


# CSSBFM:

## A Brain Foundation Model Based on Cross-Scale Spatiotemporal Structure for Heterogeneous EEG Tasks

**Xiao Chen[1] Lianghua He[1] Haiyang Lu[1]**

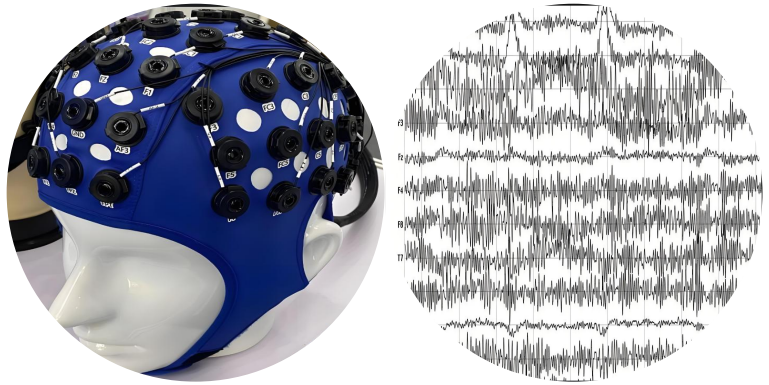
[1] Tongji University

Paper 1663

# Background

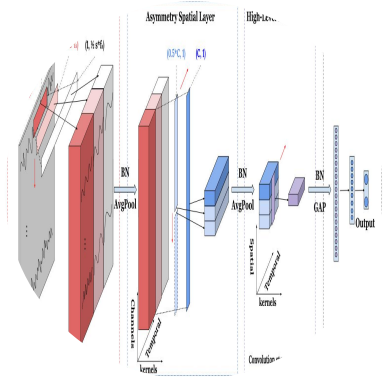
- EEG(electroencephalography) :

  Non-invasive, Easy to acquire for Neuroscience & BCI



- Deep Learning(CNNs/RNNs/GNNs) :

  strong task-specific performance




- Foundation models :

  revolutionize NLP & CV

  unlock robust, transferable EEG representations



# Challenges in EEG Foundation Models

- Fixed-Window Tokenization :

  Destroys local spatiotemporal coupling patterns

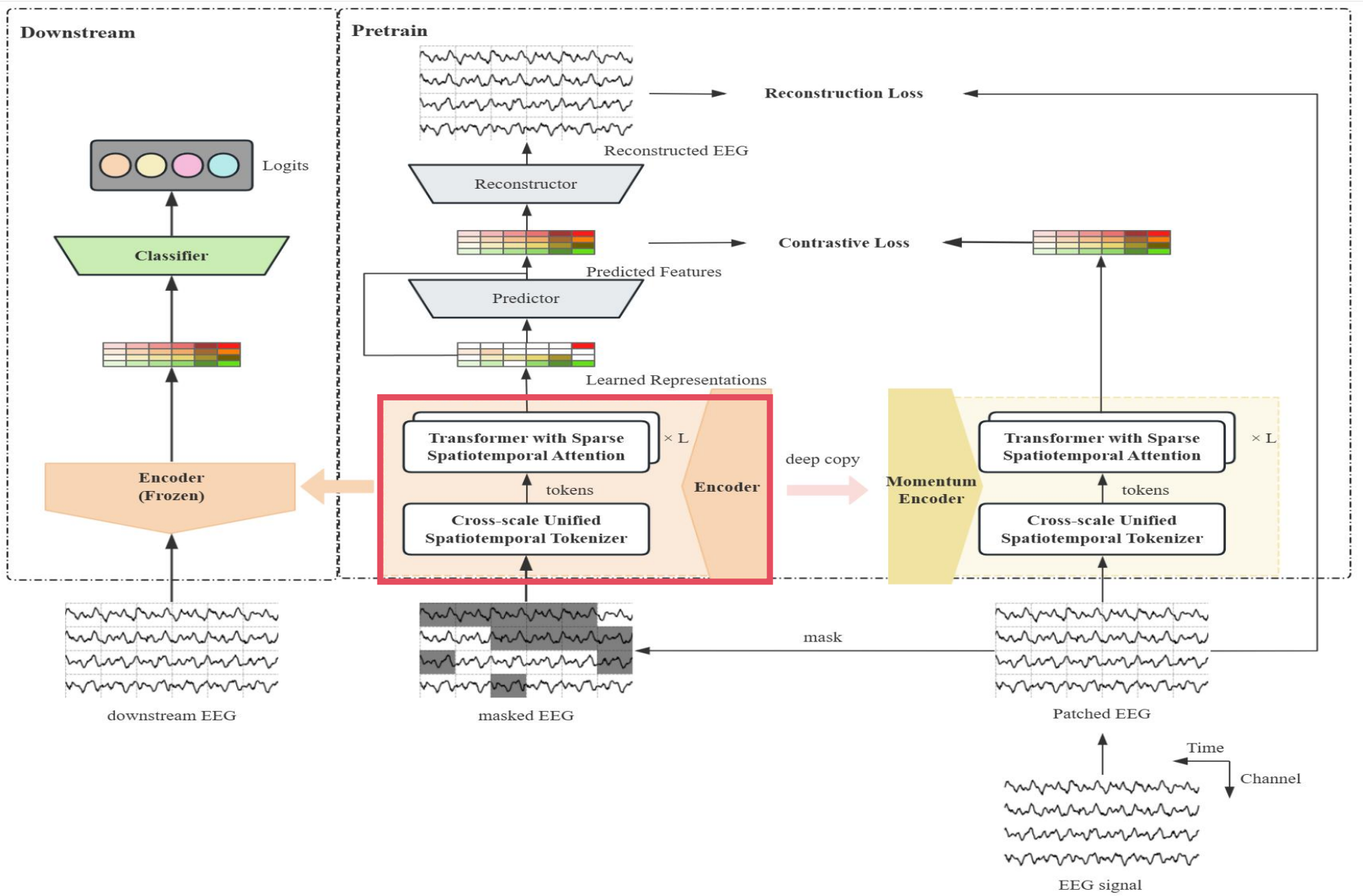
- Dense Full Attention :

