 55% yields the best overall performance, providing stable results across both fine-tuned and frozen settings and eight datasets (Mumtaz, TUAB, ISRUC, HMC, BCIC2020-3, TUEV, PhysioNetMI, and Faced). In contrast, random masking

Table 17: Effect of 2nd loss during pretraining and finetuning. The reported metric is balanced accuracy. Best results per dataset are in bold.

Dataset	LP		FT	
	No 2nd loss	+ 2nd loss	No 2nd loss	+ 2nd loss
Mumtaz	0.818 ± 0.043	0.920 ± 0.018	0.818 ± 0.043	<b>0.922 ± 0.018</b>
TUAB	0.797 ± 0.004	0.802 ± 0.005	0.803 ± 0.003	<b>0.810 ± 0.005</b>
ISRUC	0.699 ± 0.006	0.625 ± 0.003	<b>0.777 ± 0.002</b>	0.770 ± 0.002
HMC	0.598 ± 0.008	0.591 ± 0.005	0.713 ± 0.011	<b>0.723 ± 0.005</b>
BCIC2020-3	0.234 ± 0.009	0.237 ± 0.008	0.390 ± 0.017	<b>0.481 ± 0.008</b>
TUEV	0.442 ± 0.060	0.520 ± 0.005	0.533 ± 0.024	<b>0.623 ± 0.011</b>
PhysioNetMI	0.379 ± 0.058	0.533 ± 0.019	0.563 ± 0.011	<b>0.583 ± 0.009</b>
Faced	0.220 ± 0.008	0.233 ± 0.004	0.302 ± 0.016	<b>0.410 ± 0.004</b>
Avg.	0.523	0.558	0.612	<b>0.665</b>

Table 18: Performance comparison across different masking ratios (0.25, 0.55, 0.75) between block masking strategy and random masking, evaluated for full fine-tuning versus frozen embeddings. We display the average balanced accuracy on the small model over eight downstream tasks.

Masking Ratio	Frozen		Full Fine-Tuning	
	Random	Block	Random	Block
0.25	0.523	0.513	0.612	0.602
0.55	0.550	<b>0.558</b>	0.643	<b>0.665</b>
0.75	0.519	0.546	0.606	0.655

favors smaller ratios (25%), but its unstructured nature leads to highly redundant inputs, making the reconstruction task artificially easier. These findings align with ablation results reported in Cbramod, Labram, and BIOT.

Table 19: Ablation study on PhysioNetMI and Mental Arithmetic datasets. The reported metric is balanced accuracy, with the average computed across both tasks, with the Base model.

\*Note that the learnable positional encoding matches the baseline, but does not allow for the extension to larger time windows or unseen spatial configurations.

Ablated component	PhysionetMI	Mental Arithmetic	Average
Learnable PE*	<b>0.650 ± 0.0113</b>	0.752 ± 0.0421	0.701 ± 0.0218
MLP4D	0.637 ± 0.0056	0.717 ± 0.0425	0.677 ± 0.0214
Position noise	0.628 ± 0.0084	0.692 ± 0.0665	0.660 ± 0.0335
Dropout block masking	0.645 ± 0.0155	0.678 ± 0.0521	0.662 ± 0.0272
Temporal block masking	0.646 ± 0.0155	0.723 ± 0.0422	0.685 ± 0.0225
Base Performance	0.6480 ± 0.0140	<b>0.7660 ± 0.0355</b>	<b>0.707 ± 0.0191</b>

Table 19 presents an ablation study on two downstream tasks to assess the contribution of each component in our SSL pipeline. All components appear to contribute positively to performance. The “Learnable PE” line is not a true ablation, but rather a variant using learnable positional embeddings, where a separate embedding is learned for each electrode and time index observed during pretraining. Although this approach performs well, it is limited to the spatial and temporal configurations seen during training (approximately 400 unique electrode names, over 10-second windows) and does not generalize to longer sequences or unseen electrode layouts, unlike REVE’s 4D positional encoding.

Table 20 presents an ablation study on the choice of activation and normalization functions, an important design factor in transformer-based foundation models. We compare GEGLU + RMSNorm, GELU + RMSNorm, and GEGLU + LayerNorm configurations during pretraining, and report downstream performance after fine-tuning on three datasets using the REVE-Small model.

Table 20: Ablation study on activation functions and normalization layers (GEGLU vs. GELU, RMSNorm vs. LayerNorm). We report downstream balanced accuracy after pretraining the REVE-Small model with each configuration.

