CaMBRAIN: Real-time, Continuous EEG Inference with Causal State Space Models

Anonymous authors
Paper under double-blind review

Abstract

Electroencephalography (EEG) signals are a critical, non-invasive method to monitor electrical brain activity. They play a major role in neuromuscular and biomechanical settings, as they are used in clinical settings to monitor sleep, seizure activity, muscular response, etc. However, the length of EEGs is limitless - they can span anywhere from a couple seconds to multiple hours. This serves as a major hurdle for existing deep learning methods due to two major factors: **(1)** existing EEG models are predominantly built upon the attention mechanism, incurring quadratic scaling as the sequence length increases, and **(2)** raw EEG signals must be processed in a sliding-window fashion due to current models' requirements of fixed-length input, preventing global understanding across the entire signal as a whole. To this extent, we propose CaMBRAIN, a **Ca**usal, **M**amba-based EEG model for real-time, streaming inference of **brain** EEG signals. We argue that bidirectional approaches to processing EEG signals are counterintuitive and needlessly expensive due to the causal, unidirectional nature of EEG. Thus, we propose a unidirectional, causal, state space model (SSM) for EEG understanding. However, training such a model is not straightforward, as crucial EEG events can be extremely brief - occurring within fractions of a second - yet may be separated by long intervals spanning minutes. While current EEG methods use self-supervised objectives to optimize for signal reconstruction, these objectives are not well suited for streaming state space models, they fail to explicitly train the hidden state to retain the salient long-range context needed for streaming inference. Therefore, we introduce a multi-stage self-supervised training pipeline specifically tailored to encourage long-range memory retention and strong performance on EEG signals, while preserving the linear-time complexity of state space models. CaMBRAIN achieves state-of-the-art (SOTA) results on multiple EEG benchmarks with 6–158$\times$ less sustained compute than EEG foundation model baselines, enabling long-range, continuous inference over variable-length EEG signals.

1 Introduction

Electroencephalography (EEG) is a widely used, non-invasive technique for measuring electrical activity in the brain (Biasiucci et al., 2019), playing a central role in clinical applications such as seizure detection (Tran et al., 2022; Tzallas et al., 2009) and cognitive behavioral analysis (Koelstra et al., 2011; Badr et al., 2024). Unlike many modalities in computer vision or natural language processing, EEG is inherently a continuous, high-frequency, multichannel temporal signal, often spanning minutes to hours of recording. This unique structure presents both an opportunity and a challenge: while EEG contains rich temporal dynamics across multiple frequency channels, effectively modeling these intricate, long-range dependencies efficiently remains difficult.

Recent advances in deep learning have significantly improved EEG analysis, with transformer-based architectures emerging as the dominant paradigm due to their strong performance in capturing temporal dependencies (Wen et al., 2023). However, this success comes with two critical limitations that hinder their applicability to real-world EEG settings. First, transformers incur quadratic computational complexity with respect to sequence length, making them fundamentally inefficient for long-duration EEG signals. Second, due to this unfavorable scaling, existing approaches typically rely on a sliding-window inference strategy where overlapping segments of a long signal are processed independently and their predictions are aggregated (Fig. 1). While effective on a small scale, this approach introduces substantial redundancy and computational overhead by repeatedly recomputing

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.

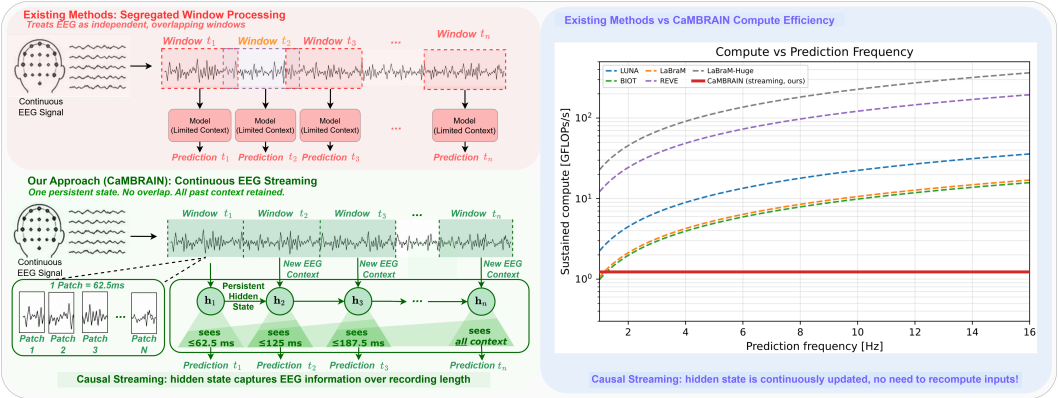

Figure 1: Existing methods (top) operate on overlapping sliding windows, repeatedly recomputing features over shared regions and limiting context to short segments. In contrast, CaMBRAIN (bottom) enables continuous EEG inference by maintaining a persistent hidden state that incrementally compresses all past context, eliminating redundant computation and enabling efficient, global temporal EEG understanding in real time (right). **Window sizes not-to-scale.**

overlapping regions. More importantly, it prevents the model from maintaining a persistent representation of long-range context, since each window is processed in isolation rather than as part of a continuous signal history.

These limitations are particularly problematic in real-time or continuous EEG monitoring, where both efficiency and long-range temporal reasoning are essential. In such settings, EEG should not be treated as a collection of independent segments, but as a continuous signal where each new observation builds on prior context. A model deployed in this regime must process incoming EEG without looking ahead at future signal, and must do so at a cost that does not grow with how long the recording has been running; equivalently, the entire relevant past must be retained in a compact memory that updates as the signal continues. This matches how clinicians monitor EEG at the bedside, and the need for continuous, efficient EEG processing has been acknowledged in prior work (El Ouahidi et al., 2025; Tegon et al., 2025).

In this work, we introduce CaMBRAIN, a causal state space model for EEG that enables real-time, continuous streaming inference. Built on a Mamba-3 architecture, our model processes EEG signals in consecutive 62.5 ms patches, even across windows, to update a persistent hidden state that compactly summarizes all past information. Unlike sliding-window approaches, CaMBRAIN does not revisit previous inputs; instead, it leverages the inherent property of state space models to encode long-range temporal context from previous patches directly within the hidden state, allowing efficient and scalable inference over arbitrarily long signals. This design aligns naturally with the causal and unidirectional nature of EEG, avoiding the need for bidirectional processing and redundant computation.

However, architectural choices alone do not fully address the demands of continuous EEG inference. The pretraining objective also shapes what the hidden state retains, and prior EEG foundation models (Fu et al., 2024; Banville et al., 2021; Mohammadi Foumani et al., 2024) have largely relied on self-supervised objectives. While such reconstruction-based objectives can capture local signal structure, they inherently prioritize learning short-range, signal fidelity at the masked regions rather than learning to compress and retain globally relevant context. This is misaligned with the demands of continuous inference, where the objective is also to selectively retain and propagate the most salient information forward in time. In the context of state space models, this distinction is critical: the hidden state must function as a compact, information-rich summary of the entire history, rather than a mechanism for reconstructing local signal details. To address this, we propose a representation-level self-supervised learning framework inspired by JEPA-style training, adapted to the challenges of continuous, causal EEG modeling. By combining multi-step autoregressive prediction with masked latent prediction under a strict unidirectional constraint, our approach explicitly encourages the model to encode predictive, informative, and long-range temporal features within its hidden state. We summarize our contributions as the following:

- We reformulate EEG inference as causal streaming: the model reads the signal once, in order, updating a persistent hidden state with each new patch instead of re-encoding fixed windows.
- We propose CaMBRAIN, a causal model that realizes this formulation in real time at constant cost per patch, and is competitive with or better than windowed EEG foundation models at a fraction of their compute.
- We introduce a self-supervised training method that teaches the hidden state to predict upcoming EEG representations, rather than reconstruct the raw signal, so it keeps the long-range context that streaming needs.

2 Related Works

2.1 Self-Supervised Learning in EEG

Deep learning has become increasingly effective for EEG analysis (Craik et al., 2019), with recent work shifting from task-specific supervised models toward large-scale self-supervised and foundation models. Early approaches such as BENDR (Kostas et al., 2021) adapted contrastive sequence pretraining to EEG, while BIOT (Yang et al., 2023a) introduced a transformer-based framework for cross-dataset biosignal learning. Other works focus on scaling EEG pretraining objectives: LaBraM (Jiang et al., 2024) uses vector-quantized neural spectrum prediction, while REVE (El Ouahidi et al., 2025) studies large-scale pretraining across heterogeneous EEG setups. In parallel, recent models have improved architectural flexibility for EEG, with CBraMod (Wang et al., 2024) modeling spatial and temporal dependencies through cross-attention and LUNA (Döner et al., 2025) introducing topology-agnostic channel representations for variable EEG montages. These methods have improved transfer across EEG datasets and downstream tasks, including abnormal EEG detection (Obeid & Picone, 2016), event detection (Obeid & Picone, 2016), sleep staging (Alvarez-Estevez & Rijsman, 2021), and cognitive-state prediction (Zyma et al., 2019; Mumtaz et al., 2017; Shoeb, 2009). However, despite their strong representation learning capabilities, most existing EEG models still operate on fixed-length segments and are typically applied to long recordings through sliding-window inference. This formulation treats a continuous EEG as a collection of independent or overlapping windows, requiring redundant computation over shared regions and limiting the model's ability to maintain a persistent representation of the entire signal history. In contrast, our work focuses on continuous EEG inference, where each new EEG patch updates a persistent hidden state without revisiting previous inputs. This setting is important for real-time clinical and wearable applications, where models must produce frequent predictions while preserving long-range context over long recordings.

