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

Berkay Döner [1] **Thorir Mar Ingolfsson** [1] **Luca Benini** [1] **Yawei Li** [1]

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. Pre-trained on TUEG and Siena ($>$ 21,000 hours of raw EEG across diverse montages) using a masked signal reconstruction task, LUNA transfers effectively to four downstream tasks: abnormality detection, artifact detection, slowing classification, and emotion recognition. It demonstrates 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 GPU memory use by up to **10**$\times$. Code and pre-trained models can be found at `https://github.com/pulp-bio/BioFoundation`.

1. Introduction

Electroencephalography (EEG) provides vital non-invasive brain activity insights for diagnostics and neuroscience. Deep neural networks have advanced EEG analysis via end-to-end learning, moving beyond handcrafted pipelines (Craik et al., 2019). Transformers now rival traditional techniques by jointly modeling temporal and cross-electrode correlations (Song et al., 2023; Wen et al., 2023).

Despite this progress, a crucial bottleneck remains: EEG corpora exhibit significant topological heterogeneity. Varying electrode counts and placements hinder model transferability across montages, leading to performance degradation. For instance, cross-dataset evaluations show accuracy drops of up to 14 pp for motor-imagery decoders (Xu et al., 2020), 13–15 pp for emotion-recognition models (Yang et al., 2023; Yi et al., 2023). Current solutions to this problem are limited. Approaches include training models per montage or using only shared electrodes, discarding up to 80% of data (Lin et al., 2023). Flattening the electrode and time axes into long sequences leads to 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), which is prohibitive for dense caps (Yang et al., 2023). This shows the need for a single, montage-agnostic architecture that scales efficiently.

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. We pre-train LUNA using a masked-patch reconstruction objective on TUEG (Obeid & Picone, 2016) and SIENA (Detti et al., 2020) (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.** An encoder that projects arbitrary-sized channel sets into a fixed latent space.
- **Linear-in-channels complexity.** Patch-wise temporal attention that decouples complexity from electrode count.
- **State-of-the-art accuracy-efficiency trade-off.** LUNA achieves strong results across a range of EEG benchmarks, showing significant capabilities while reducing FLOPs and GPU memory usage on high-density EEG recordings.

2. Related Work

Self-Supervised Learning Strategies in EEG SSL is key for EEG foundation models. BENDR (Kostas et al., 2021) pioneered this by adapting masked prediction with a contrastive loss. Later work includes masked spectrogram prediction (Wang et al., 2023), vector-quantized patch prediction (Chen et al., 2024; Jiang et al., 2024), and raw signal reconstruction (Wang et al., 2025).

Modeling Spatial Structure and Topology in EEG Several strategies have been explored in the literature to capture

[1]Integrated Systems Laboratory, ETH Zürich, Zürich, Switzerland. Correspondence to: Yawei Li <yawli@iis.ee.ethz.ch>.

Proceedings of the 1st ICML Workshop on Foundation Models for Structured Data, Vancouver, Canada. 2025.

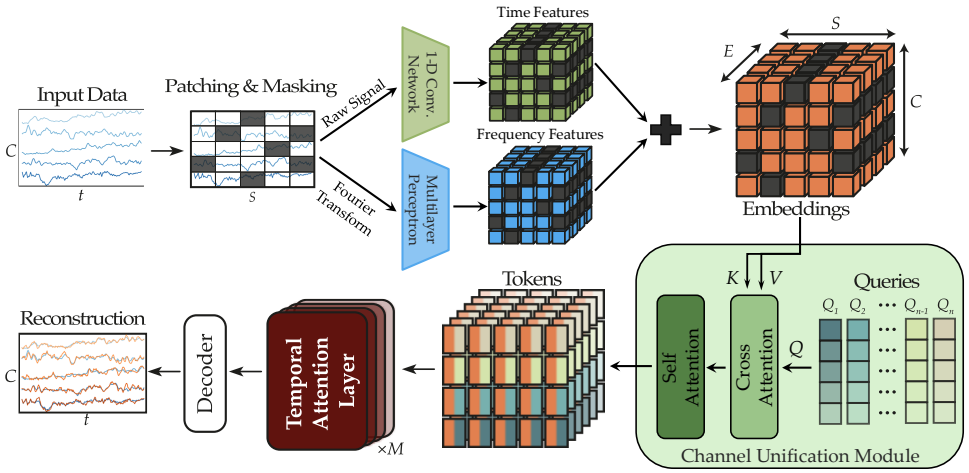

Figure 1: Overview of LUNA. EEG signals ($B \times C \times T$) are segmented into patches and embedded. Channel-Unification Module maps channel-wise features into a fixed-size latent space using learned queries (Q). Patch-wise Temporal Attention processes this latent sequence. The decoder generates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

the spatial relationships between electrodes:

Channel Independence: Some models (BrainBERT (Wang et al., 2023), EEGFormer (Chen et al., 2024)) process channels independently initially, handling variable counts but deferring cross-channel interaction modeling.