  Introduces spurious correlations in noisy EEG data

- Common-Channel Strategy :

  Discards non-shared channels → loss of brain region information

# Overview of our CSSBFM



# Whole-Brain Topology & Functional Partitioning

- 58-Channel Whole-Brain Topology :

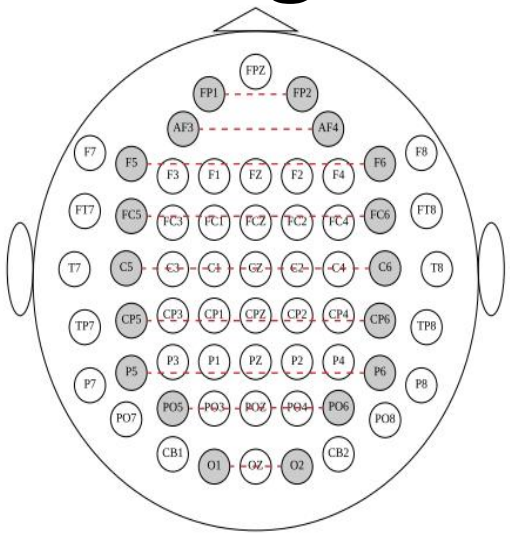
  Based on 10-10 system

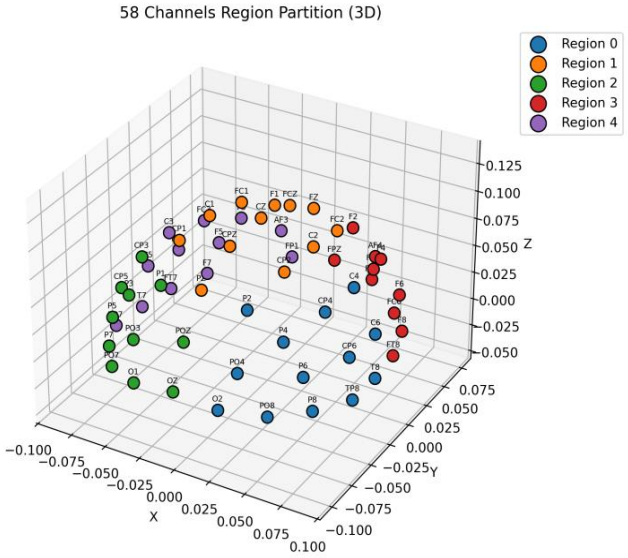
- Neuroscience-Guided Connectivity :

