S-CEReBrO: Breaking the Memory Barrier in Continuous EEG Monitoring

Glenn Anta Bucagu[1], Thorir Mar Ingolfsson[1], Yawei Li[1,2], and Luca Benini[1,3]

[1] ETH Zurich, Zurich, Switzerland
[2] Nanyang Technological University, Singapore
[3] University of Bologna, Bologna, Italy
gbucagu@ethz.ch

Abstract. Foundation models offer a promising paradigm for Electroencephalography (EEG) analysis, leveraging generalizable representations from vast unlabeled datasets. Yet, Transformer-based architectures face a critical bottleneck: global attention mechanisms couple the attention memory state to the signal duration, causing memory overflow during continuous monitoring. To address this, we introduce **S-CEReBrO** (Streaming CEReBrO), an evolution of the CEReBrO architecture designed for continuous monitoring. Our novel *Windowed Alternating Attention* mechanism factorizes attention computation into fixed-size spatiotemporal windows, guaranteeing constant KV cache memory as only the active window requires resident attention maps. Empirical scaling analysis confirms that windowed alternating attention can process signals **100**× longer than full self-attention and **3**× longer than low-rank linear attention. Compared to low-rank linear attention on long contexts, windowed alternating attention requires **55%** of the memory while increasing inference throughput by **2.1**×. Pre-trained on >25,000 hours of recordings from >12,000 subjects, S-CEReBrO achieves state-of-the-art performance on **7 of 11** downstream tasks, with up to **60**% fewer parameters. This work represents a significant step toward the realization of efficient, generalizable, and continuous EEG monitoring. An accompanying code repository is available [4].

Keywords: EEG · Foundation Models · Efficient Neural Networks.

1 Introduction

Electroencephalography (EEG) non-invasively records the brain's electrical activity by capturing voltage fluctuations at the scalp. As the field shifts toward continuous, long-term monitoring for applications like seizure detection, the resulting data, structured as C channels $\times$ T timesteps, poses a scaling challenge. While hardware limits C to a fixed range (typically 2–64), T grows indefinitely during prolonged observation. To process this high-dimensional data, EEG foundation models have emerged as a powerful alternative to supervised methods,

[4] https://github.com/pulp-bio/biofoundation

Table 1. Comparison of Transformer-based EEG foundation models. S-CEReBrO is the only architecture that achieves strictly linear attention complexity and a constant KV cache memory footprint relative to signal duration (T).

Model	Params (M)	Attention Mechanism	Attention Complexity	Peak Memory
BIOT [30]	3.2	Linear	$\mathcal{O}(CT)$	$\mathcal{O}(T)$
CBraMod [28]	4.0	Criss-Cross	$\mathcal{O}(CT(C+T))$	$\mathcal{O}(T)$
LaBraM [5]	5.8–369	Full Self-Attention	$\mathcal{O}((CT)^2)$	$\mathcal{O}(T)$
LUNA [4]	7.0–311	Perceiver	$\mathcal{O}(T(C+T))$	$\mathcal{O}(T)$
CEReBrO [3]	2.4	Alternating	$\mathcal{O}(CT(C+T))$	$\mathcal{O}(T)$
S-CEReBrO (Ours)	**2.4**	**Windowed Alternating**	$\mathcal{O}(CT)$	$\mathcal{O}(1)$

Peak Memory: Scaling of the KV cache memory relative to total signal duration T.

leveraging large-scale pre-training to learn general representations. Although the Transformer architecture currently dominates this field [5, 28, 30] and shows promise for resource-constrained inference [3], a critical memory bottleneck remains. Wearable devices cannot usually store the whole history of Transformer key and value (KV) states for large T. This memory-bound limitation prevents high-performance foundation models from sustaining the processing required for continuous EEG monitoring.

