# LUNA: Efficient and Topology-Agnostic Foundation Model for EEG Signal Analysis

**Berkay Döner**[1] **Thorir Mar Ingolfsson**[1] **Luca Benini**[1] **Yawei Li**[1]*
[1]Integrated Systems Laboratory, ETH Zürich, Switzerland

## Abstract

Electroencephalography (EEG) offers a non-invasive lens into human brain activity, but building large-scale models is hampered by *topological heterogeneity*: each public EEG data defines its own electrode layout, limiting generalization. We introduce **LUNA** (**L**atent **U**nified **N**etwork **A**rchitecture), a self-supervised foundation model that reconciles disparate electrode geometries while scaling linearly—not quadratically—with channel count. LUNA compresses multi-channel EEG into a fixed-size, topology-agnostic latent space via *learned queries* and cross-attention. Downstream transformer blocks then operate exclusively on this latent representation using patch-wise temporal self-attention, decoupling computation from electrode count. Pre-trained on TUEG and Siena ($>$ 21,000 hours of raw EEG across diverse montages) using a masked-patch reconstruction objective, LUNA transfers effectively to four downstream tasks: abnormality detection, artifact rejection, slowing classification, and emotion recognition. It demonstrates highly competitive performance across several benchmarks, achieving state-of-the-art results on TUAR and TUSL, e.g., **0.921 AUROC** on TUAR, while reducing FLOPs by **300**$\times$ and trimming GPU memory use by up to **10**$\times$. Critically, these gains are consistent across all evaluated electrode configurations. Code and pre-trained models will be released upon publication.

## 1 Introduction

Electroencephalography (EEG) provides deep insight into brain activity without requiring invasive procedures, and plays a crucial role in clinical diagnostics, cognitive neuroscience, and human-computer interaction. In recent years, deep neural networks have significantly advanced EEG analysis, shifting from handcrafted pipelines to end-to-end learning systems [1]. Transformer-based models now rival traditional signal processing techniques by jointly modelling long-range temporal dynamics and cross-channel correlations [2, 3].

Despite this progress, *a fundamental bottleneck remains*: EEG corpora exhibit significant *topological heterogeneity*. Electrode count and placement vary widely across public and private datasets, making it difficult to transfer models across montages. This limitation manifests in pronounced performance degradation during cross-dataset evaluation. For example, motor-imagery decoders lose up to 14 percentage points (pp) in accuracy when transferring from PhysioNet to KU datasets [4], while state-of-the-art emotion-recognition models such as BIOT and MMM exhibit 13–15 pp drops between SEED and DEAP montages [5, 6]. Similarly, patient-to-patient transfer in stereotactic EEG (sEEG) remains an unsolved challenge, with naive models performing near chance without explicit spatial encoding [7].

Existing approaches offer limited solutions to this problem. Some train bespoke models for each montage, while others retain only shared electrodes—discarding up to **80%** of available data [8].

*Correspondence to `yawli@iis.ee.ethz.ch`.

Preprint. Under review.

More general approaches that flatten channels and time into long sequences incur quadratic self-attention complexity, $\mathcal{O}\big((S \cdot C)^2\big)$ where $S$ is the number of time segments and $C$ is the number of electrodes (channels), rapidly exhausting memory on dense caps [5]. These challenges underscore the need for a **single, montage-agnostic architecture that scales efficiently with electrode count**.

**LUNA** (**L**atent **U**nified **N**etwork **A**rchitecture) directly addresses this gap. Our key innovation is a topology-invariant encoder that maps arbitrary electrode layouts into a fixed latent space via learned queries and cross-attention. Temporal self-attention layers then operate exclusively on this latent space, decoupling computational cost from the number of electrodes. We pre-train LUNA using a masked-patch reconstruction objective on TUEG [9] and SIENA [10] (over *21,000* hours of raw EEG data), and fine-tune on four downstream benchmarks spanning abnormality and artifact detection, slowing classification, and emotion recognition.

The key contributions of this work are the following:

- **Topology-invariant encoder.** A learnt query / cross-attention module that projects arbitrary-sized channel sets into a fixed latent space.
- **Linear-in-channels complexity.** Patch-wise temporal attention that decouples FLOPs and memory from electrode count.
- **State-of-the-art accuracy-efficiency trade-off.** LUNA achieves strong results across a range of EEG benchmarks, demonstrating significant capabilities with balanced accuracies of **81.57%** on TUAB and **39.18%** on SEED-V [11], and AUROC scores of **0.921** on TUAR and **0.802** on TUSL, while reducing FLOPs by **300**$\times$ and GPU memory footprint by up to **10**$\times$ on high-density EEG recordings. Crucially, these gains hold across diverse electrode configurations, confirming LUNA's generalization capability.

## 2 Related Work

To contextualize our contributions, this section discusses relevant state-of-the-art methodologies that we will compare against. We focus on advancements in self-supervised learning for time series, the emergence of foundation models for physiological signals, and existing approaches to managing variable input structures, especially concerning topological heterogeneity in the EEG domain and computational efficiency.

### 2.1 Self-Supervised Learning Strategies in EEG

Foundation models for EEG primarily rely on self-supervised learning (SSL) to leverage large unlabeled datasets. Masked signal modeling is a dominant paradigm. BENDR [12] pioneered this for EEG by adapting masked prediction concepts from speech, applying a contrastive objective to predict masked convolutional features. Subsequent models refined this: BrainBERT [13] performs masked prediction on channel-independent spectrograms for intracranial electroencephalography (iEEG); EEGFormer [14] and LaBraM [15] predict vector-quantized (VQ) representations of masked patches, learning discrete codebooks; CBraMod [16] directly reconstructs masked raw signal patches. LUNA employs a similar masked reconstruction objective but applies it after projecting channel information into a unified latent space, requiring the decoder to reconstruct channel-specific details from this compressed representation.

### 2.2 Modeling Spatial Structure and Topology Variation in EEG

Capturing the spatial relationships between EEG channels is vital but complicated by varying electrode counts and layouts across datasets. Several strategies have been explored in the literature:
**Channel Independence:** Early approaches and models like BrainBERT [13] and EEGFormer [14] process each channel's data independently before potentially combining them later. While inherently handling varying channel numbers, this neglects early modeling of cross-channel interactions.
**Fixed-Topology Spatial Modeling:** Models like Brant [17] use dedicated spatial encoders alongside temporal ones but assume a consistent channel configuration, limiting cross-dataset generalization. Graph Neural Networks (GNNs) [18] explicitly model spatial relationships using a predefined adjacency graph, but require mechanisms to handle dynamically changing graph structures when topologies vary. LUNA avoids pre-defined graphs or fixed structures.

