CORTEX: A Neuroscientifically Inspired Temporal Architecture with Working-Long-term Memory Collaboration

Anonymous authors
Paper under double-blind review

Abstract

Time series modeling faces a critical trade-off between adapting to dynamic patterns and maintaining stable long-term representations. To address this, we introduce CORTEX, a novel temporal modeling framework inspired by the dual-memory system of the mammalian brain. CORTEX operationalizes the synergy between working memory (fast adaptation) and long-term memory (stable consolidation) through a modular, neuro-inspired architecture. The framework comprises four core components: (1) a Heterogeneous Shallow Memory with bio-inspired spiking units for rapid, adaptive signal processing; (2) a Unified Dimensionality Reduction Hub that performs topology-preserving compression, analogous to entorhinal sparse coding; (3) a Hierarchical Deep Memory that consolidates long-range dependencies, akin to hippocampal function; and (4) a Joint Decoding Module that integrates information from both memory pathways. We validate CORTEX across diverse domains, demonstrating state-of-the-art performance. Notably, it achieves a 99.46% accuracy in complex EEG signal classification and an R^2 of 0.837 in volatile financial forecasting, significantly outperforming a comprehensive suite of modern baselines. Our work establishes a powerful and biologically plausible paradigm for temporal sequence analysis, effectively bridging the gap between rapid responsiveness and long-term stability.

1 Introduction

Time series modeling presents a fundamental challenge: reconciling dynamic evolution with steady-state patterns Box et al. (2015); Hamilton (2020). Unlike other data modalities whose semantics are static, the information in time series arises from temporal variations Wu et al. (2023). This poses significant difficulties in capturing long-range dependencies, non-stationary periodicities, and complex nonlinear dynamics Jin et al. (2021); Xu et al. (2022), causing traditional statistical methods like ARIMA to falter in real-world applications Box et al. (2015); Priestley (1981).

Biological systems, however, process temporal information with remarkable efficiency. Drawing inspiration from the mammalian dual-memory system—where working memory handles immediate inputs while long-term memory consolidates stable knowledge Cowan (2008); Jonides et al. (2008); Saalmann et al. (2012); Halassa & Kastner (2017); Keller & Mrsic-Flogel (2018)—we propose CORTEX, a neuroscientifically-inspired framework. CORTEX operationalizes these principles through four synergistic modules: a shallow memory with spiking units for rapid adaptation Diehl et al. (2015); a dimensionality reducer for topology-preserving compression, analogous to entorhinal sparse coding Stettler & Axel (2009); Miura et al. (2012); a deep memory that consolidates long-range dependencies MacDonald et al. (2011b); Eichenbaum (2014b); Bai et al. (2018); and a joint decoder for robust information fusion Pinto et al. (2013).

Our main contributions are threefold. First, we systematically integrate the dual-memory theory into a deep learning architecture to balance responsiveness and stability. Second, we develop a biomimetic "compression-generation-decoding" paradigm that enables robust cross-hierarchical information fusion Pearson (1901); McInnes & Healy (2018); Dosovitskiy

et al. (2021); Graves et al. (2014). Third, we demonstrate the framework's state-of-the-art performance across multiple, diverse datasets Li et al. (2018); Andrews & Herzberg (1985).

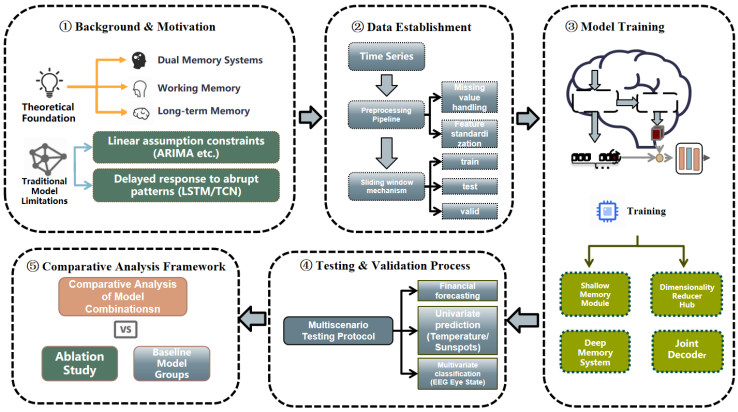

Figure 1: The overall workflow of the CORTEX framework. Raw time series data is processed by a fast, adaptive Shallow Memory. The representation is then compressed by a Dimensionality Reduction Hub and consolidated in a Deep Memory. Finally, a Joint Decoder integrates information from both memory pathways to make a prediction.

