NEUROSTREAM: A VIDEO-INSPIRED CAUSAL SPATIOTEMPORAL ENCODER FOR STREAMING EEG

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper presents a hybrid-architecture foundation model for multichannel EEG that models spatiotemporal features separately. EEG foundation models have developed rapidly in recent years, and most adopt ViT-like paradigms: they treat continuous EEG signals as static patch grids, divide the signal into patches (tokens), flatten them into a sequence, and feed the sequence into a Transformer. This approach requires users to carefully set the temporal window size for each task to obtain the best performance, and the window size cannot be changed while the model is operating. Unlike static images, EEG is structurally more similar to video, with distinct temporal and spatial dimensions. Inspired by work on streaming video understanding, we regard continuous EEG signals as short frames arriving over time and process their spatial and temporal dimensions differently: within each frame, a Transformer models spatial relationships among electrode channels, while a causal Mamba state-space module maintains cross-frame temporal states for each channel. When a new frame arrives, the model only needs to perform incremental computation. It can accommodate EEG signals of different lengths and, through learnable state-transition parameters, form task-related effective context ranges, thereby reducing reliance on preset window lengths. We construct a self-supervised pretraining task on large-scale EEG datasets: next-frame amplitude-spectrum prediction. Given historical EEG frames, the model predicts the amplitude spectrum of each channel in the next frame, frame by frame, enabling it to learn cross-frame frequency-domain changes in continuous EEG signals. We conduct experiments on seizure detection, sleep staging, and emotion recognition. The results show that the proposed pretrained model transfers effectively to different downstream tasks, achieving a **balanced accuracy above 95%** and an **AUROC of 0.99**, particularly on the cross-subject CHB-MIT seizure-monitoring task.

1 INTRODUCTION

Electroencephalography (EEG), a noninvasive electrophysiological monitoring technique with high temporal resolution (Edelman et al., 2025), is widely used in clinical critical-care monitoring, overnight sleep analysis, real-time seizure monitoring, and brain–computer interfaces (BCIs) (Lotte et al., 2018; Carìa, 2025). Deep learning has also improved the interpretation of brain activity from EEG.

EEG datasets vary in channel configurations and acquisition protocols (Jurcak et al., 2007), requiring conventional supervised models to be trained from scratch for each dataset. Meanwhile, the cost of EEG acquisition and annotation limits the size of individual datasets and the number of subjects they cover. Substantial inter-subject variability further constrains cross-subject generalization (Tran et al., 2026). Building on the success of self-supervised learning and large-scale pretraining in natural language processing and computer vision (Devlin et al., 2019; Chen et al., 2021), EEG foundation models have emerged as a promising research paradigm. Representative models, including BIOT (Yang et al., 2023), LaBraM (Jiang et al., 2024), and EEGPT (Wang et al., 2024), are first pretrained on large EEG corpora such as the Temple University Hospital EEG (TUH) corpus (Obeid & Picone, 2016), and then fine-tuned on specific downstream datasets. Such pretraining helps address the limited amount of data available in any single downstream dataset. Following the Vision Transformer (ViT) (Dosovitskiy et al., 2021) paradigm for static images, these models typically ar-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

range EEG as a two-dimensional time–channel patch grid, add temporal and channel embeddings to encode patch positions, and flatten the patches into a token sequence for a Transformer. This representation allows a model to accept inputs of different sizes and supports training on EEG collected with different specifications. EEG foundation models have consequently achieved strong results on a range of clinical and BCI tasks (Yang et al., 2025; Lu et al., 2026).

Despite this progress, ViT-based EEG models still pose challenges for long recordings. Because self-attention cost grows quadratically with sequence length, continuous recordings are commonly divided into shorter windows that are processed independently, preventing information from passing between windows. The window size must be carefully chosen for each task; both overly long and overly short windows can hurt performance, and long windows require more computation and memory and, when assigned a single label, reduce temporal resolution (Ouyang et al., 2022; Didaci et al., 2024). In addition, the learned temporal-embedding table is fixed during fine-tuning, which constrains the window length available at inference time.

Clinical EEG is often acquired as an unbounded, continuous stream: monitoring in neonatal intensive-care units can last tens of hours or even days (Shoeb, 2009), while overnight sleep recordings typically span 8–10 hours (Kemp et al., 2000; Khalighi et al., 2016). EEG can therefore be viewed as a video-like sequence of multichannel sensor snapshots, where electrodes form the spatial dimension and successive short segments form the temporal dimension. For video, TimeSformer introduced Divided Space-Time Attention (Bertasius et al., 2021): it first applies temporal attention to patches at the same spatial location across frames, and then spatial attention among patches within each frame. CBraMod adopts a related spatiotemporal modeling idea for EEG (Wang et al., 2025). However, temporal attention still grows quadratically with the number of frames. Streaming-video approaches such as video-SALMONN S address long inputs by maintaining a bounded memory state that is updated as each new frame arrives, enabling processing under a fixed memory budget (Sun et al., 2026).

Motivated by these observations, we propose NeuroStream, a streaming EEG encoder that separates spatial and temporal modeling. Electrodes distributed over the scalp exhibit dense spatial dependencies associated with brain-region co-activation and changing functional connectivity, including event-related desynchronization (ERD) during motor imagery (He et al., 2025; Sun et al., 2025; Peng et al., 2023; Pfurtscheller & da Silva, 1999). We therefore use spatial attention to model interactions among the channel tokens within each frame. Across frames, EEG rhythms evolve over time and context flows causally; we use a recurrent state-space model (SSM) to update each channel's temporal representation. Its fixed-size hidden state stores historical information for each channel, so processing a new frame does not require rereading all previous frames. For fixed channel and hidden dimensions, this permits incremental inference with computation that scales linearly with the number of frames. Because temporal position is carried by the recurrent state, the model does not require a learned temporal-position table and can process variable-duration inputs. State updates and decay also allow the model to learn which historical information to retain, reducing reliance on a discrete, task-specific window (Hochreiter & Schmidhuber, 1997; Cho et al., 2014; Gu et al., 2022; Gu & Dao, 2024).

To align pretraining with this streaming design, we introduce Causal Next-Frame Amplitude Spectrum Prediction. Given preceding EEG frames from the TUH corpus (Obeid & Picone, 2016), the model updates its state frame by frame and predicts the amplitude spectrum of each channel in the next frame. This causal objective encourages the encoder to learn the cross-frame evolution of EEG spectral content. We evaluate transfer on representative short- and long-duration tasks, including seizure monitoring (Gómez et al., 2020), sleep staging (Aboalayon et al., 2016; Keenan & Magalang, 2026), emotion recognition (Zheng & Lu, 2015), motor-imagery classification (Tarara et al., 2025; Zhang et al., 2021), and P300 detection (Won et al., 2022).

Our contributions are:

- We introduce a streaming EEG foundation model with bounded, frame-wise state updates that supports variable-duration inference and learns task-relevant effective context through state transitions.
- We design a hybrid architecture with within-frame spatial Transformer attention and across-frame causal SSM updates, separately modeling channel interactions and temporal dynamics.

- We propose next-frame amplitude-spectrum prediction, a self-supervised objective aligned with streaming computation.
- We evaluate transfer on long- and short-duration EEG tasks, including training samples lasting minutes to tens of minutes and settings where evaluation sequences exceed the training-segment length.

2 METHOD

NeuroStream treats EEG as a causal stream of multichannel frames, modeling channel interactions within each frame and temporal dynamics across frames. It supports variable-duration inference with computation linear in frame count.

2.1 NOTATION

Let each frame contain S samples from C EEG channels, so $\mathbf{x}_t \in \mathbb{R}^{C\times S}$ and $\mathbf{x}_t^c \in \mathbb{R}^S$ is channel c in frame t. We use d for hidden dimension, N for the number of spatiotemporal blocks, T for frames per segment, and omit batch dimension B. After frame encoding, channel tokens are $\mathbf{u}_t^c \in \mathbb{R}^d$ and $\mathbf{u}_t \in \mathbb{R}^{C\times d}$. Block ℓ maps $\mathbf{u}_t^{(\ell-1)}$ to $\mathbf{u}_t^{(\ell)}$; its Transformer and Mamba outputs are $\boldsymbol{\alpha}_t^{(\ell)}$ and $\boldsymbol{\beta}_t^{(\ell)}$, and its causal state is $\mathbf{h}_t^{(\ell)} \in \mathbb{R}^{C\times d_s}$, with channel state $\mathbf{h}_{t,c}^{(\ell)} \in \mathbb{R}^{d_s}$. Here, $\mathrm{Conv1d}$ denotes one-dimensional convolution, MHA multi-head attention, and FFN a feed-forward network. $\mathrm{MambaStep}$ denotes a single causal Mamba state update, FFT the fast Fourier transform, and $\mathcal{L}_{\mathrm{SmoothL1}}$ the Smooth L1 loss.

2.2 PROBLEM FORMULATION: STREAMING CAUSAL ENCODING

Streaming causal encoding learns a mapping f_θ that produces the current-frame representation $\mathbf{y}_t$ and updated state $\mathbf{h}_t$ from $\mathbf{x}_t$ and the preceding state $\mathbf{h}_{t-1}$:

$$(\mathbf{y}_t, \mathbf{h}_t) = f_\theta(\mathbf{x}_t, \mathbf{h}_{t-1}). \tag{1}$$

The encoder must maintain bounded state independent of t, use only current and past frames ($\mathbf{x}_{\le t}$) to ensure causality, and scale as $O(C^2T)$ rather than quadratically with frame count. Bounded causal state enables continuous, arbitrary-duration EEG encoding when future signals are unavailable.