Fixed-Topology Spatial Modeling: Others like Brant (Zhang et al., 2023) use spatial encoders assuming consistent configurations. GNNs (Tang et al., 2022) model spatial relations via predefined graphs, requiring adaptation for varying topologies. LUNA avoids such fixed structures.

Joint Spatio-Temporal Attention: LaBraM (Jiang et al., 2024) flattens channel and patch dimensions into one sequence to learn spatio-temporal dependencies, incurring quadratic $\mathcal{O}((S \cdot C)^2)$ complexity. CBraMod (Wang et al., 2025) and CEReBrO (Dimofte et al., 2025) use alternating spatial and temporal attention mechanisms, reducing complexity to $\mathcal{O}(max(S^2, C^2))$. LUNA projects channels to the latent space before temporal attention.

Explicit Topology Mapping: MMM (Yi et al., 2023) maps channels to predefined regions using hand-engineered features. PopT (Chau et al., 2025) aggregates pre-computed temporal features using 3D electrode coordinates. These are not fully end-to-end or use external information.

3. Methodology

LUNA adopts an encoder-decoder architecture that transforms EEG signals from heterogeneous montages into a unified latent representation (Figure 1).

3.1. Encoder

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

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 (Wu & He, 2019), GELU (Hendrycks & Gimpel, 2016)) encodes local temporal features similar to state-of-the-art methods such as LaBraM and CBraMod. **Frequency Embedding:** The magnitude and phase from each patch's Fourier transform are passed through an MLP. These representations are summed to obtain patch features.

Channel Positional Encoding To encode electrode locations, we apply NeRF-inspired sinusoidal encoding (Mildenhall et al., 2021) to normalized 3D electrode coordinates, followed by an MLP projection. This yields positional features, which are added to patch features.

Channel-Unification Module 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. The cross-attention mechanism then computes the output representation $\mathbf{A}_{\text{out}} \in \mathbb{R}^{(B \cdot S) \times Q \times E}$:

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

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

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

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 $\mathbf{X}'_{\text{unified}} \in \mathbb{R}^{B\times S\times (Q\cdot E)}$. A stack of Transformer encoder processes these blocks with Rotary Positional Embeddings (RoPE) (Su et al., 2024) to capture temporal dependencies efficiently.

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

3.2. Decoder

There are 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 $E_{dec} \in \mathbb{R}^{C\times E}$ attend to E_{out} via cross-attention. 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 $E_{agg} \in \mathbb{R}^{1\times (Q\cdot E)}$ attends to E_{out} to produce a pooled output.

3.3. Training Objectives

LUNA is pre-trained with the combination of two losses:

Reconstruction Loss A Smooth L1 loss is applied to both masked and visible patches: $L_{rec} = \bar{S}_M + \alpha \cdot \bar{S}_{\neg M}$ where $\bar{S}_M$ and $\bar{S}_{\neg M}$ are the average SmoothL1 losses on masked (M) and non-masked sets and $\text{SmoothL1}(x, \hat{x}) = 0.5(x-\hat{x})^2$ if $|x - \hat{x}| < \beta$, else $\beta|x - \hat{x}| - 0.5\beta^2$, with $\beta = 1$.