2.2 State Space Models

State space models (SSMs) have recently emerged as efficient alternatives to attention for long-sequence modeling. Structured SSMs such as S4 (Gu et al., 2022) and S5 (Smith et al., 2023) showed that continuous-time dynamical systems can be parameterized for stable and efficient sequence learning, while Mamba (Gu & Dao, 2023) introduced selective state spaces that improve content-dependent modeling while retaining linear-time scaling. These properties make SSMs a natural fit for EEG, since EEG is a long, causal, temporally ordered signal with relevant events that may be sparse and separated by long intervals. Recent EEG models have begun to explore this direction: EEGMamba (Wang et al., 2025) and FEMBA (Tegon et al., 2025) apply bidirectional Mamba architectures to EEG foundation modeling. However, these methods typically operate on fixed-length windows and often rely on bidirectional scans or reconstruction-based pretraining, which limits their suitability for EEG streaming inference. CaMBRAIN instead uses a unidirectional causal Mamba-3 encoder whose hidden state is explicitly preserved across incoming EEG patches, enabling efficient real-time updates while avoiding both quadratic attention and bidirectional dependence on future context.

3 Methodology

We present **CaMBRAIN** (left, Fig. 2), a causal EEG foundation model for continuous streaming inference. Unlike prior approaches that rely on bidirectional processing and reconstruction-based objectives, our method operates in a unidirectional, forward-only manner and prioritizes learning representations rather than reconstructing signals for optimal streaming performance. Our method is built on two key principles: **(1)** causal, linear-time sequence modeling to enable continuous EEG

streaming, and **(2)** representation-level self-supervised learning to encourage the SSM hidden state to capture and retain salient EEG context over long temporal horizons. We first provide some background on state space models and EEG signals, followed by our three stages of training: (i) causal predictive pretraining (Sec. 3.2.1), (ii) latent prediction learning for long-horizon memory (Sec. 3.2.2), and (iii) downstream fine-tuning (Sec. 3.2.3).

3.1 PRELIMINARIES

EEG Patching. We represent an EEG segment as a multichannel time series $X \in \mathbb{R}^{C\times T}$, where C denotes the number of EEG channels and T denotes the number of temporal samples. Rather than operating directly on raw samples, EEGs are partitioned into non-overlapping patches spanning all channels and P time steps, producing $G_T = \lfloor T/P \rfloor$ patches. Each patch is then embedded using a transformer (Döner et al., 2025), yielding EEG tokens $\mathbf{X} = \{x_1, \ldots, x_{G_T}\}$, where each $x_t \in \mathbb{R}^d$ summarizes the EEG content at temporal index t across all channels. We refer to these embeddings as *EEG tokens*.

State Space Models. Structured state space models have emerged as an efficient alternative for sequence modeling, offering near-linear complexity with respect to sequence length while capturing long-range dependencies. At a high level, SSMs model a time-dependent input signal as the evolution of a latent dynamical system. Given an input sequence $x(t) \in \mathbb{R}^L$, an SSM maintains a latent state $h(t) \in \mathbb{R}^N$ that evolves over time and produces an output $y(t) \in \mathbb{R}^L$. To operate on discrete sequences, the continuous dynamics are discretized using a learnable step size Δ, yielding:

$$h_t = \bar{\mathbf{A}} h_{t-1} + \bar{\mathbf{B}} x_t, \qquad y_t = \bar{\mathbf{C}} h_t.$$

where $\bar{\mathbf{A}} \in \mathbb{R}^{N\times N}$ governs state transitions, and $\bar{\mathbf{B}}$ and $\bar{\mathbf{C}}$ denote input and output projections, respectively. After discretization, the same SSM can be evaluated either recurrently or as a global convolution. The recurrent form updates a hidden state one step at a time (left), while unrolling the recurrence yields a convolutional form (right):

$$\begin{aligned} h_t &= \bar{\mathbf{A}} h_{t-1} + \bar{\mathbf{B}} x_t, & \bar{\mathbf{K}} &= (\bar{\mathbf{C}}\bar{\mathbf{B}}, \bar{\mathbf{C}}\bar{\mathbf{A}}\bar{\mathbf{B}}, \cdots, \bar{\mathbf{C}}\bar{\mathbf{A}}^t\bar{\mathbf{B}}), \\ y_t &= \bar{\mathbf{C}} h_t, & y &= x * \bar{\mathbf{K}}. \end{aligned}$$

Thus, during training the sequence can be processed efficiently in parallel using the convolutional form, while at inference the recurrent form updates the hidden state online. This formulation naturally defines a *causal* model, parameterized by $\theta = \{\bar{\mathbf{A}}, \bar{\mathbf{B}}, \bar{\mathbf{C}}, \Delta\}$ where the hidden state at time step t depends only on past and current inputs $\{x_1, \ldots, x_t\}$. As a result, SSMs are well-suited for streaming settings, enabling autoregressive inference with constant memory and per-step compute. Due to the rich spectral structure across multiple frequency bands in EEG signals, we leverage a Mamba-3 architecture (Lahoti et al., 2026) for our causal encoder, which employs complex-valued state transitions to better capture oscillatory dynamics and phase relationships. In addition to this, we leverage SSM's ability to compress previous context into an accessible hidden state to achieve near real-time, continuous streaming inference of EEG signals with accurate long-range capabilities. Because each of these properties has direct implications for EEG modeling, we summarize the underlying recurrences and their EEG intuition in Appendix A.6.

3.2 CAMBRAIN

A key advantage of CaMBRAIN is its inherited ability from state space modeling to maintain and update a persistent hidden state, allowing new EEG samples to be processed continuously while retaining relevant historical information (see Tab. 3a). During inference, CaMBRAIN processes incoming EEG in 62.5ms patches, using its hidden state to make real-time predictions with context beyond the current patch. In contrast, prior EEG models (Kostas et al., 2021; Yang et al., 2023a; Song et al., 2022) operate on fixed-length windows, where each segment is processed independently. While such models can capture temporal structure within a window, they do not maintain an explicit state across windows and therefore require repeated reprocessing of overlapping windows to gain longer-context understanding of an EEG. Additionally, their reliance on bidirectional attention limits their applicability in strictly causal settings, as future context is required during inference. However, architecture alone does not guarantee effective streaming: the hidden state must also be trained to

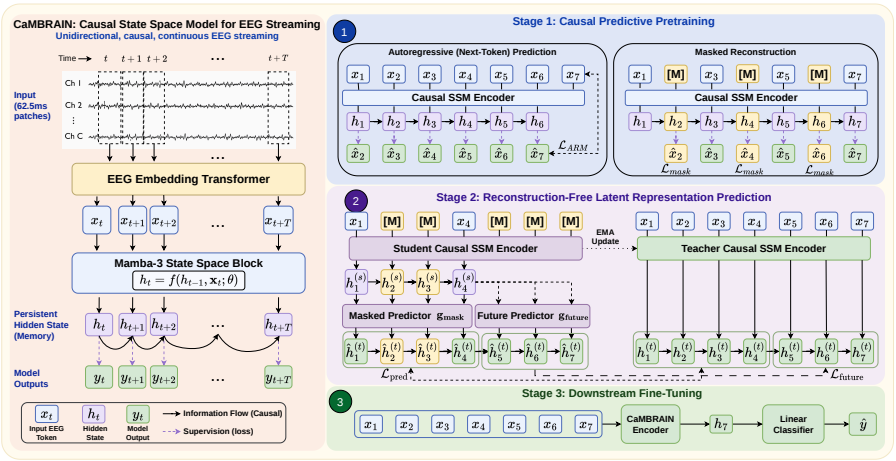

Figure 2: Overview of **CaMBRAIN**. CaMBRAIN is a causal state space model that processes EEG in 62.5 ms patches using a unidirectional state space model with a persistent hidden state. We train it with a three-stage pipeline: (1) causal predictive pretraining with autoregressive and masked reconstruction objectives to learn local temporal structure, (2) decoder-free latent JEPA training using a student–teacher framework to enforce long-range consistency and memory retention in the hidden state, and (3) downstream fine-tuning for task-specific prediction. This design enables real-time, continuous inference over arbitrarily long EEG recordings.

retain previous context and produce future predictions. We therefore design CaMBRAIN around both a causal streaming architecture and a training objective that encourages long-range memory retention. An overview of CaMBRAIN is provided in Fig. 2, followed by its training formulation below.

3.2.1 Stage 1: Causal Predictive Pretraining

The goal of Stage 1 is to instill general temporal structure in the encoder under a strict causal constraint, before any representation-level training. While one of our aims is to perform large-scale self-supervised learning without reconstruction, it is important that the model is first instilled with some general temporal understanding of EEGs for better performance (see Sec. 4.3). We pretrain the encoder using two complementary objectives applied in a strictly causal, forward-only manner: autoregressive modeling (ARM) and masked reconstruction (top-right, Fig. 2).

Autoregressive Modeling (ARM). Given a sequence of EEG token embeddings $\{x_t\}_{t=1}^{G_T}$, the model produces hidden states $\{h_t\}_{t=1}^{G_T}$, where each h_t depends only on past inputs $x_{\leq t}$. The autoregressive objective predicts the embedding of the next patch from the current hidden state, $\hat{x}_{t+1} = g(h_t)$, where $g(\cdot)$ is a learnable prediction head. The ARM loss is defined as:

$$\mathcal{L}_{\text{ARM}} = \sum_{t=1}^{G_T-1} \ell(\hat{x}_{t+1}, x_{t+1}),$$

where the target embeddings $\{x_{t+1}\}$ are obtained from the unmasked signal. While this next-patch prediction objective encourages the model to capture the temporal dynamics and evolution of EEG signals, it primarily enforces consistency across time and does not explicitly encourage the learning of robust local representations. To complement this, we combine ARM with a masked modeling objective.