2.3 PATCH TOKENIZATION

As shown in Figure 1, a shared one-dimensional convolution maps each channel patch $\mathbf{x}_t^c$ to a d-dimensional token:

$$\mathbf{u}_{t,c}^{(0)} = \mathrm{Conv1d}(\mathbf{x}_t^c) \in \mathbb{R}^d. \tag{2}$$

We add a learnable electrode-position embedding $\mathbf{e}_{\mathrm{channel}} \in \mathbb{R}^{C\times d}$ to initialize the channel tokens:

$$\mathbf{u}_t^{(0)} = [\mathbf{u}_{t,1}^{(0)}, \ldots, \mathbf{u}_{t,C}^{(0)}] + \mathbf{e}_{\mathrm{channel}} = \mathbf{u}_t + \mathbf{e}_{\mathrm{channel}}. \tag{3}$$

This embedding identifies electrode positions and is adapted to each downstream channel configuration during fine-tuning.

2.4 SPATIO-TEMPORAL BLOCK

Each spatiotemporal block serially combines within-frame **Spatial Attention** with an across-frame, channel-wise **Temporal SSM** (Figure 1).

Branch 1: Within-frame Spatial Transformer. For input $\mathbf{u}_t^{(\ell-1)} \in \mathbb{R}^{C\times d}$, we reshape to $[B\cdot T, C, d]$, normalize each frame's channel tokens with RMSNorm (Zhang & Sennrich, 2019), and project them to queries, keys, and values for multi-head attention:

$$\boldsymbol{\alpha}_t^{(\ell)} = \mathrm{MHA}(Q, K, V) \in \mathbb{R}^{C\times d}. \tag{4}$$

$$\boldsymbol{\alpha}_t^{(\ell)} = \mathrm{FFN}(\boldsymbol{\alpha}_t^{(\ell)}) \in \mathbb{R}^{C\times d}. \tag{5}$$

Attention over the C tokens captures channel dependencies and co-activation within each frame, with complexity $O(C^2)$ independent of sequence length.

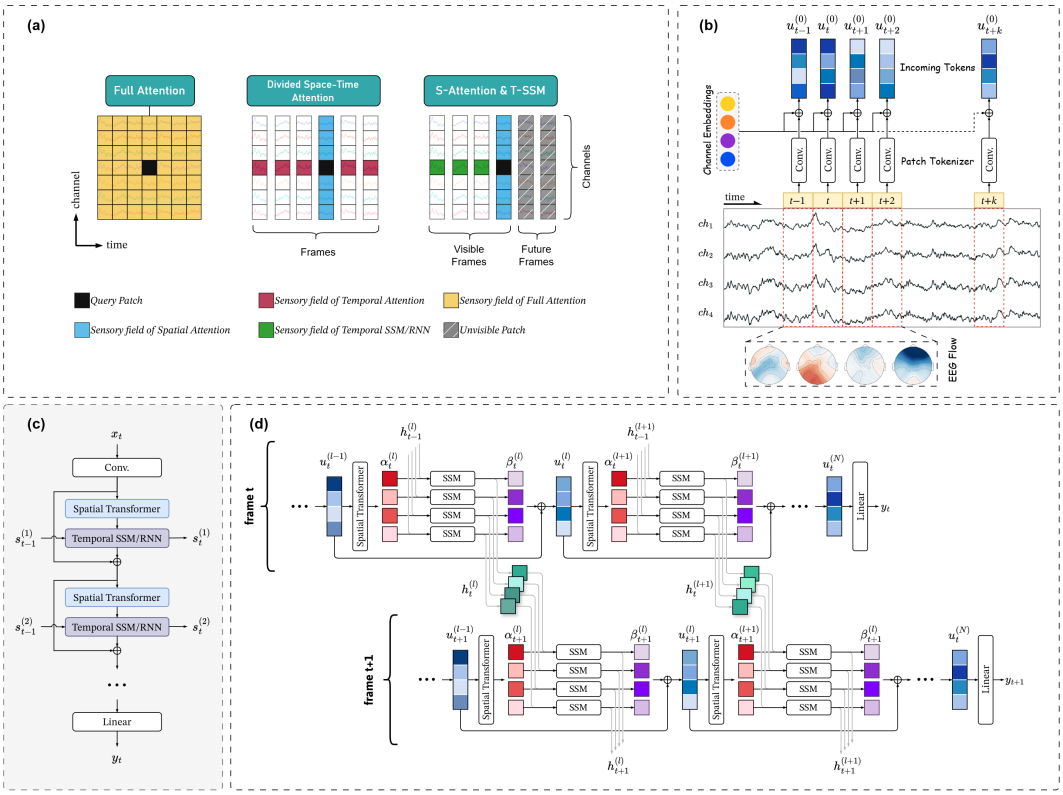


Figure 1: Architecture of NeuroStream and its causal streaming data path. (a) Comparison of Full Attention, Divided Space-Time Attention, and the Spatial-Attention + Temporal-SSM design proposed here; (b) EEG-frame arrival, channel embedding, and frame-wise state updates; (c) stacking of spatiotemporal blocks; (d) within-frame cross-channel Transformer and across-frame channel-wise SSM.

Branch 2: Across-frame Channel-wise Temporal SSM. To model temporal evolution, we reshape the Transformer output to $[B \cdot C, T, d]$ and apply a causal Mamba recurrence independently to each channel:

$$(\boldsymbol{\beta}_{t,c}^{(\ell)}, \mathbf{h}_{t,c}^{(\ell)}) = \mathrm{MambaStep}_c(\boldsymbol{\alpha}_{t,c}^{(\ell)}, \mathbf{h}_{t-1,c}^{(\ell)}), \quad c = 1, \ldots, C. \tag{6}$$

Each update uses only the current input and preceding state, ensuring causality; states across layers form the model's recurrent state.

The spatial Transformer and temporal SSM form a residual block:

$$\mathbf{u}_t^{(\ell)} = \mathbf{u}_t^{(\ell-1)} + \boldsymbol{\beta}_t^{(\ell)}. \tag{7}$$

The block output $\mathbf{u}_t^{(\ell)}$ feeds the next block. Across N blocks, state has size $N \times C \times d_s$; single-step inference uses $O(C)$ state, independent of processed-frame count.

2.5 Pretraining: Next-Frame Amplitude-Spectrum Prediction

We train on unlabeled EEG with GPT-style next-frame amplitude-spectrum prediction (Radford et al., 2018) (Figure 2). Unlike encoder-output targets, which may risk representation collapse as in Neuro-GPT (Cui et al., 2024), our target is the EEG signal's own frequency structure and teaches causal prediction of spectral evolution.

Data augmentation. We randomly mask approximately 30% of channel–frame tokens with a learnable mask vector $[\mathrm{M}]$. Masks vary across epochs and affect inputs only, not targets, encouraging prediction from incomplete channel information and improving robustness.

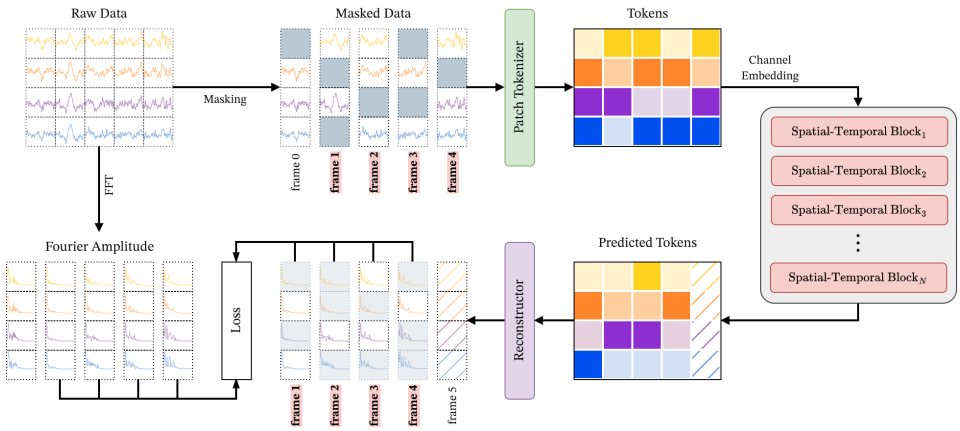


Figure 2: Next-frame FFT amplitude-spectrum pretraining pipeline: input frames are tokenized into channel–frame tokens and randomly masked before being sent to the causal encoder; the model predicts the amplitude spectrum of each channel in the next frame, frame by frame, and the loss is calculated after alignment with the target shifted by one time step.

Pretraining objective. Following frequency-domain pretraining in BrainBERT (Wang et al., 2023), EEGFormer (Chen et al., 2024), LBLM (Zhou et al., 2025), and LaBraM (Jiang et al., 2024), we predict the next-frame amplitude spectrum. Given $\{\mathbf{x}_1, \ldots, \mathbf{x}_t\}$, the target for channel c is:

$$\mathbf{A}_{t+1}^{c} = \mathrm{FFT}(\mathbf{x}_{t+1}^{c}) \in \mathbb{R}^{S/2+1}. \tag{8}$$

The target is the one-sided FFT amplitude spectrum with $S/2 + 1$ bins. A token-wise linear head maps each encoder output $\mathbf{u}_t^{(N),c}$ to the corresponding next-frame spectrum $\widehat{\mathbf{A}}_{t+1}^{c}$.