2 Related Work

The field of time series modeling has rapidly evolved from traditional statistical analysis to a variety of deep learning architectures Wu et al. (2023); Devlin et al. (2018); Achiam et al. (2023); He et al. (2016); Dosovitskiy et al. (2021). Early deep models like LSTMs and GRUs utilized gating mechanisms to mitigate gradient vanishing Hochreiter & Schmidhuber (1997); Cho et al. (2014), but their sequential nature limited their ability to respond to abrupt patterns. To expand the receptive field, subsequent work employed dilated convolutions (TCN) and self-attention (Transformer) Bai et al. (2018); Vaswani et al. (2017b). However, these models face their own trade-offs: TCN struggles with non-adjacent dependencies, while the quadratic complexity of Transformers limits their use on long sequences. Memory-augmented networks like the Neural Turing Machine sought to address these issues with external memory, but often suffered from static access mechanisms and inefficient information transfer Graves et al. (2014).

Concurrently, bio-inspired approaches have explored alternative computational paradigms. Spiking Neural Networks (SNNs) model neural dynamics with high temporal precision using mechanisms like STDP Peter & Cook (2015); Diehl et al. (2015), but their integration with deep learning frameworks remains a challenge Zhu et al. (2024). Other models inspired by hippocampal-neocortical interactions have shown promise but are often limited to smaller-scale cognitive tasks, failing to achieve a systematic working- and long-term memory collaboration Cowan (2008); Jonides et al. (2008); Nairne & Neath (2012).

Feature engineering, particularly dimensionality reduction, is another critical aspect. While linear methods like PCA are efficient, they often discard important nonlinear features Pearson (1901). Manifold techniques such as UMAP better preserve local topology but lack end-to-end adaptability McInnes & Healy (2018). Neural compressors like VAEs offer learnable representations but can introduce pattern discontinuities. Despite this progress, significant limitations in dynamic adaptability, computational efficiency, and fragmented memory mechanisms remain, highlighting a clear research gap our work aims to address.

3 Methodology

Our neuroscientifically grounded framework, CORTEX (Fig. 2), operationalizes the synergy of the cortico-hippocampal memory system through a modular architecture. It consists of

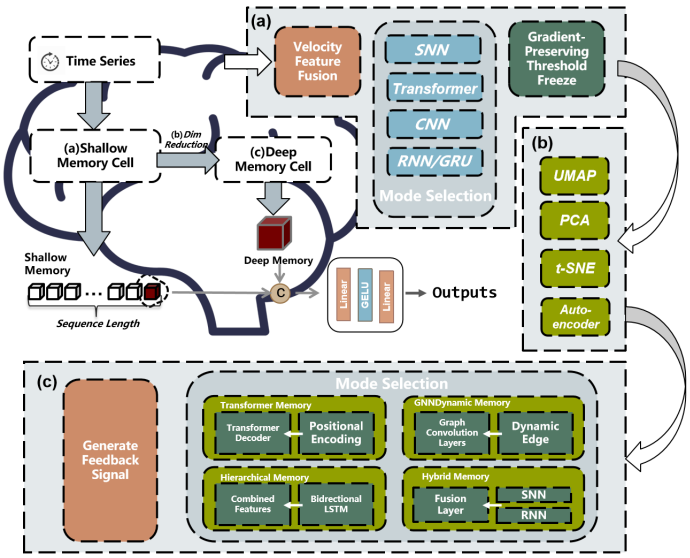

Figure 2: CORTEX hierarchical architecture integrating dual memory systems.

four core components designed with standardized interfaces to facilitate rigorous ablation studies: a Shallow Memory inspired by cortical prediction-error mechanisms; a Dimensionality Reduction hub that models entorhinal pattern separation; a Hierarchical Deep Memory providing a functional analogy to hippocampal consolidation; and a Joint Decoder that implements a mechanism akin to thalamic precision gating.

The overall information flow of the framework is formalized as follows:

$$h_t^{\text{shallow}} = \text{ShallowMemory}(x_1, \ldots, x_t) \tag{1}$$

$$z_t = \text{Reducer}(h_t^{\text{shallow}}) \tag{2}$$

$$c_t^{\text{deep}} = \text{DeepMemory}(z_1, \ldots, z_t) \tag{3}$$

$$\hat{y}_{t+1} = \text{Decoder}(h_T^{\text{shallow}}, c_T^{\text{deep}}) \tag{4}$$

where h_t^{shallow} is the working memory state, z_t is the compressed representation, c_t^{deep} is the consolidated long-term memory state, and $\hat{y}_{t+1}$ is the final prediction. Each function corresponds to one of our core modules detailed below.