Query Specialization Loss To promote diverse latent space, we penalize similarity in query-channel affinity matrices by minimizing the mean value of off-diagonal elements:

$$\mathcal{L}_{\text{spec}} = \frac{\lambda_{\text{spec}}}{B' \cdot Q \cdot (Q-1)} \sum_{b'=1}^{B'} \sum_{i=1}^{Q} \sum_{j=1, j\neq i}^{Q} \left((\mathbf{A}_{\text{affinity}} \mathbf{A}_{\text{affinity}}^{T})_{b',i,j} \right)^2$$

4. Results

4.1. Experimental Setup

Datasets We pre-train LUNA on Temple University Hospital EEG Corpus (TUEG) and the Siena Scalp EEG Database, spanning recordings with 20, 22, and 29 channels, amounting to over 21,900 hours of EEG data (see Table 4). Downstream evaluations cover four diverse benchmarks: **TUAB** (Obeid & Picone, 2016): Abnormal EEG detection (binary classification), **TUAR** (Obeid & Picone, 2016): Artifact detection (multi-class classification) **TUSL** (Obeid & Picone, 2016): Slowing event classification (4-class classification). **SEED-V** (Liu et al., 2022): Emotion recognition (5-class classification), with unseen 62-channel topology.

Baselines We compare LUNA against supervised and self-supervised methods including ContraWR (Yang et al., 2021), CNN-Transformer (Peh et al., 2022), FFCL (Li et al., 2022), EEGNet (Lawhern et al., 2018), EEG-GNN (Tang et al., 2022), Transformer (Song et al., 2021), LaBraM, CBraMod, FEMBA (Tegon et al., 2025), CEReBrO, EEGFormer, BIOT (Yang et al., 2023), BENDR, BrainBERT, and EEG2Rep (Foumani et al., 2024). LUNA is evaluated in three sizes: Base (7M), Large (43M), and Huge (311M).

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.

Table 1: Performance comparison on TUAB.

Model	Size	Bal. Acc. (%) ↑	AUC-PR ↑	AUROC ↑
Supervised Models				
ContraWR	1.6M	77.46 ± 0.41	0.8421 ± 0.0140	0.8456 ± 0.0074
CNN-Transformer	3.2M	77.77 ± 0.22	0.8433 ± 0.0039	0.8461 ± 0.0013
FFCL	2.4M	78.48 ± 0.38	0.8448 ± 0.0065	0.8569 ± 0.0051
ST-Transformer	3.2M	79.66 ± 0.23	0.8521 ± 0.0026	0.8707 ± 0.0019
Self-supervised Models				
BENDR	0.39M	76.96 ± 3.98	-	0.8397 ± 0.0344
BrainBERT	43.2M	-	0.8460 ± 0.0030	0.8530 ± 0.0020
EEGFormer-Base	2.3M	-	0.8670 ± 0.0020	0.8670 ± 0.0030
BIOT	3.2M	79.59 ± 0.57	0.8692 ± 0.0023	0.8815 ± 0.0043
EEG2Rep	-	80.52 ± 2.22	-	0.8843 ± 0.0309
FEMBA-Huge	386M	81.82 ± 0.16	0.9005 ± 0.0017	0.8921 ± 0.0042
CEReBrO	85.15M	81.67 ± 0.23	0.9049 ± 0.0026	0.8916 ± 0.0038
LaBraM-Base	5.9M	81.40 ± 0.19	0.8965 ± 0.0016	0.9022 ± 0.0009
LaBraM-Huge	369.8M	**82.58 ± 0.11**	0.9204 ± 0.0011	**0.9162 ± 0.0016**
CBraMod	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 0.92 on TUAR and AUROC 0.80 on TUSL, outperforming other models.

Table 2: Performance comparison on TUAR and TUSL.

Model	Size	TUAR		TUSL	
		AUROC ↑	AUC-PR ↑	AUROC ↑	AUC-PR ↑
Supervised Models					
EEGNet	-	0.75 ± 0.01	0.43 ± 0.03	0.64 ± 0.01	0.35 ± 0.01
EEG-GNN	-	0.84 ± 0.02	0.49 ± 0.01	0.72 ± 0.01	0.38 ± 0.00
Self-supervised Models					
BrainBERT	43.2M	0.75 ± 0.01	0.35 ± 0.01	0.59 ± 0.01	0.35 ± 0.00
EEGFormer-Large	3.2M	0.85 ± 0.00	0.48 ± 0.01	0.68 ± 0.01	0.39 ± 0.00
FEMBA-Base	47.7M	0.90 ± 0.01	**0.56 ± 0.00**	0.73 ± 0.01	0.29 ± 0.01
FEMBA-Large	77.8M	0.91 ± 0.00	0.52 ± 0.00	0.71 ± 0.01	0.28 ± 0.01
LUNA-Base	7M	0.90 ± 0.01	0.50 ± 0.01	0.76 ± 0.02	0.30 ± 0.00
LUNA-Large	43M	0.92 ± 0.00	0.51 ± 0.01	0.77 ± 0.01	0.29 ± 0.02
LUNA-Huge	311.4M	**0.92 ± 0.01**	0.53 ± 0.01	**0.80 ± 0.01**	0.29 ± 0.01

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 LUNA's performance lags behind leading methods like CBraMod by 2-3 pp. This suggests gener-

alizing zero-shot to vastly different, high-density layouts remains challenging.

