哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

# EEGPT: Pretrained Transformer for Universal and Reliable Representation of EEG Signals


Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, Haifeng Li*

Faculty of Computing, HIT


2024.11.12

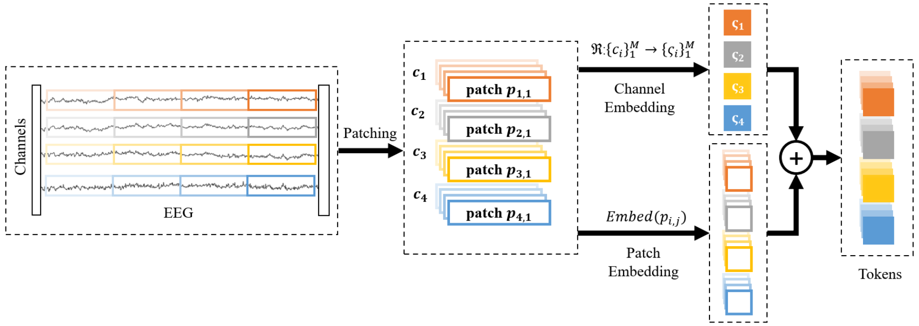
# Background

Electroencephalogram (EEG) data analysis faces numerous challenges, specifically:

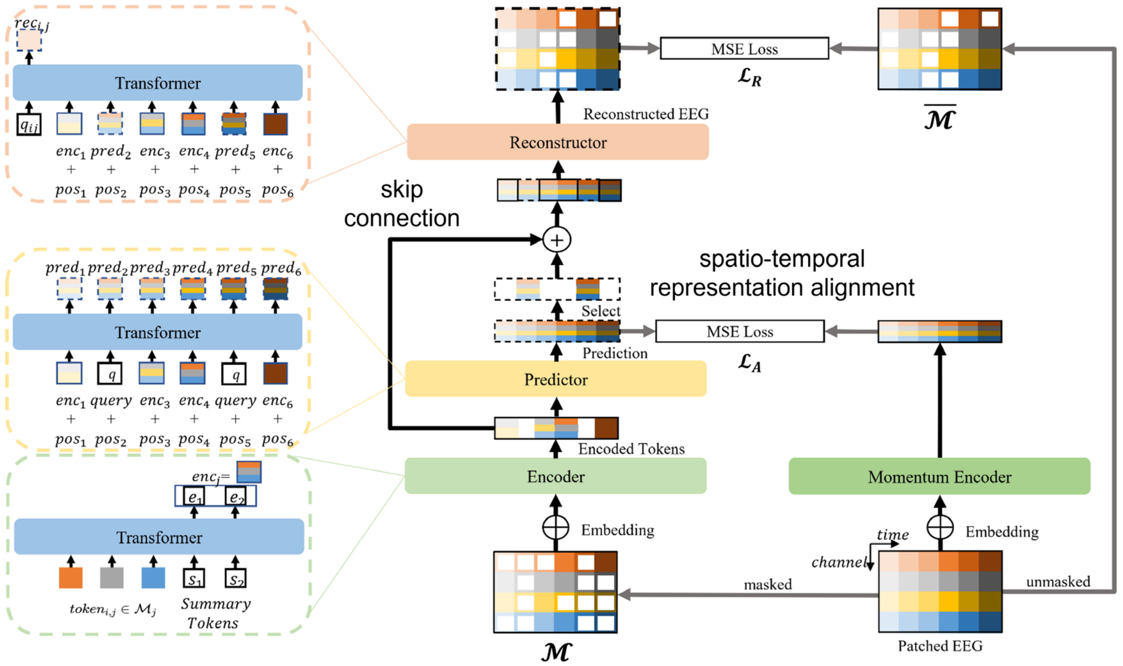
- **Low signal-to-noise**
- **High inter-subject variability**
- **Inherently task-dependent variations**
- **Channel mismatch**

We introduces the **EEG Pretrained Transformer (EEGPT)**, a novel, universal model with over 10 million parameters. By training on a wide-ranging dataset, model universality is enhanced. Improvements to the model structure increase its compatibility even with missing channels and enhance the quality of representations.

