DIVER-0 : A Fully Channel Equivariant EEG Foundation Model

Danny Dongyeop Han[1], Ahhyun Lucy Lee[1], Taeyang Lee[1], Yonghyeon Gwon[1], Sebin Lee[1], Seong Jin Lee[1], David Keetae Park[2], Shinjae Yoo[2], Chun Kee Chung[3*], Jiook Cha[1*]

SNU CONNECTOME LAB

[1] Seoul National University [2] Brookhaven National Laboratory [3] Seoul National University Hospital

ICML International Conference On Machine Learning

Introduction

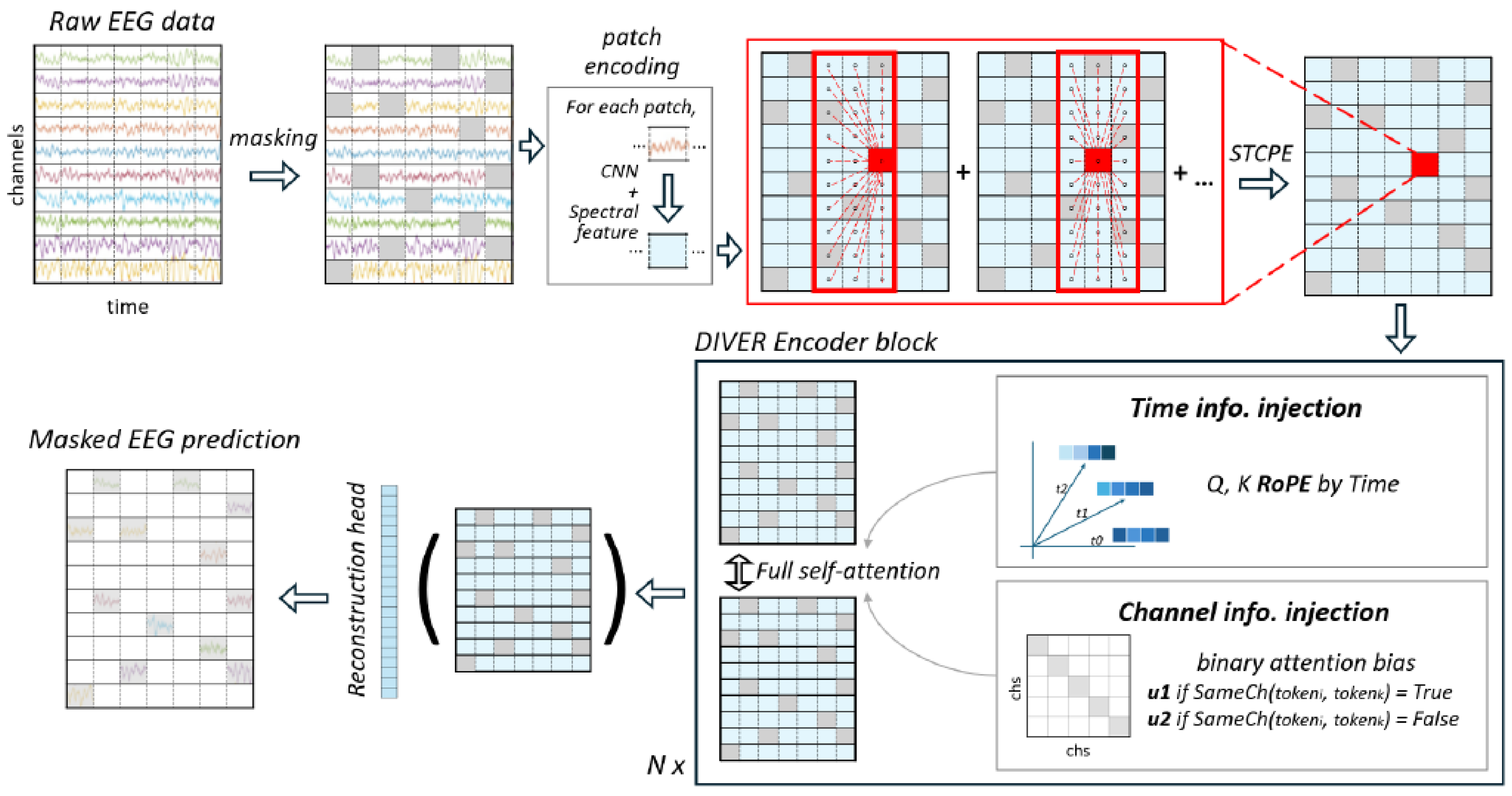
Motivation

- EEG (Electroencephalography) records brain electrical activity through electrodes, widely utilized in Brain-Computer Interface.
- Current EEG decoding models have two critical limitations:
  1. **Restrictive modeling of spatio-temporal brain dynamics**
     They cannot learn brain's distributed dynamics and adapt to unseen electrode configurations. (e.g., criss-cross attention [1])
  2. **Poor handling of channel information**
     They are vulnerable to channel permutation and have limited generalizability across electrode configurations.

Contribution

- **DIVER transformer block**
  → Unified spatio-temporal attention, combining **rotary positional embedding (RoPE)** for time relationship and **binary attention bias** for electrode differentiation inspired by Woo et al. (2024) [2].
- **Sliding Temporal Conditional Positional Encoding (STCPE)**