Table 3: Performance comparison on SEED-V.

Model	Size	Bal. Acc. (%) ↑	Cohen's Kappa ↑	Weighted F1 ↑
Supervised Models				
ContraWR	1.6M	0.3546 ± 0.0105	0.1905 ± 0.0188	0.3544 ± 0.0121
CNN-Transformer	3.2M	0.3678 ± 0.0078	0.2072 ± 0.0183	0.3642 ± 0.0088
FFCL	2.4M	0.3641 ± 0.0092	0.2078 ± 0.0201	0.3645 ± 0.0132
ST-Transformer	3.5M	0.3052 ± 0.0072	0.1083 ± 0.0121	0.2833 ± 0.0105
Self-supervised Models				
BIOT	3.2M	0.3837 ± 0.0187	0.2261 ± 0.0262	0.3856 ± 0.0203
LaBraM-Base	5.8M	0.3976 ± 0.0138	0.2386 ± 0.0209	0.3974 ± 0.0111
CBraMod	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 better efficiency compared to full and alternating attention models when the number of channels or patches is increased (while keeping the other fixed). As shown in Figure 2, LUNA's patch-wise attention enables thousands of temporal patches without the quadratic cost faced by LaBraM. Also, Figure 3 shows that LUNA maintains near-constant compute cost when channel count increases, outperforming CBraMod's scaling for dense EEG.

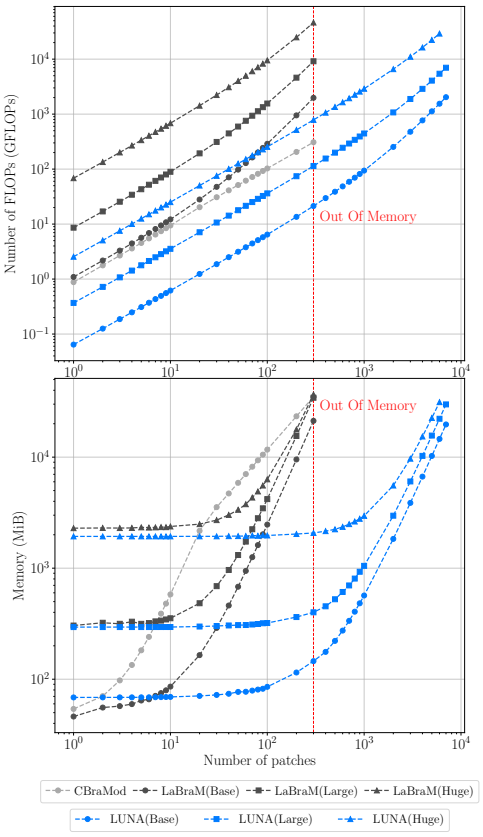

Figure 2: Scaling with number of patches.

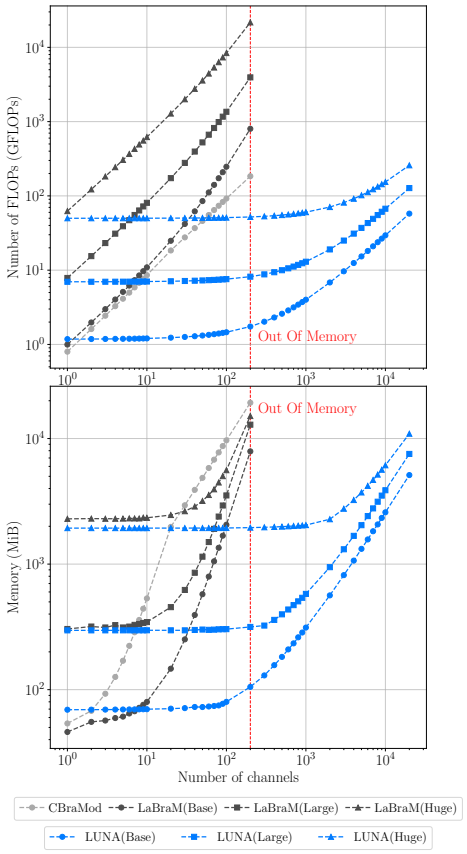

Figure 3: Scaling with number of channels.