# ➢ Method



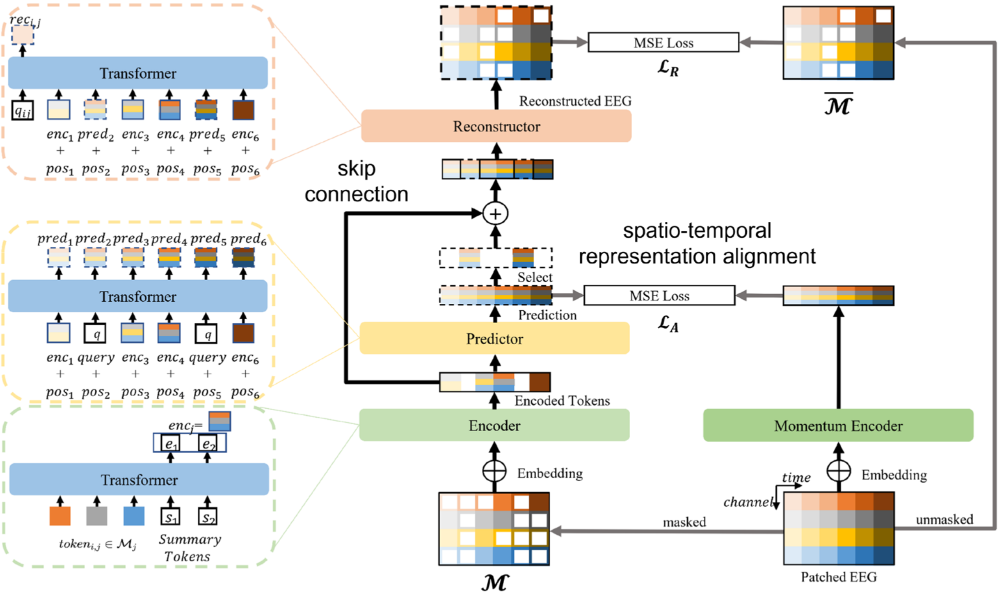

Patchify & Local spatio-temporal embedding.




The EEGPT Structure for Pretraining.

- **Patchify** the input EEG signal as $p_{i,j}$ through masking 50% time and 80% channel patches, splitting into the $\mathcal{M}$ part and the $\overline{\mathcal{M}}$ part.
- **Embedding** $p_{i,j}$ to $token_{i,j}$ by local spatio-temporal embedding.
- **Encoder** extracts feature $\boldsymbol{enc_j}$ consisting of $\{e_i\}_{i=1}^{S}$ for each time $j$ in the $\mathcal{M}$ part.
- **Predictor** predicts $\boldsymbol{pred_j}$, aligning with the Momentum Encoder output $\boldsymbol{menc_j}$.
- **Reconstructor** generates $\boldsymbol{rec_{i,j}}$ to reconstruct the EEG signal in the $\overline{\mathcal{M}}$ part.

# Method




**The EEGPT Structure for Pretraining.**

Based on the masked autoencoder, we introduces a **spatio-temporal representation alignment** to explicitly represent $\boldsymbol{z}$, changing target function from:

$$\min_{\theta,\phi} \mathbb{E}_{x\sim\mathcal{D}} \mathcal{H}(d_\phi(z), x \odot (1-M)), z = f_\theta(x \odot M)$$

to:

$$\min_{\theta,\phi} \mathbb{E}_{x\sim\mathcal{D}} \underbrace{\mathcal{H}(d_\phi(z), x \odot (1-M))}_{\mathcal{L}_R} + \underbrace{\mathcal{H}(z, f_\theta(x))}_{\mathcal{L}_A}, z = f_\theta(x \odot M)$$

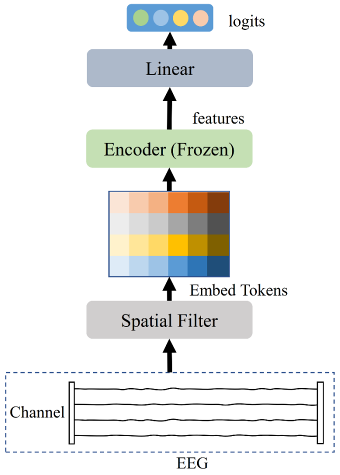
Where $\boldsymbol{M}$ is mask matrix, $\mathcal{H}$ is similarity measure, and $f_\theta, d_\phi$ are the encoder and decoder with parameters $\theta$ and $\phi$, respectively. This method results in improved encoding quality and generalization.