Masked Reconstruction. To aid the model in capturing local structure, we additionally apply a masked reconstruction objective which focuses on reconstructing missing information from the causal context provided by the encoder hidden states. Masking is performed at the token level after EEG embedding. Specifically, a subset of EEG tokens are selected, and all values within each selected token are masked with zeros. The masked tokens are passed through the encoder, and a decoder $d(\cdot)$ reconstructs the original EEG signal from the masked tokens: $\hat{X} = d(h_{1:G_T})$. Given the set of masked positions $\mathcal{M}$, the reconstruction loss is applied only on masked regions:

$$\mathcal{L}_{\text{mask}} = \ell(\hat{X}, X; \mathcal{M}).$$

The overall Stage 1 objective combines both losses:

$$\mathcal{L}_{\text{Stage1}} = \lambda_{\text{ARM}}\mathcal{L}_{\text{ARM}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}}, \tag{1}$$

where λ_{ARM} and λ_{mask} control the relative weighting of the two objectives.

While Stage 1 instills strong local temporal structure in the encoder, it trains a single property of the model: predicting and reconstructing patches from their immediate causal context. As previously mentioned, these objectives do not explicitly encourage the model to compress and retain the most salient information over long temporal horizons - an especially important attribute for continuous streaming inference. Moreover, prior approaches require a decoder even at downstream time, increasing inference overhead. To address this, we propose a latent prediction framework for EEG streaming in Stage 2 that directly optimizes representation quality without decoder-based reconstruction.

3.2.2 Stage 2: Reconstruction-free Causal Latent Representation Refinement

Our second stage of training tasks the model to predict target representations produced by a separate network (middle-right, Fig. 2), directly encouraging the hidden state to encode informative features of the EEG signal. Stage 1 provides a useful initialization by training the encoder to recover raw EEG content, encouraging sensitivity to local waveform structure. However, raw-signal reconstruction is a limited objective for streaming EEG decoding: it mainly rewards fidelity to short-range signal details, much of which in EEG is noise, artifact or drift, and it is poorly suited for supervising what the hidden state should preserve over long horizons. We therefore adopt a JEPA-style (Assran et al., 2023) training framework, in which the teacher observes the unmasked input signal, while the student operates on a partially masked view and learns to predict the corresponding teacher representations through two heads: a masked predictor g_{mask} for masked latent modeling and a future predictor g_{future} for multi-step future latent prediction. The teacher is implemented as an exponential moving average (EMA) of the student and is not updated through backpropagation, providing a stable and slowly evolving target.

Building on this idea, we adapt the framework to streaming EEG decoding in four ways that distinguish it from JEPA. First, student and teacher are both causal, so the target at position t depends only on patches $1, \ldots, t$; JEPA's encoders attend over the whole input, which in EEG would let the later course of an event, such as a seizure's evolution, shape the representation of its onset, information an online decoder never has. Second, each prediction is read from a single student hidden state $h_t^{(s)}$ by a lightweight head rather than by a predictor network attending over all visible tokens; because EEG events are sparse and can be minutes apart, the information needed to decode them must persist in the state itself, and supervising the state directly pushes it there. Third, besides masked positions, the student predicts the next K teacher latents; under within-input masking, a masked EEG patch is typically flanked by visible neighbors and, because EEG is strongly correlated over short lags, can often be filled in by interpolation, whereas forecasting from a causal state requires encoding the slower dynamics of the signal. Fourth, Stage 2 starts from the Stage-1 encoder rather than from random initialization, so the teacher's first targets already reflect EEG waveform structure rather than random projections of a noisy signal. Our process is implemented as follows:

Student–Teacher Framework. We maintain two networks with identical architectures: a student network with parameters θ and a teacher network with parameters ϕ. The teacher is initialized as a copy of the student and is updated as an exponential moving average (EMA) of the student parameters:

$$\phi \leftarrow \tau\phi + (1-\tau)\theta,$$

where $\tau \in [0, 1)$ is a momentum parameter that is gradually increased during training. The teacher network is not updated via backpropagation and serves as a stable target for representation learning. The student processes a randomly masked version of the input EEG signal, while the teacher receives the unmasked signal. Both networks produce sequences of latent representations $\{h_t^{(s)}\}_{t=1}^{G_T}$ and $\{h_t^{(t)}\}_{t=1}^{G_T}$. At masked positions $\mathcal{M}$, the student is trained to predict the corresponding teacher representations using the masked predictor g_{mask}:

$$\hat{h}_t^{(t)} = g_{\text{mask}}(h_t^{(s)}), \quad t \in \mathcal{M},$$

and is optimized with a Smooth L1 loss:

$$L_{\text{pred}} = \sum_{t \in \mathcal{M}} \text{SmoothL1}\left(\hat{h}_t^{(t)}, h_t^{(t)}\right).$$

This asymmetry forces the student to infer missing local EEG content from its available causal context, encouraging the hidden state to encode informative representations from partial observations. However, this objective alone does not enforce that the hidden state encodes information predictive of future EEG dynamics.

Multi-Step Autoregressive Prediction. While masked latent prediction encourages the model to infer missing EEG content, it does not explicitly enforce that the hidden state captures the temporal dynamics of the signal. In particular, accurately reconstructing masked regions does not guarantee that the model can predict future EEG activity from its current hidden state, especially over extended time horizons. To address this, we extend the autoregressive objective to multiple future steps. Given a hidden state $h_t^{(s)}$, the model predicts the next K latent representations from a single hidden state using a separate future predictor g_{future}:

$$\hat{h}_{t+1:t+K}^{(t)} = g_{\text{future}}(h_t^{(s)}),$$

where g_{future} is a lightweight prediction head shared across time steps. The targets are the corresponding future teacher representations, which provide a stable reference without exposing future information to the student encoder:

$$L_{\text{future}} = \sum_{t=1}^{G_T-K} \sum_{k=1}^{K} \ell\left(\hat{h}_{t+k}^{(t)}, h_{t+k}^{(t)}\right).$$

By requiring the model to predict multiple future representations from a single hidden state, this objective encourages the hidden state to encode information that is predictive of the long-range temporal dynamics of the EEG signal, rather than relying solely on local context. The final Stage 2 loss combines both objectives:

$$L_{\text{Stage2}} = L_{\text{pred}} + L_{\text{future}}.$$

3.2.3 Stage 3: Downstream Fine-Tuning

After pretraining, we adapt the model to downstream EEG tasks using supervised fine-tuning (bottom-right, Fig. 2). Given an input EEG sequence, the pretrained encoder produces a sequence of latent representations. Unlike prior window-based or attention-based EEG models, which typically require pooling or aggregating token representations across a fixed input segment, CaMBRAIN can use just the final hidden state for classification. In particular, successive 62.5 ms EEG patches can be processed online by updating the hidden state without revisiting past inputs, and predictions can be produced directly from the current hidden state without storing or pooling over previous tokens. We therefore pass the final hidden state to a lightweight linear classification head for task-specific prediction. Following previous works (Döner et al., 2025; El Ouahidi et al., 2025), we fine-tune the entire model end-to-end using standard supervision. Full implementation details are given in Sec. A.2

4 Experiments

4.1 Datasets and Evaluation Protocol

We evaluate CaMBRAIN on a diverse set of EEG benchmarks spanning artifact detection, abnormality classification, mental stress detection, and seizure detection, including TUAR (Obeid & Picone, 2016), TUAB (Obeid & Picone, 2016), MAT (Zyma et al., 2019), and CHB-MIT (Shoeb, 2009). These datasets cover a range of clinical and real-world EEG settings, varying in subject populations, recording conditions, and label distributions. This diversity allows us to assess performance across tasks, populations, and recording setups (splits in Appendix A.3). Following prior work, we treat each dataset as a downstream task and fine-tune our pretrained model accordingly. Importantly, our training pipeline is designed to learn general-purpose EEG representations, and downstream fine-tuning serves to evaluate how well these representations transfer across tasks and domains. We report standard evaluation metrics for EEG classification, including area under the ROC curve (AUROC) and area under the precision-recall curve (AUPR), which are common in EEG tasks such as seizure detection and artifact classification.

4.2 Results

Despite avoiding overlapping windows and redundant recomputation, CaMBRAIN achieves competitive or superior performance across all benchmarks. On TUAR artifact detection and CHB-MIT

seizure detection (Table 1), CaMBRAIN obtains the best AUROC among compared methods, reaching 0.936 on TUAR and best mean of 0.921 on CHB-MIT. These gains are especially notable for CHB-MIT, where seizure events are highly imbalanced and long-range temporal context is important.

We further evaluate CaMBRAIN on TUAB abnormal EEG detection and MAT mental stress detection (Table 2). CaMBRAIN remains competitive with substantially larger EEG foundation models and achieves the best MAT balanced accuracy and AUROC. Overall, these results suggest that persistent-state streaming and representation-level training provide strong cross-task EEG representations while avoiding the redundant computation of sliding-window inference. look at A.10 for evaluation protocol and splits.

Compute efficiency at deployment. Table 1b reports sustained compute at the 16 Hz streaming update rate (62.5 ms patches), the load a real-time monitor must support. CaMBRAIN's 1.23 GFLOPs/s is between 6.1× (CBraMod) and 158× (REVE) below every foundation-model baseline while matching or exceeding their AUROC. The gap is structural, not size-driven: CBraMod is 9× smaller than CaMBRAIN yet needs 6× more sustained compute, because attention costs $\mathcal{O}(T)$ per query where the SSM step is $\mathcal{O}(1)$, and persistent state avoids recomputing overlapping windows. A state space backbone alone does not remove the second cost: EEGMamba, also built from Mamba blocks and 11× smaller than CaMBRAIN, re-encodes its window at every update and needs 7.60 GFLOPs/s (6.1× CaMBRAIN), while CaMBRAIN also scores higher on CHB-MIT (AUROC 0.921 vs. 0.908; AUC-PR 0.389 vs. 0.358).

Overall, the results demonstrate that CaMBRAIN benefits from both its streaming architecture and representation-level training strategy. CaMBRAIN avoids the need for overlapping window-based processing and the associated redundant computation. Furthermore, the combination of masked latent prediction and multi-step autoregressive objectives encourages the hidden state to capture predictive, long-range temporal structure, rather than local signal fidelity.