To bridge this gap, we introduce **S-CEReBrO** (Streaming CEReBrO), an evolution of the CEReBrO architecture [3]. S-CEReBrO utilizes a novel *Windowed Alternating Attention* mechanism that interleaves spatial and temporal attention blocks, restricting interactions to local neighborhoods in time and space. This allows our model to process long signals while maintaining a constant resident memory. We summarize our contributions as follows:

- **Locality-Driven Attention:** We introduce a Windowed Alternating Attention that maintains a constant peak resident memory by evicting stale tokens from the computational buffer. While total computational complexity scales linearly ($\mathcal{O}(CT)$) to process the full signal, our approach eliminates the cumulative memory overflow inherent to global attention. This mechanism can handle 14-hour continuous monitoring, with 100× the capacity of full self-attention, while needing 55% of the memory of low-rank linear attention and achieving 2.1× higher inference throughput.
- **Cross-Task SoTA Performance:** We demonstrate that global attention is often redundant for EEG analysis. S-CEReBrO outperforms global attention baselines on 7 of 11 public benchmarks, spanning clinical diagnostics, BCI, and state monitoring. It achieves these results while using up to 60% fewer parameters than existing models.

2 Related Works

Although Transformer-based EEG foundation models excel at learning generalized representations, the high computational cost of their global attention mecha-

nisms prevents them from processing the long contexts required for continuous monitoring. Comparisons of the current state-of-the-art are available in Table 1.

Models like LaBraM [5] achieve high performance but inherit the quadratic complexity $\mathcal{O}((CT)^2)$ of vanilla Transformers. BIOT [30] approximates attention via low-rank projections [29] to achieve $\mathcal{O}(CT)$ complexity. CBraMod [28] employs Criss-Cross attention to separate temporal and channel attention across different heads, with $\mathcal{O}(T(C+T))$ complexity. LUNA [4] maps high-dimensional EEG inputs onto a fixed-size set of latent variables, yielding $\mathcal{O}(T(C+T))$ complexity. CEReBrO [3] uses alternating attention to separate spatial and temporal attention in separate layers. However, all these models maintain a global receptive field, meaning their resident memory scales with T. In contrast, S-CEReBrO enforces algorithmic locality through *Windowed Alternating Attention*, ensuring a resident memory state that is independent from T. While windowed attention mechanisms have been popularized by Large Language Models to handle long context [1], these methods are not directly transferable to neural signals. Unlike text, which adheres to a 1D linear grammar, EEG signals exhibit a complex 2D non-linear structure. Competing efficient attention mechanisms lack the necessary spatiotemporal priors for bidirectional EEG inference. For example, Longformer [1] utilizes 1D windows that break the 2D EEG structure, and its bidirectional form still requires an $O(T)$ KV cache. Similarly, Performer [2] lacks a spatiotemporal prior by relying on content-agnostic random features, while its bidirectional kernel sum materializes the entire sequence, reinstating the $O(T)$ KV cache bottleneck.

3 Method

S-CEReBrO reconciles the long context requirements of continuous EEG monitoring with the constraints of real-world hardware. We use a novel *Windowed Alternating Attention* mechanism with linear complexity $O(CT)$ and a constant KV cache memory. S-CEReBrO learns useful representations via masked autoencoding. Figure 1 summarizes the pipeline. We detail each component below.

3.1 Spatiotemporal Tokenization

Given a raw EEG recording $\mathbf{X} \in \mathbb{R}^{C \times T}$, we tokenize the signal into a sequence of patch embeddings:

1. **Patching:** We apply a sliding window of length L and stride S to each channel, yielding a patch tensor $\mathbf{P} \in \mathbb{R}^{C \times N_p \times L}$, where $N_p = \lfloor \frac{T-L}{S} \rfloor + 1$ [19].
2. **Feature Projection:** A shared convolutional encoder projects each raw patch $\mathbf{P}_{c,i}$ into a latent embedding $\mathbf{E}_{c,i} \in \mathbb{R}^{d_e}$.
3. **Spatiotemporal Embeddings:** We preserve spatiotemporal topology using two learnable embeddings: an index-based matrix $\mathbf{W}_{\text{pos}} \in \mathbb{R}^{N_p \times d_e}$ for temporal order, and a continuous spatial embedding $\mathbf{W}_{\text{chan},c} \in \mathbb{R}^{d_e}$ obtained by mapping 3D electrode coordinates (x, y, z) via a shared MLP [4].