**Joint Spatio-Temporal Attention:** LaBraM [15] flattens channel and patch dimensions into one long sequence, allowing a standard Transformer to learn spatio-temporal dependencies simultaneously. However, this incurs $\mathcal{O}((SC)^2)$ complexity, scaling quadratically with both sequence length/patches (S) and channels (C). CBraMod [16] and CEReBrO [19] use alternating or parallel spatial and temporal attention mechanisms, reducing complexity to $\mathcal{O}(max(S^2, C^2))$ but still scaling quadratically with the dominant dimension. BIOT [5] uses linear attention after flattening, improving efficiency but potentially limiting modeling capacity. LUNA differs significantly by performing channel unification first before applying temporal attention with quadratic complexity only on the patch dimension and the much smaller latent dimension Q.
**Explicit Topology Mapping:** Some methods explicitly map varying topologies to a canonical representation. MMM [6] maps channels to predefined anatomical regions but relies on hand-engineered features (Differential Entropy) rather than raw signals. PopT [20] aggregates pre-computed channel-independent temporal features using 3D electrode coordinates. While achieving topology invariance, these methods are not fully end-to-end or rely on external information (regions). LUNA learns an end-to-end mapping from raw signals using learned queries without requiring pre-defined structures.

### 2.3 Learned Queries and Efficient Attention for Set Abstraction

LUNA's core mechanism for topology unification draws inspiration from architectures designed for permutation-invariant processing of set-structured data. Set Transformer [21] introduced the concept of using a small set of learnable inducing points (queries) and an Induced Set Attention Block to summarize information from a larger input set via cross-attention, reducing the complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(M \cdot N)$. PerceiverIO [22] further developed this mechanism, demonstrating its power in creating a fixed-size latent bottleneck capable of handling diverse, variable-sized inputs across different modalities (images, text) and enabling flexible decoding via task-specific output queries.

LUNA adapts this principle specifically for EEG topology invariance. We treat the set of EEG channel features at a given time interval (patch) as the input set. By applying cross-attention between the channel features (as keys/values) and a small number (Q) of learned queries, LUNA projects the variable-channel input onto a fixed-size latent space ($\mathbb{R}^{Q \times E}$). This projection is permutation-invariant with respect to the input channels, thus achieving topology agnosticism. Furthermore, it improves computational efficiency, as the complexity of this step scales linearly with the number of channels.

## 3 Methodology

Developing generalizable foundation models for EEG is hindered by two primary obstacles: the **topological heterogeneity** of EEG montages (varying channel counts and layouts) and the **computational complexity** of attention mechanisms. Standard models struggle with diverse input channel configurations, limiting data aggregation and generalizability. Furthermore, transformer-based approaches often face prohibitive $\mathcal{O}((C \cdot S)^2)$ or $\mathcal{O}(\max(C^2, S^2))$, as discussed in the section 2.2, complexity when processing $C$ channels and $S$ temporal patches. This limits their applicability to high-density EEG or long recordings.

LUNA addresses these challenges using a smaller latent space. Firstly, Channel-Unification Module (Sec . 3.1) employs learned queries and cross-attention to project variable-channel features into a fixed-dimension latent space, achieving topology invariance. Secondly, by unifying channel information into a compact set of $Q$ queries ($Q \ll C$) before temporal processing, LUNA significantly reduces computational demands. This design enables efficient and scalable processing of heterogeneous EEG data, paving the way for more robust foundation models. LUNA adopts an encoder-decoder architecture that transforms EEG signals from heterogeneous montages into a unified latent representation, enabling topology-agnostic modeling and efficient downstream decoding (Figure 1).

### 3.1 Encoder

The encoder comprises three key modules that transform the input EEG into a topology-agnostic latent representation: patch feature extraction, channel unification, and patch-wise temporal modeling.

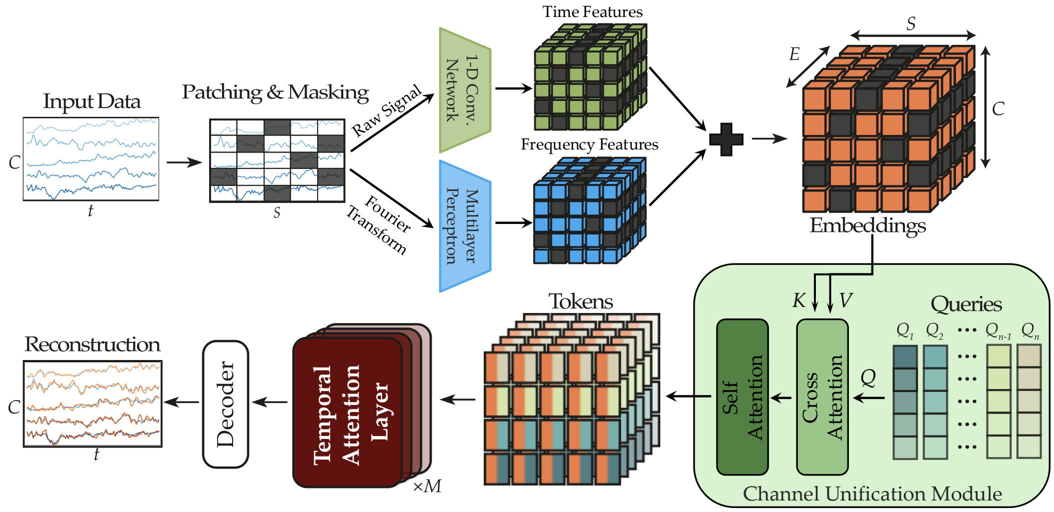



Figure 1: Overview of LUNA. EEG signals ($B \times C \times T$) are segmented into patches and embedded. Channel-Unification Module maps channel-wise features into a fixed-size latent space using learned queries ($Q$). Patch-wise Temporal Attention processes this latent sequence. The decoder generates task-specific outputs.

**Patch Feature Extraction** Given raw EEG $x \in \mathbb{R}^{B\times C\times T}$ (Batch $B$, Channels $C$, Time $T$), we segment each channel into $S = T/P$ non-overlapping temporal patches of size $P$. These patches are embedded via two parallel pathways:

**Temporal Embedding:** A 1D convolutional network (with GroupNorm [23], GELU [24]) encodes local temporal features similar to state-of-the-art methods such as LaBraM[15] and CBraMod [16], **Frequency Embedding:** The magnitude and phase from each patch's Fourier transform are projected through an MLP. These representations are summed to obtain patch features $x_{features}$.

**Channel Positional Encoding** To encode electrode locations, we apply NeRF-inspired sinusoidal encoding [25] to normalized 3D electrode coordinates, followed by an MLP projection. This yields $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{B\times C\times E}$, which is added to $x_{\text{features}}$.

During pre-training, a random subset of tokens is masked using a learnable embedding.

**Channel-Unification Module** To handle varying channel counts ($C$) across recordings, we introduce a cross-attention module that maps patch-wise features into a fixed latent space. Specifically, Q learned queries $\mathbf{Q}_{\text{learn}} \in \mathbb{R}^{Q\times E}$, which are learnable parameters of the model, initialized orthogonally to encourage diverse representations and optimized through backpropagation during training, cross-attend to patch features.