Table 1: Performance comparison on TUAR artifact detection and CHB-MIT seizure detection.

(a) TUAR artifact detection (MMC Objective).

Model	Size	AUROC ↑	AUC-PR ↑
Supervised Models			
EEGNet (Lawhern et al., 2018)	—	0.752	0.433
EEG-GNN (Demir et al., 2021)	—	0.837	0.488
GraphS4mer (Tang et al., 2023)	—	0.833	0.461
Self-supervised Models			
BrainBERT (Wang et al., 2023)	43M	0.753	0.350
EEGFormer-B (Chen et al., 2024)	2.3M	0.847	0.483
EEGFormer-L (Chen et al., 2024)	3.2M	0.852	0.483
LUNA-B (Döner et al., 2025)	7M	0.902	0.495
LUNA-H (Döner et al., 2025)	311M	0.921	0.528
CaMBRAIN	**37M**	**0.936**	**0.565**

(b) CHB-MIT seizure detection. GFLOPs/s reports sustained compute at the 16 Hz streaming update rate; lower is better.

Model	Size	AUROC ↑	AUC-PR ↑	GFLOPs/s ↓
Supervised Models				
EEGNet (Lawhern et al., 2018)	—	0.805	0.191	—
ST-Transformer (Song et al., 2021)	3.5M	0.824	0.142	—
Self-supervised Models				
BENDR (Kostas et al., 2021)	0.39M	0.863	0.307	—
BIOT (Yang et al., 2023a)	3.2M	0.876	0.328	15.80
LaBraM (Jiang et al., 2024)	5.8M	0.868	0.329	16.99
CBraMod (Wang et al., 2024)	4.0M	0.889	0.369	7.51
LUNA-L (Döner et al., 2025)	43M	0.896	0.316	35.85
EEGMamba (Wang et al., 2025)	3.3M	0.908	0.358	7.60
REVE (El Ouahidi et al., 2025)	72M	0.908	0.380	194.79
CaMBRAIN	**37M**	**0.921**	**0.389**	**1.23**

Table 2: Comparison on TUAB abnormal EEG detection and MAT mental stress detection.

(a) TUAB abnormal EEG detection.

Model	Size	AUROC ↑	AUC-PR ↑
Supervised Models			
SPaRCNet (Jing et al., 2023)	0.8M	0.867	0.841
ContraWR (Yang et al., 2023b)	1.6M	0.845	0.842
CNN-Transformer (Peh et al., 2022)	3.2M	0.846	0.843
ST-Transformer (Song et al., 2021)	3.5M	0.870	0.852
Self-supervised Models			
BrainBERT (Wang et al., 2023)	43.2M	0.853	0.846
EEGFormer (Chen et al., 2024)	2.3M	0.867	0.867
BIOT (Yang et al., 2023a)	3.2M	0.881	0.879
FEMBA (Tegon et al., 2025)	47.7M	0.882	0.889
CEReBrO (Dimofte et al., 2025)	85.2M	0.891	0.904
LaBraM (Jiang et al., 2024)	5.8M	0.902	0.896
CBraMod (Wang et al., 2024)	4.0M	**0.915**	**0.922**
LUNA (Döner et al., 2025)	7M	0.886	0.895
CaMBRAIN	**37M**	0.867	0.876

(b) MAT mental stress detection.

Method	AUROC ↑	Bal. Acc. ↑
EEGNet (Lawhern et al., 2018)	0.732	0.677
EEGConformer (Song et al., 2022)	0.742	0.680
SPaRCNet (Jing et al., 2023)	0.741	0.687
ContraWR (Yang et al., 2023b)	0.733	0.663
CNN-Transformer(Peh et al., 2022)	0.725	0.677
FFCL (Li et al., 2022)	0.733	0.679
ST-Transformer (Song et al., 2021)	0.713	0.663
BIOT (Yang et al., 2023a)	0.753	0.687
LaBraM-B (Jiang et al., 2024)	0.772	0.690
CBraMod (Wang et al., 2024)	0.790	0.725
REVE-B (El Ouahidi et al., 2025)	0.845	0.766
CaMBRAIN	**0.876**	**0.778**

4.3 Ablations

4.3.1 Ablation on Hidden State

To exhibit the value of persistent memory, we compare persistent state inference, where CaMBRAIN carries its hidden state across the full EEG recording, against windowed inference, where the state is reset every 5 s to mimic conventional windowed evaluation. As shown in Table 3a, persistent state inference consistently improves both onset localization and patch-level classification, a $+0.015$ increase in AUROC. These gains confirm that CaMBRAIN benefits not only from efficient inference, but from allowing the hidden state to compress and retain long-range EEG history. This supports our central claim: **treating EEG as a continuous stream provides useful temporal context that is lost when models process each window independently.**

4.3.2 Ablation on Pretraining Stages

We further ablate the impact of our self-supervised pretraining strategy in Table 3b. The first row corresponds to standard downstream fine-tuning on TUAR from random initialization, which yields the weakest performance. Initializing from Stage 1 pretraining alone provides a modest improvement (+0.60 AUROC, +2.80 AUPR), indicating that reconstruction-based pretraining improves general EEG feature extraction. In contrast, our full two-stage approach achieves substantially larger gains (+2.30 AUROC, +10.9 AUPR over standard fine-tuning). These results show that while Stage 1 provides a useful initialization, Stage 2 is critical for learning representations that transfer effectively to downstream artifact detection. Unlike reconstruction-based objectives, which emphasize local signal fidelity, our Stage 2 training directly optimizes latent representations across time, encouraging the hidden state to encode invariant and temporally relevant information necessary for continuous EEG inference.

Table 3: Ablations on our proposed pretraining stages and the impact of a persistent hidden state for continuous EEG inference evaluation.

(a) Value of persistent hidden-state memory during continuous CHB-MIT evaluation. Streaming inference carries the hidden state across the full recording, while windowed inference resets the state every 5 s. Scores scaled by 100

Metric	Continuous	Windowed	Δ
Seizure onset			
Mean prob. at onset (↑)	25.2	11.8	**+13.5**
Peak prob. near onset (↑)	38.3	27.6	**+10.8**
Prob. at onset patch (↑)	21.6	6.20	**+15.4**
Sparse patch classification			
AUROC (↑)	88.9	87.4	**+1.50**
BAC, $T = 0.5$ (↑)	72.3	66.6	**+6.30**

(b) Pretraining ablation results on TUAR MCC 5-class artifact detection, with scores scaled by 100. The first row shows standard downstream fine-tuning, the second adds Stage 1 pretraining, and the final row uses both Stage 1 and Stage 2. Strongest performance is obtained only when both stages are combined.

Stage 1	Stage 2	AUROC ↑	AUPR ↑
✗	✗	93.7	75.0
✓	✗	94.3 **(+0.60)**	77.8 **(+2.80)**
✓	✓	**96.0** **(+2.30)**	**85.9** **(+10.9)**

5 Conclusion

We introduced CaMBRAIN, a causal state space model for EEG that enables real-time, continuous streaming inference over arbitrarily long signals. By leveraging a Mamba-3 architecture, CaMBRAIN maintains a persistent hidden state that compactly summarizes all past EEG context, eliminating the need for sliding-window processing and redundant computation. In addition, we proposed a JEPA-style, representation-level self-supervised training framework that encourages the model to encode predictive, invariant, and long-range temporal information within its hidden state, rather than focusing on local signal reconstruction. Across a diverse set of EEG benchmarks spanning artifact detection, abnormality classification, seizure detection, and mental stress detection, CaMBRAIN achieves strong performance while providing substantial improvements in computational efficiency. These results demonstrate that modeling EEG as a continuous stream, combined with representation-level learning, leads to more effective and scalable solutions compared to existing approaches. Overall, this work highlights the importance of aligning both model architecture and training objectives with the inherent properties of EEG signals.

AI USE STATEMENT

In this work, we used generative AI tools only to assist with writing and grammar in the manuscript text. We have not used generative AI tools for any task that requires disclosure under the ICLR 2027 AI Policy for Authors: generating synthetic data, developing theoretical models or conceptual frameworks, formulating or proving mathematical claims, proposing or refining hypotheses, designing or giving feedback on research methodology or experiments, implementing methods, translation, cleaning or reformatting datasets, qualitative or thematic data analysis, or interpreting results. We have reviewed all AI-assisted text, and we take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only publicly available, de-identified EEG corpora (TUEG, TUAR, and TUAB from the Temple University Hospital EEG corpus (Obeid & Picone, 2016), CHB-MIT (Shoeb, 2009), and MAT (Zyma et al., 2019)), accessed under their respective data use terms; no new human-subject data were collected. CaMBRAIN is a research model and has not been clinically validated. Its outputs, including seizure-detection probabilities, should not be used for clinical decision-making without prospective validation and regulatory review.

REPRODUCIBILITY STATEMENT

Architecture, optimization hyperparameters for both pretraining stages and fine-tuning, the pretraining corpus, and hardware are given in Appendix A.2. Dataset splits and the full preprocessing pipeline, including causal robust quartile normalization, are described in Appendix A.3. The streaming inference procedure is given in Algorithm 1, and streaming-parallel numerical equivalence is verified in Appendix A.5. Per-seed means and standard deviations for all downstream results are reported in Table 8.

REFERENCES

Diego Alvarez-Estevez and Roselyne M Rijsman. Inter-database validation of a deep learning approach for automatic sleep scoring. *PloS one*, 16(8):e0256111, 2021.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15619–15629, 2023.

Yara Badr, Usman Tariq, Fares Al-Shargie, Fabio Babiloni, Fadwa Al Mughairbi, and Hasan Al-Nashash. A review on evaluating mental stress by deep learning using eeg signals. *Neural Computing and Applications*, 36(21):12629–12654, 2024.

Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. Uncovering the structure of clinical eeg signals with self-supervised learning. *Journal of Neural Engineering*, 18(4):046020, 2021.

Andrea Biasiucci, Benedetta Franceschiello, and Micah M Murray. Electroencephalography. *Current Biology*, 29(3):R80–R85, 2019.

Yuqi Chen, Kan Ren, Kaitao Song, Yansen Wang, Yifan Wang, Dongsheng Li, and Lili Qiu. Eegformer: Towards transferable and interpretable large-scale eeg foundation model. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024.

Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. Deep learning for electroencephalogram (eeg) classification tasks: a review. *Journal of neural engineering*, 16(3):031001, 2019.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

Andac Demir, Toshiaki Koike-Akino, Ye Wang, Masaki Haruna, and Deniz Erdogmus. Eeg-gnn: Graph neural networks for classification of electroencephalogram (eeg) signals. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1061–1067. IEEE, 2021.

Alexandru Dimofte, Glenn Anta Bucagu, Thorir Mar Ingolfsson, Xiaying Wang, Andrea Cossettini, Luca Benini, and Yawei Li. Cerebro: Compact encoder for representations of brain oscillations using efficient alternating attention. *arXiv preprint arXiv:2501.10885*, 2025.

Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. Luna: Efficient and topology-agnostic foundation model for eeg signal analysis. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. Reve: A foundation model for eeg-adapting to any setup with large-scale pretraining on 25,000 subjects. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Zanhao Fu, Huaiyu Zhu, Yisheng Zhao, Ruohong Huan, Yi Zhang, Shuohui Chen, and Yun Pan. Gmaeeg: A self-supervised graph masked autoencoder for eeg representation learning. *IEEE Journal of Biomedical and Health Informatics*, 28(11):6486–6497, 2024.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*, 2024.

Jin Jing, Wendong Ge, Shenda Hong, Marta Bento Fernandes, Zhen Lin, Chaoqi Yang, Sungtae An, Aaron F Struck, Aline Herlopian, Ioannis Karakis, et al. Development of expert-level classification of seizures and rhythmic and periodic patterns during eeg interpretation. *Neurology*, 100(17): e1750–e1762, 2023.

Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

Aakash Lahoti, Kevin Li, Berlin Chen, Caitlin Wang, Aviv Bick, J Zico Kolter, Tri Dao, and Albert Gu. Mamba-3: Improved sequence modeling using state space principles. In *The Fourteenth International Conference on Learning Representations*, 2026.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

Hongli Li, Man Ding, Ronghua Zhang, and Chunbo Xiu. Motor imagery eeg classification algorithm based on cnn-lstm feature fusion network. *Biomedical signal processing and control*, 72:103342, 2022.

Navid Mohammadi Foumani, Geoffrey Mackellar, Soheila Ghane, Saad Irtza, Nam Nguyen, and Mahsa Salehi. Eeg2rep: enhancing self-supervised eeg representation through informative masked inputs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5544–5555, 2024.

Wajid Mumtaz, Likun Xia, Syed Saad Azhar Ali, Mohd Azhar Mohd Yasin, Muhammad Hussain, and Aamir Saeed Malik. Electroencephalogram (eeg)-based computer-aided technique to diagnose major depressive disorder (mdd). *Biomedical Signal Processing and Control*, 31:108–115, 2017.

Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in neuroscience*, 10:196, 2016.

Wei Yan Peh, Yuanyuan Yao, and Justin Dauwels. Transformer convolutional neural networks for automated artifact detection in scalp eeg. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3599–3602. IEEE, 2022.

Ali Hossam Shoeb. *Application of machine learning to epileptic seizure onset detection and treatment*. PhD thesis, Massachusetts Institute of Technology, 2009.

Jimmy TH Smith, Andrew Warrington, and Scott Linderman. Simplified state space layers for sequence modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie. Transformer-based spatial-temporal feature learning for eeg decoding. *arXiv preprint arXiv:2106.11170*, 2021.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022.

Siyi Tang, Jared Dunnmon, Liangqiong Qu, Khaled Kamal Saab, Tina Baykaner, Christopher Lee-Messer, and Daniel Rubin. Modeling multivariate biosignals with graph neural networks and structured state space. In *ICLR 2023 Workshop on Time Series Representation Learning for Health*, 2023.

Anna Tegon, Thorir Mar Ingolfsson, Xiaying Wang, Luca Benini, and Yawei Li. Femba: Efficient and scalable eeg analysis with a bidirectional mamba foundation model. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–7. IEEE, 2025.

Ly V Tran, Hieu M Tran, Tuan M Le, Tri TM Huynh, Hung T Tran, and Son VT Dao. Application of machine learning in epileptic seizure detection. *Diagnostics*, 12(11):2879, 2022.

Alexandros T Tzallas, Markos G Tsipouras, and Dimitrios I Fotiadis. Epileptic seizure detection in eegs using time–frequency analysis. *IEEE transactions on information technology in biomedicine*, 13(5):703–710, 2009.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Brainbert: Self-supervised representation learning for intracranial recordings. In *The Eleventh International Conference on Learning Representations*, 2023.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. *arXiv preprint arXiv:2412.07236*, 2024.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Shijian Li, and Gang Pan. Eegmamba: An eeg foundation model with mamba. *Neural Networks*, pp. 107816, 2025.

Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: a survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 6778–6786, 2023.

Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36:78240–78260, 2023a.

Chaoqi Yang, Cao Xiao, M Brandon Westover, and Jimeng Sun. Self-supervised electroencephalogram representation learning for automatic sleep staging: model development and evaluation study. *JMIR AI*, 2(1):e46769, 2023b.

Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1):14, 2019.

A Appendix

A.1 Limitations

CaMBRAIN's design and evaluation leave several directions open for future work. First, while CaMBRAIN sets the new best result on TUAR, MAT, and CHB-MIT, it remains competitive with rather than ahead of the strongest foundation models on TUAB abnormal-EEG detection (Table 2a). Closing this gap is a natural next step. Second, the scale of our pretraining corpus $\sim$21k hours of TUEG is substantially smaller than that of recent EEG foundation models such as REVE (El Ouahidi et al., 2025), which pretrains on a much larger and more heterogeneous corpus spanning many sites and montages. Scaling CaMBRAIN's two-stage objective to corpora of comparable size, and across more diverse acquisition setups, is a direction which may yield further gains, particularly on tasks where TUEG underrepresents the target distribution.

A.2 Implementation Details

Architecture. CaMBRAIN has 36.7 M parameters with model dimension $d_{\text{model}} = 704$, 4 Mamba-3 blocks, state size $d_{\text{state}} = 64$, head dimension $d_{\text{head}} = 64$ (giving 11 heads), and SwiGLU FFN expansion ratio 4. The channel embedder is a cross-attention module inspired from (Döner et al., 2025) with $Q = 4$ learnable queries and $H = 4$ attention heads. Each block applies pre-RMSNorm followed by a residual connection: $z \leftarrow z + \textsc{MambaStep}(\textsc{RMSNorm}(z))$ then $z \leftarrow z + \textsc{SwiGLU}(\textsc{RMSNorm}(z))$. Patches are 16 samples (62.5 ms at 256 Hz) and a learned cyclic positional embedding indexed by $\tau \bmod 80$ is added at the input of the encoder, where 80 is the patch count of a 5 s pretraining window. Classification reads the last-token hidden state through a linear head.

Pretraining Stage 1 (Hybrid Causal Loss). The encoder is initialized as a strictly forward-only Mamba-3 stack with a Conv2d input adapter (22 channels, kernel 1×16). Two losses are combined: causal autoregressive prediction of next-patch latents (ARM) and masked raw-EEG reconstruction with a lightweight decoder. Mask ratio is 0.4, loss weights $\lambda_{\text{ARM}} = \lambda_{\text{mask}} = 0.5$. We use the LAMB optimizer with peak learning rate 1×10^{-3}, weight decay 0.01, cosine schedule, 5 warmup epochs, batch size 1024 (grad-accumulated), and bf16-mixed precision.

Pretraining Stage 2 (EMA-JEPA). The student loss is the sum of (i) multi-step latent prediction with horizon $K = 4$ and (ii) block-masked latent prediction against an EMA teacher with momentum $\tau_{\text{ema}} = 0.9999$ ramped from 0.99 over the first 5% of training. Both losses are Smooth-L_1. We use AdamW with peak learning rate 1×10^{-4}, weight decay 0.05, $\beta = (0.9, 0.999)$, and the same cosine schedule. Both pretraining stages are strictly causal: target latents at position t are produced by the teacher conditioned only on patches $\{1, \ldots, t\}$.

Pretraining corpus. Stage 1 and Stage 2 both use the Temple University Hospital EEG corpus (TUEG, $\sim$21k hours), preprocessed with the pipeline described in Section A.3. Subjects appearing in any downstream evaluation set (TUAR, TUAB) are held out from pretraining.

Finetuning. We fine-tune the entire model end-to-end with AdamW, layer-wise learning-rate decay 0.75 (encoder relative to head), cosine schedule with 5% warmup, weight decay 0.05. Per-task loss is cross-entropy for multiclass tasks (TUAR), binary cross-entropy for binary tasks (TUAB, MAT, CHB-MIT). Three random seeds ({42, 43, 44}) per task; we report mean values in the main text.

Hardware. Pretraining on 8×H100 PCIe; finetuning and all inference benchmarks on a single H100 NVIDIA GPU. Streaming inference is implemented in PyTorch with the official `mamba_ssm` package (Lahoti et al., 2026) compiled with CUDA 13.

A.3 Datasets and Preprocessing

Table 4 summarizes the four downstream datasets. All recordings undergo a uniform preprocessing pipeline before normalization: a 4th-order zero-phase Butterworth bandpass between 0.1 and 75 Hz, a 60 Hz notch filter for line interference, and resampling to 256 Hz. For TUH-family datasets

(TUAR, TUAB) we form a 22-channel bipolar double-banana montage by differencing the standard longitudinal pairs. CHB-MIT is delivered in bipolar form (16 channels) and is left as-is. MAT uses the original 19-channel 10–20 unipolar montage.