  → **Positional encoding using transformer** that preserves both temporal translation equivariance and channel permutation equivariance.

Methods



a. Patch Encoding

- Patch-wise CNNs and Fast Fourier Transform (FFT)

b. Sliding Temporal Conditional Positional Encoding (STCPE)

- A transformer block slides across the temporal dimension.
- At each time step, the transformer processes patches from all channels simultaneously.

c. DIVER Encoder

- **RoPE** encodes temporal relationship between timepoints by rotating embeddings.
- **Binary attention bias** informs the model when two electrodes represent the same channel.

d. Attention mechanism

$$\begin{aligned} E_{ij,mn} &= (\mathbf{W}^Q \boldsymbol{x}_{i,m})^T \mathbf{R}_{i-j} (\mathbf{W}^K \boldsymbol{x}_{j,n}) \\ &\quad + u^1 * \mathbb{1}_{\{m=n\}} + u^2 * \mathbb{1}_{\{m \neq n\}}, \\ A_{ij,mn} &= \frac{\exp\{E_{ij,mn}\}}{\sum_{k,o} \exp\{E_{ik,mo}\}} \end{aligned} \tag{1}$$

where $\mathbf{W}^Q \boldsymbol{x}_{i,m}, \mathbf{W}^K \boldsymbol{x}_{j,n} \in \mathbb{R}^{d_h}$ are the query and key vectors, $\mathbf{R}_{i-j} \in \mathbb{R}^{d_h \times d_h}$ is the rotary projection matrix, $u^1, u^2 \in \mathbb{R}$ are learnable scalars that can be different in each head, $\mathbb{1}_{\{cond\}} = 1$ if $cond$ is true, and $\mathbb{1}_{\{cond\}} = 0$ otherwise.

- The model is pretrained using masked patch reconstruction.