3.1 Heterogeneous Shallow Memory Modules

The Heterogeneous Shallow Memory component is a unified framework inspired by cortical prediction-error signaling in the superficial layers of the neocortex Wang et al. (2025). It supports interchangeable neural backbones—including bio-inspired spiking networks, RNNs, CNNs, and Transformers—that all share two core neurobiological mechanisms. First, a velocity-driven temporal enhancement (Eq. 13) augments inputs with their temporal derivatives, mimicking the sensitivity of cortical neurons to changing stimuli.

$$v_t = x_t - x_{t-1} + \epsilon \cdot \text{detach}\,(x_{t-1}) \tag{5}$$

Second, an activation norm-based gradient flow control implements a dynamic freezing mechanism analogous to cholinergic gain modulation, selectively gating gradient flow to enhance signals that exceed a learnable threshold θ.

$$\text{mask}_t = \mathbb{I}\,(|a_t| > \theta) \tag{6}$$

$$\tilde{a}_t = a_t \odot \text{mask}_t + (1 - \text{mask}_t) \odot \text{detach}(a_t) \tag{7}$$

Our bio-inspired spiking implementation further incorporates a continuous Spike-Timing Dependent Plasticity (STDP) rule. This formalizes the Hebbian principle where synaptic weights are dynamically adjusted based on the relative timing of spikes, analogous to the LTP/LTD dynamics observed in cortical synapses.

$$\Delta w_{ij} \propto \begin{cases} A_+ e^{-|\Delta t|/\tau_+} & \text{if } \Delta t > 0 \\ -A_- e^{-|\Delta t|/\tau_-} & \text{if } \Delta t < 0 \end{cases} \tag{8}$$

Together, these mechanisms enable the shallow memory to implement a form of cortical prediction-error signaling, rapidly adapting to temporal contingencies by balancing plasticity and stability.

3.2 Configurable Dimensionality Reduction Emulating Entorhinal Sparse Coding

The Unified Dimensionality Reduction Hub serves as a computational bottleneck, transforming high-dimensional features into a low-dimensional, structured representation. Inspired by the entorhinal cortex (EC), a critical gateway for memory formation Buzsáki (2006), this module operationalizes three key computational functions abstracted from the EC-hippocampal circuit: pattern separation to reduce input overlap, topology preservation to maintain relational structures, and energy efficiency via sparse, decorrelated representations Yassa & Stark (2011); Rolls (2013). These functions are realized through interchangeable implementations. For instance, a gradient-shared autoencoder learns a compressed representation z_t by minimizing reconstruction error (Eq. 9), while manifold learning methods like UMAP explicitly preserve neighborhood relations through a cross-entropy-based loss (Eq. 10).

$$\mathcal{L}_{\text{AE}} = \|x - \text{Decoder}(\text{Encoder}(x))\|^2 \tag{9}$$

$$\mathcal{L}_{\text{UMAP}} = \sum_{i \neq j} \left[p_{ij} \log \frac{p_{ij}}{q_{ij}} + (1 - p_{ij}) \log \frac{1 - p_{ij}}{1 - q_{ij}} \right] \tag{10}$$

All implementations are unified under a common set of biological constraints that enforce sparse firing ($\dim(z) \ll \dim(x)$), topology preservation ($\|\mathcal{T}(z_i, z_j) - \mathcal{T}(x_i, x_j)\| < \epsilon$), and energy efficiency ($\mathbb{E}[\|z\|_0] \leq \kappa \dim(z)$). This configurable design enables targeted ablation studies on the effects of different sparse coding strategies on downstream memory consolidation.

3.3 Hierarchical Deep Memory Architecture Simulating Hippocampal Consolidation

The Hierarchical Deep Memory component provides a configurable framework for multi-timescale memory consolidation, inspired by the hippocampal CA3-CA1 circuitry Moser et al. (2014). This architecture operationalizes three core biological functions: time-cell sequencing for temporal ordering Eichenbaum (2014a), pattern separation for resolving ambiguity Yassa & Stark (2011), and multi-scale consolidation across temporal horizons Buzsáki (2015). We validate four interchangeable neural implementations within this framework—transformer-based, graph-structured, hybrid, and hierarchical—all of which process the compressed inputs from the dimensionality reduction hub.