Causal Sliding Robust Quartile Normalization (RQN). At each sample time t and for each channel c, we compute the median $m_{c,t}$ and interquartile range $\mathrm{IQR}_{c,t} = Q_3 - Q_1$ over the past $W = 1280$ samples (a 5 s causal sliding buffer at 256 Hz), and normalize

$$\hat{x}_{c,t} = \frac{x_{c,t} - m_{c,t}}{\mathrm{IQR}_{c,t} + \varepsilon},$$

with $\varepsilon = 10^{-6}$. Only samples in $[\max(0,\, t - W + 1),\, t]$ enter the statistics, so the normalization is strictly causal: a sample at time t never depends on a sample at $t' > t$. Compared to per-window z-scoring (Döner et al., 2025), quantile-based statistics are insensitive to high-amplitude artifacts (eye blinks, muscle bursts, electrode pops) that dominate the channel-wise mean and standard deviation but not the quartiles. During parallel-mode processing (pretraining and finetuning, where the model sees complete 5 s windows), we use standard window-level RQN: a single median and IQR per channel per window. During streaming inference, the median and IQR are maintained over a circular buffer of the most recent W samples and updated each patch.

Table 4: Downstream datasets and splits. TUAR uses a file-level random split (train/val/test = 80/10/10) following Döner et al. (2025) and Chen et al. (2024). All other test sets are subject-disjoint from training: TUAB uses the official TUH train/evaluation split; CHB-MIT uses patients chb01–chb19 for training, chb20–chb21 for validation, and chb22–chb24 for testing; MAT uses a 26/5/5 subject split.

Dataset	Subjects	Channels	Task	Window	Class balance
TUAR (Obeid & Picone, 2016)	213	22 (bipolar)	6-class artifact (MMC)	5 s	artifact families
TUAB (Obeid & Picone, 2016)	2,329	22 (bipolar)	binary normal/abnormal	5 s	∼50/50
MAT (Zyma et al., 2019)	36	19 (10–20)	binary rest/stress	5 s	balanced
CHB-MIT (Shoeb, 2009)	24	16 (bipolar)	binary seizure	5 s	0.42% positive

A.4 Persistent vs. Windowed Streaming: Single-Recording Visualization

Figure 3 illustrates the persistent-vs-windowed ablation on chb22_20, a representative recording from the CHB-MIT test set. The middle panel shows that the persistent-state model begins climbing toward the seizure threshold approximately 15 s before annotated onset and saturates near probability 1.0 throughout the ictal period. The windowed model, restricted to a 5 s context window, remains near zero until onset and exhibits a substantial probability dip (down to ~ 0.3) during a segment of high-amplitude ictal noise around $t = 8$–13 s — a regime where the persistent hidden state, having integrated tens of seconds of pre-ictal context, maintains its prediction. The bottom panel shows per-block hidden-state norms across the recording. Norms remain stable across the full 45 s window with no drift, indicating that the long-range integration the persistent-state model exploits does not come at the cost of unbounded state magnitude growth. Per-recording AUROC on chb22_20: persistent 0.96, windowed 0.93.

A.5 Streaming-Parallel Equivalence

Mamba's parallel scan and recurrent step are mathematically equivalent on a complete input (Dao & Gu, 2024; Lahoti et al., 2026). We verify this numerically for the full CaMBRAIN pipeline (encoder + head) across 30 random input seeds and the 3 CHB-MIT-fine-tuned checkpoints from our main results.

The probability-space difference is below the resolution of any downstream metric in this paper.

A.6 Mamba-3 Innovations and EEG Intuition

We adopt three innovations from Mamba-3 (Lahoti et al., 2026) unchanged in our encoder. None is novel to this work or designed for EEG, but each targets a structural property of streaming clinical EEG. Full derivations are in Lahoti et al. (2026); below we summarize the EEG intuition for each.

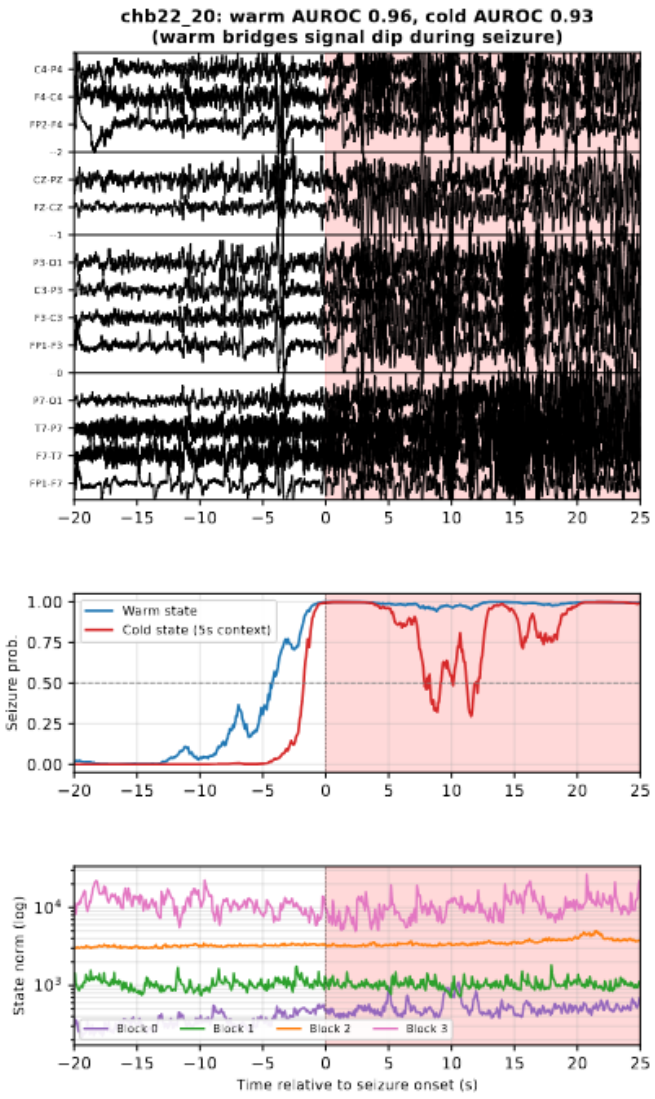

Figure 3: Persistent vs. windowed streaming on chb22_20. **Top**: raw EEG, with the annotated seizure period shaded. **Middle**: per-patch seizure probability for persistent streaming (blue, labeled "warm"; hidden state carried from $t = 0$ of recording) vs. windowed streaming (red, labeled "cold"; hidden state reset every 5 s). The persistent model reaches threshold before onset and bridges the windowed model's mid-seizure probability dip. **Bottom**: log hidden-state norms for the four Mamba-3 blocks; no drift over 45 s.

Table 5: Streaming-parallel equivalence on real fine-tuned weights. 30 input seeds $\times$ 3 checkpoints $=$ 7,200 forward passes.

Metric	Value
Max logit difference	5.9×10^{-3}
Max probability difference	3.1×10^{-4}
Class-label agreement	100%

Exponential-trapezoidal discretization. Mamba-1/2 use a first-order (Euler) integration rule with local truncation error $O(\Delta_t^2)$; Mamba-3 uses a second-order (trapezoidal) rule with truncation error $O(\Delta_t^3)$, blending $B_t x_t$ with $B_{t-1} x_{t-1}$ via a learned data-dependent weight $\lambda_t \in [0, 1]$ (Lahoti et al.,

2026, Proposition 1). The accuracy gain matters more for continuous physical signals than for discrete-token sequence modeling: in language, $B_t x_t$ is a learned embedding of a token and the integral being discretized is a modeling abstraction; in EEG, $B_t x_t$ is a projection of an actual sampled voltage and the discretization approximates a real continuous-time integral. Higher-order accuracy on the state-input integral is a closer fit to the data-generating process.

Complex-valued state transitions. Real-eigenvalue SSMs cannot natively track rotational hidden-state dynamics (Lahoti et al., 2026, Sec. 3.2), which is the relevant regime for any signal whose information is partly carried in oscillatory phase. Mamba-3's complex-valued state transition is implemented via a data-dependent rotary embedding on the B and C projections (Lahoti et al., 2026, Proposition 3): $h_t = e^{\Delta_t A_t} h_{t-1} + \big(\prod_{i \le t} \mathbf{R}_i^\top\big) \Delta_t B_t x_t$, where $\mathbf{R}_t$ is a block-diagonal of 2×2 rotations parameterized by learned, data-dependent angles. EEG is a canonical multi-band oscillatory signal — delta, theta, alpha, beta, gamma — and phase relationships across channels and bands carry information that real-eigenvalue exponential decay alone cannot represent. The data-dependent rotation angle lets the SSM advance phase by an amount that depends on the current input, tracking oscillations whose instantaneous frequency varies with the neural state. We test whether the trained state actually carries downstream-relevant long-horizon information in our hidden-state ablation in Sec. 4.3, where we compare seizure-onset prediction from a state warmed across the full pre-recording context versus one initialized from a 5 s cold start.

MIMO parameterization. SISO SSMs are memory-bound at decode time: the recurrent step's arithmetic intensity is $\Theta(1)$, leaving tensor cores idle while the state read dominates wall-clock latency (Lahoti et al., 2026, Table 2a). The MIMO generalization extends B and C to rank R, raising arithmetic intensity to $\Theta(R)$ at unchanged state size and approximately unchanged decode latency (Lahoti et al., 2026, Sec. 3.3). The streaming setting that motivates this paper is decode-only — one recurrent step per 62.5 ms patch, in perpetuity — so the per-step expressivity gain at fixed wall-clock latency translates directly into a more capable streaming model than a SISO Mamba-3 at the same latency budget would allow. We report measured per-patch latency in Appendix A.8.