<b>Dataset</b>	<b>GEGLU + RMSNorm</b>	<b>GELU + RMSNorm</b>	<b>GEGLU + LayerNorm</b>
BCIC-IV-2a	<b>0.581 ± 0.012</b>	0.560 ± 0.018	0.537 ± 0.018
TUEV	<b>0.623 ± 0.011</b>	0.592 ± 0.010	0.577 ± 0.034
PhysioNetMI	0.583 ± 0.009	<b>0.586 ± 0.009</b>	0.559 ± 0.007
Avg.	<b>0.596</b>	0.579	0.558

The GEGLU + RMSNorm combination achieves the best average performance (0.596), outperforming the others on BCIC-IV-2a and TUEV. GELU + RMSNorm performs similarly but only leads on PhysioNetMI. In contrast, GEGLU + LayerNorm consistently underperforms, highlighting the effectiveness of RMSNorm over LayerNorm and the benefits of gated activations like GEGLU in this context.

## E Additional results

This section presents supplementary experiments that further support the main results, focusing on few-shot performance and evaluation under reduced-electrode configurations.

### E.1 Sparse setups

Table 21: Performance of REVE-Base under sparse input configurations. Balanced accuracy is reported for PhysionetMI (Left–Right) and imagined speech tasks as the number of EEG channels is progressively reduced.

<b>Channels</b>	<b>PhysionetMI L-R</b>	<b>Speech</b>
64	0.824 ± 0.008	0.565 ± 0.016
32	0.808 ± 0.007	0.490 ± 0.094
16	0.787 ± 0.008	0.469 ± 0.014
8	0.781 ± 0.006	0.294 ± 0.063
4	0.728 ± 0.009	0.258 ± 0.019
2	0.700 ± 0.025	0.228 ± 0.006
1	0.660 ± 0.019	0.209 ± 0.008

Table 21 reports REVE-Base’s performance under increasingly sparse input configurations. On the Physionet MI L-R task, accuracy degrades gracefully from 0.824 with 64 channels to 0.660 with a single channel, demonstrating robustness to reduced spatial coverage. In contrast, the imagined speech task is more sensitive to channel sparsity, with performance dropping from 0.565 to 0.258 with four channels and 0.209 with one, close to random chance. These results confirm that while REVE generalizes well under limited input, tasks requiring broad spatial information remain more challenging.

### E.2 Few-shot experiments

We conducted few-shot (FS) experiments to simulate realistic BCI usage scenarios. Tasks were constructed from the BCI IV-2a dataset using two motor imagery classes (Left–Right). For each subject, multiple inductive FS runs were performed. In each run,  $N$  labeled samples per class (“shots”) were randomly selected within a session for training, while the remaining samples from both sessions were used for evaluation.

Classification was done using a Nearest Class Mean (NCM) classifier. Each configuration was repeated 20 times per subject, and we report the average balanced accuracy across subjects and runs. We evaluated two configurations of REVE-Base:

- REVE-Base (PT): directly after self-supervised pretraining, with no further supervised adaptation.
- REVE-Base (XFT): after cross-dataset fine-tuning on multiple labeled Left–Right MI datasets ((Schirrmeister et al., 2017), (Cho et al., 2017), (Goldberger et al., 2000), (Lee et al., 2019), (Yi et al., 2014)). REVE’s 4D positional encoding enables joint training across diverse electrode configurations without requiring channel alignment or selection.

Table 22: Few-shot performance of REVE-Base on BCI IV-2a dataset

N-shots	1	2	5	10	20
REVE-Base (PT)	$0.588 \pm 1.45$	$0.601 \pm 0.001$	$0.652 \pm 0.013$	$0.688 \pm 0.010$	$0.723 \pm 0.010$
REVE-Base (XFT)	$0.605 \pm 1.12$	$0.645 \pm 0.009$	$0.705 \pm 0.009$	$0.768 \pm 0.009$	$0.817 \pm 0.004$

Table 22 shows that REVE-Base achieves competitive accuracy even without supervised adaptation, demonstrating that its pretrained embeddings can be effectively leveraged for downstream BCI tasks. After cross-dataset fine-tuning, performance improves consistently across all shot counts, with gains reaching +10% at 20 shots. This indicates that REVE transfers well across subjects and datasets, while benefiting from minimal supervised adaptation. Such generalization is uncommon among BCI embedding models, which typically require task or subject-specific retraining.

## F Experiment details

### F.1 Compute resources

We include details about the compute nodes that were used for pretraining.