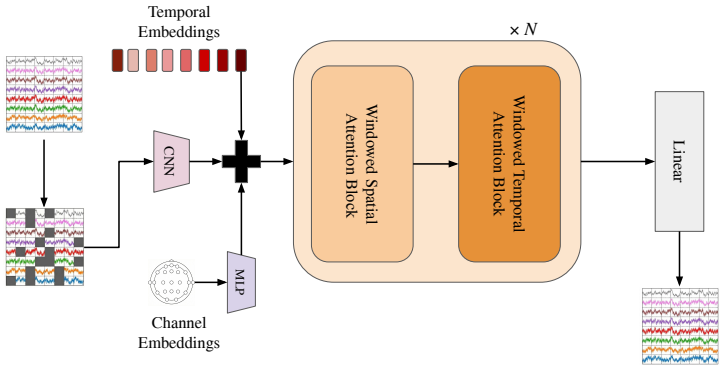


Fig. 1. Overview of the S-CEReBrO architecture and masked autoencoding pipeline, featuring N composite attention blocks, where each comprises a WSA and a WTA layer.

The final input tensor to the Transformer, $\mathbf{H}^{(0)} \in \mathbb{R}^{C \times N_p \times d_e}$, is computed as:

$$\mathbf{H}^{(0)}_{c,i} = \mathbf{E}_{c,i} + \mathbf{W}_{\text{pos},i} + \mathbf{W}_{\text{chan},c} \tag{1}$$

where $\mathbf{E}_{c,i}$ is the projected patch embedding, $\mathbf{W}_{\text{pos},i}$ is the temporal embedding for index i, and $\mathbf{W}_{\text{chan},c}$ is the spatial embedding for channel c.

3.2 Windowed Temporal Attention (WTA)

WTA layers model temporal dependencies independently for each channel c. To maintain a constant resident memory, each query $\mathbf{q}_{c,i}$ attends only to keys within a local neighborhood $\mathcal{N}_t$ of size w_t.

$$\text{WTA}(\mathbf{H}_{c,i}) = \text{Softmax}\left(\frac{\mathbf{q}_{c,i}(\mathbf{K}_{c,\mathcal{N}_t(i)})^T}{\sqrt{d_e}}\right)\mathbf{V}_{c,\mathcal{N}_t(i)} \tag{2}$$

where the neighborhood $\mathcal{N}_t(i,l)$ for a time step i at layer l is defined as the set of valid indices:

$$\mathcal{N}_t(i,l) = \{k \in [0, T-1] \mid k = i + (j \cdot d_l + s_l)\}^5 \tag{3}$$

for $j \in [-\lfloor \frac{w_t-1}{2} \rfloor, \ldots, \lceil \frac{w_t-1}{2} \rceil]$. Layer-specific temporal dilation (d_l) and shift factors (s_l) expand the model's context without increasing its memory footprint. Although individual WTA windows remain local, they capture long-range dependencies by alternating with WSA layers. These WSA layers act as temporal bridges; by mixing channel features, they allow information from isolated temporal windows to interact indirectly as they propagate through the model.

[5] We follow 0-indexed notation used in standard deep learning frameworks.

Table 2. Key Hyperparameters for S-CEReBrO.

Hyperparameter	**Value**
Spatial Window Size (w_s)	$\min(C, 7)$
Temporal Window Size (w_t)	$\min(T, 5)$
Dilation Factors (d_l)	$\{1, 2, 4\}$
Shift Factors (s_l)	$\{0, 1, -1, 2, -2\}$
Number of Attention Block Pairs (N)	3
Embedding Dimension (d_e)	180
Number of Attention Heads (h)	5

3.3 Windowed Spatial Attention (WSA)

WSA layers facilitate cross-channel information exchange at a fixed time index i. Attention computation is restricted to a local w_s-sized neighborhood $\mathcal{N}_s$.

$$\text{WSA}(\mathbf{H}_{c,i}) = \text{Softmax}\left(\frac{\mathbf{q}_{c,i}(\mathbf{K}_{\mathcal{N}_s(c,l),i})^T}{\sqrt{d_e}}\right)\mathbf{V}_{\mathcal{N}_s(c,l),i} \tag{4}$$

where the neighborhood $\mathcal{N}_s(c, l)$ for a channel c at layer l is defined as the set of valid indices:

$$\mathcal{N}_s(c, l) = \{n \in [0, C-1] \mid n = c + (m \cdot d_l + s_l)\} \tag{5}$$