We form targets from an unmasked copy of each frame sequence and feed a masked copy to the model. Because output at frame t predicts the target at $t + 1$, we pair outputs 1 through $T - 1$ with targets 2 through T and average the loss over pairs and channels:

$$\mathcal{L}_{\text{pretrain}} = \frac{1}{(T-1)C} \sum_{t=1}^{T-1} \sum_{c=1}^{C} \mathcal{L}_{\text{SmoothL1}} \left(\widehat{\mathbf{A}}_{t}^{c}, \mathbf{A}_{t+1}^{c} \right). \tag{9}$$

2.6 Downstream Task Adaptation

For downstream tasks, we replace the pretraining head with a three-layer MLP and use the flattened final-frame encoder output $\mathbf{u}_t^{(N)}$ at the task's temporal granularity to produce logits $\mathbf{y}_t$.

3 Experiments

3.1 Experimental Setup

Datasets. Pretraining uses TUSZ, TUAR, TUEP, and TUAB from the Temple University Hospital EEG Corpus (Obeid & Picone, 2016), with the 19 channels common to all four. Downstream evaluation covers CHB-MIT (seizure monitoring) (Shoeb, 2009; Pollard et al., 2026; Guttag, 2010), Sleep-EDFx Sleep Telemetry (sleep staging) (Kemp et al., 2000; Pollard et al., 2026), SEED (emotion recognition) (Zheng & Lu, 2015; Duan et al., 2013), BNCI2014004 (motor imagery) (Tangermann et al., 2012; Aristimunha et al., 2026a), and BNCI2014009 (P300 detection) (Aristimunha et al., 2026a; Riccio et al., 2013). All signals are resampled to 200 Hz and converted to μV; dataset descriptions are provided in Appendix A, and the full list of pretraining channels is in Appendix G.

Model details and pretraining setup. We use hidden dimension $d = 200$, $N = 6$ blocks (10 attention heads and FFN dimension 800 per block), and SSM state dimension $d_s = 64$. Inputs are divided into non-overlapping 1-s frames ($S = 200$ at 200 Hz). Pretraining uses 30-s segments

Table 1: Overview of the pretraining and downstream evaluation datasets. The first four TUH subsets are used for unsupervised pretraining without their task labels. "Input length" denotes the duration of the segment contained in a single training or evaluation sample.

Paradigms	Dataset	Input length	Channels	Task / labels
Seizure	TUSZ	30 s	19	Pretraining
Artifact	TUAR	30 s	19	Pretraining
Events	TUEP	30 s	19	Pretraining
Abnormal	TUAB	30 s	19	Pretraining
Seizure	CHB-MIT	300/600 s	16	Seizure / non-seizure
Sleep	Sleep-EDFx	0.5/1 h	2	W, N1, N2, N3, REM
Emotion	SEED	10 s	62	Positive, neutral, negative
Motor imagery	BNCI2014004	4 s	3	Left / right hand
P300	BNCI2014009	1 s	16	Target / non-target

($T = 30$), batch size 1024, initial learning rate 10^{-4} with a OneCycle scheduler (Smith & Topin, 2018), and 40 epochs. The pretrained model has approximately 5.0M parameters; further pretraining details are provided in Appendix B.

Baselines. We compare supervised EEGNet (Lawhern et al., 2018), ShallowNet (Schirrmeister et al., 2017), EEGConformer (Song et al., 2023), and SPaRCNet (Jing et al., 2023) with pretrained BIOT (Yang et al., 2023), LaBraM (Jiang et al., 2024), LUNA (Döner et al., 2025), and CBraMod (Wang et al., 2025). We use public Braindecode implementations (Schirrmeister et al., 2017; Aristimunha et al., 2026b), evaluate them under our common settings, and initialize pretrained models from released weights. LSTM (Hochreiter & Schmidhuber, 1997; Cho et al., 2014) and Mamba-3 (Lahoti et al., 2026) are additional streaming-inference baselines.

Metrics. We report balanced accuracy (BalAcc), AUROC, and average precision (AP) for binary tasks, and Cohen's kappa and weighted F1 for multiclass tasks. Random baselines are 0.5 for AUROC and the positive-class prevalence for AP. Higher is better; results are mean $\pm$ standard deviation over three runs.

3.2 Results

Long-duration tasks. To test long-sequence capability, we evaluate CHB-MIT and Sleep-EDFx with inputs substantially longer than conventional settings. NeuroStream updates its state continuously and can retain or decay historical information independently of total sequence length.

Table 2 reports results for continuous 300-s training samples (300 one-second labels), using subjects 1–19 for training, 20–21 for validation, and 22–23 for testing (Yang et al., 2023). Positive samples comprise 3.2‰ of the test set. Streaming models (LSTM, Mamba-3, and NeuroStream) produce one output per second and are tested at both 300 and 600 s; other baselines use 1-s windows because full-length inputs exceed available resources. NeuroStream reaches BalAcc/AUROC of 0.9552/0.9888 at 300 s and 0.9589/0.9920 at 600 s, outperforming all baselines. With 1-s samples, its BalAcc is 0.878 versus 0.821 for the best baseline (Appendix D, Table 11). LSTM remains near chance at both longer sequence lengths (Table 2).

Table 3 reports results on 1800-s segments (60 labels at 30-s resolution), with subjects 01–02 and 04–17 for training, 18–20 for validation, and 21, 22, and 24 for testing. Streaming models ingest each full sample; we also test 3600-s inputs, while other baselines use 30-s windows. NeuroStream achieves BalAcc of 0.7501/0.7592 at 1800/3600 s, above CBraMod (0.7257), with other metric point estimates stable across lengths. With 30-s samples, its BalAcc is 0.74 versus 0.75 for the longer-sample setting (Appendix D, Table 12).

Short-duration tasks. We use SEED, BNCI2014004, and BNCI2014009 as representative short-duration tasks. Unlike clinical EEG collected for long-term monitoring, these datasets primarily represent BCI scenarios, in which events are brief and their timing and order are controlled. On SEED emotion recognition (Table 4), we use five-fold cross-validation with subject-wise partitions. NeuroStream improves over the strongest baseline, CBraMod, by 1.7, 2.0, and 2.6 percentage points

Table 2: CHB-MIT seizure-detection results. NeuroStream, LSTM, and Mamba-3 are evaluated using 300-s and 600-s test settings; other baselines are evaluated on 1-s subwindows.

Model	Param.	BalAcc	AUROC	AP
EEGNet	1.5k	0.6570±0.0616	0.7443±0.0240	0.0146±0.0015
ShallowNet	0.27M	0.7561±0.0234	0.9269±0.0114	0.1711±0.0958
EEGConformer	0.23M	0.8033±0.0772	0.9115±0.0291	0.0907±0.0536
SPaRCNet	0.69M	0.8386±0.0067	0.9155±0.0055	0.0317±0.0057
LSTM	0.87M			
– 300 s		0.5060±0.0182	0.5129±0.0124	0.0032±0.0003
– 600 s		0.5071±0.0173	0.5196±0.0098	0.0032±0.0003
Mamba-3	0.66M			
– 300 s		0.5351±0.0706	0.6230±0.0661	0.1780±0.0708
– 600 s		0.5344±0.0726	0.6225±0.0650	0.1785±0.0704
BIOT	3.2M	0.7524±0.0115	0.8429±0.0142	0.2438±0.0090
LaBraM	5.8M	0.7557±0.0298	0.8896±0.0269	0.1489±0.0141
CBraMod	4.9M	0.7874±0.0140	0.9113±0.0090	0.0647±0.0165
Proposed	7.7M			
– 300 s		0.9552±0.0184	0.9888±0.0032	0.4099±0.0537
– 600 s		**0.9589**±0.0147	**0.9920**±0.0025	**0.4728**±0.0737

Table 3: Sleep-EDFx sleep-staging results (mean ± standard deviation over three repeated runs). NeuroStream, LSTM, and Mamba-3 are evaluated using 1800-s and 3600-s test settings; other baselines are evaluated on 30-s subwindows.

Model	Param.	BalAcc	Weighted F1	Cohen's Kappa
EEGNet	16k	0.3753±0.0770	0.2403±0.0125	0.1148±0.0451
ShallowNet	83k	0.6157±0.0201	0.5396±0.0595	0.3780±0.0578
EEGConformer	4.2M	0.5952±0.0406	0.6371±0.0416	0.4931±0.0700
SPaRCNet	1.4M	0.5153±0.0439	0.2322±0.1021	0.1987±0.0606
LSTM	0.15M			
– 1800 s		0.4372±0.0258	0.1939±0.0315	0.1462±0.0127
– 3600 s		0.4378±0.0256	0.1926±0.0303	0.1460±0.0117
Mamba-3	0.30M			
– 1800 s		0.6210±0.0160	0.5845±0.0093	0.4205±0.0142
– 3600 s		0.6249±0.0157	0.5921±0.0099	0.4270±0.0149
BIOT	3.2M	0.7171±0.0070	**0.7746**±0.0059	**0.6602**±0.0069
LaBraM	5.8M	0.7150±0.0059	0.7272±0.0143	0.5979±0.0164
CBraMod	5.0M	0.7257±0.0092	0.7543±0.0208	0.6322±0.0283
Proposed	5.5M			
– 1800 s		0.7501±0.0136	0.6306±0.0324	0.5112±0.0317
– 3600 s		**0.7592**±0.0058	0.6459±0.0214	0.5247±0.0184

in balanced accuracy (BalAcc), weighted F1, and Cohen's kappa, respectively. For motor-imagery classification on BNCI2014004 and event-related potential detection on BNCI2014009 (Table 5), we use subject-level three-fold and five-fold cross-validation, respectively. Overall, supervised baselines perform better than pretrained EEG foundation-model baselines on these datasets. NeuroStream trails the best supervised model, ShallowNet, by 2.85 percentage points in BalAcc on BNCI2014004, and SPaRCNet by 0.86 points on BNCI2014009; it nevertheless remains ahead of the other EEG foundation-model baselines on both tasks.