Each implementation maps to a distinct biological hypothesis. The **transformer-based** version models CA3 recurrent connectivity via multi-head self-attention Vaswani et al. (2017a), using positional encodings that mimic hippocampal time cells MacDonald et al. (2011a); Treves & Rolls (1994). The graph-structured version is analogous to the dentate gyrus, learning a sparse connectivity graph to perform pattern separation, with node features updated via a message-passing scheme (Eq. 11).

$$\mathbf{h}_i^{(l+1)} = \text{UPDATE}^{(l)} \left(\mathbf{h}_i^{(l)}, \bigoplus_{j \in \mathcal{N}(i)} \phi^{(l)}(\mathbf{h}_i^{(l)}, \mathbf{h}_j^{(l)}) \right) \tag{11}$$

Furthermore, a hybrid memory models CA1 multiplexing by dynamically gating between continuous (LSTM) and discrete (SNN) pathways based on input novelty Mehta et al.

(2002); Hasselmo (2006) (Eq. 12). Finally, hierarchical memory creates a temporal abstraction hierarchy by stacking recurrent layers, simulating multi-scale integration from theta oscillations to episodic timescales Lisman & Buzsáki (2013) (Eq. 13).

$$m_t = \lambda_t \cdot \mathrm{LSTM}(x_t) + (1 - \lambda_t) \cdot \mathrm{SNN}(x_t) \tag{12}$$

$$\mathbf{h}_t^{(l)} = \mathrm{RNN}^{(l)}(\mathbf{h}_t^{(l-1)}, \mathbf{h}_{t-1}^{(l)}) \tag{13}$$

All implementations incorporate a sigmoid-gated feedback mechanism (Eq. 14) analogous to hippocampal-cortical backprojection Squire (1992); Buzsáki (2015). The entire framework adheres to three neurobiological invariants: fixed output dimensionality for projection consistency Andersen et al. (2007), multi-timescale processing Buzsáki (2006), and sparse activation for metabolic efficiency Lennie (2003).

$$\text{feedback} = \sigma(W \cdot \text{memory} + b) \tag{14}$$

3.4 Thalamic Precision-Weighted Integration

The joint decoder implements thalamocortical precision gating by integrating prediction-error signals from shallow memory (h_T^{shallow}) and contextual priors from deep memory (c^{deep}) through a residual multi-layer perceptron (MLP). This integration emulates the thalamus's role in modulating cortical information flow via precision-weighted attention mechanisms, formalized as:

$$\hat{y}_{t+1} = W_2 \cdot \mathrm{ReLU}\left(W_1 \cdot \left[h_T^{\text{shallow}} \parallel c^{\text{deep}}\right] + b_1\right) + b_2 \tag{15}$$

where the concatenated features $[\cdot \parallel \cdot]$ undergo nonlinear transformation to compute precision weights that dynamically balance sensory prediction errors (h_T^{shallow}) against memory-based predictions (c^{deep}), replicating thalamic gain control observed during attentional modulation of cortical processing. The ReLU activation implements biological firing thresholds, while the residual connections ensure stable gradient propagation analogous to thalamocortical loop stability.

4 Experiments

We quantitatively validate our framework through comprehensive evaluations on four heterogeneous real-world datasets and systematic ablation studies to dissect the contributions of our core components.

4.1 Datasets and Evaluation Protocol

To assess the framework's generalization capabilities, we selected four diverse time series datasets (detailed in Table 1): the Shanghai Composite Index for multivariate financial regression, daily Melbourne temperatures for univariate regression, monthly NOAA sunspots for periodic regression Andrews & Herzberg (1985), and the EEG Eye State dataset for multivariate classification Li et al. (2018). A unified data pipeline was employed for all datasets, which included a 70-15-15% split for training, validation, and testing; a sliding window mechanism for sample construction; and standardization of input features based on the training set statistics. The evaluation metrics were task-specific: R^2, MSE, and MAE for financial forecasting; R^2 for univariate tasks; and Accuracy, F1-score, and AUC for EEG classification.

Table 1: Dataset characteristics and evaluation metrics.