# Experiments

## Comparative Study

| Datasets | Methods | Balanced Accuracy | Cohen's Kappa | Weighted F1 / AUROC |
|---|---|---|---|---|
| BCIC-2A | BENDR | 0.4899±0.0070 | 0.3199±0.0094 | 0.4836±0.0076 |
| | BIOT | 0.4590±0.0196 | 0.2787±0.0261 | 0.4282±0.0289 |
| | LaBraM | 0.5613±0.0052 | 0.4151±0.0069 | 0.5520±0.0052 |
| | Ours | **0.5846±0.0070** | **0.4462±0.0094** | **0.5715±0.0051** |
| BCIC-2B | BENDR | 0.7067±0.0011 | 0.4131±0.0022 | 0.7854±0.0029 |
| | BIOT | 0.6409±0.0118 | 0.2817±0.0236 | 0.7095±0.0141 |
| | LaBraM | 0.6851±0.0063 | 0.3703±0.0125 | 0.7576±0.0067 |
| | Ours | **0.7212±0.0019** | **0.4426±0.0037** | **0.8059±0.0032** |
| Sleep-EDFx | BENDR | 0.6655±0.0043 | 0.6659±0.0043 | 0.7507±0.0029 |
| | BIOT | 0.6622±0.0013 | 0.6461±0.0017 | 0.7415±0.0010 |
| | LaBraM | 0.6771±0.0022 | 0.6710±0.0006 | 0.7592±0.0005 |
| | Ours | **0.6917±0.0069** | **0.6857±0.0019** | **0.7654±0.0023** |
| KaggleERN | BENDR | 0.5672±0.0020 | 0.1461±0.0037 | 0.6030±0.0044 |
| | BIOT | 0.5118±0.0089 | 0.0297±0.0224 | 0.5495±0.0167 |
| | LaBraM | 0.5439±0.0029 | 0.0944±0.0066 | 0.5693±0.0052 |
| | Ours | **0.5837±0.0064** | **0.1882±0.0110** | **0.6621±0.0096** |
| PhysioP300 | BENDR | 0.6114±0.0118 | 0.2227±0.0237 | 0.6588±0.0163 |
| | BIOT | 0.5485±0.0325 | 0.0968±0.0647 | 0.5308±0.0333 |
| | LaBraM | 0.6477±0.0110 | 0.2935±0.0227 | 0.7068±0.0134 |
| | Ours | **0.6502±0.0063** | **0.2999±0.0139** | **0.7168±0.0051** |

**Comparative experiments.**




**Linear-probing method.**

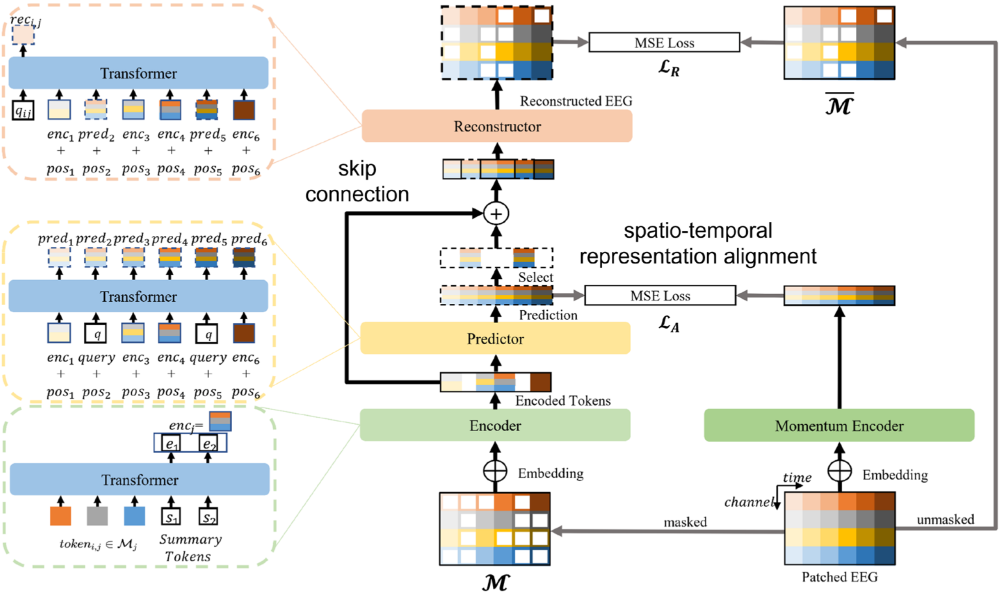
In our comparative study, our EEGPT model outperformed BENDR, BIOT, and LaBraM across various EEG datasets. Notably, EEGPT showed significant accuracy gains on **motor imagery (BCIC-2A: +9.4%, BCIC-2B: +1.5%)** and **sleep stage detection (Sleep-EDFx: +2.6%)** over BENDR. Despite using only an additional linear layer for fine-tuning, our model's enhanced feature extraction capability was evident. It also surpassed BIOT and LaBraM in **ERP-type tasks on KaggleERN (+7.2% over BIOT, +2.6% over BENDR)** and **PhysioP300 (+10.2% over BIOT, +3.9% over BENDR).** Our model's universal feature learning across paradigms addresses key challenges in EEG channel adaptability and representation quality, offering a robust solution for EEG data analysis.