Overall, these results show that NeuroStream performs well with both short-duration inputs and the much longer sequences used in the clinical monitoring tasks above.

3.3 Ablation Studies

Spatial and temporal components. On Sleep-EDFx and CHB-MIT, we compare the full model with (i) removing spatial attention, i.e., $\boldsymbol{\alpha}_t^{(\ell)} = \mathbf{u}_t^{(\ell-1)}$, and (ii) replacing Mamba-3 with an LSTM

Table 4: The results of different methods on SEED.

Model	Params.	BalAcc	Weighted F1	Cohen's Kappa
EEGNet	5.1k	0.4120±0.0015	0.3950±0.0076	0.1180±0.0023
ShallowNet	0.12M	0.5590±0.0046	0.5453±0.0055	0.3385±0.0070
EEGConformer	1.5M	0.5525±0.0112	0.5391±0.0070	0.3287±0.0168
SPaRCNet	1.2M	0.5637±0.0122	0.5595±0.0129	0.3456±0.0184
BIOT	3.2M	0.4856±0.0036	0.4704±0.0056	0.2284±0.0055
LUNA	7.1M	0.5553±0.0088	0.5443±0.0109	0.3329±0.0132
CBraMod	5.3M	0.5733±0.0060	0.5663±0.0067	0.3600±0.0090
Proposed	15M	**0.5907**±0.0002	**0.5861**±0.0015	**0.3860**±0.0003

Table 5: The results of different methods on BNCI2014004/BNCI2014009.

	BNCI2014004			BNCI2014009		
Model	BalAcc	AUROC	AP	BalAcc	AUROC	AP
EEGNet	0.5620±0.0061	0.6189±0.0237	0.6216±0.0255	0.8160±0.0038	0.8935±0.0016	0.6827±0.0069
ShallowNet	**0.6627**±0.0009	**0.7277**±0.0014	**0.7327**±0.0032	0.7628±0.0004	0.8442±0.0013	0.6042±0.0065
EEGConformer	0.6485±0.0080	0.7126±0.0063	0.7139±0.0091	0.7886±0.0024	0.8600±0.0021	0.6064±0.0083
SPaRCNet	0.5910±0.0103	0.6401±0.0161	0.6326±0.0207	**0.8250**±0.0023	**0.8995**±0.0030	**0.7057**±0.0042
BIOT	0.5150±0.0024	0.5041±0.0088	0.5120±0.0188	0.5403±0.0030	0.5528±0.0032	0.2058±0.0045
LaBraM	0.5215±0.0021	0.5222±0.0039	0.5185±0.0083	0.7327±0.0049	0.8035±0.0023	0.5175±0.0056
CBraMod	0.6112±0.0091	0.6534±0.0160	0.6478±0.0140	0.7960±0.0023	0.8703±0.0016	0.6236±0.0015
Proposed	0.6342±0.0038	0.6932±0.0012	0.6872±0.0020	0.8164±0.0032	0.8921±0.0045	0.6765±0.0068

(Figure 3). The full model performs best in every experiment. Both variants reduce CHB-MIT BalAcc by approximately 12 percentage points, to near SPaRCNet's level, supporting the contributions of spatial attention and Mamba-3 temporal modeling.

Pretraining and task-head size. For each frame $\mathbf{x}_t$, the convolutional tokenizer and N spatiotemporal encoder blocks produce $\mathbf{u}_t^{(N)} \in \mathbb{R}^{C\times d}$. We flatten this tensor into a Cd-dimensional vector and feed it to the task-specific three-layer MLP classifier. Thus, the first linear layer has Cd input units and, for a fixed hidden width, its parameter count grows linearly with the number of channels C. On datasets with many EEG channels, the classification head can substantially increase the total fine-tuned model size, potentially making it several times larger than the pretrained encoder. Table 6 compares MLP hidden sizes 200/400/800 on Sleep-EDFx, CHB-MIT, and BNCI2014009. Hidden size has little effect, whereas pretrained initialization consistently improves performance over random initialization. Thus, a smaller task head can reduce computation while retaining performance when pretrained weights are used.

Table 6: Ablation results for different MLP hidden dimensions, with and without pretrained weights. Cell colors encode column-wise mean performance (darker indicates higher values): red, green, and pink correspond to Sleep-EDFx BalAcc, CHB-MIT BalAcc, and BNCI2014009 AP, respectively.

Pretraining	MLP hidden size	Sleep-EDFx BalAcc	CHB-MIT BalAcc	BNCI2014009 AP
Yes	800	**0.7592**±0.0058	**0.9589**±0.0147	**0.6872**±0.0020
	400	0.7359±0.0032	0.9509±0.0038	0.6861±0.0055
	200	0.7483±0.0059	0.9528±0.0260	0.6775±0.0009
No	800	0.7028±0.0202	0.7481±0.0075	0.5517±0.0024
	400	0.7262±0.0415	0.7519±0.0546	0.5634±0.0254
	200	0.7150±0.0290	0.7448±0.0909	0.5809±0.0213

4 Conclusion

We view the continuous EEG stream as a sequence of video-like frames and draw on spatiotemporal modeling in video understanding to propose NeuroStream, an EEG foundation model with a hybrid

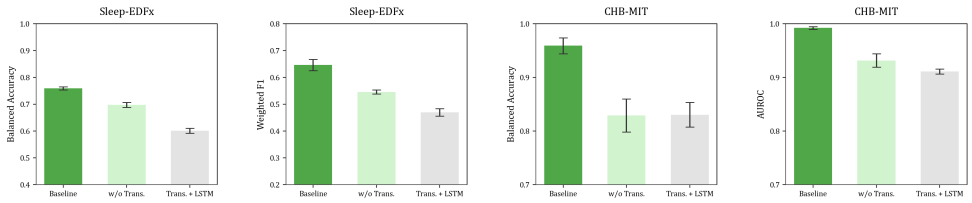


Figure 3: Ablation results for the spatiotemporal module. We compare the full model, removal of the Spatial Transformer (w/o Trans.), and replacement of Temporal Mamba with an LSTM (Trans. + LSTM).

Spatial Attention and Temporal SSM architecture. Spatial Attention models spatial relationships among EEG channels, while Temporal SSM maintains a fixed-size history for each channel to model long-range temporal dependencies across frames. This spatiotemporal decomposition enables the model to process continuous inputs frame by frame and handle EEG signals of varying lengths. By controlling updates to historical states, the model learns task-relevant effective context and reduces reliance on preset input-window sizes. To match NeuroStream's streaming input/output properties, we also design a self-supervised pretraining method: next-frame amplitude-spectrum prediction. We evaluate the model on multiple long- and short-duration downstream datasets. The results show strong performance on EEG analysis tasks with different durations and stable behavior in sequence-length extension experiments. Ablations further validate the effectiveness of the proposed within-frame Spatial Attention, cross-frame Temporal SSM, and pretraining strategy. In future work, we will explore this framework for long-duration time-series modeling with other multichannel sensors and multimodal data.

Ethics Statement

This work is a secondary analysis of previously collected EEG datasets and involved no new participant recruitment, direct interaction, or data collection. All datasets were obtained lawfully through public channels. For datasets requiring an application or permission, including the TUH and SEED datasets, we obtained authorization from the respective providers before access. We complied with the applicable access and use requirements and did not attempt to re-identify participants.

Reproducibility Statement

The main text and appendices describe the model architecture, datasets, preprocessing, data splits, training settings, and evaluation metrics. An anonymized source-code package, including the implementation and the scripts and configuration needed to reproduce the reported experiments, will be submitted as supplementary material.

AI Use Statement

Generative AI tools (Codex) were used to translate the Chinese draft into English; edit the manuscript for clarity and consistency; check terminology, references, and cross-references; and provide feedback on the framing of the long-sequence experiments and the presentation and interpretation of results. These tools were not used to collect or generate EEG data, train or evaluate models, or calculate the reported metrics. The authors reviewed AI-assisted material and are responsible for verifying all claims, citations, and numerical results against the source documents and experiment records, and for the final content of the paper.

REFERENCES

K. A. I. Aboalayon, M. Faezipour, W. S. Almuhammadi, and S. Moslehpour. Sleep Stage Classification Using EEG Signal Analysis: A Comprehensive Survey and New Investigation, 2016. Entropy, vol. 18, no. 9, pp. 272. DOI: 10.3390/e18090272.

B. Aristimunha et al. Mother of All BCI Benchmarks, 2026a. Mother of all BCI Benchmarks. (). Python. DOI: 10.5281/zenodo.10034223.

B. Aristimunha et al. Braindecode: toolbox for decoding raw electrophysiological brain data with deep learning models, 2026b. (). Python. Zenodo. DOI: 10.5281/ZENODO.17699192.