Let the input to this module be the tensor $\mathbf{X}_{token} \in \mathbb{R}^{B\times(C\cdot S)\times E}$, representing the spatially-aware features for $B$ samples, $S$ patches per channel, and feature dimension $E$. We first reshape this tensor to $\mathbf{X}' \in \mathbb{R}^{(B\cdot S)\times C\times E}$ to treat each patch instance across the batch independently while isolating the channel dimension for attention. The cross-attention mechanism then computes the output representation $\mathbf{A}_{\text{out}} \in \mathbb{R}^{(B\cdot S)\times Q\times E}$:

$$\mathbf{A}_{\text{out}} = \text{MultiHeadAttention}(\mathbf{Q}, \mathbf{X}', \mathbf{X}') \tag{1}$$

A feed-forward network (FFN) with residual connection refines the outputs, followed by $L$ Transformer encoder layers operating on the query dimension $Q$.

$$\mathbf{X}_{\text{unified}} = \text{TransformerEncoder}(\mathbf{A}_{\text{out}} + \text{FFN}(\mathbf{A}_{\text{out}})) \tag{2}$$

The result $\mathbf{X}_{\text{unified}} \in \mathbb{R}^{(B\cdot S)\times Q\times E}$ decouples further processing from the original electrode layout.

**Patch-wise Temporal Encoder** The unified representations are reshaped into temporal sequences $\mathbf{X}'_{\text{unified}} \in \mathbb{R}^{B\times S\times(Q\cdot E)}$. These are processed by a stack of Transformer encoder blocks with Rotary

Positional Embeddings (RoPE) [26] to capture temporal dependencies efficiently. A key advantage of this encoding approach is that each of the $S$ temporal tokens in $\mathbf{X}'_{\text{unified}}$ now encapsulates richer, aggregated information from multiple input channels, rather than representing a single channel's segment. Furthermore, by not tokenizing each channel independently for temporal processing, the effective sequence length for the temporal Transformers is reduced from $S \cdot C$ to just $S$, leading to significant reductions in computational complexity and memory requirements.

$$E_{\text{out}} = \text{TemporalEncoder}(\mathbf{X}'_{\text{unified}})$$

### 3.2 Decoder

LUNA supports two decoding strategies depending on the task: reconstruction (pre-training) and classification (fine-tuning).

**Reconstruction Head (Pre-training)** For masked patch reconstruction, $C$ learned decoder queries $E_{dec} \in \mathbb{R}^{B\times C\times E}$ attend to $E_{out}$ via cross-attention, producing channel-specific representations $E_{dec}$. A linear projection recovers the patch values $\hat{x} \in \mathbb{R}^{B\times(C\cdot S)\times P}$.

**Classification Head (Fine-tuning)** For downstream tasks, a single aggregation query $E_{agg} \in \mathbb{R}^{B\times 1\times(Q\cdot E)}$ attends to $E_{out}$ to produce a pooled representation, which is passed to an MLP for classification.

### 3.3 Training Objectives

LUNA is pre-trained with a masked reconstruction loss and an auxiliary query specialization loss.

**Reconstruction Loss** A Smooth L1 loss is applied to both masked and visible patches:

$$L_{rec} = \frac{1}{N_{masked}} \sum_{i\in M} \text{SmoothL1}(x_{orig_i}, x_{recons_i}) + \alpha \cdot \frac{1}{N_{visible}} \sum_{i\notin M} \text{SmoothL1}(x_{orig_i}, x_{recons_i})$$

and SmoothL1$(x, \hat{x})$ = $0.5(x-\hat{x})^2$ if $|x-\hat{x}| < \beta$, else $\beta|x-\hat{x}| - 0.5\beta^2$, with $\beta = 1$.

**Query Specialization Loss** To promote diversity among queries, we penalize similarity in query-channel affinity matrices by minimizing the mean value of off-diagonal elements:

$$\mathcal{L}_{\text{spec}} = \frac{\lambda_{\text{spec}}}{B' \cdot Q \cdot (Q-1)} \sum_{b'=1}^{B'} \sum_{i=1}^{Q} \sum_{j=1, j\neq i}^{Q} \left((\mathbf{A}_{\text{affinity}} \mathbf{A}^T_{\text{affinity}})_{b',i,j}\right)^2$$

## 4 Results

### 4.1 Experimental Setup

**Datasets** We pre-train LUNA on a combined corpus of Temple University Hospital EEG Corpus (TUEG) [9] and the Siena Scalp EEG Database [10], spanning recordings with 20, 22, and 29 channels amounting to over 21,900 hours of EEG data (see Table 9). Downstream evaluations cover four diverse benchmarks: **TUAB** [9]: Abnormal EEG detection (binary classification), **TUAR** [9]: Artifact detection (multi-class classification) **TUSL** [9]: Slowing event classification (4-class classification). **SEED-V** [11]: Emotion recognition (5-class classification), with unseen 62-channel topology. All subjects and recordings from the downstream evaluation datasets (TUAB, TUAR, TUSL, SEED-V) were strictly excluded from this pre-training set to ensure fair evaluation of generalization. For LUNA, the input EEG is segmented into patches, consisting of 40 timestamps. For most datasets, EEG recordings are sliced into non-overlapping 5-second segments to form individual training/evaluation samples. SEED-V dataset uses its default 1-second sample duration.

**Fine-tuning and Data Splits** For the TUAB dataset, we use the official train-test split. As the TUSL and TUAR datasets lack official test splits, we implement an 80%/10%/10% randomized split for training, validation, and testing. For SEED-V, fifteen trials are divided equally into train, validation, and test sets for each session. For the TUAR dataset, we adopt a multiclass classification approach, restricting to 5 distinct artifact types in a single-label setting, similar to EEGFormer [14]. We optimize binary cross-entropy loss for TUAB and cross-entropy loss for other datasets. We report the mean and standard deviation of results obtained across three different random seeds.

**Preprocessing** We apply a minimal, standardized preprocessing pipeline to all EEG data. Signals are first bandpass filtered between 0.1 Hz and 75 Hz. A notch filter (50Hz or 60Hz) is applied to remove power-line interference. All signals are then resampled to 256 Hz. For TUEG, TUAB, TUAR, and TUSL datasets, signals are converted to a bipolar montage; Siena and SEED-V are processed in unipolar format. Finally, each channel within each sample is normalized using z-score normalization.

**Computational Environment** All experiments were conducted on a cluster of eight NVIDIA A100 GPUs, using Python 3.11.6 and PyTorch 2.4.1 with CUDA 12.1. Training utilizes 'bf16' mixed-precision. Detailed hyperparameters for pre-training and fine-tuning are provided in Appendix A.3.

**Baselines and Variants** We compare against state-of-the-art supervised and self-supervised methods, including transformer-based architectures such as LaBraM [15], CBraMod [16], EEGFormer [14], and BIOT [5]. LUNA is evaluated in three configurations: Base (7M), Large (43M), and Huge (311M parameters). Model size is increased by expanding the depth of the Patch-wise Temporal Encoder, the hidden embedding dimension $E$, and the number/size of queries $Q$ in the Channel-Unification Module. Key architectural settings are detailed in Appendix A.1.

### 4.2 Downstream Task Performance

**Abnormal EEG Detection (TUAB)** LUNA demonstrated competitive performance on TUAB (Table 1). LUNA-Huge achieves AUROC of 0.8957 and AUPR of 0.9029, surpassing most self-supervised baselines and approaching large-scale models like LaBraM and CBraMod. Notably, LUNA maintains strong performance despite being significantly smaller, highlighting its efficiency.