# Experiments

## Ablation Study

| Variants | $\mathcal{L}_A$ | $\mathcal{L}_R$ | BCIC-2A-BAC | BCIC-2B-AUROC | KaggleERN-AUROC |
|---|---|---|---|---|---|
| A: w/o $\mathcal{L}_A$ | 37.13 | 0.57 | 0.5287±0.0086 | 0.7264±0.0381 | 0.5752±0.0164 |
| B: w/o LN | 0.15 | 0.002 | 0.5567±0.0088 | 0.7920±0.0012 | 0.5891±0.0227 |
| C: w/o skip | 0.12 | 0.56 | 0.5796±0.0011 | 0.7702±0.0122 | 0.6356±0.0296 |
| D: with all | 0.24 | 0.56 | **0.5846±0.0070** | **0.8059±0.0032** | **0.6621±0.0096** |

Ablation experiments.




The EEGPT Structure for Pretraining.

In our ablation study (as above table), we found that:

**A**: Removing alignment loss ($\mathcal{L}_A$) led to a **6-9% drop** in downstream task performance, despite similar reconstruction loss.

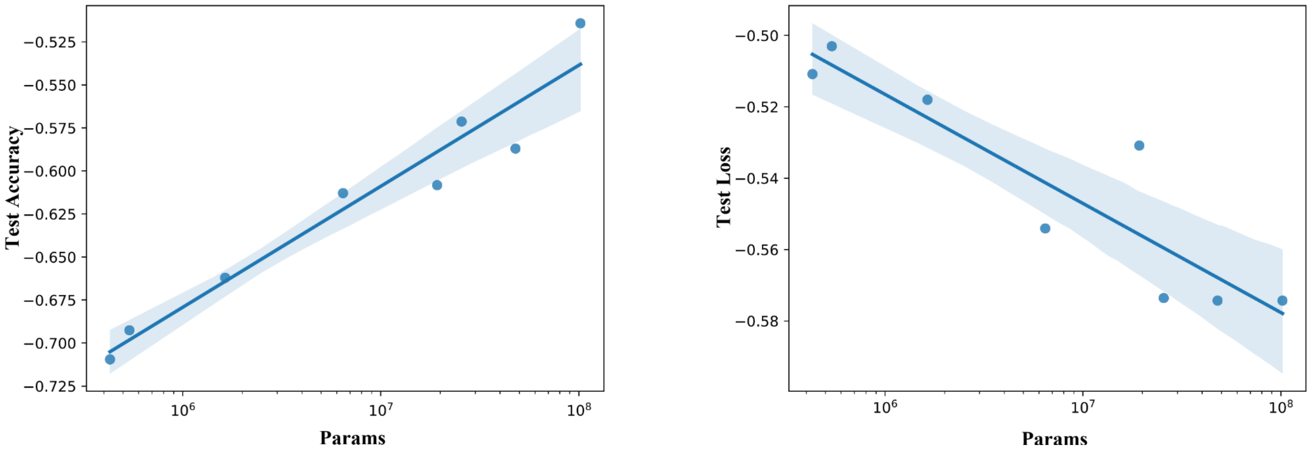
**B**: Removing layer normalization (**LN**) increased vulnerability to extreme values and covariate shift, reducing downstream performance by **3-7%.**

**C**: Without **skip connection**, had lower alignment loss but **1-3% lower** downstream task performance.

These findings underscore the effectiveness of our dual self-supervised approach, enhancing EEG representation quality through spatio-temporal alignment.