for $m \in [-\lfloor \frac{w_s-1}{2} \rfloor, \ldots, \lceil \frac{w_s-1}{2} \rceil]$. Layer-specific dilation (d_l) and shift factors (s_l) stagger the receptive fields, enabling distant but physiologically related channels to interact across the model's depth. This design efficiently approximates global spatial relationships without the high memory overhead of distance-based maps or dynamic clusters. Beyond substantial efficiency gains (see Section 4.5), our attention mechanism introduces a beneficial spatiotemporal prior that translates into superior performance on several downstream EEG tasks (Section 4.2). Key hyperparameters are presented in Table 2. We set $w_t = 5$ so the cascaded WTA layers cover the entire pre-training context (Section 4). Shift factors $\{0,\pm1,\pm2\}$ tile these fields without gaps across dilations $\{1,2,4\}$. We set $w_s = 7$ to cover standard 10-20 EEG montages.

3.4 Algorithmic Distinction and Computational Complexity

The difference between windowed alternating attention and standard alternating attention lies in the *management of the computational state*. Alternating attention [3] requires retaining the key-value states for all N_p temporal positions and C spatial positions, leading to a $\mathcal{O}(N_p) \propto \mathcal{O}(T)$ resident memory footprint that eventually exceeds hardware limits as $T \to \infty$. Its total computational complexity is $\mathcal{O}(CN_p(C + N_p)) \propto \mathcal{O}(CT(C + T))$. In contrast, Windowed Alternating Attention enforces algorithmic locality. By evicting stale tokens outside the fixed

neighborhoods $\mathcal{N}_t(i,l), \mathcal{N}_s(c,l)$ from the computational buffer, the resident memory requirement remains constant relative to T. Information from these evicted tokens is preserved through depth because the WSA layers mix channels at each timestep prior to eviction. Since w_t and w_s are hyperparameters fixed at design time, the total computational complexity is $\mathcal{O}(CN_p(w_t + w_s)) \propto \mathcal{O}(CT)$. Current state-of-the-art EEG foundation models are typically forced into arbitrary fixed-window chunking to mitigate memory bottlenecks, which treats the EEG signal as a series of isolated, independent segments. In contrast, S-CEReBrO employs First In First Out (FIFO)-based streaming to maintain a continuous, stateful receptive field. This approach preserves the long-range temporal dependencies essential for continuous monitoring that are otherwise lost in segmented batch processing.

3.5 Pre-training

Following [3], we employ Masked Autoencoding (MAE) for pre-training. Input sequences are tokenized into patches $\mathbf{P}$, with a fixed percentage randomly replaced by a shared learnable [MASK] token. S-CEReBrO processes these tokens, and a linear projection layer reconstructs the original patches as $\hat{\mathbf{P}}$. The pre-training objective minimizes the weighted sum of reconstruction losses for masked ($\mathcal{M}$) and visible ($\overline{\mathcal{M}}$) positions:

$$\mathcal{L} = \mathcal{L}_{\mathrm{m}} + \alpha \mathcal{L}_{\mathrm{v}}, \text{ where } \mathcal{L}_{\{\mathrm{m,v}\}} = \frac{1}{|\mathcal{K}|} \sum_{(c,i)\in\mathcal{K}} \|\mathbf{P}_{c,i} - \hat{\mathbf{P}}_{c,i}\|_2^2 \text{ for } \mathcal{K} \in \{\mathcal{M}, \overline{\mathcal{M}}\} \quad (6)$$

4 Experiments

4.1 Pre-training Corpus

S-CEReBrO is pre-trained on over 25,000 hours of EEG from more than 12,000 subjects, aggregating diverse clinical (TUEG [20]) and commercial datasets (SEED series [14, 32, 33], BOAS [15], SleepEDFx [7], BCI-NER [22], GWD [9]). This corpus spans heterogeneous montages (2–62 channels) and referencing schemes, ensuring robustness across recording conditions. All signals are resampled to 200 Hz and processed in 30-second windows. To prevent downstream data leakage, subjects from the TUAB dataset are strictly excluded.