  Local: physical distance prior

  Global: cross-hemisphere functional prior

- Spectral Clustering → 5 Functional Regions



# Cross-scale Unified Spatiotemporal Tokenizer (CUST)

- Unified Spatiotemporal Graph :

  Integrates spatial & temporal structure

$$\widehat{G}_{(\tau)} = \begin{pmatrix} G & I & 0 & \cdots \\ I & G & I & \cdots \\ 0 & I & G & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

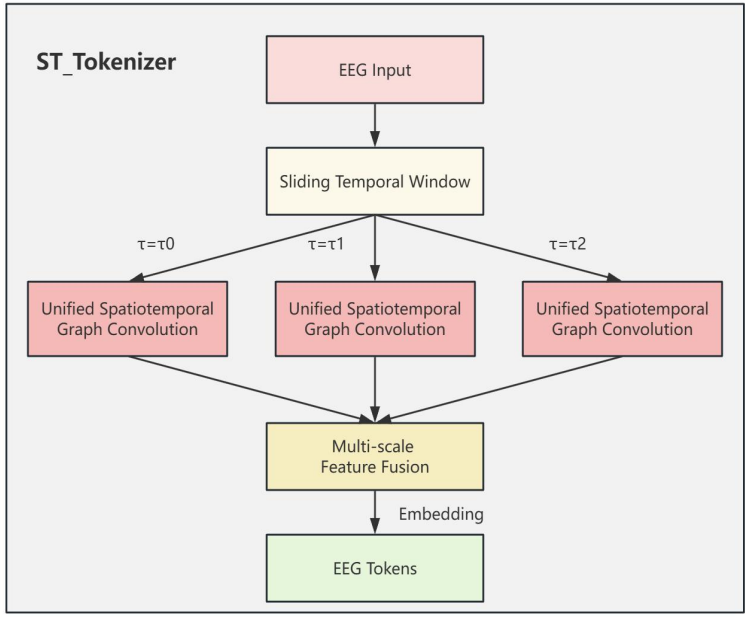
- Decoupled Multi-Hop Convolution :

  Exact k-hop aggregation (K=3)

$$H_\tau = \sigma\left(\sum_{k=1}^{K} \widetilde{D}_{(\tau,k)}^{-\frac{1}{2}} \left(\widehat{G}_{(\tau,k)} + G_{(k)}^{res}\right) \widetilde{D}_{(\tau,k)}^{-\frac{1}{2}} X W_k\right)$$

- Multi-Scale Parallel Tokenization:

  Captures patterns at different scales



# Cross-region Sparse Spatiotemporal Attention (CSSA)

- Spatial: Cross-Region Sparse Attention

  Guided by functional brain partitions

  Each electrode attends only to relevant regions

$$Attention_s(c) = \sum_{c' \in N_s(c)} Softmax\left(\frac{Q_t^c \cdot (K_t^{c'})^\top}{\sqrt{d_k}}\right) V_t^{c'}$$

- Temporal: Index-Aligned Sparse Attention

  Time axis partitioned into T_r regions

  Cross-region alignment by intra-region offset

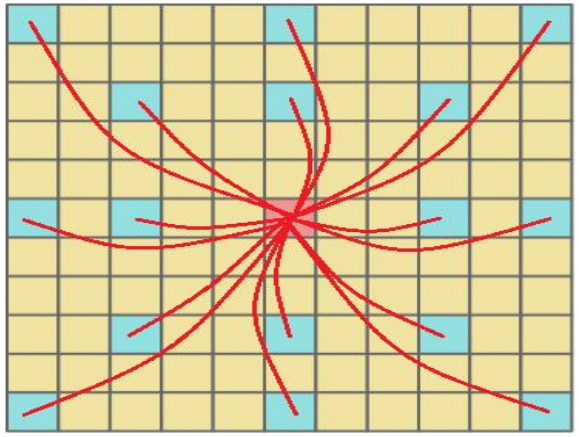
$$Attention_t(t) = \sum_{t' \in N_t(t)} Softmax\left(\frac{Q_c^t \cdot (K_c^{t'})^\top}{\sqrt{d_k}}\right) V_c^{t'}$$