Experimental Results

1. Performance comparison on downstream EEG tasks

Methods	FACED, 9-class Bal. Acc.(%)	FACED, 9-class Kappa(%)	FACED, 9-class F1(%)	PhysioNet-MI, 4-class Bal. Acc. (%)	PhysioNet-MI, 4-class Kappa(%)	PhysioNet-MI, 4-class F1(%)
EEGNet	40.9 ± 1.2	33.4 ± 2.5	41.2 ± 1.4	58.1 ± 1.3	44.7 ± 2.0	58.0 ± 1.2
SPaRCNet	46.7 ± 1.6	39.8 ± 2.9	47.3 ± 1.3	59.3 ± 1.5	45.6 ± 2.3	59.4 ± 1.5
ST-Transformer	48.1 ± 0.8	41.4 ± 1.3	48.0 ± 1.0	60.4 ± 0.8	47.1 ± 2.0	60.5 ± 0.8
EEGConformer	45.6 ± 1.3	38.6 ± 1.9	45.1 ± 1.1	60.5 ± 1.0	47.4 ± 1.7	60.6 ± 1.0
LaBraM-Base	52.7 ± 1.1	47.0 ± 1.9	52.9 ± 1.0	61.7 ± 1.2	49.1 ± 1.9	61.8 ± 1.4
CBraMod	55.1 ± 0.9	50.4 ± 1.2	56.2 ± 0.9	**64.2 ± 0.9**	**52.2 ± 1.7**	**64.3 ± 1.0**
DIVER 10% (Ours)	**59.2 ± 0.8**	**54.0 ± 0.9**	**59.6 ± 0.7**	62.8 ± 0.5	50.4 ± 0.7	62.9 ± 0.5

Performance values for baseline methods are quoted from Wang et al. (2024) [1] using identical preprocessing and evaluation protocols. DIVER results are from our experiments. All metrics are reported as mean ± standard deviation across 5 random seeds. DIVER 10% indicate that the model was pretrained using 10% of the full TUEG dataset.

2. Ablation study of pretrained model components

Methods	FACED, 9-class Bal. Acc. (%)	FACED, 9-class Kappa(%)	FACED, 9-class F1(%)	PhysioNet-MI, 4-class Bal. Acc. (%)	PhysioNet-MI, 4-class Kappa(%)	PhysioNet-MI, 4-class F1(%)
DIVER 10% (Ours)	**59.2 ± 0.8**	**54.0 ± 0.9**	**59.6 ± 0.7**	62.8 ± 0.5	50.4 ± 0.7	62.9 ± 0.5
w/o patch-wise CNN encoding	57.3 ± 0.9	51.7 ± 1.0	57.4 ± 0.9	61.9 ± 0.5	49.3 ± 0.6	62.1 ± 0.5
w/o spectral embedding	58.0 ± 1.1	52.6 ± 1.3	58.3 ± 1.2	**63.4 ± 0.5**	**51.1 ± 0.6**	**63.5 ± 0.5**
w/o STCPE	58.4 ± 1.0	53.0 ± 1.1	58.6 ± 1.0	62.8 ± 0.5	50.5 ± 0.6	63.0 ± 0.5
STCPE → ACPE	58.7 ± 0.9	53.4 ± 1.1	59.0 ± 1.0	62.4 ± 0.3	49.9 ± 0.4	62.5 ± 0.3
w/o RoPE	57.7 ± 0.6	52.2 ± 0.7	58.1 ± 0.6	62.6 ± 0.4	50.2 ± 0.5	62.8 ± 0.4
w/o Binary attention bias	58.1 ± 1.0	52.6 ± 1.1	58.4 ± 1.0	62.8 ± 0.5	50.4 ± 0.7	63.0 ± 0.6
DIVER → Vanilla block	55.3 ± 1.9	49.5 ± 2.2	55.5 ± 1.9	61.6 ± 1.5	48.8 ± 1.9	61.7 ± 1.5
DIVER → CBraMod block	57.0 ± 0.6	51.5 ± 0.6	57.4 ± 0.5	63.1 ± 0.8	50.8 ± 1.0	63.2 ± 0.8
CBraMod 10%	56.5 ± 0.8	51.0 ± 1.0	56.9 ± 0.8	62.4 ± 0.6	49.9 ± 0.8	62.6 ± 0.7