- Compute Type: GPU-accelerated nodes
- GPU Model: NVIDIA A100
- CPU Model: Intel Cascade Lake SP 6248
- CPU Cores per Node: 40 cores
- Total Memory per Node: 192 GB
- Storage: Access to a shared full-flash parallel file system based on IBM Spectrum Scale
- Job Scheduler: Slurm

We also estimate the number of floating-point operations (FLOPs) required to train the REVE-Base model, following the formulation from Chowdhery et al. (2023):

$$\tau = \frac{D \cdot (6N + 12LHQT)}{P \cdot \eta}$$

where  $\tau$  denotes the training time (in seconds),  $D = 60k \times 3600 \times 1.1 \times 68 \times 17$  is the total number of tokens seen during pretraining (corresponding to 60k hours of EEG, an overlap coefficient of 1.1, 68 average channels, and 17 epochs),  $N = 72M$  is the number of model parameters,  $L = 23$  the number of encoder-decoder layers,  $H = 8$  the number of attention heads,  $Q = 64$  the head dimension, and  $T = 68 \times 11$  the average number of tokens per sequence (channels  $\times$  patches).

The peak throughput is  $P = 312$  TFLOPs at half precision, achievable on A100 GPUs, and the model FLOPs utilization is set to  $\eta = 0.5$  (50%).

This configuration yields an estimated 260 A100 GPU hours for a single pretraining run. The formula can be directly adapted for other model sizes or hardware configurations.

### F.2 Use of Existing Assets

We used Python (Python Software Foundation License), and some associated libraries for the implementation:

1. PyTorch (BSD-3 License)
2. NumPy (NumPy license)
3. scikit-learn (BSD license)
4. Pandas (BSD 3-Clause License)
5. Hugging Face's Accelerate (Apache License 2.0)

## 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 claims in the abstract and introduction accurately reflect the paper's contributions. The introduction clearly states the goals of REVE: building a foundation model for EEG that generalizes across datasets, durations, and electrode configurations. These claims are supported by:

- A novel 4D positional encoding (Section 2.2), validated by transfer to unseen setups.
- Pretraining on 92 datasets (Section 3.1), the largest EEG corpus to date.
- Extensive evaluations across 10 downstream tasks, showing consistent gains in full fine-tuning and linear probing (Section 4, Tables 2–16).

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

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

Answer: [Yes]

Justification: The paper includes a dedicated Limitations and Future Work section outlining key constraints of REVE, such as fixed input duration requirements, positional encoding limitations, and the limited dataset curation and selection. We also acknowledge that while scaling effects are observed, identifying precise scaling laws remains future work. These points reflect a clear understanding of the method's current boundaries and opportunities for improvement.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 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: [NA]

Justification: The paper does not include theoretical results, as it is focused on applications of a foundation model for EEG and does not delve into theoretical proofs or assumptions.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

### 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 provides a comprehensive description of the model architecture, the training data sources, and the routines for both pretraining and fine-tuning. The hyperparameters of the model are reported.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

- (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 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 authors provide full access to the code required for reproducing the experiments. Detailed instructions are included, outlining the necessary commands and environment settings to faithfully reproduce the results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 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, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper provides all necessary details regarding the experimental setting, including the data splits, the hyperparameters, and the type of optimizer used. These details are provided in the main text, with further specifics available in the supplemental material and the released code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 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 balanced accuracy as the primary metric, along with the mean and standard deviation. These metrics are used to match the baselines, providing a measure of variability in the results. This results in a 68% CI under normality assumption.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 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 discusses the use of NVIDIA A100 GPUs while estimating the amount of GPU hours used for each experiment.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research aligns with the NeurIPS Code of Ethics by ensuring responsible and ethical practices in all aspects of the research process, as discussed in ethical considerations section.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 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 both the potential positive and negative societal impacts of the work. On the positive side, the model can greatly benefit healthcare by improving the accuracy and efficiency of EEG-based applications such as brain-computer interfaces and diagnostic tools. On the negative side, the model's decoder, which could potentially reconstruct raw EEG data, poses a privacy risk. To mitigate this, the decoder is not being released, thus reducing the potential for misuse in generating sensitive or private information. The paper emphasizes the responsible and ethical use of the technology, with awareness of its potential risks.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 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 (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: To mitigate privacy risks, the decoder of the MAE model, which could reconstruct raw EEG data, is not being released. This safeguard reduces the potential for misuse while allowing responsible access to the model's embeddings.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 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: See the section about existing assets in the appendix.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

## 13. New assets

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

Answer: [Yes]

Justification: An anonymized repository containing the code for the model, its pretraining and fine-tuning is released.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

## 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: [NA]

Justification: The paper does not involve new crowdsourcing experiments or direct research with human subjects. We use pre-existing EEG datasets, and as such, there are no instructions or compensation details to report. The datasets used have been ethically sourced, with the original collection protocols ensuring participant consent and privacy in line with ethical guidelines.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

**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 (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve new research with human subjects, as it relies on pre-existing EEG datasets. Therefore, no potential risks to participants were incurred, and no new IRB approvals or equivalent reviews were required. The datasets used have been ethically sourced, with the original studies obtaining necessary participant consent and privacy protections.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

**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? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.