4.2 Downstream Tasks

To evaluate S-CEReBrO, we selected 11 downstream benchmarks spanning clinical neurology (seizure detection, sleep staging) and consumer BCI (motor imagery, emotion recognition). Summarized in Table 3, these tasks comprise binary, multi-class, and regression objectives with heterogeneous sampling rates (128–1000 Hz), montages (6–64 channels), and durations (4–30 s). We employ subject-stratified splits to ensure zero overlap between sets. All benchmarks utilize subjects and channel configurations unseen during pre-training.

Table 3. Overview of downstream EEG decoding tasks and datasets used for fine-tuning.

Decoding Task	Dataset	Frequency(Hz)	Channels	Duration(s)	Subjects	Samples	Objective
Seizure Detection	CHB-MIT [24]	256	16	10	22	326,993	2-class
	Neonate [27]	256	19	5	79	80,458	2-class
Motor Imagery Classification	PhysioNet-MI [23]	160	64	4	109	9,837	4-class
	SHU-MI [16]	250	32	4	32	11,988	2-class
Abnormal Classification	TUAB [20]	250	22	10	2,383	409,455	2-class
Mental Workload Classification	STEW [12]	128	14	4	48	6,660	3-class
Vigilance Estimation	SEED-VIG [17]	200	17	8	23	20,355	Regression
Sleep Stage Classification	ISRUC [8]	200	6	30	100	89,240	5-class
Mental Disorder Diagnosis	Mumtaz [18]	256	19	5	62	7,143	2-class
Mental Stress Detection	MentalArithmetic [34]	500	20	5	36	1,707	2-class
Emotion Recognition	SEED-V [13]	1000	62	4	20	117,744	5-class

4.3 Baselines and Performance Metrics

We compare S-CEReBrO against leading non-foundational architectures (EEGNet [10], EEGConformer [26], SPaRCNet [6], CNN-Transformer [21], ST-Transformer [25], ContraWR [31] and FFCL [11]). To reflect practical deployability on resource-constrained hardware, we restrict our EEG foundation model baselines to the highest performing models with fewer than 10M parameters. These are BIOT [30], LaBraM-Base [5] and CBraMod [28]. Model checkpoints are selected based on the optimal validation score. We use AUROC for binary classification, weighted F1-score for multi-class classification, and NRMSE for regression.

4.4 Cross-Task Evaluation

As detailed in Table 4, S-CEReBrO achieves state-of-the-art performance on 7 of 11 tasks. It outperforms both specialized baselines and larger foundation models despite its compact 2.4M parameter count. The most significant gains occur in clinical and behavioral monitoring. Specifically, in seizure detection (CHB-MIT, Neonate) and mental workload assessment (STEW, SEED-VIG), S-CEReBrO excels. These are popular use cases for continuous monitoring, where model efficiency and accuracy are paramount. Table 5 confirms that the localized inductive bias of S-CEReBrO is particularly effective for cognitive monitoring. On the STEW and SEED-VIG benchmarks, S-CEReBrO achieves state-of-the-art performance, surpassing every model including its global counterpart (CEReBrO).

4.5 Computational Complexity and Scaling Analysis

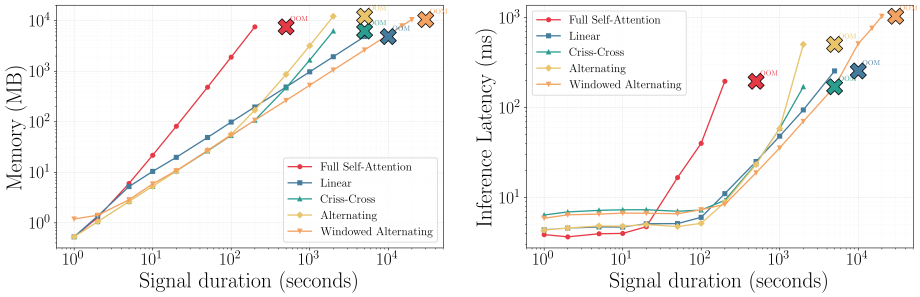
We conduct scaling experiments to evaluate windowed alternating attention against established attention baselines for long-form EEG monitoring, with results illustrated in Figure 2. At $T = 15,000$ seconds (4 hours), windowed alternating attention requires 55% of the memory and achieves 2.1× higher inference throughput than the second-best method (linear attention). Moreover, windowed alternating attention is the only mechanism to avoid a memory limit, successfully processing signals up to 50,000 seconds (14 hours); a duration 100× longer than for full self-attention, 10× longer than for alternating attention, and 3× longer than for linear attention.