Table 1: Performance comparison on TUAB abnormal EEG detection.

| Model | Size | Bal. Acc. (%) ↑ | AUC-PR ↑ | AUROC ↑ |
|---|---|---|---|---|
| *Supervised Models* | | | | |
| SPaRCNet [27] | 0.8M | 78.96 ± 0.18 | 0.8414 ± 0.0018 | 0.8676 ± 0.0012 |
| ContraWR [28] | 1.6M | 77.46 ± 0.41 | 0.8421 ± 0.0140 | 0.8456 ± 0.0074 |
| CNN-Transformer [29] | 3.2M | 77.77 ± 0.22 | 0.8433 ± 0.0039 | 0.8461 ± 0.0013 |
| FFCL [30] | 2.4M | 78.48 ± 0.38 | 0.8448 ± 0.0065 | 0.8569 ± 0.0051 |
| ST-Transformer [31] | 3.2M | 79.66 ± 0.23 | 0.8521 ± 0.0026 | 0.8707 ± 0.0019 |
| *Self-supervised Models* | | | | |
| BENDR [12] | 0.39M | 76.96 ± 3.98 | - | 0.8397 ± 0.0344 |
| BrainBERT [13] | 43.2M | - | 0.8460 ± 0.0030 | 0.8530 ± 0.0020 |
| EEGFormer-Base [14] | 2.3M | - | 0.8670 ± 0.0020 | 0.8670 ± 0.0030 |
| BIOT [5] | 3.2M | 79.59 ± 0.57 | 0.8692 ± 0.0023 | 0.8815 ± 0.0043 |
| EEG2Rep [32] | - | 80.52 ± 2.22 | - | 0.8843 ± 0.0309 |
| FEMBA-Huge [33] | 386M | 81.82 ± 0.16 | 0.9005 ± 0.0017 | 0.8921 ± 0.0042 |
| CEReBrO [19] | 85.15M | 81.67 ± 0.23 | 0.9049 ± 0.0026 | 0.8916 ± 0.0038 |
| LaBraM-Base [15] | 5.9M | 81.40 ± 0.19 | 0.8965 ± 0.0016 | 0.9022 ± 0.0009 |
| LaBraM-Huge [15] | 369.8M | **82.58 ± 0.11** | 0.9204 ± 0.0011 | **0.9162 ± 0.0016** |
| CBraMod [16] | 69.3M | 82.49 ± 0.25 | **0.9221 ± 0.0015** | 0.9156 ± 0.0017 |
| **LUNA-Base** | 7M | 80.63 ± 0.08 | 0.8953 ± 0.0016 | 0.8868 ± 0.0015 |
| **LUNA-Large** | 43M | 80.96 ± 0.10 | 0.8986 ± 0.0005 | 0.8924 ± 0.0010 |
| **LUNA-Huge** | 311.4M | 81.57 ± 0.11 | 0.9029 ± 0.0014 | 0.8957 ± 0.0011 |

**Artifact and Slowing Detection (TUAR and TUSL)** LUNA delivers state-of-the-art results on TUAR and TUSL (Table 2). LUNA-Huge achieves AUROC 0.921 on TUAR, outperforming FEMBA-Large and other methods. On TUSL, LUNA-Huge reaches AUROC 0.802, the highest among all compared models.

Table 2: Performance comparison on TUAR (artifact detection) and TUSL (slowing event classification).

| Model | Size | TUAR | | TUSL | |
|---|---|---|---|---|---|
| | | **AUROC ↑** | **AUC-PR ↑** | **AUROC ↑** | **AUC-PR ↑** |
| *Supervised Models* | | | | | |
| EEGNet [34] | - | 0.752 ± 0.006 | 0.433 ± 0.025 | 0.635 ± 0.015 | 0.351 ± 0.006 |
| EEG-GNN [18] | - | 0.837 ± 0.022 | 0.488 ± 0.015 | 0.721 ± 0.009 | 0.381 ± 0.004 |
| GraphS4mer [35] | - | 0.833 ± 0.006 | 0.461 ± 0.024 | 0.632 ± 0.017 | 0.359 ± 0.001 |
| *Self-supervised Models* | | | | | |
| BrainBERT [13] | 43.2M | 0.753 ± 0.012 | 0.350 ± 0.014 | 0.588 ± 0.013 | 0.352 ± 0.003 |
| EEGFormer-Base [14] | 2.3M | 0.847 ± 0.014 | 0.483 ± 0.026 | 0.713 ± 0.010 | **0.393 ± 0.003** |
| EEGFormer-Large [14] | 3.2M | 0.852 ± 0.004 | 0.483 ± 0.014 | 0.679 ± 0.013 | 0.389 ± 0.003 |
| FEMBA-Base [33] | 47.7M | 0.900 ± 0.010 | **0.559 ± 0.002** | 0.731 ± 0.012 | 0.289 ± 0.009 |
| FEMBA-Large [33] | 77.8M | 0.915 ± 0.003 | 0.521 ± 0.001 | 0.714 ± 0.007 | 0.282 ± 0.010 |
| **LUNA-Base** | 7M | 0.902 ± 0.011 | 0.495 ± 0.010 | 0.767 ± 0.023 | 0.301 ± 0.003 |
| **LUNA-Large** | 43M | 0.918 ± 0.003 | 0.505 ± 0.010 | 0.771 ± 0.006 | 0.293 ±0.021 |
| **LUNA-Huge** | 311.4M | **0.921 ± 0.011** | 0.528 ± 0.012 | **0.802 ± 0.005** | 0.289 ±0.008 |

**Emotion Recognition on Unseen Montage (SEED-V)** The SEED-V benchmark tests generalization to a novel 62-channel montage, distinct from pre-training data. Results in Table 3 show that while LUNA effectively operates on this unseen topology, its performance (e.g., Bal. Acc.) lags behind leading methods like CBraMod by 2-3 pp. This suggests a trade-off inherent in LUNA's design: while its query-based unification enables efficient, topology-agnostic processing across common montage variations (as demonstrated on TUAB/TUAR/TUSL), generalizing zero-shot to vastly different, high-density layouts remains challenging, possibly due to positional encoding constraints. Despite this gap, LUNA shows positive scaling from Base to Large models, underscoring its potential.

Table 3: Performance comparison on SEED-V emotion recognition (5-classes).