# Experiments

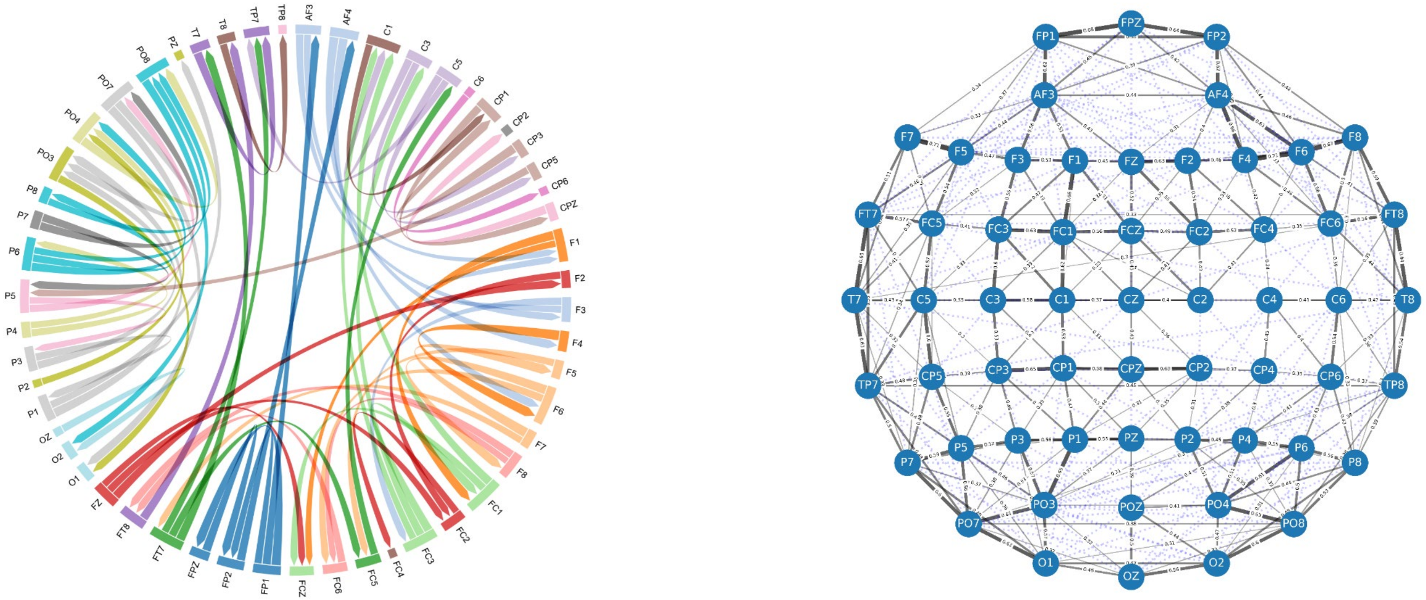
## Scaling Study




**The impact of scaling model size on performance of BCIC-2A.**

We explored the impact of model size and summary tokens on pretraining loss and downstream task performance using 8 model variants. Results from the **BCIC-2A** dataset showed that **as model size and summary tokens increased, reconstruction loss decreased, and task accuracy improved.**

# Visualization




**The illustration of the model's learned channel relationships after pretraining.**

We illustrates the model's learned channel relationships after pretraining.

Left figure depicts **channel similarities (cosine similarity > 0.5)** with clusters indicating positional relationships.

Right figure maps actual electrode positions, showing higher similarity between close channels **(solid lines for >0.3, dashed for 0.1-0.3)** and notable similarity between distant, opposite electrodes.

哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

# Thanks

Please refer to our full paper for detailed methodologies and results.

The code for this paper is available at https://github.com/BINE022/EEGPT

Guangyu Wang

wangguangyu@stu.hit.edu.cn