4.4. 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 LUNA learns flexible, data-driven spatial unification.

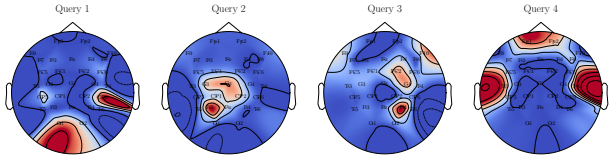

Figure 4: Visualization of the attention patterns of queries.

5. Conclusion

We introduced LUNA, a self-supervised foundation model that unifies diverse EEG electrode layouts into a fixed, montage-agnostic latent space. LUNA achieves competitive performance on different benchmarks, with significant FLOPs/memory reductions across all tested configurations. However, LUNA's performance on unseen SEED-V topologies indicates sensitivity, likely due to pre-trained positional encodings, showing a need for enhanced spatial generalization or hybrid embeddings in future work.

References

Chau, G., Wang, C., Talukder, S. J., Subramaniam, V., Soedarmadji, S., Yue, Y., Katz, B., and Barbu, A. Population transformer: Learning population-level representations of neural activity. In *The Thirteenth International Conference on Learning Representations*, January 2025.

Chen, Y., Ren, K., Song, K., Wang, Y., Wang, Y., Li, D., and Qiu, L. EEGFormer: Towards transferable and interpretable large-scale EEG foundation model. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, February 2024.

Craik, A., He, Y., and Contreras-Vidal, J. L. Deep learning for electroencephalogram (EEG) classification tasks: a review. *Journal of Neural Engineering*, 16(3):031001, April 2019. ISSN 1741-2552. doi: 10.1088/1741-2552/ab0ab5.

Detti, P., Vatti, G., and Zabalo Manrique de Lara, G. 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/.

Dimofte, A., Bucagu, G. A., Ingolfsson, T. M., Wang, X., Cossettini, A., Benini, L., and Li, Y. CEReBrO: Compact encoder for representations of brain oscillations using efficient alternating attention, January 2025.

Foumani, N. M., Mackellar, G., Ghane, S., Irtza, S., Nguyen, N., and Salehi, M. 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, pp. 5544–5555. ACM, August 2024. doi: 10.1145/3637528.3671600.

Hendrycks, D. and Gimpel, K. Gaussian error linear units (gelus), June 2016.

Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., Henaff, O. J., Botvinick, M., Zisserman, A., Vinyals, O., and Carreira, J. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*, January 2022.

Jiang, W., Zhao, L., and liang Lu, B. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, January 2024.

Kostas, D., Aroca-Ouellette, S., and Rudzicz, F. 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. ISSN 1662-5161. doi: 10.3389/fnhum.2021.653659.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, July 2018. ISSN 1741-2552. doi: 10.1088/1741-2552/aace8c.

Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., and Teh, Y. W. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, June 2019.

Li, H., Ding, M., Zhang, R., and Xiu, C. Motor imagery EEG classification algorithm based on cnn-lstm feature fusion network. *Biomedical Signal Processing and Control*, 72:103342, February 2022. ISSN 1746-8094. doi: 10.1016/j.bspc.2021.103342.

Lin, X., Chen, J., Ma, W., Tang, W., and Wang, Y. EEG emotion recognition using improved graph neural network with channel selection. *Computer Methods and Programs in Biomedicine*, 231:107380, April 2023. ISSN 0169-2607. doi: 10.1016/j.cmpb.2023.107380.

Liu, W., Qiu, J.-L., Zheng, W.-L., and Lu, B.-L. 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. ISSN 2379-8939. doi: 10.1109/tcds.2021.3071170.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. NeRF: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, December 2021. ISSN 0001-0782. doi: 10.1145/3503250.

Obeid, I. and Picone, J. The temple university hospital EEG data corpus. *Frontiers in Neuroscience*, 10:196, May 2016. ISSN 1662-453X. doi: 10.3389/fnins.2016.00196.