G. Bertasius, H. Wang, and L. Torresani. Is Space-Time Attention All You Need for Video Understanding?, 2021. arXiv: arXiv:2102.05095. DOI: 10.48550/arXiv.2102.05095.

A. Carìa. Integrating Large Language Models and Brain Decoding for Augmented Human-Computer Interaction: A Prototype LLM-P3-BCI Speller, 2025. In: Advances in Information and Communication, K. Arai, (eds.), Cham: Springer Nature Switzerland, 2025, pp. 403–416. DOI: 10.1007/978-3-031-84457-7_25.

X. Chen, S. Xie, and K. He. An Empirical Study of Training Self-Supervised Vision Transformers, 2021. arXiv: arXiv:2104.02057. DOI: 10.48550/arXiv.2104.02057.

Y. Chen et al. EEGFormer: Towards Transferable and Interpretable Large-Scale EEG Foundation Model, 2024. arXiv: arXiv:2401.10278. DOI: 10.48550/arXiv.2401.10278.

S. Chevallier et al. The largest EEG-based BCI reproducibility study for open science: the MOABB benchmark, 2024. URL `https://universite-paris-saclay.hal.science/hal-04537061`.

K. Cho et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation, 2014. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), A. Moschitti, B. Pang and W. Daelemans, (eds.), Doha, Qatar: Association for Computational Linguistics, , pp. 1724–1734. DOI: 10.3115/v1/D14-1179.

J. Connolly, H. Edwards, B. Bourgeois, S. T. Treves, A. Shoeb, and J. Guttag. CHB-MIT, 2026. NEMAR. DOI: 10.82901/nemar.nm000110.

W. Cui et al. Neuro-GPT: Towards A Foundation Model For EEG, 2024. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), , pp. 1–5. DOI: 10.1109/ISBI56570.2024.10635453.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), J. Burstein, C. Doran and T. Solorio, (eds.), Minneapolis, Minnesota: Association for Computational Linguistics, , pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

L. Didaci, S. M. Pani, C. Frongia, and M. Fraschini. How Time Window Influences Biometrics Performance: An EEG-Based Fingerprint Connectivity Study, 2024. Signals, vol. 5, no. 3, pp. 597–604. DOI: 10.3390/signals5030033.

B. Döner, T. M. Ingolfsson, L. Benini, and Y. Li. LUNA: Efficient and Topology-Agnostic Foundation Model for EEG Signal Analysis, 2025. arXiv: arXiv:2510.22257. DOI: 10.48550/arXiv.2510.22257.

A. Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021. arXiv: arXiv:2010.11929. DOI: 10.48550/arXiv.2010.11929.

R.-N. Duan, J.-Y. Zhu, and B.-L. Lu. Differential entropy feature for EEG-based emotion classification, 2013. In: 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER), , pp. 81–84. DOI: 10.1109/NER.2013.6695876.

B. J. Edelman et al. Non-Invasive Brain-Computer Interfaces: State of the Art and Trends, 2025. IEEE Rev. Biomed. Eng., vol. 18, pp. 26–49, 2025. DOI: 10.1109/RBME.2024.3449790.

C. Gómez, P. Arbeláez, M. Navarrete, C. Alvarado-Rojas, M. Le Van Quyen, and M. Valderrama. Automatic seizure detection based on imaged-EEG signals through fully convolutional networks, 2020. Sci. Rep., vol. 10, no. 1, pp. 21833. DOI: 10.1038/s41598-020-78784-3.

A. Gu and T. Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces, 2024. arXiv: arXiv:2312.00752. DOI: 10.48550/arXiv.2312.00752.

A. Gu, K. Goel, and C. Ré. Efficiently Modeling Long Sequences with Structured State Spaces, 2022. arXiv: arXiv:2111.00396. DOI: 10.48550/arXiv.2111.00396.

J. Guttag. CHB-MIT scalp EEG database, 2010. PhysioNet. DOI: 10.13026/C2K01R.

H. He, X. Xiao, J. Yue, M. Xu, and D. Ming. Research on the Relationship between Resting-State Spontaneous Electroencephalography and Task-Evoked Electroencephalography, 2025. Sheng Wu Yi Xue Gong Cheng Xue Za Zhi, vol. 42, no. 3, pp. 620–627. DOI: 10.7507/1001-5515.202501021.

S. Hochreiter and J. Schmidhuber. Long Short-Term Memory, 1997. Neural Comput., vol. 9, no. 8, pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

V. Jayaram and A. Barachant. MOABB: trustworthy algorithm benchmarking for BCIs, 2018. J. Neural Eng., vol. 15, no. 6, pp. 066011. DOI: 10.1088/1741-2552/aadea0.

W.-B. Jiang, L.-M. Zhao, and B.-L. Lu. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI, 2024. arXiv: arXiv:2405.18765. DOI: 10.48550/arXiv.2405.18765.

J. Jing et al. Development of Expert-Level Classification of Seizures and Rhythmic and Periodic Patterns During EEG Interpretation, 2023. Neurology, vol. 100, no. 17, pp. e1750–e1762. DOI: 10.1212/WNL.0000000000207127.

V. Jurcak, D. Tsuzuki, and I. Dan. 10/20, 10/10, and 10/5 systems revisited: Their validity as relative head-surface-based positioning systems, 2007. NeuroImage, vol. 34, no. 4, pp. 1600–1611. DOI: 10.1016/j.neuroimage.2006.09.024.

B. T. Keenan and U. J. Magalang. The AASM sleep staging paradigm gains data-driven support, 2026. Sleep, vol. 49, no. 2, pp. zsaf343. DOI: 10.1093/sleep/zsaf343.

B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG, 2000. IEEE Trans. Biomed. Eng., vol. 47, no. 9, pp. 1185–1194. DOI: 10.1109/10.867928.

S. Khalighi, T. Sousa, J. M. Santos, and U. Nunes. ISRUC-Sleep: A comprehensive public dataset for sleep researchers, 2016. Comput. Methods Programs Biomed., vol. 124, pp. 180–192. DOI: 10.1016/j.cmpb.2015.10.013.

A. Lahoti et al. Mamba-3: Improved Sequence Modeling using State Space Principles, 2026. arXiv: arXiv:2603.15569. DOI: 10.48550/arXiv.2603.15569.

V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces, 2018. J. Neural Eng., vol. 15, no. 5, pp. 056013. DOI: 10.1088/1741-2552/aace8c.

F. Lotte et al. A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update, 2018. J. Neural Eng., vol. 15, no. 3, pp. 031005. DOI: 10.1088/1741-2552/aab2f2.

Z. Lu et al. OmniEEG-Bench: A Standardized Evaluation Benchmark for EEG Foundation Models, 2026. arXiv: arXiv:2606.00815. DOI: 10.48550/arXiv.2606.00815.

I. Obeid and J. Picone. The Temple University Hospital EEG Data Corpus, 2016. Front. Neurosci., vol. 10, pp. 196. DOI: 10.3389/fnins.2016.00196.

D. Ouyang, Y. Yuan, G. Li, and Z. Guo. The Effect of Time Window Length on EEG-Based Emotion Recognition, 2022. Sensors, vol. 22, no. 13, pp. 4939. DOI: 10.3390/s22134939.

X. Peng et al. Robust dynamic brain coactivation states estimated in individuals, 2023. Sci. Adv., vol. 9, no. 3, pp. eabq8566. DOI: 10.1126/sciadv.abq8566.

G. Pfurtscheller and F. H. Lopes da Silva. Event-related EEG/MEG synchronization and desynchronization: basic principles, 1999. Clin. Neurophysiol., vol. 110, no. 11, pp. 1842–1857. DOI: 10.1016/S1388-2457(99)00141-8.

T. Pollard et al. PhysioNet as a global platform for biomedical research, 2026. Nat. Health, pp. 1–4. DOI: 10.1038/s44360-026-00096-z.

A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training, 2018. 2018.

A. Riccio et al. Attention and P300-based BCI performance in people with amyotrophic lateral sclerosis, 2013. Front. Hum. Neurosci., vol. 7. DOI: 10.3389/fnhum.2013.00732.

R. T. Schirrmeister et al. Deep learning with convolutional neural networks for EEG decoding and visualization, 2017. Hum. Brain Mapp., vol. 38, no. 11, pp. 5391–5420, 2017. DOI: 10.1002/hbm.23730.

A. H. Shoeb. Application of machine learning to epileptic seizure onset detection and treatment, 2009. URL `https://dspace.mit.edu/handle/1721.1/54669`. Thesis, Massachusetts Institute of Technology, 2009.

L. N. Smith and N. Topin. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates, 2018. arXiv: arXiv:1708.07120. DOI: 10.48550/arXiv.1708.07120.

Y. Song, Q. Zheng, B. Liu, and X. Gao. EEG Conformer: Convolutional Transformer for EEG Decoding and Visualization, 2023. IEEE Trans. Neural Syst. Rehabil. Eng., vol. 31, pp. 710–719, 2023. DOI: 10.1109/TNSRE.2022.3230250.

G. Sun et al. video-SALMONN S: Memory-Enhanced Streaming Audio-Visual LLM, 2026. arXiv: arXiv:2510.11129. DOI: 10.48550/arXiv.2510.11129.

S. Sun et al. Atypical hierarchical brain connectivity in autism: Insights from stepwise causal analysis using Liang information flow, 2025. NeuroImage, vol. 310, pp. 121107. DOI: 10.1016/j.neuroimage.2025.121107.