Table 4. Comparison of S-CEReBrO against state-of-the-art baselines across downstream tasks. We report mean ± standard deviation across five random seeds.

Dataset	Metric	Best Supervised Model	BIOT [30] (3.2M)	CBraMod [28] (4.0M)	LaBraM [5] (5.8M)	S-CEReBrO (Ours) (2.4M)
CHB-MIT	AUROC ↑	86.62 ± 0.82	80.84 ± 4.43	84.55 ± 0.74	79.70 ± 5.64	**87.45 ± 1.93**
Neonate	AUROC ↑	81.45 ± 4.96	79.78 ± 5.47	76.63 ± 6.58	79.78 ± 3.30	**85.76 ± 2.49**
PhysioNet-MI	Weighted F1 ↑	60.62 ± 0.95	43.10 ± 1.36	48.95 ± 1.87	49.95 ± 4.55	**60.93 ± 0.95**
SHU-MI	AUROC ↑	64.31 ± 0.82	66.16 ± 1.67	64.58 ± 1.01	65.70 ± 1.50	**66.48 ± 0.61**
SEED-V	Weighted F1 ↑	26.08 ± 1.80	24.38 ± 2.75	26.32 ± 1.04	22.74 ± 1.67	**28.16 ± 0.71**
STEW	Weighted F1 ↑	49.15 ± 6.56	46.87 ± 2.63	49.09 ± 3.15	46.77 ± 1.73	**52.90 ± 2.56**
SEED-VIG	NRMSE ↓	1.01 ± 0.03	0.95 ± 0.04	0.94 ± 0.05	0.99 ± 0.04	**0.93 ± 0.03**
TUAB	AUROC ↑	88.97 ± 0.33	88.56 ± 0.74	**89.36 ± 0.46**	88.30 ± 0.40	89.30 ± 0.25
ISRUC	Weighted F1 ↑	77.19 ± 1.05	77.90 ± 1.46	**78.93 ± 0.72**	77.29 ± 0.83	78.23 ± 0.59
Mumtaz	AUROC ↑	97.81 ± 0.83	**99.01 ± 0.90**	98.69 ± 0.27	95.74 ± 0.77	98.23 ± 0.17
Mental Arithmetic	AUROC ↑	**75.66 ± 2.77**	70.88 ± 6.77	72.23 ± 3.16	62.84 ± 2.75	68.97 ± 1.71

Table 5. Performance on STEW (3-class mental workload classification) and SEED-VIG (regression on vigilance estimates).

Methods	Model Size	STEW, 3-class			SEED-VIG, Regression		
		Balanced Acc. ↑	Cohen's κ ↑	Weighted F1 ↑	NRMSE ↓	RMSE ↓	Pearson ↑
EEGNet	0.003M	49.58 ± 2.37	0.2556 ± 0.0321	48.99 ± 2.56	1.0753 ± 0.0375	0.1884 ± 0.0066	0.2741 ± 0.0430
EEGConformer	0.55M	49.97 ± 7.06	0.2910 ± 0.0683	49.15 ± 6.56	1.5028 ± 0.1056	0.2633 ± 0.0185	0.2837 ± 0.0595
SpaRCNet	0.79M	45.49 ± 3.99	0.1952 ± 0.0564	46.14 ± 3.88	1.3257 ± 0.1841	0.2323 ± 0.0323	0.2658 ± 0.0932
CNN-Transformer	1.6M	50.89 ± 2.01	0.2643 ± 0.0366	48.04 ± 1.35	1.0095 ± 0.0306	0.1772 ± 0.0056	0.4284 ± 0.0417
ContraWR	3.2M	50.47 ± 3.58	0.2630 ± 0.0312	48.69 ± 4.79	1.0740 ± 0.0645	0.1882 ± 0.0113	0.4083 ± 0.0358
ST-Transformer	2.4M	47.81 ± 4.59	0.2434 ± 0.0779	46.04 ± 4.03	1.1680 ± 0.1290	0.2047 ± 0.0226	0.3108 ± 0.0336
FFCL	3.5M	47.91 ± 4.16	0.2397 ± 0.0843	45.89 ± 6.48	1.0587 ± 0.0771	0.1855 ± 0.0135	0.3751 ± 0.0331
BIOT	3.2M	47.61 ± 1.75	0.2367 ± 0.0263	46.87 ± 2.63	0.9528 ± 0.0426	0.1670 ± 0.0075	0.4253 ± 0.0683
CBraMod	4.0M	48.30 ± 3.44	0.2498 ± 0.0550	49.09 ± 3.15	0.9369 ± 0.0461	0.1642 ± 0.0181	**0.5189 ± 0.0307**
LaBraM	5.8M	49.42 ± 1.45	0.2530 ± 0.0107	46.77 ± 1.73	0.9895 ± 0.0370	0.1734 ± 0.0065	0.4508 ± 0.0112
CEReBrO	2.4M	52.15 ± 3.58	0.2949 ± 0.0503	49.80 ± 3.74	0.9328 ± 0.0237	0.1636 ± 0.0142	0.4501 ± 0.0366
S-CEReBrO	2.4M	**55.29 ± 1.66**	**0.3162 ± 0.0267**	**52.90 ± 2.56**	**0.9284 ± 0.0316**	**0.1522 ± 0.0228**	0.4710 ± 0.0222