Peh, W. Y., Yao, Y., and Dauwels, J. Transformer convolutional neural networks for automated artifact detection in scalp EEG, July 2022.

Song, Y., Jia, X., Yang, L., and Xie, L. Transformer-based spatial-temporal feature learning for EEG decoding, June 2021.

Song, Y., Zheng, Q., Liu, B., and Gao, X. EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, December 2023. ISSN 1534-4320. doi: 10.1109/tnsre.2022.3230250.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, February 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.127063.

Tang, S., Dunnmon, J., Saab, K. K., Zhang, X., Huang, Q., Dubost, F., Rubin, D., and Lee-Messer, C. Self-supervised graph neural networks for improved electroencephalographic seizure analysis. In *International Conference on Learning Representations*, January 2022.

Tegon, A., Ingolfsson, T. M., Wang, X., Benini, L., and Li, Y. FEMBA: Efficient and scalable EEG analysis with a bidirectional mamba foundation model, February 2025.

Wang, C., Subramaniam, V., Yaari, A. U., Kreiman, G., Katz, B., Cases, I., and Barbu, A. BrainBERT: Self-supervised representation learning for intracranial recordings. In *The Eleventh International Conference on Learning Representations*, February 2023.

Wang, J., Zhao, S., Luo, Z., Zhou, Y., Jiang, H., Li, S., Li, T., and Pan, G. CBraMod: A criss-cross brain foundation model for EEG decoding. In *The Thirteenth International Conference on Learning Representations*, January 2025.

Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., and Sun, L. Transformers in time series: A survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI-2023, pp. 6778–6786. International Joint Conferences on Artificial Intelligence Organization, August 2023. doi: 10.24963/ijcai.2023/759.

Wu, Y. and He, K. Group normalization. *International Journal of Computer Vision*, 128(3):742–755, July 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01198-w.

Xu, L., Xu, M., Ke, Y., An, X., Liu, S., and Ming, D. Cross-dataset variability problem in EEG decoding with deep learning. *Frontiers in Human Neuroscience*, 14:103, April 2020. ISSN 1662-5161. doi: 10.3389/fnhum.2020.00103.

Yang, C., Xiao, C., Westover, M. B., and Sun, J. Self-supervised electroencephalogram representation learning for automatic sleep staging: Model development and evaluation study. *JMIR AI*, 2:e46769, July 2021. ISSN 2817-1705. doi: 10.2196/46769.

Yang, C., Westover, M., and Sun, J. BIOT: Biosignal transformer for cross-data learning in the wild. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 78240–78260. Curran Associates, Inc., September 2023.

Yi, K., Wang, Y., Ren, K., and Li, D. Learning topology-agnostic EEG representations with geometry-aware modeling. In *Thirty-seventh Conference on Neural Information Processing Systems*, September 2023.

Zhang, D., Yuan, Z., Yang, Y., Chen, J., Wang, J., and Li, Y. Brant: Foundation model for intracranial neural signal. In *Thirty-seventh Conference on Neural Information Processing Systems*, September 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. Experimental Details

Datasets We use publicly available EEG datasets, provided in 4.

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 The pre-trained models are fine-tuned on the downstream tasks to evaluate the performance, similar to the baselines. During fine-tuning, all model layers are trained with a layer-wise learning rate decay, meaning earlier layers have progressively lower learning rates. 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 (Chen et al., 2024). 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. More details are reported in the Table 6.

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. 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.2. Model Setup Details

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

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 query self-attention layers		3	3	3
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 β			(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	

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 β	(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.2.1. Complexity Analysis

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

Table 4: 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

Table 7: Complexity Breakdown of LUNA Encoder Stages.

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

Table 8: Attention Complexity Comparison.

Method	Bottleneck Complexity
LUNA	$O(B \cdot S^2 \cdot Q \cdot E)$ *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)	$O(B \cdot S^2 \cdot C \cdot E)$
Alternating Attention (Channels)	$O(B \cdot S \cdot C^2 \cdot E)$

A.3. Detailed Literature Review

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.

A.3.1. 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 (Kostas et al., 2021) 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 (Wang et al., 2023) performs masked prediction on channel-independent spectrograms for intracranial electroencephalography (iEEG); EEGFormer (Chen et al., 2024) and LaBraM (Jiang et al., 2024) predict vector-quantized (VQ) representations of masked patches, learning discrete codebooks; CBraMod (Wang et al., 2025) 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.