A.7 Spectral Band Ablation on Seizure

We apply a 4th-order zero-phase Butterworth band-stop filter to the CHB-MIT test set, in turn for each canonical EEG band, and re-evaluate the seed-42 checkpoint on the filtered input. Delta-band removal causes the only large drop (17 AUROC points); other bands lie within seed-level noise. The result is consistent with the clinical role of slow-wave activity as a scalp-EEG seizure marker, and suggests the model has learned a frequency-relevant feature without explicit spectral supervision.

Table 6: Spectral band ablation on CHB-MIT (seed 42). Δ AUROC is the change from the unperturbed baseline.

Band	Range (Hz)	AUROC	Δ
Real test	—	0.935	—
Delta	0.5–4	0.766	$\mathbf{-0.169}$
Theta	4–8	0.906	-0.029
Alpha	8–13	0.937	$+0.002$
Beta	13–30	0.935	-0.000
Gamma	30–75	0.900	-0.035

A.8 Latency and Memory at Streaming Inference

Single-patch streaming latency on a single H100 PCIe is 4.97 ms (mean over 1,000 patches, 50-patch warm-up, random weights), and the persistent state size is $\sim$ 720 KB. Both are constant in context length by construction: each step processes one 16-sample patch and updates a fixed-size hidden state, with no buffer that grows with the input. We verified constancy at context lengths from 5 s to 4 hours. Bidirectional baselines such as LUNA-Base scale linearly in latency and grow in memory with context length (Döner et al., 2025) and exceed available VRAM at hour-scale inputs on the same hardware.

Table 7: Per-component breakdown of the streaming step. FLOPs are analytical; state is the persistent memory carried across patches.

Component	FLOPs (M)	State (KB)
SwiGLU FFNs ($\times 4$)	47.60	—
Mamba-3 SSM step ($\times 4$)	25.78	720
Channel Embedding	3.36	—
RMSNorms + classifier head	0.02	—
Total per step	76.76	720

Algorithm 1 CaMBRAIN Streaming Inference

Require: Patch stream $\{x_1, x_2, \dots\}$, $x_t \in \mathbb{R}^{C \times P}$
Ensure: Per-patch prediction p_t

function RESET
  $h_\ell \leftarrow \mathbf{0}$ for $\ell = 1, \dots, L$ ▷ per-block SSM state
  $\tau \leftarrow 0$ ▷ position counter
end function

function STEP(x_t)
  $z \leftarrow \text{CHANNELEMBEDDER}(x_t)$
  $z \leftarrow z + \mathbf{p}_{\tau \bmod 80}$ ▷ cyclic position embedding
  for $\ell = 1, \dots, L$ **do**
    $z, h_\ell \leftarrow z + \text{MAMBASTEP}(\text{RMSNORM}(z), h_\ell)$
    $z \leftarrow z + \text{SWIGLU}(\text{RMSNORM}(z))$
  end for
  $p_t \leftarrow \sigma\big(\mathbf{W}_{\text{cls}}^\top z\big)$ ▷ classify from last token
  $\tau \leftarrow \tau + 1$; **return** p_t
end function

Constants for our model: $L = 4$ blocks, $C = 16$, 19 or 22 channels, $P = 16$ samples per patch.

Table 8: CaMBRAIN downstream results, mean $\pm$ std across three random seeds.

Dataset	Metric	Mean $\pm$ Std
TUAR (MMC, 6-class)	AUROC	0.9362 ± 0.0028
	AUPR	0.5651 ± 0.0125
TUAB (binary)	AUROC	0.8668 ± 0.0061
	AUPR	0.8757 ± 0.0046
	BAC ($t = 0.5$)	0.7770 ± 0.0117
	BAC (Youden)	0.7953 ± 0.0065
MAT (binary)	AUROC	0.8758 ± 0.0222
	AUPR	0.7345 ± 0.0518
	BAC	0.7778 ± 0.0028
CHB-MIT (binary)	AUROC	0.9205 ± 0.0169
	AUPR	0.3888 ± 0.0282

A.9 Causal Preprocessing Checks

The zero-phase filter of Appendix A.3 is offline benchmark construction, as in prior work, and is not part of the streaming inference path. To measure its effect, we rebuilt CHB-MIT from the raw EDF files with a strictly causal filter: a 0.1–75 Hz 8th-order Butterworth bandpass (design order 4, band type), pre-warped and bilinear-transformed at 256 Hz and run as four cascaded biquads in a single forward pass, and a 60 Hz notch ($Q = 30$) run as one biquad with zeros on the unit circle at $84.375°$ and $-84.375°$ and poles at radius 0.9757. Each biquad output depends only on $x[n]$, $x[n-1]$, $x[n-2]$, $y[n-1]$ and $y[n-2]$, with no time-reversed pass, so the filter has zero look-ahead but nonzero group delay, from 46 ms at 1 Hz to 4.5 ms at 10–20 Hz. The reported results normalize with window-level quartile statistics, which are not causal within a 5 s window, while streaming uses the causal sliding RQN of Appendix A.3; "strictly causal" in this paper refers to the streaming path. Table 9 compares the offline and causal variants of both steps; the end-to-end CHB-MIT rerun under causal normalization covers a single seed (42).

Table 9: Causal preprocessing checks. The filter checks and both AUROC reruns are on CHB-MIT; arrows go from the offline variant used for the reported results to the causal variant.

Step	Check	Result
Filter	Spectral-feature correlation, chb23_09 (4 h)	0.99953
	Bandpower error, chb23_09	$< 0.03\%$
	AUROC (3 seeds)	$0.9205 \rightarrow 0.9185$ (-0.002)
Normalization	Input correlation	99.97%
	AUROC (seed 42 only)	$0.9305 \rightarrow 0.9225$ (-0.008)

A.10 Dataset Splits and Subject Leakage

Table 10 gives, for each dataset, the split, the unit that is disjoint across training, validation and test, and the overlap with the TUEG pretraining corpus, which is none for every dataset. For TUAR, the main results (Table 1a) use a recording-level split, under which 12 of the 30 test subjects also appear in training. On a subject-disjoint TUAR split, AUROC is 2.9 points lower than under the recording-level split (Table 11).

Table 10: Per-dataset splits. The disjoint unit is the unit that does not cross the training, validation and test sets; the last column gives the overlap with the TUEG pretraining corpus. For CHB-MIT, chb21 is a later recording of the chb01 patient, so validation is not fully patient-disjoint from training; the test patients (chb22–chb24) are.

Dataset	Split (train / val / test)	Disjoint unit	TUEG overlap
TUAB	official TUH train / eval	subject	none, held out
TUAR (main results)	80 / 10 / 10% by recording	recording	none, held out
TUAR (subject-disjoint)	167 / 20 / 22 by subject	subject	none, held out
MAT	26 / 5 / 5 by subject	subject	none, other institution
CHB-MIT	chb01–19 / 20–21 / 22–24	subject/patient	none, other institution

Table 11: TUAR AUROC under the two split protocols.

TUAR protocol	AUROC
Recording-level, as reported (all baselines)	0.936
Subject-disjoint	0.907

Evaluation protocol. All scores in Tables 1 and 2 are window-level, for CaMBRAIN as for the baselines. Although CaMBRAIN can emit a prediction every 62.5 ms, on the benchmarks it makes one prediction per evaluation window, read from the hidden state at the window's last patch, and every metric is computed over the labelled windows of the test set **This has been done to make sure the**

fairness of comparision. Each score therefore ranks the same unit as the baseline scores, a labelled EEG window, and AUROC and AUC-PR are threshold-free, so no score depends on a tuned decision threshold. On TUAR and TUAB we follow the protocol of Döner et al. (2025): signals resampled to 256 Hz, non-overlapping 5 s windows, the official TUAB split, and an 80/10/10 random TUAR split, which we draw at the recording level rather than the sample level, so that no test recording contributes windows to training. Sampling at 256 Hz rather than the 200 Hz used by some baselines changes the number of samples in a window, not the span of signal it covers. This is because of pre-training design choices of other algorithm and does not reflect a difference in evaluation as it is a fair live setting.

The maintained state adds to what a window-level score shows. Carrying the state across windows instead of resetting it every 5 s raises the seizure probability at onset and the patch-level AUROC on CHB-MIT (Table 3a), and at full sensitivity it lowers false alarms from 44.0 to 10.5 per hour (Appendix A.11).

A.11 Event-Level Seizure Detection

Patch-level AUROC does not show how a seizure monitor behaves on events, so Table 12 scores the CHB-MIT test set at the event level over all three held-out patients (chb22–chb24) and all 26 annotated onsets, with the hidden state either retained across the recording or reset. Each configuration is scored at the threshold that attains full sensitivity, so the two are compared at a matched operating point: both detect all 26 onsets, and the persistent state does so at 10.5 false alarms per hour against 44.0 with resets. Between 2 and 20 false alarms per hour, the persistent model holds 77–100% sensitivity, while the reset configuration has no operating point in that range. With the state retained, the median detection delay is 10 s; across the test set, 12% of detections precede the annotated onset.

Table 12: Event-level seizure detection on the CHB-MIT test set (patients chb22–chb24, 26 onsets), at the threshold attaining full sensitivity.

Event-level, 26 onsets	With state	Without state
Sensitivity	100% (26/26)	100% (26/26)
False alarms / h	10.5	44.0
Per-patient AUROC (22/23/24)	0.982 / 0.910 / 0.889	0.978 / 0.887 / 0.874

A.12 Memory Timescales of the Persistent State

To locate the timescale of the useful context, we reset the hidden state at fixed intervals while streaming the three CHB-MIT test patients (26 onsets, 3 seeds; Table 13). Keeping the state gives the highest onset probabilities, while the finite horizons of 5, 7 and 10 s do not separate from one another, so the benefit comes from context older than 10 s; resets shorter than 5 s or between 10 s and the full recording (e.g., 30 s or several minutes) were not tested. Because the state is an explicit recurrence, its decay can be read out directly: the per-layer memory half-lives on CHB-MIT are 36.3, 24.6, 18.6 and 0.02 s, and the median half-life on real recordings is 18.7 s, against 0.19 s for the pretrained-only encoder. Permuting the time axis within each patch preserves variance and amplitude but destroys temporal morphology, and AUROC falls from 0.94 to 0.21; the result is below chance, so predictions invert rather than degrade, and the model does not rely on amplitude statistics alone.