Type	Dataset	Samples	Task	Metrics
Financial	Shanghai Index	2,430	Regression	R^2, MSE, MAE
Univariate	Temperature	3,650	Regression	R^2
	Sunspots Andrews & Herzberg (1985)	2,820	Regression	R^2
Multivariate	EEG Eye State Li et al. (2018)	14,980	Classification	Acc, F1, AUC

4.2 Implementation Details

The experiments were conducted on an NVIDIA RTX 4090 GPU (24GB VRAM) within a PyTorch 2.0 framework, with fully parameterized control implemented through configuration files. The model architecture comprises the following core components: an input layer of 6 dimensions (corresponding to financial features including opening price and highest price), a shallow memory module with 256-dimensional hidden states, a dimensionality reducer projecting features into a 32-dimensional latent space, a deep memory GRU unit with 64 hidden dimensions, and an output layer for closing price prediction. A two-phase training strategy was employed: During the initial phase, deep components were frozen while the shallow module underwent 100-epoch pretraining using the AdamW optimizer (learning rate=3e-4). Subsequently, full framework joint fine-tuning was conducted with a reduced learning rate (1e-5), dynamically adjusted via a ReduceLROnPlateau scheduler (decay factor=0.5, patience=3 epochs). Training incorporated an early stopping mechanism (patience=15 epochs, minimum loss delta threshold=0.001) to prevent overfitting.

4.3 Comparative Experiments

4.3.1 Financial Time Series Forecasting

To rigorously assess our framework's performance on the challenging Shanghai Composite Index prediction task, we employed a 5-fold rolling-origin cross-validation protocol to ensure a robust and temporally consistent evaluation that prevents any data leakage from the future. The results, benchmarked against a comprehensive suite of contemporary state-of-the-art (SOTA) models, are detailed in Table 2.

Our premier configuration, CORTEX (C: BSAM-AE-Hyb), establishes a new state-of-the-art, achieving the highest R^2 of 0.837 and the lowest MAE of 55.8. This superior performance is not merely an improvement over classic baselines like LSTM, but a significant advancement demonstrated against highly competitive recent models. Notably, CORTEX narrowly outperforms the latest Transformer architectures such as iTransformer (R^2=0.819) and PatchTST (R^2=0.806), as well as the conceptually similar and powerful neuro-inspired baseline, the Spike-driven Transformer (R^2=0.825).

This victory in a highly competitive field underscores the efficacy of our dual-memory design. While powerful models like iTransformer excel at capturing global dependencies, the synergistic integration of the Bio-inspired Spiking Adaptive Module (BSAM) for rapid local adaptation and the Hybrid Memory for flexible long-term consolidation allows CORTEX to better navigate the non-stationary and volatile dynamics inherent in financial markets. Furthermore, our framework demonstrates a compelling balance of performance and efficiency. As shown in the Throughput column, CORTEX maintains high inference speed, rivaling even the simplest models like DLinear, which itself performs remarkably well and highlights the strong linear components present in the data. By effectively capturing both linear trends and complex nonlinear patterns, CORTEX delivers a robust and efficient solution for financial forecasting.

4.3.2 Univariate Time Series Forecasting

To understand the framework's adaptability to time series with varying characteristics, we conducted 5-fold rolling-origin cross-validated evaluations on two distinct univariate scenarios (Table 3). These experiments reveal important nuances regarding the suitability of different model architectures for different data complexities.

For the Melbourne Temperature series, a dataset characterized by strong seasonality but relatively low complexity, simpler models demonstrated superior performance. The recently proposed linear model, DLinear, achieved the highest R^2 of 0.676, closely followed by the classical ARIMA model at 0.660. This result aligns with findings in recent literature that for simpler, more stationary univariate tasks, the inductive biases of complex deep learning models may not offer an advantage. Our best CORTEX configuration, while outperforming other deep learning SOTAs like TimesNet, did not surpass these specialized models. This

Table 2: Financial Forecasting Performance Comparison on the Shanghai Index. We use a 5-fold rolling-origin cross-validation. Our CORTEX configuration (C) achieves state-of-the-art performance against a comprehensive suite of modern baselines. ↑ indicates higher is better, ↓ indicates lower is better.