M. Tangermann et al. Review of the BCI Competition IV, 2012. Front. Neurosci., vol. 6. DOI: 10.3389/fnins.2012.00055.

P. Tarara, I. Przybył, J. Schöning, and A. Gunia. Motor imagery-based brain-computer interfaces: an exploration of multiclass motor imagery-based control for Emotiv EPOC X, 2025. Front. Neuroinformatics, vol. 19. DOI: 10.3389/fninf.2025.1625279.

X.-T. Tran et al. Inter- and Intra-Subject Variability in EEG: A Systematic Survey, 2026. arXiv: arXiv:2602.01019. DOI: 10.48550/arXiv.2602.01019.

L. van der Maaten and G. Hinton. Visualizing Data using t-SNE, 2008. J. Mach. Learn. Res., vol. 9, no. 86, pp. 2579–2605, 2008.

C. Wang et al. BrainBERT: Self-supervised representation learning for intracranial recordings, 2023. arXiv: arXiv:2302.14367. DOI: 10.48550/arXiv.2302.14367.

G. Wang, W. Liu, Y. He, C. Xu, L. Ma, and H. Li. EEGPT: Pretrained Transformer for Universal and Reliable Representation of EEG Signals, 2024. In: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2024, pp. 39249–39280. DOI: 10.52202/079017-1239.

J. Wang et al. CBraMod: A Criss-Cross Brain Foundation Model for EEG Decoding, 2025. arXiv: arXiv:2412.07236. DOI: 10.48550/arXiv.2412.07236.

K. Won, M. Kwon, M. Ahn, and S. C. Jun. EEG Dataset for RSVP and P300 Speller Brain-Computer Interfaces, 2022. Sci. Data, vol. 9, no. 1, pp. 388. DOI: 10.1038/s41597-022-01509-w.

C. Yang, M. B. Westover, and J. Sun. BIOT: Cross-data Biosignal Learning in the Wild, 2023. arXiv: arXiv:2305.10351. DOI: 10.48550/arXiv.2305.10351.

L. Yang, Q. Sun, A. Li, and M. M. V. Hulle. Are EEG Foundation Models Worth It? Comparative Evaluation with Traditional Decoders in Diverse BCI Tasks, 2025. URL https://openreview.net/forum?id=5Xwm8e6vbh. In Proceedings of the Fourteenth International Conference on Learning Representations.

B. Zhang and R. Sennrich. Root Mean Square Layer Normalization, 2019. arXiv: arXiv:1910.07467. DOI: 10.48550/arXiv.1910.07467.

H. Zhang, X. Zhao, Z. Wu, B. Sun, and T. Li. Motor imagery recognition with automatic EEG channel selection and deep learning, 2021. J. Neural Eng., vol. 18. DOI: 10.1088/1741-2552/abca16.

W.-L. Zheng and B.-L. Lu. Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks, 2015. IEEE Trans. Auton. Ment. Dev., vol. 7, no. 3, pp. 162–175. DOI: 10.1109/TAMD.2015.2431497.

J. Zhou et al. Pretraining Large Brain Language Model for Active BCI: Silent Speech, 2025. arXiv: arXiv:2504.21214. DOI: 10.48550/arXiv.2504.21214.

A Dataset Description

The datasets used in this work, their preprocessing details, and the experimental configurations are described below.

A.1 TUSZ

The Temple University Hospital EEG Seizure Corpus is part of the Temple University EEG (TUEG) Corpus and is used for seizure detection; its total duration is approximately 1,474.78 hours. TUSZ contains EEG signals with manually annotated seizure events (onset, offset, channels, and seizure type). TUEG data mostly contain 19–22 electrode channels specified by the international 10–20 system, are sampled at 256 Hz, and use μV as the unit. The preprocessing applied to the TUEG subsets used here includes 0.3–75 Hz band-pass filtering, 60 Hz power-line filtering, and resampling to 200 Hz.

A.2 TUAR

The Temple University Hospital EEG Artifact Corpus is a subset of TUEG and contains annotations for five different artifact types. Its total recording duration is approximately 99.98 hours; the recordings contain 19–22 electrode channels and are sampled at 256 Hz.

A.3 TUEP

The Temple University Hospital EEG Epilepsy Corpus is a subset of TUEG curated by certified neurologists. It contains 100 participants with epilepsy and 100 participants without epilepsy. The dataset was jointly developed by the National Institutes of Health (NIH) and several other collaborators. Its total recording duration is approximately 626.76 hours; the recordings contain 19–22 electrode channels and are sampled at 256 Hz.

A.4 TUAB

The Temple University Hospital EEG Abnormal Corpus is a corpus of EEG recordings labeled as normal or abnormal. Its total recording duration is approximately 1,142 hours; the recordings contain 19–22 electrode channels and are sampled at 256 Hz.

Table 7: Seizure and normal EEG durations in the CHB-MIT training, validation, and test splits; training-set durations are reported before undersampling.

Class	Training set	Validation set	Test set
Seizure	9,572 s	505 s	638 s
Normal	3,011,428 s	212,700 s	199,200 s

A.5 CHB-MIT

The CHB-MIT Scalp EEG Database contains continuous EEG recordings from 23 pediatric subjects with intractable epilepsy, acquired at Children's Hospital Boston. The recordings were sampled at 256 Hz using bipolar montages based on the international 10–20 electrode system (Guttag, 2010; Connolly et al., 2026). The original voltage unit is V, and the labels have two classes (seizure or non-seizure) at one-second resolution. We downsample the data to 200 Hz, convert the unit to μV, and use the 16 bipolar channels from the BIOT paper: `FP1-F7`, `F7-T7`, `T7-P7`, `P7-O1`, `FP2-F8`, `F8-T8`, `T8-P8`, `P8-O2`, `FP1-F3`, `F3-C3`, `C3-P3`, `P3-O1`, `FP2-F4`, `F4-C4`, `C4-P4`, and `P4-O2`.

The positive and negative samples in this dataset are highly imbalanced: seizure samples account for only 0.31% of the total test-set duration. We calculate the durations of seizure and normal samples in the training, validation, and test sets, as shown in Table 7. To prevent the model from favoring the majority class during training, we undersample the training set: each continuous 300-s segment is retained only if it contains at least one seizure label; otherwise, it is discarded. After undersampling, the positive-to-negative sample ratio in the training set is approximately 1:4.5.

Our experiments include two settings. (1) We divide the data into 300-s samples, each corresponding to 300 labels; the training batch size is 32, the maximum learning rate is 10^{-4}, and the model is trained for 30 epochs. For NeuroStream, LSTM, and Mamba-3, test samples also include sequences of 600 s. (2) We divide the training, validation, and test data into 1-s samples, each corresponding to one label; the training batch size is 64, the maximum learning rate is 10^{-4}, and the model is trained for 30 epochs.

A.6 SLEEP-EDFX

The Sleep-EDF database contains 197 whole-night polysomnographic recordings, including EEG, EOG, chin EMG, and event markers. The EEG data contain two bipolar channels, `Fpz-Cz` and `Pz-Oz`. Professionals scored sleep every 30 s according to the Rechtschaffen and Kales sleep-staging standard. The scored patterns consist of stages W, R, 1, 2, 3, 4, M (movement time), and ? (not scored). Sleep-EDFx includes two studies: (1) the Sleep Cassette Study, which investigates the effect of age on sleep in healthy White participants aged 25–101 who were not taking sleep-related medication; and (2) the Sleep Telemetry Study, which investigates the effects of temazepam on 22 White men and women taking no other medication. The EEG was sampled at 100 Hz and measured in V. We use data from the Sleep Telemetry Study.

Our preprocessing includes resampling to 200 Hz, converting the unit to μV, and applying a 0.3–30 Hz band-pass filter. For the labeled data, we follow the AASM scoring manual (Keenan & Magalang, 2026) and relabel the sleep patterns into five stages: wake (W), stage 1 (N1), stage 2 (N2), stage 3 (N3), and rapid eye movement (REM) sleep.

Our experiments include two settings. (1) We divide the data into 1800-s (0.5-h) samples, each corresponding to 60 labels; the training batch size is 32, the maximum learning rate is 10^{-4}, and the model is trained for 30 epochs. For NeuroStream, LSTM, and Mamba-3, test samples also include sequences of 3600 s (1 h). (2) We divide the training, validation, and test data into 30-s samples, each corresponding to one label; the training batch size is 64, the maximum learning rate is 10^{-4}, and the model is trained for 30 epochs.

A.7 SEED

The Shanghai Jiao Tong University Emotion EEG dataset (Zheng & Lu, 2015) records EEG from 15 subjects while they watch 15 Chinese movie clips with different emotion labels (positive, neutral, and negative). Each recording lasts approximately four minutes and contains 62 channels. The original data have been downsampled to 200 Hz, band-pass filtered from 0 to 75 Hz, and are measured in μV. We divide the data into 10-s segments; the batch size is 64, the maximum learning rate is 10^{-4}, and training lasts 30 epochs.

A.8 BNCI2014004

The BNCI 2014-004 Motor Imagery dataset (BCI Competition IV Dataset 2b) (Tangermann et al., 2012) contains EEG data from nine subjects performing two motor-imagery tasks (left hand and right hand). Each subject completed 120 trials. The recordings include three bipolar EEG channels (C3, Cz, and C4) and three electrooculography (EOG) channels. The signals were sampled at 250 Hz and band-pass filtered from 0.5 to 100 Hz and notch-filtered at 50 Hz. We obtain this dataset using the MOABB package (Aristimunha et al., 2026a; Jayaram & Barachant, 2018; Chevallier et al., 2024) and resample the EEG to 200 Hz. Each experimental sample is the EEG data between $t = 2$ s and $t = 6$ s after the start of each trial. The batch size is 512, the maximum learning rate is 5×10^{-4}, and training lasts 30 epochs.