A.3.2. 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 (Wang et al., 2023) and EEGFormer (Chen et al., 2024) 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 (Zhang et al., 2023) use dedicated spatial encoders alongside temporal ones but assume a consistent channel configuration, limiting cross-dataset generalization. Graph Neural Networks (GNNs) (Tang et al., 2022) 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 (Jiang et al., 2024) 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 (Wang et al., 2025) and CEReBrO (Dimofte et al., 2025) 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 (Yang et al., 2023) 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 (Yi et al., 2023) maps channels to predefined anatom-

Table 9: 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 ± 0.002	0.895 ± 0.002	0.902 ± 0.011	0.495 ± 0.010
Unification Module:				
- Region-based Attention	0.883 ± 0.001 (↓ 0.004)	0.892 ± 0.002 (↓ 0.003)	0.896 ± 0.001 (↓ 0.006)	0.509 ± 0.006 (↑ 0.014)
Other Components:				
- w/o Query Specialization Loss	0.884 ± 0.003 (↓ 0.003)	0.892 ± 0.002 (↓ 0.003)	0.895 ± 0.005 (↓ 0.007)	0.498 ± 0.010 (↑ 0.003)
- w/o Frequency Features	0.876 ± 0.012 (↓ 0.011)	0.883 ± 0.005 (↓ 0.012)	0.893 ± 0.011 (↓ 0.009)	0.490 ± 0.011 (↓ 0.005)

ical regions but relies on hand-engineered features (Differential Entropy) rather than raw signals. PopT (Chau et al., 2025) 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.

A.4. 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 (Lee et al., 2019) 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 (Jaegle et al., 2022) 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.

A.5. Additional Quantitative Results

A.5.1. Ablation Studies

We validate the impact of LUNA's key design choices on TUAB and TUAR (Table 9).

Learned Queries vs. Fixed Regions Replacing learned queries with predefined spatial regions (similar to what MMM (Yi et al., 2023) 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.

A.5.2. Training Curves

The pre-training loss curves for LUNA-Base are shown in 5. The reconstruction loss drops show an 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.

A.6. Additional Visualizations

Pre-trained Representations t-SNE visualizations (Figure 6 and Figure 7) 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.

Reconstruction Examples Figures 8, 9, and 10 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.

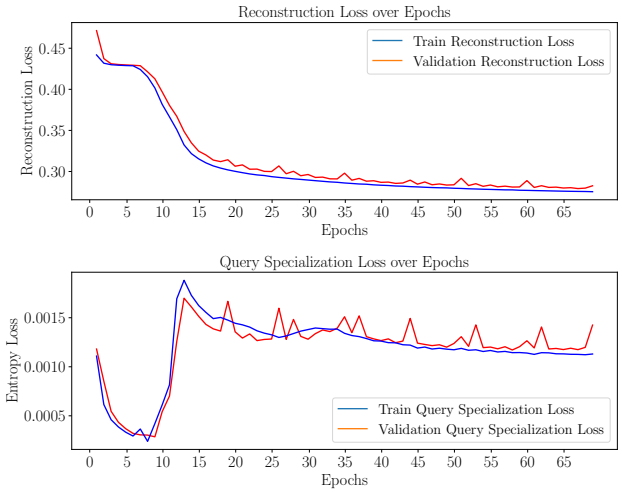

Figure 5: Loss curves during pre-training for LUNA-Base (Reconstruction and Query Specialization Loss).

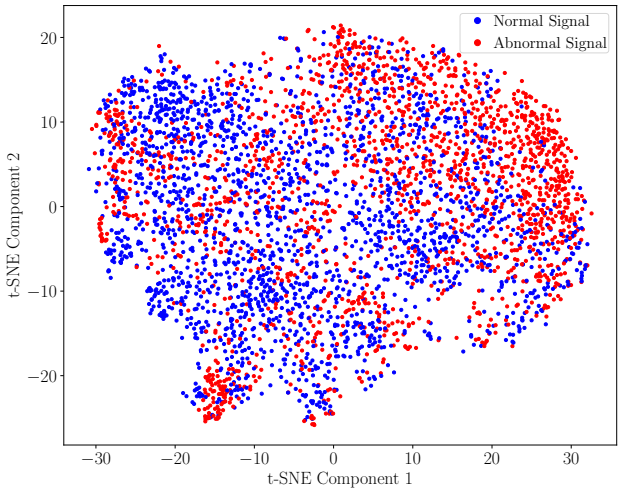

Figure 6: TUAB dataset (Normal vs. Abnormal Signal).