(a) Memory usage vs. signal duration **(b)** Inference latency vs. signal duration

Fig. 2. Memory usage and inference latency vs. signal duration T. For windowed alternating attention, total memory and runtime scale linearly with T, while the KV cache memory remains constant. Crosses denote Out of Memory (OOM) thresholds on an NVIDIA A100 GPU with 40 GB of RAM.

5 Conclusion

We propose S-CEReBrO, a foundation model that mitigates memory bottlenecks in continuous EEG monitoring. *Windowed Alternating Attention* maintains a constant KV cache memory independent of the signal length, processing signals up to 100× longer than self-attention and 3× longer than low-rank linear attention. This localized spatiotemporal inductive bias drives efficiency and accuracy, achieving state-of-the-art performance on 7 of 11 tasks with up to 60% fewer parameters than existing foundation models. Ultimately, S-CEReBrO paves the way towards high-fidelity, continuous, real-time EEG analysis. Future work will focus on developing long-form benchmarks to evaluate hour-scale streaming, and improving sample efficiency for better adaptation to exceptionally small datasets.

Acknowledgements. This work was supported by a grant from the Swiss National Supercomputing Centre (CSCS) under project IDs lp12 and lp160 on Alps, the ETH Future Computing Laboratory (EFCL), and a donation from Huawei Technologies.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer (2020)
2. Choromanski, K.M., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J.Q., Mohiuddin, A., Kaiser, L., Belanger, D.B., Colwell, L.J., Weller, A.: Rethinking Attention with Performers. In: International Conference on Learning Representations (2021)
3. Dimofte, A., Bucagu, G.A., Ingolfsson, T.M., Wang, X., Cossettini, A., Benini, L., Li, Y.: CEReBrO: Compact Encoder for Representations of Brain Oscillations Using Efficient Alternating Attention. arXiv:2501.10885 [cs.LG] (2025)
4. Döner, B., Ingolfsson, T.M., Benini, L., Li, Y.: LUNA: Efficient and Topology-Agnostic Foundation Model for EEG Signal Analysis. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
5. Jiang, W., Zhao, L., liang Lu, B.: Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI. In: The Twelfth International Conference on Learning Representations (2024)
6. Jing, J., et al.: Development of Expert-Level Classification of Seizures and Rhythmic and Periodic Patterns During EEG Interpretation. Neurology **100**(17), e1750–e1762 (Apr 2023), pMID: 36878708; PMCID: PMC10136013
7. Kemp, B., et al.: Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. IEEE Transactions on Biomedical Engineering **47**(9), 1185–1194 (2000)
8. Khalighi, S., Sousa, T., Santos, J.M., Nunes, U.: ISRUC-sleep: A comprehensive public dataset for sleep researchers. Computer Methods and Programs in Biomedicine **124**, 180–192 (2016)