3. Channel Permutation analysis on FACED dataset

Finetuning	Methods	Pretrain: Intact Bal. Acc.(%)	Intact Kappa(%)	Intact F1(%)	Pretrain: Permute Bal. Acc.(%)	Permute Kappa(%)	Permute F1(%)
Intact	**DIVER 10% (Ours)**	**59.2 ± 0.9**	**54.0 ± 0.9**	**59.6 ± 0.7**	59.6 ± 0.8	54.4 ± 0.9	59.9 ± 0.8
Intact	w/o STCPE	58.4 ± 1.0	53.0 ± 1.1	58.6 ± 1.0	58.4 ± 0.8	52.9 ± 0.9	58.5 ± 0.8
Intact	STCPE → ACPE	56.7 ± 1.1	51.2 ± 1.3	57.2 ± 1.1	**60.1 ± 0.9**	**54.8 ± 1.1**	**60.3 ± 0.9**
Intact	w/o Binary attention bias	58.1 ± 1.0	52.6 ± 1.1	58.4 ± 1.0	58.1 ± 1.3	52.7 ± 1.4	58.5 ± 1.2
Intact	DIVER → Vanilla block	55.3 ± 1.9	49.5 ± 2.2	55.5 ± 1.9	56.3 ± 0.7	50.7 ± 0.9	56.6 ± 0.8
Intact	DIVER → CBraMod block	57.0 ± 0.6	51.5 ± 0.6	57.4 ± 0.5	56.8 ± 0.6	51.2 ± 0.7	57.1 ± 0.6
Intact	CBraMod 10%	56.5 ± 0.8	51.0 ± 1.0	56.9 ± 0.8	58.7 ± 0.4	53.4 ± 0.4	59.0 ± 0.5
Permute	**DIVER 10% (Ours)**	59.1 ± 0.4	53.8 ± 0.5	59.4 ± 0.3	59.0 ± 0.7	53.6 ± 0.8	**59.3 ± 0.7**
Permute	w/o STCPE	58.7 ± 0.4	53.3 ± 0.4	58.9 ± 0.3	**59.1 ± 0.4**	**53.6 ± 0.4**	58.9 ± 0.2
Permute	STCPE → ACPE	**60.1 ± 0.9**	**54.8 ± 1.0**	**60.3 ± 0.8**	58.8 ± 0.4	53.4 ± 0.5	59.1 ± 0.4
Permute	w/o Binary attention bias	58.6 ± 1.2	53.2 ± 1.4	58.9 ± 1.3	58.7 ± 0.2	53.3 ± 0.2	59.0 ± 0.2
Permute	DIVER → Vanilla block	57.6 ± 1.7	51.8 ± 1.8	57.9 ± 0.2	55.6 ± 0.8	49.9 ± 0.9	55.9 ± 0.7
Permute	DIVER → CBraMod block	57.0 ± 0.8	51.4 ± 0.9	57.3 ± 0.9	57.1 ± 0.6	51.6 ± 0.7	57.4 ± 0.6
Permute	CBraMod 10%	56.5 ± 0.5	50.9 ± 0.6	56.8 ± 0.5	58.0 ± 0.9	52.5 ± 1.1	58.2 ± 1.0

* 'w/o': our model without the corresponding component.
'→' : replacement of our model's component with the alternative component.

Conclusion

- We present **DIVER-0**, a novel EEG foundation model with two key innovations: **DIVER transformer block** and **Sliding Temporal Conditional Positional Encoding (STCPE)**.
- Experiments demonstrate that DIVER-0 achieves **competitive performance** across representative BCI tasks while maintaining strict permutation equivariance.
- DIVER-0's consistent performance across all channel permutation demonstrates the value of maintaining **strict permutation equivariance for cross-dataset generalization.**
- As a future work, we will conduct 1) **channel position injection** while still being channel permutation equivariant and 2) **massive scaling** in data size, model size, and compute.

Reference

[1] 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. arXiv preprint arXiv:2412.07236, 2024.

[2] Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., and Sahoo,D. Unified training of universal time series forecasting transformers. International Conference on Machine Learning, 2024.

[3] Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. Neurocomputing, 568:127063, 2024.