A.9 BNCI2014009

The BNCI 2014-009 P300 dataset (Riccio et al., 2013) contains EEG data from 10 subjects using a P300 speller system with both grid-speller and geo-speller paradigms. We obtain the grid-speller paradigm data from the MOABB package (Aristimunha et al., 2026a; Jayaram & Barachant, 2018; Chevallier et al., 2024). The data include 16 EEG channels, sampled at 256 Hz and band-pass filtered from 0.1 to 20 Hz. There are two labels (target and non-target), with a sample ratio of 1:5. Each experimental sample is 1 s of EEG beginning at stimulus onset. The batch size is 512, the maximum learning rate is 5×10^{-4}, and training lasts 30 epochs.

B Hyperparameters for Pre-training

Table 8 lists the NeuroStream model structure and pretraining hyperparameters. Other settings not listed use the default values of PyTorch or related dependencies. We also train two variants: NeuroStream-Lite, with approximately 2.7M parameters, and NeuroStream-Plus, with approximately 10M parameters. The parameter-count differences arise from changing the number of spatiotemporal blocks and the FFN dimension in the Transformer branch.

C Results of Different-Size Variants

We compare NeuroStream and its two variants of different sizes on multiple downstream datasets; the results are shown in Tables 9 and 10. On Sleep-EDFx, NeuroStream-Plus performs substantially better than the other two models (BalAcc, weighted F1, and Cohen's kappa are 0.7746, 0.6775, and 0.5664, respectively). NeuroStream-Lite performs substantially worse than the two larger models on CHB-MIT and Sleep-EDFx (AP of 0.3313 on CHB-MIT and Cohen's kappa of 0.4344 on Sleep-EDFx). The metrics of the three variants are relatively close on the other datasets.

D Other Experiments on CHB-MIT and Sleep-EDFx

On CHB-MIT, in addition to the results in Table 2, we set up another experiment in which all models are trained using 1-s samples (one label per sample), with batch size 64, maximum learning rate 10^{-4}, and 30 training epochs. The results are shown in Table 11. Some baselines perform well in this experiment, with ShallowNet achieving the highest balanced accuracy among the baselines, 0.8210. Although our model's balanced accuracy decreases from 0.9589 to 0.8780, it remains higher than that of all baselines.

Table 8: Network architecture and pretraining hyperparameters for NeuroStream Base, Lite, and Plus.

Component	Hyperparameter	Setting
Patch tokenizer	Input channels	{1, 8, 8}
	Output channels	{8, 8, 8}
	Kernel size	{15, 3, 3}
	Stride / padding	{8, 1, 1} / {7, 1, 1}
Spatiotemporal blocks	Number of layers	6 (Base), 4 (Lite), 12 (Plus)
Spatial Transformer	Hidden dimension	200
	FFN dimension / heads	800 (Base/Plus), 400 (Lite) / 10
Temporal Mamba	State dimension	64
	Head dimension / chunk size	64 / 16
	Expansion ratio	2.56
Pretraining	Batch size	1024
	Maximum learning rate / scheduler	10^{-4} / OneCycle
	Optimizer	AdamW
	Epochs	40

Table 9: Results of different NeuroStream model sizes on the long-duration CHB-MIT and Sleep-EDFx settings.

	CHB-MIT			Sleep-EDFx		
Variant	BalAcc	AUROC	AP	BalAcc	Weighted F1	Cohen's Kappa
Base	0.9528±0.0260	**0.9934**±0.0003	**0.5064**±0.0406	0.7483±0.0059	0.6175±0.0262	0.4995±0.0201
Lite	0.9417±0.0213	0.9856±0.0039	0.3313±0.0557	0.7090±0.0163	0.5617±0.0136	0.4344±0.0188
Plus	**0.9599**±0.0065	0.9916±0.0018	0.4729±0.1422	**0.7746**±0.0331	**0.6775**±0.0483	**0.5664**±0.0561

On Sleep-EDFx, in addition to the results in Table 3, we set up another experiment in which all models are trained using 30-s samples (one label per sample), with batch size 64, maximum learning rate 10^{-4}, and 30 training epochs. The results are shown in Table 12.

E Model Feature Visualizations on Downstream Datasets

Here, we apply the NeuroStream encoder to several downstream datasets to extract feature tensors $\mathbf{u}_t^{(N)}$, then use t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten & Hinton, 2008) to map them to two-dimensional space for visualization. We conduct experiments under three conditions: (1) before pretraining, using a randomly initialized model; (2) after pretraining, using a model with pretrained weights loaded but not yet fine-tuned on the downstream dataset; and (3) after fine-tuning, loading the pretrained weights into NeuroStream and further fine-tuning it on the corresponding downstream dataset. The visualizations are shown in Figure 4. Among these results, the features extracted by the randomly initialized model have the highest entropy (degree of disorder), with samples from different classes mixed together irregularly. After pretraining, the disorder in the feature visualizations decreases; in particular, Figure 4(b) and (h) show a more evident separation between different classes, demonstrating the effectiveness of our pretraining. Fine-tuning the model further improves the clustering.

F Complexity Comparison

We provide theoretical estimates and simulation experiments for three architectures: Full Attention, Divided Space-Time Attention, and the Spatial Attention + Temporal SSM architecture used by NeuroStream. Let C denote the number of EEG channels, T the number of frames, and d the model hidden dimension.

Table 10: Results of different NeuroStream model sizes on the short-duration BNCI2014004 and BNCI2014009 tasks.

	BNCI2014004			BNCI2014009		
Variant	BalAcc	AUROC	AP	BalAcc	AUROC	AP
Base	0.6328±0.0052	0.6883±0.0028	0.6775±0.0009	0.8167±0.0022	0.8940±0.0009	0.6756±0.0088
Lite	**0.6370**±0.0018	**0.6966**±0.0032	**0.6934**±0.0054	**0.8175**±0.0044	**0.8945**±0.0045	**0.6861**±0.0086
Plus	0.6193±0.0029	0.6704±0.0059	0.6610±0.0046	0.8134±0.0052	0.8884±0.0050	0.6694±0.0107

Table 11: Baseline comparison on CHB-MIT when training and evaluation use 1-s samples.

Model	Params.	BalAcc	AUROC	AP
EEGNet	1.5k	0.8029±0.0258	0.9593±0.0095	**0.3931**±0.0334
ShallowNet	0.27M	0.8210±0.0179	**0.9755**±0.0010	0.3809±0.0297
EEGConformer	0.23M	0.7994±0.0389	0.9017±0.0813	0.1526±0.0345
SPaRCNet	0.69M	0.7210±0.0413	0.9411±0.0275	0.2487±0.0649
BIOT	3.2M	0.6901±0.0108	0.9174±0.0051	0.0527±0.0077
LaBraM	5.8M	0.7733±0.0419	0.9080±0.0331	0.1628±0.0461
CBraMod	4.9M	0.8097±0.0362	0.9404±0.0025	0.1681±0.0059
Proposed	7.7M	**0.8780**±0.0254	0.9721±0.0038	0.3913±0.0184

F.1 Full Attention

For a full-attention architecture, all $C \times T$ tokens perform pairwise attention calculations. Consider a single query token: it must interact with all $C \times T$ keys or values, with complexity $O(CTd)$. Ignoring the hidden dimension, this is:

$$O(CT). \tag{10}$$

There are $C \times T$ query tokens in the entire layer, so the total complexity is:

$$O(CT) \times O(CT) = O(C^2T^2). \tag{11}$$

Strictly speaking, a complete Transformer block also includes QKV projections, an output projection, and an FFN, with computational complexity $O(CTd^2)$. However, d is generally fixed, so the complexity of these components grows linearly with input-sequence length and can be omitted here.

F.2 Divided Space-Time Attention

The Divided Space-Time Attention architecture has two branches, Spatial Attention and Temporal Attention, which act on the spatial and temporal dimensions, respectively.