9. Kobler, R.J., Sburlea, A.I., Müller-Putz, G.R.: Tuning Characteristics of Low-Frequency EEG to Positions and Velocities in Visuomotor and Oculomotor Tracking Tasks. Scientific Reports (2018)
10. Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J.: EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. Journal of Neural Engineering **15**(5), 056013 (Jul 2018)
11. Li, H., Ding, M., Zhang, R., Xiu, C.: Motor imagery EEG classification algorithm based on CNN-LSTM feature fusion network. Biomedical Signal Processing and Control **72**, 103342 (February 2022)
12. Lim, W.L., Sourina, O., Wang, L.: STEW: Simultaneous Task EEG Workload Dataset (2018)
13. Liu, W., Qiu, J.L., Zheng, W.L., 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 (2021)
14. Liu, W., et al.: Identifying Similarities and Differences in Emotion Recognition with EEG and Eye Movements among Chinese, German, and French People. Journal of Neural Engineering **19**(2), 026012 (2022)
15. López-Larraz, E., et al.: Bitbrain Open Access Sleep Dataset (2025)
16. Ma, J., Yang, B., Qiu, W., Li, Y., Gao, S., Xia, X.: A large EEG dataset for studying cross-session variability in motor imagery brain-computer interface. Scientific Data **9**(1), 531 (2022)
17. Min, J., Wang, P., Hu, J.: Driver fatigue detection through multiple entropy fusion analysis in an EEG-based system. PLoS One **12**(12), e0188756 (2017)
18. Mumtaz, W.: MDD Patients and Healthy Controls EEG Data (2016), dataset. doi:10.6084/m9.figshare.4244171.v2
19. Nie, Y., et al.: A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In: The Eleventh International Conference on Learning Representations (2023)
20. Obeid, I., Picone, J.: The Temple University Hospital EEG Data Corpus. Frontiers in Neuroscience **10**, 196 (2016)
21. Peh, W.Y., et al.: Transformer Convolutional Neural Networks for Automated Artifact Detection in Scalp EEG (2022)
22. Perrin, M., Maby, E., Daligault, S., Bertrand, O., Mattout, J.: Objective and subjective evaluation of online error correction during P300-based spelling. Advances in Human-Computer Interaction **2012**(4) (2012)
23. Schalk, G., McFarland, D.J., Hinterberger, T., Birbaumer, N., Wolpaw, J.R.: BCI2000: A general-purpose brain-computer interface (BCI) system. IEEE Transactions on Biomedical Engineering **51**(6), 1034–1043 (2004)
24. Shoeb, A.H.: Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment. Ph.D. thesis, Massachusetts Institute of Technology (2009)
25. Song, Y., Jia, X., Yang, L., Xie, L.: Transformer-based Spatial-Temporal Feature Learning for EEG Decoding (2021)
26. Song, Y., Zheng, Q., Liu, B., Gao, X.: EEGConformer: Convolutional Transformer for EEG Decoding and Visualization. IEEE Transactions on Neural Systems and Rehabilitation Engineering **31**, 710–719 (2023)
27. Stevenson, N.J., Tapani, K., Lauronen, L., Vanhatalo, S.: A dataset of neonatal EEG recordings with seizure annotations. Scientific Data **6**, 190039 (Mar 2019)
28. Wang, J., Zhao, S., Luo, Z., Zhou, Y., Jiang, H., Li, S., Li, T., Pan, G.: CBraMod: A Criss-Cross Brain Foundation Model for EEG Decoding. In: The Thirteenth International Conference on Learning Representations (2025)

29. Wang, S., et al.: Linformer: Self-attention with linear complexity (2020)
30. Yang, C., Westover, M.B., Sun, J.: BIOT: Cross-data Biosignal Learning in the Wild (2023)
31. Yang, C., et al.: Self-supervised EEG Representation Learning for Automatic Sleep Staging (2023)
32. Zheng, W.L., Liu, W., Lu, Y., Lu, B.L., Cichocki, A.: EmotionMeter: A Multimodal Framework for Recognizing Human Emotions. IEEE Transactions on Cybernetics pp. 1–13 (2018)
33. Zheng, W.L., Lu, B.L.: Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Transactions on Autonomous Mental Development **7**(3), 162–175 (2015)
34. Zyma, I., et al.: Electroencephalograms during mental arithmetic task performance. Data **4**(1), 14 (2019)