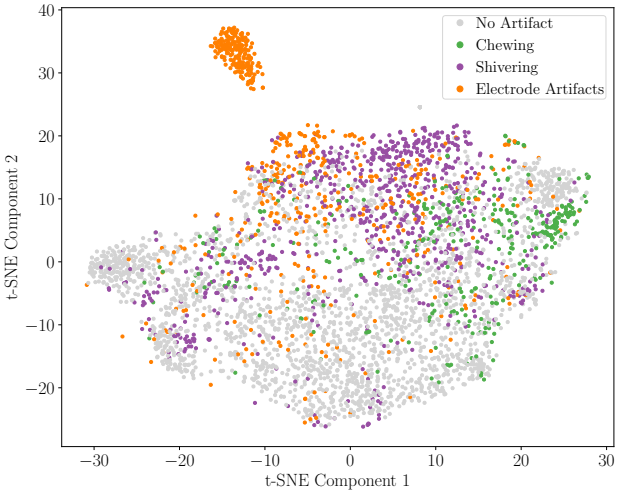

Figure 7: TUAR dataset (Artifact Types).

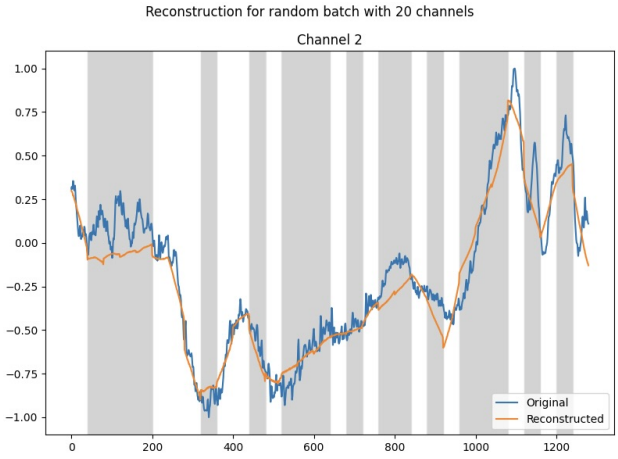

Figure 8: Example reconstruction on input with 20 channels (masked regions in gray).

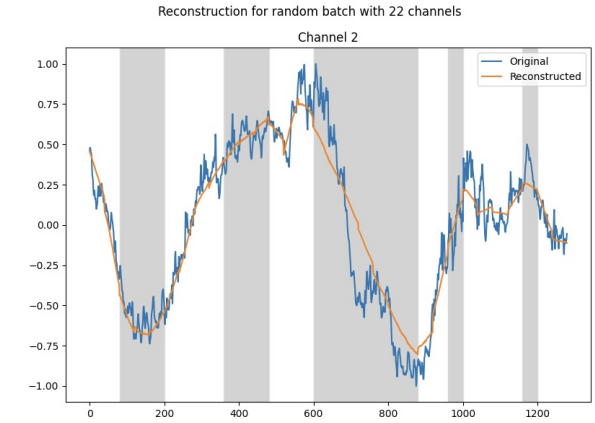

Figure 9: Example reconstruction on input with 22 channels (masked regions in gray).

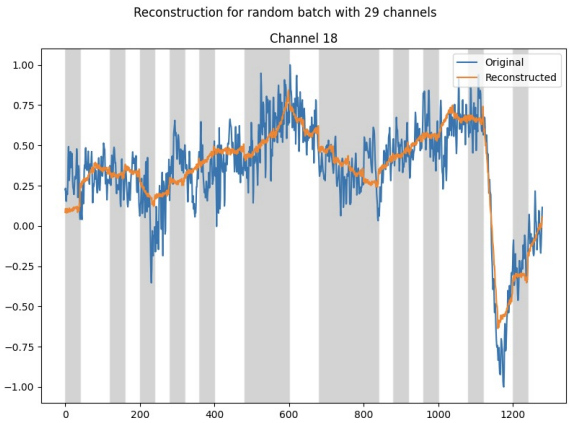

Figure 10: Example reconstruction on input with 29 channels (masked regions in gray).