Spatial Attention. There are C channel tokens within each frame. A query token only needs to interact with the C tokens in the same frame, with complexity $O(Cd)$, or $O(C)$ when ignoring the hidden dimension. The full EEG signal has T frames and $C \times T$ tokens, so the total complexity is:

$$O(CT) \times O(C) = O(C^2T). \tag{12}$$

Temporal Attention. Each channel has T tokens along the temporal axis. A query token needs to interact with the T tokens in the same channel, with complexity $O(T)$. The EEG signal has C channels and $C \times T$ tokens, so the total complexity is:

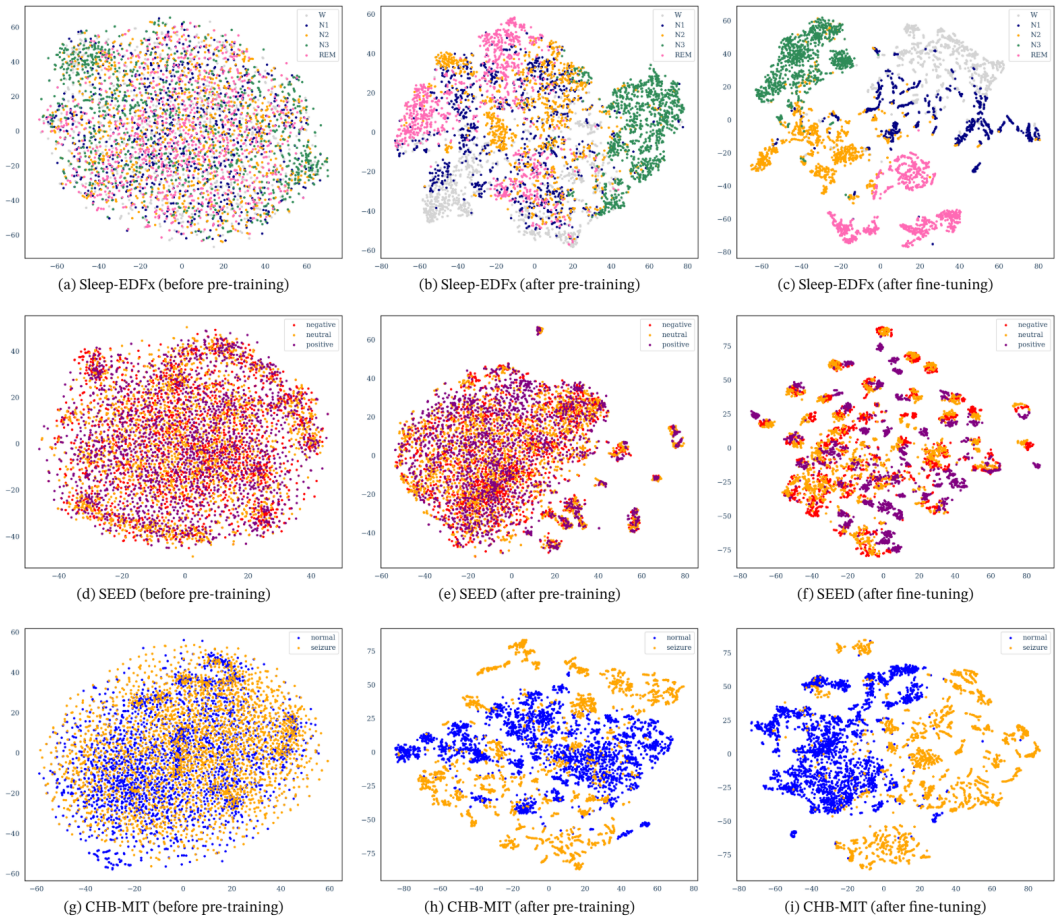
$$O(CT) \times O(T) = O(CT^2). \tag{13}$$

Therefore, the total complexity of Divided Space-Time Attention is:

$$O(C^2T + CT^2). \tag{14}$$

Table 12: Baseline comparison on Sleep-EDFx when training and evaluation use 30-s windows.

Model	Params.	BalAcc	Weighted F1	Cohen's Kappa
EEGNet	16k	0.5707±0.0352	0.5658±0.0518	0.4007±0.0569
ShallowNet	83k	0.6894±0.0262	0.7078±0.0418	0.5659±0.0524
EEGConformer	4.2M	0.6212±0.0426	0.6493±0.0468	0.4989±0.0588
SPaRCNet	1.4M	0.6772±0.0368	0.7230±0.0314	0.6081±0.0489
BIOT	3.2M	0.7196±0.0095	0.7408±0.0161	0.6581±0.0883
LaBraM	5.8M	0.7257±0.0182	0.7434±0.0286	0.6188±0.0369
CBraMod	5.0M	**0.7549**±0.0043	**0.7736**±0.0235	**0.6780**±0.0145
Proposed	5.5M	0.7389±0.0037	0.6799±0.0048	0.5479±0.0040



(a) Sleep-EDFx (before pre-training) (b) Sleep-EDFx (after pre-training) (c) Sleep-EDFx (after fine-tuning)

(d) SEED (before pre-training) (e) SEED (after pre-training) (f) SEED (after fine-tuning)

(g) CHB-MIT (before pre-training) (h) CHB-MIT (after pre-training) (i) CHB-MIT (after fine-tuning)

Figure 4: Qualitative t-SNE visualizations of features from Sleep-EDFx (a–c), SEED (d–f), and CHB-MIT (g–i). Columns show features before pretraining, after pretraining, and after downstream fine-tuning; colors indicate class labels.

F.3 Spatial Attention + Temporal SSM

The spatial component of this architecture remains within-frame Spatial Attention, with complexity $O(C^2T)$. The temporal component is changed to a channel-wise Temporal SSM. At a single time step, for each channel, the SSM only needs:

- the current token;
- the state from the preceding time step.

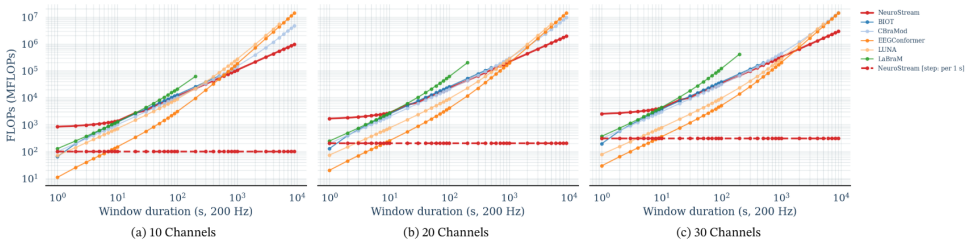


Figure 5: Measured FLOPs versus EEG input duration for 10, 20, and 30 channels at 200 Hz on log–log axes. Both whole-sequence and frame-wise NeuroStream inference are shown.

Thus, for fixed hidden and state dimensions, the cost of a single update does not grow with history length T, and the temporal complexity per update is $O(1)$. The EEG signal has C channels and T time steps, so the total SSM complexity is:

$$O(CT) \times O(1) = O(CT). \tag{15}$$

Therefore, the overall complexity of Spatial Attention + Temporal SSM can be written as:

$$O(C^2T + CT). \tag{16}$$

F.4 Simulation Experiments

We select Transformer models with several different architectures and simulate EEG windows with different channel counts and durations as their inputs. Figures 5 and 6 show their FLOPs and memory-use curves for different input sizes on logarithmic axes. Due to limits on the size of internal positional-embedding tables, LaBraM and BIOT fail when input duration exceeds 200 s and 500 s, respectively. For models whose computational complexity grows linearly with temporal-window length, the asymptotic slopes of the FLOPs and memory curves should approach 1; for models whose computational complexity grows quadratically with temporal-window length, the slopes should approach 2. In Figure 5, NeuroStream's curve is noticeably higher than those of the other models at short durations. This is because the Mamba-3 kernel requires sequence length to align with the chunk size; we zero-pad the frame sequence inside the model, making the computation for 1–16 s similar and flattening the curve in the short-duration region. We also report FLOPs and memory use per frame when NeuroStream performs single-step inference. Table 13 shows the tail slopes (FLOPs and memory exponents) of the FLOPs and memory curves for the 30-channel setting.

NeuroStream's FLOPs and memory exponents both approach 1, consistent with our theoretical derivation that computational complexity is linear in EEG-signal duration. EEGConformer uses a Full Attention architecture with theoretical spatiotemporal complexity $O(C^2T^2)$; its measured FLOPs exponent is 1.92, close to the theoretical value. However, its memory exponent is close to 1 because its implementation uses a FlashAttention kernel, which computes attention in tiles on on-chip SRAM and does not write the complete attention matrix back to GPU memory, reducing space complexity to linear. BIOT is a foundation model using linear attention, so its computational complexity is linear in input size and its FLOPs and memory exponents both approach 1. The FLOPs and memory exponents of the other models approach 2.

G The Channels Used During Pre-training

During pretraining, we use the 19 channels common to TUSZ, TUAR, TUEP, and TUAB, namely: Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, and O2.

G.1 Experimental Environment

We conducted the experiments on a computer equipped with an NVIDIA 3090 GPU, with Python 3.11, PyTorch 2.10, CUDA 12, and other software dependencies.

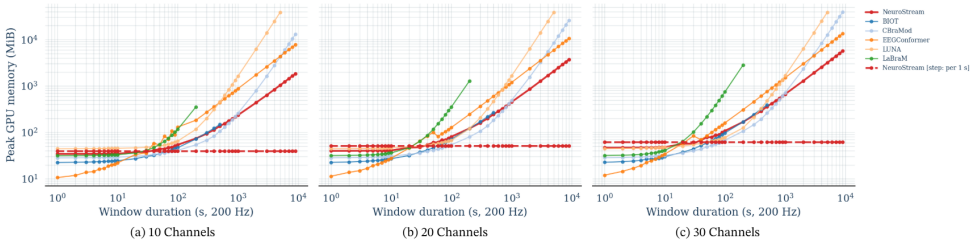


Figure 6: Peak GPU memory versus EEG input duration for 10, 20, and 30 channels at 200 Hz on log–log axes.

Table 13: Tail slopes of the FLOPs and peak-memory curves for different models in the 30-channel setting.

Model	Complexity	FLOPs exponent	Memory exponent
CBraMod	$O(C^2T + CT^2)$	1.57	1.81
EEGConformer	$O(C^2T^2)$	1.92	1.00
LUNA	$O(CT + T^2)$	1.80	1.92
LaBraM	$O(C^2T^2)$	1.62	1.70
BIOT	$O(CT)$	1.00	0.83
NeuroStream	$O(C^2T + CT)$	1.00	0.97