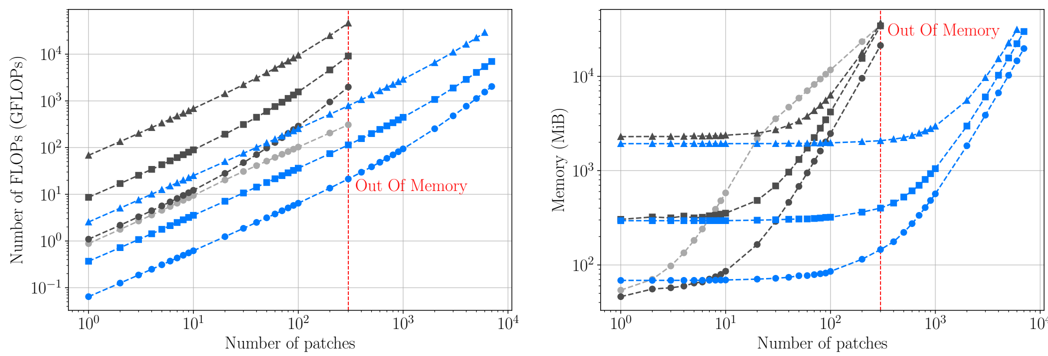
| Model | Size | Bal. Acc. (%) ↑ | Cohen's Kappa ↑ | Weighted F1 ↑ |
|---|---|---|---|---|
| *Supervised Models* | | | | |
| SPaRCNet [27] | 0.79M | 0.2949 ± 0.0078 | 0.1121 ± 0.0139 | 0.2979 ± 0.0083 |
| ContraWR [28] | 1.6M | 0.3546 ± 0.0105 | 0.1905 ± 0.0188 | 0.3544 ± 0.0121 |
| CNN-Transformer [29] | 3.2M | 0.3678 ± 0.0078 | 0.2072 ± 0.0183 | 0.3642 ± 0.0088 |
| FFCL [30] | 2.4M | 0.3641 ± 0.0092 | 0.2078 ± 0.0201 | 0.3645 ± 0.0132 |
| ST-Transformer [31] | 3.5M | 0.3052 ± 0.0072 | 0.1083 ± 0.0121 | 0.2833 ± 0.0105 |
| *Self-supervised Models* | | | | |
| BIOT [5] | 3.2M | 0.3837 ± 0.0187 | 0.2261 ± 0.0262 | 0.3856 ± 0.0203 |
| LaBraM-Base [15] | 5.8M | 0.3976 ± 0.0138 | 0.2386 ± 0.0209 | 0.3974 ± 0.0111 |
| CBraMod [16] | 14M | **0.4091 ± 0.0097** | **0.2569 ± 0.0151** | **0.4101 ± 0.0108** |
| **LUNA-Base** | 7M | 0.3730 ± 0.0098 | 0.1831 ± 0.0103 | 0.3389 ± 0.0091 |
| **LUNA-Large** | 43M | 0.3918 ± 0.0066 | 0.2073 ± 0.0045 | 0.3586 ± 0.0013 |
| **LUNA-Huge** | 311.4M | 0.3900 ± 0.0096 | 0.2037 ± 0.0103 | 0.3506 ± 0.0047 |

### 4.3 Computational Efficiency

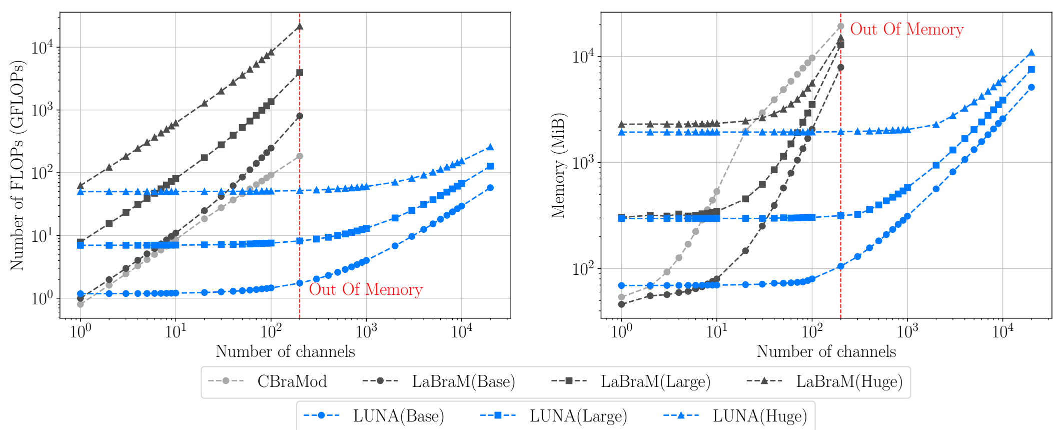
**LUNA achieves substantially better calling efficiency** compared to full and alternating attention models. As shown in Figure 2a, LUNA's patch-wise attention enables thousands of temporal patches without the quadratic cost faced by LaBraM. Likewise, Figure 2b shows that LUNA maintains near-constant compute cost when channel count increase, outperforming CBraMod's $\mathcal{O}(C^2)$ scaling for dense EEG recordings. These results confirm that LUNA decouples inference cost from input montage, making it well-suited for long recordings or high-density EEG scenarios.

### 4.4 Ablation Studies

We validate the impact of LUNA's key design choices on TUAB and TUAR (Table 4).




(a) Scaling with number of patches (Channels fixed at 22).




(b) Scaling with number of channels (Patches fixed at 20).

Figure 2: Computational cost scaling of LUNA and baseline models. (a) FLOPs and Memory usage vs. number of patches. (b) FLOPs and Memory usage vs. number of channels. LUNA demonstrates significantly better efficiency and scalability, especially compared to full attention (LaBraM), and favorable scaling compared to alternating attention (CBraMod) due to the fixed latent query space.

**Learned Queries vs. Fixed Regions** Replacing learned queries with predefined spatial regions (similar to what MMM [6] does) slightly reduces AUROC (-0.004 to -0.006), confirming that learned queries offer flexibility and adaptiveness beyond anatomical priors.

**Query Specilization Loss** Removing the specialization loss results in modest AUROC declines (-0.003 to -0.006), showing that query diversity improves robustness, especially for complex artifacts.

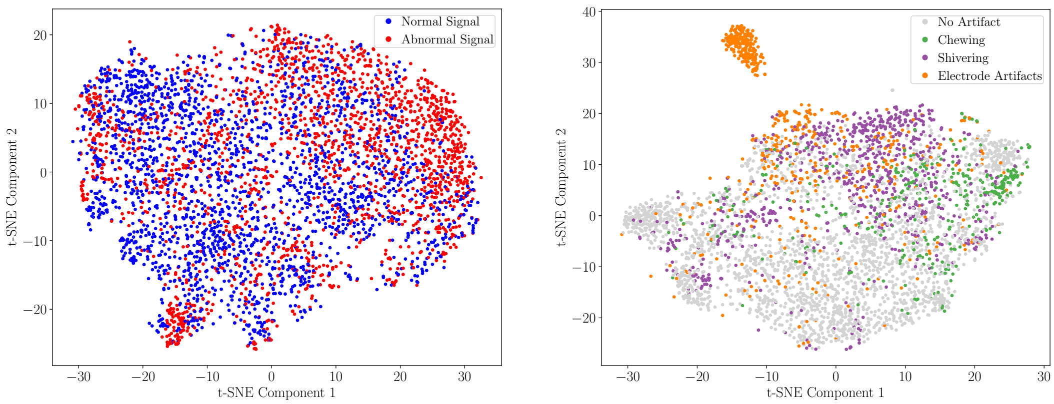
**Frequency Features** Ablating frequency embeddings leads to the largest drop (up to -0.012 AUROC), showing their complementary role to temporal features in enhancing representation quality.