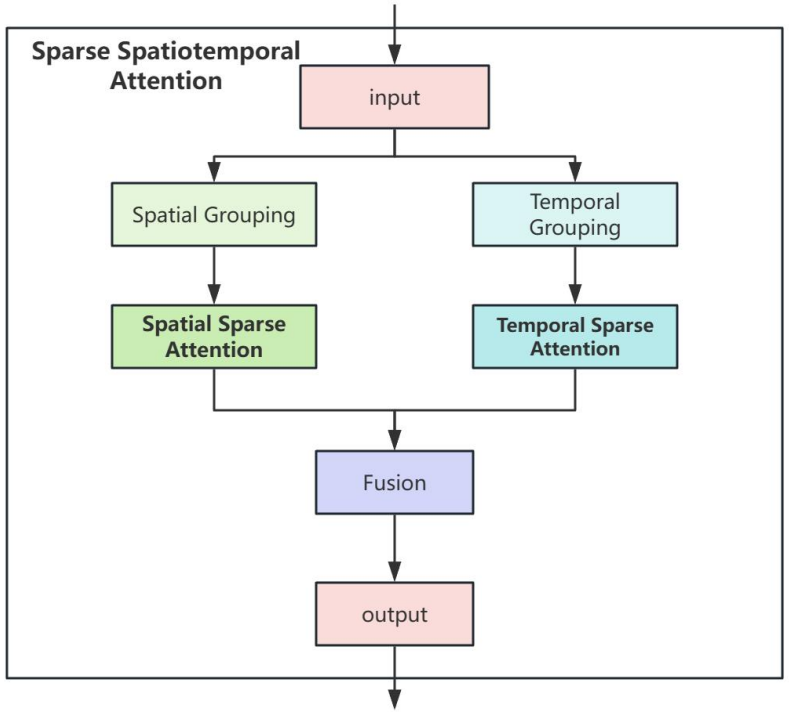
- Parallel Dual-Branch Fusion

  Spatial & temporal branches in parallel

  Learnable weighted aggregation

$$Attention_{out} = \alpha \odot Attention_s + (1-\alpha) \odot Attention_t$$





# Datasets & Experimental Setup

| | Datasets | Paradigms | Subjects | Targets |
|---|---|---|---|---|
| Pretraining | PhysioMI | MI | 109 | 4 |
| | TSU | SSVEP | 35 | 40 |
| | SEED | EMOTION | 15 | 3 |
| | M3CV | MULTI | 106 | - |
| Downstream | HGD | MI | 14 | 4 |
| | PhysioP300 | P300 | 12 | 2 |
| | SEED-IV | EMOTION | 15 | 4 |
| | SEED-V | EMOTION | 16 | 5 |
| | BCIC2020-3 | Imagined Speech | 15 | 5 |

- Baselines:

  BENDR, BIOT, LaBraM, EEGPT

- Evaluation:

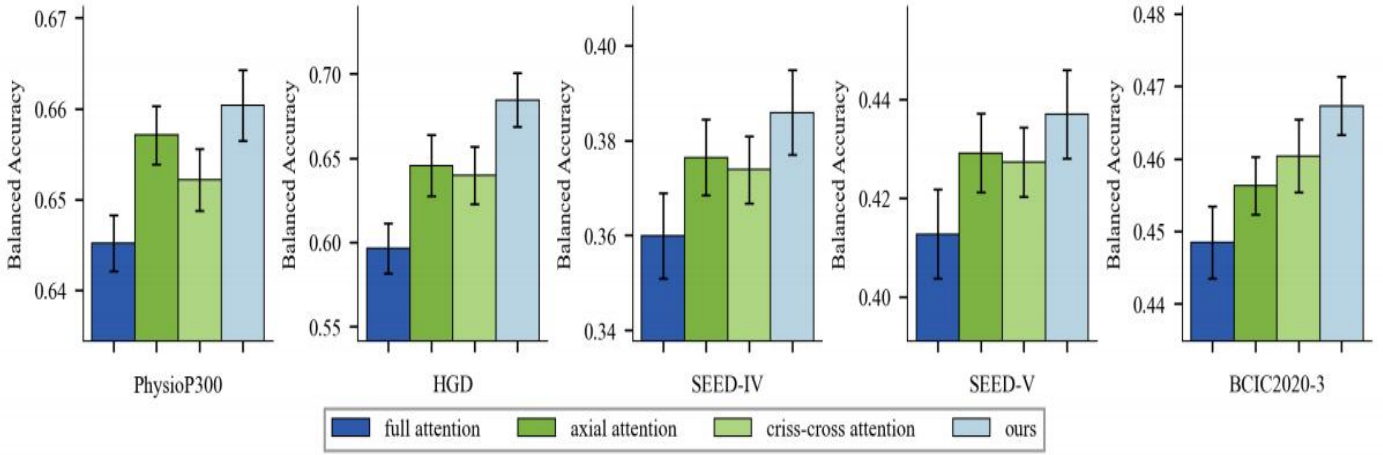
  Subject-independent split

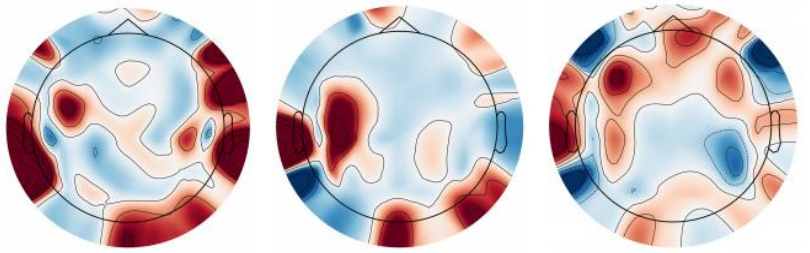
- Metrics:

  BAC, Kappa, Weighted F1/AUROC

# Main Results

| Datasets | Methods | Balanced Accuracy | Cohen's Kappa | Weighted F1/AUROC |
|---|---|---|---|---|
| PhysioP300 | BENDR | 0.6114 | 0.2227 | 0.6588 |
| | BIOT | 0.5485 | 0.0968 | 0.5308 |
| | LaBraM | 0.6477 | 0.2935 | 0.7068 |
| | EEGPT | 0.6502 | 0.2999 | 0.7168 |
| | **Ours** | **0.6604** | **0.3169** | **0.7255** |
| HGD | BENDR | 0.5501 | 0.3524 | 0.5388 |
| | BIOT | 0.5321 | 0.3002 | 0.4876 |
| | LaBraM | 0.6232 | 0.5282 | 0.6151 |
| | EEGPT | 0.5714 | 0.4286 | 0.5539 |
| | **Ours** | **0.6846** | **0.5794** | **0.6754** |
| SEED-IV | BENDR | 0.3185 | 0.1185 | 0.3087 |
| | BIOT | 0.3352 | 0.1146 | 0.3235 |
| | LaBraM | 0.3479 | 0.1274 | 0.3366 |
| | EEGPT | 0.3528 | 0.1325 | 0.3412 |
| | **Ours** | **0.3859** | **0.1783** | **0.3710** |
| SEED-V | BENDR | 0.3169 | 0.2089 | 0.3247 |
| | BIOT | 0.3653 | 0.2135 | 0.3696 |
| | LaBraM | 0.3853 | 0.2524 | 0.3881 |
| | EEGPT | 0.3910 | 0.2234 | 0.3990 |
| | **Ours** | **0.4370** | **0.2872** | **0.4402** |
| BCIC2020-3 | BENDR | 0.4193 | 0.3145 | 0.4062 |
| | BIOT | 0.4435 | 0.3280 | 0.4249 |
| | LaBraM | 0.4527 | 0.3375 | 0.4604 |
| | EEGPT | 0.4592 | 0.3648 | 0.4792 |
| | **Ours** | **0.4673** | **0.3681** | **0.4895** |

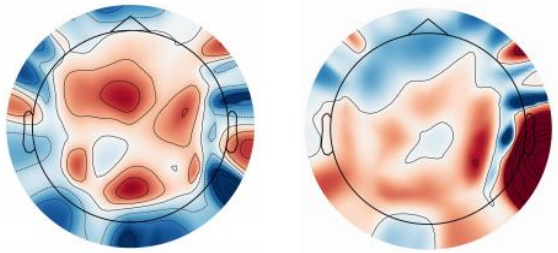


P300 ERP (PhysioNet, Target)

Motor Imagery (HGD, Right Hand)

Emotion Recognition (SEED-IV, Sad)



Emotion Recognition (SEED-V, Fear)

Imagined Speech (BCIC2020-3, 'Thank you')

# Conclusions & Future Work

- Key Takeaway :

  Explicit cross-scale spatiotemporal modeling enhances generalization and robustness of EEG foundation models.

- Future Directions :

  Multi-modal fusion → Clinical diagnosis → Personalized BCI

# Thanks

2026.10

Contact:2432123@tongji.edu.cn