Table 13: Resetting the hidden state at fixed intervals on the three CHB-MIT test patients (26 onsets), mean over 3 seeds. Probabilities $\times 100$.

Reset interval	Prob. at onset	Mean prob. over onset 5 s
keep (persistent)	27.7	34.3
10 s	10.7	19.0
7 s	15.5	23.7
5 s	11.7	18.7

A.13 Windowed Encoders with Downstream Aggregation

A windowed encoder could, in principle, use context beyond one window by aggregating its per-window features downstream. To test whether aggregation recovers what the persistent state provides, we give LUNA (Döner et al., 2025) and REVE (El Ouahidi et al., 2025) a causal exponential moving average (EMA) and a trained GRU over their own per-window latents, apply the same readouts to CaMBRAIN, and measure forecasting skill on continuous CHB-MIT (Table 14). For every encoder, the EMA and GRU scores stay close to the last-window score, and aggregation does not close the gap between encoders. The separation is therefore between encoders rather than readouts, so persistence has to be trained into the encoder rather than attached afterwards.

Table 14: Forecasting skill (mean) on continuous CHB-MIT with causal aggregation over each encoder's per-window latents.

Encoder	Last window	+ EMA	+ GRU
CaMBRAIN (persistent)	0.508	0.507	0.503
LUNA	0.478	0.477	0.477
REVE	0.470	0.468	0.463

A.14 Pretraining Ablations Across Datasets

Table 3b ablates the pretraining stages on TUAR alone; Tables 15 and 16 extend the ablation to full fine-tuning on TUAB and MAT (three seeds) and to a frozen-encoder linear probe on all four datasets. Under the probe, Stage 2 raises AUROC on every dataset, for example from 0.748 to 0.910 on CHB-MIT; under fine-tuning, its gain is large on TUAB (0.738 to 0.867) and small on MAT (0.868 to 0.876), where Stage 1 already accounts for most of the gain over training from scratch (0.641). Table 17 indicates why causality and Stage 1 alone are not enough: the Stage-1 representation is poorly conditioned and close to that of a randomly initialized encoder, while Stage 2 spreads variance over many more directions. Masked prediction supervises what the state can reconstruct rather than what it can anticipate, and Stage 2's next-patch latent prediction reaches $R^2 = 0.92$–0.97 on all four datasets, where the same architecture at random initialization fails.

Table 15: Pretraining ablation under full fine-tuning, test AUROC, mean over three seeds.

Dataset	Scratch	Stage 1	Stage 1+2
TUAB	0.717	0.738	0.867
MAT	0.641	0.868	0.876

Table 16: Pretraining ablation under linear probing of the frozen encoder, test AUROC.

Dataset	Stage 1	Stage 1+2
CHB-MIT	0.748	0.910
TUAB	0.716	0.839
MAT	0.641	0.738
TUAR	0.873	0.909

Table 17: Conditioning of the encoder representation at random initialization and after each pretraining stage: participation ratio (effective number of dimensions), share of variance on the first principal component, and number of principal components needed to reach 90% of the variance.

Metric	Random	Stage 1	Stage 1+2
Participation ratio	2.0	3.0	19.1
PC1 variance share	0.58	0.54	0.13
Dims to 90% variance	6	7	49

A.15 Linear Probing Against Baselines

Table 18 compares frozen encoders on CHB-MIT under a common linear readout; unlike the fine-tuned results in Table 1b, no encoder weights are updated. A linear readout measures the encoder rather than the head: the fit is convex on fixed features and adds no variance of its own, so with the readout held fixed, differences between scores come from the representations, whereas a higher-capacity head would mix encoder quality with head capacity. It is also uniform across models whose own heads differ in size, with REVE (El Ouahidi et al., 2025) attaching a comparatively large classification head and LUNA (Döner et al., 2025) a smaller one. On small held-out sets (three CHB-MIT patients, five MAT subjects) a higher-capacity head is more prone to overfitting, and a linear readout costs constant time per patch, which preserves the streaming property. We have not compared against MLP or attention-pooling heads.

Table 18: Linear probing of frozen encoders on CHB-MIT, test AUROC.

Frozen encoder	Params	Probe AUROC
CaMBRAIN	37 M	0.910
LUNA-Base	7 M	0.906
REVE-Base	69 M	0.890

A.16 Patch-Size Sensitivity

A patch is 16 samples at 256 Hz (62.5 ms), 2.5× finer than the 40 samples LUNA (Döner et al., 2025) uses at the same rate, which is what permits the 16 Hz prediction cadence. A 62.5 ms patch has 16 Hz spectral resolution and resolves only beta and gamma activity, so delta, theta and alpha rhythms are carried across patches by the recurrence; in the band ablation of Appendix A.7, removing delta costs 0.169 AUROC, the largest drop of any band. Table 19 varies the patch size during fine-tuning only; pretraining was not rerun at each size, because the positional-embedding period is tied to the patch count of a 5 s window and each size would need its own pretraining run. No patch size gives the highest AUROC on every dataset, and the size that does differs across datasets with no consistent direction; larger patches also lower the output rate, and 250 ms patches cap predictions at 4 Hz. An 8-sample (31 ms) patch, which gives the highest AUROC of the four sizes on MAT, would double the update rate to 32 Hz; whether the pretraining transfers to a different patch granularity is left open.

Table 19: Patch-size sensitivity, test AUROC. The patch size is varied during fine-tuning only, with no pretraining run at each size; entries come from the sweep's own fine-tuning runs and are compared within each row.

	Patch size in samples (duration)			
Dataset	8 (31 ms)	16 (62.5 ms)	32 (125 ms)	64 (250 ms)
CHB-MIT	0.878	0.886	0.894	0.857
TUAB	0.685	0.732	0.695	0.758
MAT	0.670	0.601	0.614	0.590

A.17 Label Efficiency

EEG benchmarks are small, and every model in Tables 1 and 2 is trained and evaluated under that constraint. Table 20 fine-tunes on fractions of the downstream training set, from the pretrained encoder and from scratch, with a single seed (42). The pretrained encoder scores higher than training from scratch at every fraction on all three datasets, and on TUAB and MAT it scores higher with 5% of the labels than training from scratch does with 75%.

A.18 Wall-Clock Latency and Memory

Table 21 measures per-update latency and GPU memory against LUNA-Base on the same hardware as the context grows: CaMBRAIN's stay at 4.96 ms and 229 MiB at every context length, while

Table 20: Test AUROC when fine-tuning on a fraction of the downstream training set, from the pretrained encoder (Pre.) or from scratch (Scr.), seed 42.

	CHB-MIT		TUAB		MAT	
Fraction	Pre.	Scr.	Pre.	Scr.	Pre.	Scr.
5%	0.889	0.764	0.849	0.663	0.700	0.654
10%	0.893	0.881	0.857	0.720	0.710	0.665
25%	0.907	0.849	0.859	0.732	0.676	0.653
50%	0.903	0.895	0.859	0.727	0.848	0.665
75%	0.911	0.895	0.858	0.728	0.891	0.679

LUNA-Base's grow with the context it is conditioned on, from 9.5 ms and 484 MiB at 5 s to 1824 ms and 66 GiB at 1 h. The channel count enters CaMBRAIN's per-step cost only through the 4-query embedder, which accounts for 4.4% of the step (Table 7), so the per-step cost varies by under 2% across 16 (CHB-MIT), 19 (MAT) and 22 (TUAR, TUAB) channels and the 1.23 GFLOPs/s figure is essentially the same on every dataset.

Table 21: Per-update latency and GPU memory versus context length on a single H100 PCIe (22 channels, 20 iterations). Only LUNA-Base was timed. The 4.97 ms of Appendix A.8 comes from a separate measurement.

	Latency		Memory	
Context	CaMBRAIN	LUNA-Base	CaMBRAIN	LUNA-Base
5 s	4.96 ms	9.5 ms	229 MiB	484 MiB
180 s	4.96 ms	20.6 ms	229 MiB	703 MiB
600 s	4.96 ms	97 ms	229 MiB	2.5 GiB
1 h	4.96 ms	1824 ms	229 MiB	66 GiB

A.19 Comparison with Mamba-Based EEG Models

EEGMamba (Wang et al., 2025) and FEMBA (Tegon et al., 2025) also build EEG models from Mamba blocks, but both are windowed (Table 22): each prediction encodes a fixed window with a bidirectional scan, and no state is carried from one window to the next. By streaming we mean incremental inference in which each new patch updates a persistent state at constant cost, without re-encoding earlier input; in this sense neither model streams. Re-running a windowed model on a sliding window only repeats windowed inference: every update re-encodes the whole window, its context ends at the window boundary, and aggregating window latents downstream does not recover what a trained persistent state provides (Appendix A.13). Causality alone is also not enough, since Stage 1 leaves the representation poorly conditioned; the Stage-2 objective is what trains the state to anticipate (Appendix A.14).

Table 22: Mamba-based EEG models by scan direction, pretraining, inference mode, per-update cost and context. The windowed models re-encode every window and carry no state across windows.

	EEGMamba	FEMBA	CaMBRAIN
Direction	bidirectional	bidirectional	unidirectional / causal
Pretraining	MAE	MAE	causal MAE + JEPA + ARM
Inference	windowed	windowed	streaming, persistent state
Per-update cost	re-encode window	re-encode window	one recurrent step
Context	window-bounded	window-bounded	tens of seconds