Table 4: Ablation study results (LUNA-Base) on TUAB and TUAR datasets.

| Model Configuration | TUAB AUROC | TUAB AUC-PR | TUAR AUROC | TUAR AUC-PR |
|---|---|---|---|---|
| LUNA-Base (Full Model) | 0.887 $\pm$ 0.002 | 0.895 $\pm$ 0.002 | 0.902 $\pm$ 0.011 | 0.495 $\pm$ 0.010 |
| *Unification Module:* | | | | |
| - Region-based Attention | 0.883 $\pm$ 0.001 ($\downarrow$ 0.004) | 0.892 $\pm$ 0.002 ($\downarrow$ 0.003) | 0.896 $\pm$ 0.001 ($\downarrow$ 0.006) | 0.509 $\pm$ 0.006 ($\uparrow$ 0.014) |
| *Other Components:* | | | | |
| - w/o Query Specialization Loss | 0.884 $\pm$ 0.003 ($\downarrow$ 0.003) | 0.892 $\pm$ 0.002 ($\downarrow$ 0.003) | 0.895 $\pm$ 0.005 ($\downarrow$ 0.007) | 0.498 $\pm$ 0.010 ($\uparrow$ 0.003) |
| - w/o Frequency Features | 0.876 $\pm$ 0.012 ($\downarrow$ 0.011) | 0.883 $\pm$ 0.005 ($\downarrow$ 0.012) | 0.893 $\pm$ 0.011 ($\downarrow$ 0.009) | 0.490 $\pm$ 0.011 ($\downarrow$ 0.005) |

### 4.5 Latent Space Analysis

**Pre-trained Representations** t-SNE visualizations (Figure 3) reveal that even before fine-tuning, LUNA's encoder captures task-relevant structure. Normal and abnormal EEGs form separate clusters in TUAB, while artifact classes are partially separated in TUAR, demonstrating effective pre-training.




(a) TUAB dataset (Normal vs. Abnormal Signal). (b) TUAR dataset (Artifact Types).

Figure 3: t-SNE of LUNA-Base embeddings on downstream datasets before fine-tuning.

### 4.6 Learned Query Specialization Visualization

**Query Specialization** Visual analysis of the learned queries (Figure 4) highlights their role in topology-agnostic representation. Queries exhibit distinct spatial profiles: some are localized (e.g., frontal regions), while others aggregate broader signals. This emergent specialization confirms that cross-attention learns flexible, data-driven basis functions for spatial unification.

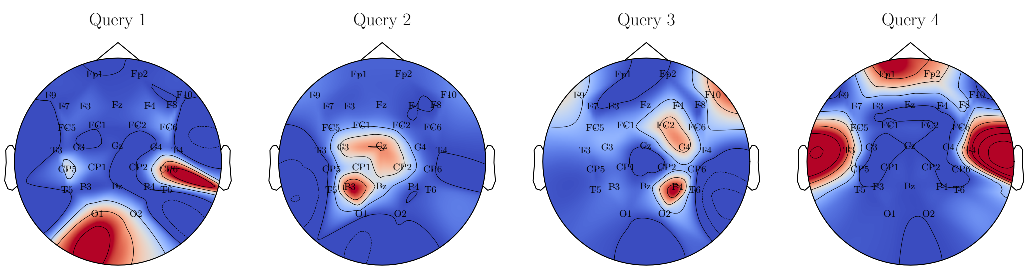



Figure 4: Visualization of the attention patterns of queries in LUNA-Base on Siena [10] topology.

## 5 Conclusion

We introduced **LUNA**, a self-supervised foundation model designed to address the challenge of topological heterogeneity in EEG analysis. By leveraging learned queries and cross-attention, LUNA unifies recordings with diverse electrode layouts into a fixed latent space, enabling montage-agnostic modeling. Through extensive experiments across abnormality detection, artifact recognition, slowing classification, and emotion recognition, we demonstrate that LUNA matches or surpasses state-of-the-art performance while offering substantial efficiency gains in FLOPs and memory usage. Critically, these benefits hold across all evaluated electrode configurations.

While LUNA achieves strong results, especially on heterogeneous montages, our analysis also reveals limitations. Performance on SEED-V suggests sensitivity to unseen channel topologies, likely stemming from reliance on positional encodings learned during pre-training. Addressing this limitation, through enhanced spatial generalization strategies or hybrid learned/geometric embeddings, is an important direction for future work.

More broadly, this work highlights the promise of topology-agnostic latent representations for scalable EEG modeling. Future extensions include exploring unified models across EEG and invasive modalities (e.g., sEEG, ECoG), integrating domain-specific priors (e.g., neurophysiological constraints), and adapting LUNA for real-time inference scenarios. Beyond technical advancements, the development of efficient, topology-invariant EEG models like LUNA could enhance neurological diagnostics and research accessibility. However, careful attention must be paid to mitigating risks such as algorithmic bias and ensuring patient data privacy for deployment. Future work should integrate ethical concerns alongside technical improvements.

## References


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

[2] 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, December 2023.

[3] Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: A survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI-2023, pages 6778–6786. International Joint Conferences on Artificial Intelligence Organization, August 2023.

[4] Lichao Xu, Minpeng Xu, Yufeng Ke, Xingwei An, Shuang Liu, and Dong Ming. Cross-dataset variability problem in EEG decoding with deep learning. *Frontiers in Human Neuroscience*, 14:103, April 2020.

[5] Chaoqi Yang, M Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 78240–78260. Curran Associates, Inc., September 2023.

[6] Ke Yi, Yansen Wang, Kan Ren, and Dongsheng Li. Learning topology-agnostic EEG representations with geometry-aware modeling. In *Thirty-seventh Conference on Neural Information Processing Systems*, September 2023.

[7] Georgios Mentzelopoulos, Evangelos Chatzipantazis, Ashwin G Ramayya, Michelle Hedlund, Vivek Buch, Kostas Daniilidis, Konrad Kording, and Flavia Vitale. Neural decoding from stereotactic EEG: accounting for electrode variability across subjects. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, September 2024.

[8] Xuefen Lin, Jielin Chen, Weifeng Ma, Wei Tang, and Yuchen Wang. EEG emotion recognition using improved graph neural network with channel selection. *Computer Methods and Programs in Biomedicine*, 231:107380, April 2023.

[9] Iyad Obeid and Joseph Picone. The temple university hospital EEG data corpus. *Frontiers in Neuroscience*, 10:196, May 2016.

[10] Paolo Detti, Giampaolo Vatti, and Garazi Zabalo Manrique de Lara. Siena scalp EEG database, 2020. This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit `http://creativecommons.org/licenses/by/4.0/`.

[11] Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):715–729, June 2022.

[12] Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15:653659, June 2021.

[13] Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. BrainBERT: Self-supervised representation learning for intracranial recordings. In *The Eleventh International Conference on Learning Representations*, February 2023.

[14] Yuqi Chen, Kan Ren, Kaitao Song, Yansen Wang, Yifan Wang, Dongsheng Li, and Lili Qiu. EEGFormer: Towards transferable and interpretable large-scale EEG foundation model. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, February 2024.

[15] 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*, January 2024.

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

[17] Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. In *Thirty-seventh Conference on Neural Information Processing Systems*, September 2023.

[18] Siyi Tang, Jared Dunnmon, Khaled Kamal Saab, Xuan Zhang, Qianying Huang, Florian Dubost, Daniel Rubin, and Christopher Lee-Messer. Self-supervised graph neural networks for improved electroencephalographic seizure analysis. In *International Conference on Learning Representations*, January 2022.

[19] Alexandru Dimofte, Glenn Anta Bucagu, Thorir Mar Ingolfsson, Xiaying Wang, Andrea Cossettini, Luca Benini, and Yawei Li. CEReBrO: Compact encoder for representations of brain oscillations using efficient alternating attention, January 2025.

[20] Geeling Chau, Christopher Wang, Sabera J Talukder, Vighnesh Subramaniam, Saraswati Soedarmadji, Yisong Yue, Boris Katz, and Andrei Barbu. Population transformer: Learning population-level representations of neural activity. In *The Thirteenth International Conference on Learning Representations*, January 2025.

[21] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pages 3744–3753. PMLR, June 2019.

[22] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J Henaff, Matthew Botvinick, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*, January 2022.

[23] Yuxin Wu and Kaiming He. Group normalization. *International Journal of Computer Vision*, 128(3):742–755, July 2019.

[24] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), June 2016.

[25] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, December 2021.

[26] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, February 2024.

[27] Jin Jing, Wendong Ge, Shenda Hong, Marta Bento Fernandes, Zhen Lin, Chaoqi Yang, Sungtae An, Aaron F. Struck, Aline Herlopian, Ioannis Karakis, Jonathan J. Halford, Marcus C. Ng, Emily L. Johnson, Brian L. Appavu, Rani A. Sarkis, Gamaleldin Osman, Peter W. Kaplan, Monica B. Dhakar, Lakshman Arcot Jayagopal, Zubeda Sheikh, Olga Taraschenko, Sarah Schmitt, Hiba A. Haider, Jennifer A. Kim, Christa B. Swisher, Nicolas Gaspard, Mackenzie C. Cervenka, Andres A. Rodriguez Ruiz, Jong Woo Lee, Mohammad Tabaeizadeh, Emily J. Gilmore, Kristy Nordstrom, Ji Yeoun Yoo, Manisha G. Holmes, Susan T. Herman, Jennifer A. Williams, Jay Pathmanathan, Fábio A. Nascimento, Ziwei Fan, Samaneh Nasiri, Mouhsin M. Shafi, Sydney S. Cash, Daniel B. Hoch, Andrew J. Cole, Eric S. Rosenthal, Sahar F. Zafar, Jimeng Sun, and M. Brandon Westover. Development of expert-level classification of seizures and rhythmic and periodic patterns during EEG interpretation. *Neurology*, 100(17):e1750–e1762, April 2023.

[28] Chaoqi Yang, Cao Xiao, M. Brandon Westover, and Jimeng Sun. Self-supervised electroencephalogram representation learning for automatic sleep staging: Model development and evaluation study. *JMIR AI*, 2:e46769, July 2021.

[29] Wei Yan Peh, Yuanyuan Yao, and Justin Dauwels. Transformer convolutional neural networks for automated artifact detection in scalp EEG, July 2022.

[30] 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, February 2022.

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

[32] 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*, KDD '24, pages 5544–5555. ACM, August 2024.

[33] Anna Tegon, Thorir Mar Ingolfsson, Xiaying Wang, Luca Benini, and Yawei Li. FEMBA: Efficient and scalable EEG analysis with a bidirectional mamba foundation model, February 2025.

[34] 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, July 2018.

[35] Siyi Tang, Jared A Dunnmon, Qu Liangqiong, Khaled K Saab, Tina Baykaner, Christopher Lee-Messer, and Daniel L Rubin. Modeling multivariate biosignals with graph neural networks and structured state space models. In Bobak J. Mortazavi, Tasmie Sarker, Andrew Beam, and Joyce C. Ho, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 209 of *Proceedings of Machine Learning Research*, pages 50–71. PMLR, June 2023.

# A Appendix

This appendix provides supplementary details regarding the model architecture, datasets, experimental settings, and additional results supporting the findings presented in the main paper.

## A.1 Model Architecture Details

The following tables show the hyperparameter setup for the pre-training and the downstream fine-tuning for LUNA.

### A.1.1 Hyperparameters for pre-training

Table 5: Hyperparameters for EEG pre-training.

| Hyperparameters | | LUNA-Base | LUNA-Large | LUNA-Huge |
|---|---|---|---|---|
| Temporal Encoder | Input channels | {1,8,8} | {1,16,16} | {1,32,32} |
| | Output channels | {16,16,16} | {24,24,24} | {32,32,32} |
| | Kernel size | | {20,3,3} | |
| | Stride | | {10,1,1} | |
| | Padding | | {9,1,1} | |
| Patch size | | | 40 | |
| Transformer encoder layers | | 8 | 10 | 24 |
| Number of queries | | 4 | 6 | 8 |
| Query size | | 64 | 96 | 128 |
| Hidden size | | 256 | 576 | 1024 |
| MLP size | | 1024 | 2304 | 4096 |
| Attention head number | | 8 | 12 | 16 |
| Batch size per GPU | | 2040 | 2040 | 720 |
| Total batch size | | 8160 | 8160 | 11520 |
| Peak learning rate | | | 1.25e-4 | |
| Minimal learning rate | | | 2.5e-7 | |
| Learning rate scheduler | | | Cosine | |
| Optimizer | | | AdamW | |
| Adam $\beta$ | | | (0.9,0.98) | |
| Weight decay | | | 0.05 | |
| Total epochs | | | 60 | |
| Warmup epochs | | | 10 | |
| Loss type | | | Smooth-L1 | |
| Non-masked region loss coefficient | | | 0.05 | |
| Query specialization loss coefficient | | | 0.8 | |
| Gradient clipping | | | 1 | |
| Mask ratio | | | 0.5 | |
| Precision | | | bf16-mixed | |

### A.1.2 Hyperparameters for downstream fine-tuning

Table 6: Hyperparameters for downstream fine-tuning.

| Hyperparameters | Values |
|---|---|
| Batch size per GPU | 512 |
| Peak learning rate | 1e-4 |
| Minimal learning rate | 5e-6 |
| Learning rate scheduler | Cosine |
| Optimizer | AdamW |
| Adam $\beta$ | (0.9,0.999) |
| Weight decay | 0.05 |
| Total epochs | 50 |
| Early stopping patience | 10 |
| Warmup epochs | 5 |
| Drop path | 0.1 (B/L) 0.2 (H) |
| Layer-wise learning rate decay | 0.5 (B) 0.8 (L/H) |
| Label smoothing (multi-class classification) | 0.1 |

### A.1.3 Complexity Analysis

The computational complexity of key attention stages and a comparison with alternatives are shown in 7 and 8.

Table 7: Complexity Breakdown of LUNA Encoder Stages.

| Stage | Input Shape | Complexity |
|---|---|---|
| Channel-Unification Module (Cross-Attn) | $(B \cdot S) \times C \times E$ | $O(B \cdot S \cdot Q \cdot C \cdot E)$ |
| Query Self-Attention | $(B \cdot S) \times Q \times E$ | $O(B \cdot S \cdot Q^2 \cdot E)$ |
| Patch-wise Attention Encoder (Self-Attn) | $B \times S \times (Q \cdot E)$ | $O(B \cdot S^2 \cdot Q \cdot E)$ |

Table 8: Attention Complexity Comparison.

| Method | Bottleneck Complexity |
|---|---|
| LUNA (Latent Space Attention) | $O(B \cdot S^2 \cdot Q \cdot E)$ *or* $O(B \cdot S \cdot Q \cdot C \cdot E)$ |
| Full-Attention (e.g., LaBraM) | $O(B \cdot S^2 \cdot C^2 \cdot E)$ |
| Alternating Attention (Patches, e.g., CBraMod) | $O(B \cdot S^2 \cdot C \cdot E)$ |
| Alternating Attention (Channels, e.g., CBraMod) | $O(B \cdot S \cdot C^2 \cdot E)$ |

## A.2 Dataset and Preprocessing Details

**Datasets Used** We use publicly available EEG datasets, provided in 9.

Table 9: Summary of Datasets Used.

| Dataset | # Subjects | # Samples (Train/Val/Test or Total) | Hours of Recordings | # Channels | Montage Used |
|---|---|---|---|---|---|
| TUEG (Pre-train) | 14,987 | 15,686,874 (Total) | 21,787.32 | 20 or 22 | Bipolar |
| Siena (Pre-train) | 14 | 101,520 (Total) | 141.0 | 29 | Unipolar |
| TUAB | 2,329 | 591,357 / 154,938 / 74,010 | 1,139.31 | 22 | Bipolar |
| TUAR | 213 | 49,241 / 5,870 / 5,179 | 83.74 | 22 | Bipolar |
| TUSL | 38 | 16,088 / 1,203 / 2,540 | 27.54 | 22 | Bipolar |
| SEED-V | 15 | 43,328 / 43,360 / 31,056 | 32.70 | 62 | Unipolar |

## A.3 Experimental Settings

**Pre-training** LUNA is pre-trained using a masked patch reconstruction task. Key hyperparameters are listed in 5.

**Computational Resources** Experiments were conducted using NVIDIA A100 GPUs. Pre-training took approximately 1 day on 8 GPUs for the base and large models and 16 GPUs for the huge model.

### A.4 Additional Quantitative Results

**Training Curves** The pre-training loss curves for LUNA-Base are shown in 5. The reconstruction loss drops shows and initial plateau then drops slowly over the epochs, while the query specialization shows a jump and then a slow decrease, indicating more orthogonal query usage over time. The initial drop of the query specialization might be due to a trivial case where a query attends to only one channel. The queries learn to attend to their own specialized areas afterwards while covering all the channels in the input.

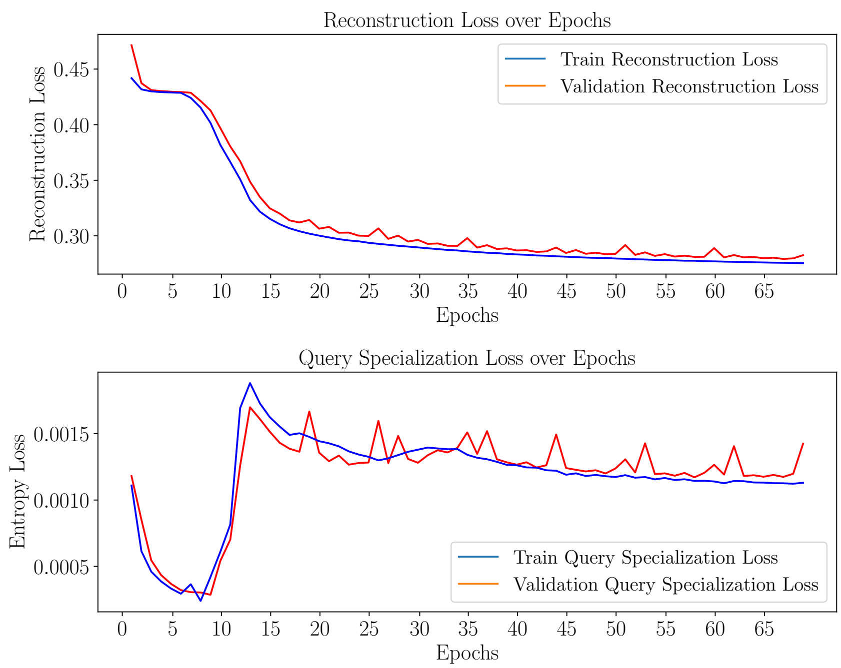



Figure 5: Loss curves during pre-training for LUNA-Base (Reconstruction and Query Specialization Loss).

### A.5 Additional Visualizations

**Reconstruction Examples** Figures 6, 7, 8 show examples of the model reconstructing masked patches (gray regions) for inputs with 20, 22, and 29 channels, respectively. The reconstructions capture the underlying signal trend and demonstrate robustness across different topologies seen during pre-training.

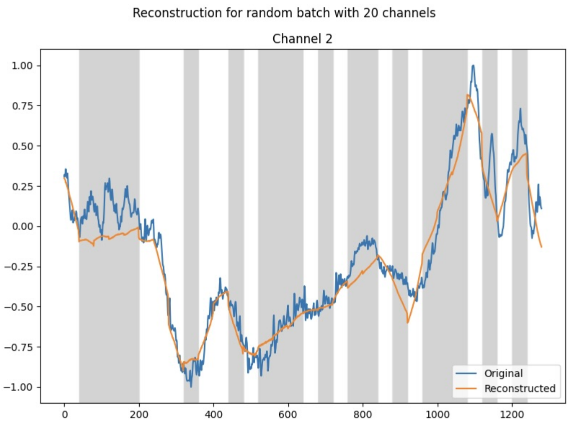



Figure 6: Example reconstruction on input with 20 channels (masked regions in gray).

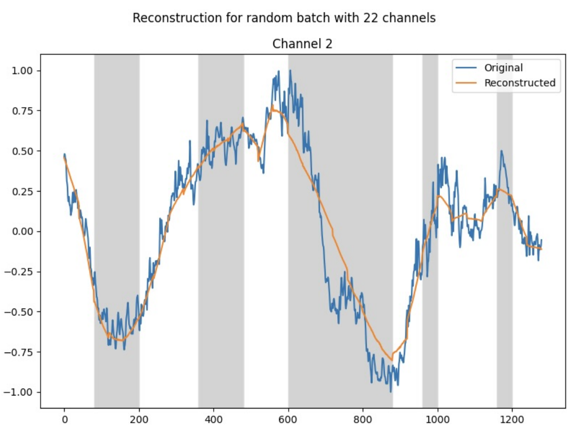



Figure 7: Example reconstruction on input with 22 channels (masked regions in gray).

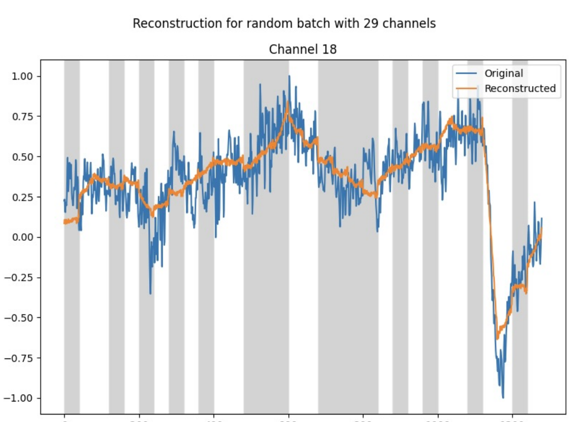



Figure 8: Example reconstruction on input with 29 channels (masked regions in gray).