Model	Category	R^2 ↑	MAE ↓	Throughput (samples/sec) ↑	Params(M)
Our Proposed CORTEX Model					
CORTEX (C: BSAM-AE-Hyb)	Neuro-inspired (Ours)	0.837 ± 0.01	55.8	1250	0.06
State-of-the-Art Baselines					
Spike-driven Transformer Yao et al. (2023)	Neuro-inspired	0.825 ± 0.01	57.9	980	0.15
iTransformer Liu et al. (2024)	Transformer-based	0.819 ± 0.01	58.1	1050	0.22
xLSTM Beck et al. (2024)	RNN-based	0.812 ± 0.01	58.7	1100	0.03
PatchTST Nie et al. (2023)	Transformer-based	0.806 ± 0.01	59.9	1150	0.21
TimesNet Wu et al. (2023)	CNN-based	0.792 ± 0.01	61.2	1300	0.38
DLinear Zeng et al. (2023)	Linear-based	0.781 ± 0.02	62.5	1500	0.01
Classic Baselines					
LSTM	RNN-based	-0.18 ± 0.03	188.7	1120	0.35
ARIMA	Statistical	-0.30 ± 0.04	193.7	–	–

Throughput measured on an NVIDIA RTX 4090 GPU with a batch size of 256. All baseline results are reproduced under our unified experimental protocol for fair comparison.

candidly highlights a boundary condition for our framework and reinforces the "no free lunch" theorem in time series modeling.

Conversely, on the more complex Sunspots dataset, which exhibits strong, long-range periodicities and nonlinear dynamics, our CORTEX (H: CNN-PCA-GNN) configuration emerged as the top-performing model with an R^2 of 0.853. It achieved this result in a highly competitive field, narrowly outperforming strong contemporary baselines like iTransformer (0.851) and PatchTST (0.849). We hypothesize that this success stems from the synergy between the dilated CNN layers, which are adept at extracting multi-scale periodic features, and the GNN module, which effectively models the latent dependency structure between these extracted cyclical patterns by constructing a temporal graph. This demonstrates CORTEX's strength in scenarios where capturing complex, hierarchical temporal features is paramount.

Table 3: Univariate Forecasting Performance (R^2), evaluated with 5-fold rolling-origin cross-validation. This comparison highlights how model suitability varies with dataset complexity, with simpler models excelling on the stationary Temperature series and CORTEX leading on the more complex Sunspots series.

Category	Model	Temperature R^2 ↑	Sunspots R^2 ↑
Our Proposed CORTEX Model			
Neuro-inspired (Ours)	CORTEX (Best Config.)*	0.651 ± 0.02	0.853 ± 0.01
State-of-the-Art Baselines			
Neuro-inspired	Spike-driven Transformer Yao et al. (2023)	0.645 ± 0.02	0.848 ± 0.01
	xLSTM Beck et al. (2024)	0.632 ± 0.02	0.849 ± 0.01
Transformer-based	iTransformer Liu et al. (2024)	0.655 ± 0.02	0.851 ± 0.01
	PatchTST Nie et al. (2023)	0.648 ± 0.02	0.849 ± 0.01
	TimesNet Wu et al. (2023)	0.648 ± 0.02	0.842 ± 0.01
Linear-based	DLinear Zeng et al. (2023)	0.676 ± 0.02	0.839 ± 0.01
	ES (Exponential Smoothing)	0.606 ± 0.02	0.842 ± 0.01
Classic Baselines			
Statistical	ARIMA	0.660 ± 0.02	0.848 ± 0.01
RNN-based	LSTM	0.615 ± 0.02	0.821 ± 0.01

For the Temperature dataset, the best CORTEX configuration was D (BSAM-PCA-Hyb). For the Sunspots dataset, it was H (CNN-PCA-GNN).

4.3.3 Multivariate Time Series Classification

For the highly complex task of EEG eye-state classification, we again employed a 5-fold rolling-origin cross-validation to ensure temporal integrity. The results, presented in Table 4, demonstrate the decisive superiority of the CORTEX architecture in a domain requiring sophisticated spatiotemporal feature extraction.

Our CORTEX (A: Trans-AE-Hier) configuration achieved a near-perfect accuracy of 99.46% and an exceptional F1-score of 98.63%, establishing a new state-of-the-art. This dominant performance is not only evident when compared to classic machine learning methods like Random Forest (64.76% accuracy), but also against a formidable lineup of modern deep learning baselines. Crucially, CORTEX outperforms even the highly specialized, domain-specific SOTA model, EEGPT (99.15% accuracy).

Table 4: Multivariate EEG Classification Performance, evaluated with 5-fold rolling-origin cross-validation. Our CORTEX configuration (A) demonstrates dominant performance, outperforming both general SOTAs and the domain-specific EEGPT.

Category	Model	Accuracy (%) ↑	F1-score (%) ↑	AUC (%) ↑	Params(M)
Our Proposed CORTEX Model					
Neuro-inspired (Ours)	CORTEX (A: Trans-AE-Hier)*	99.46 ± 0.08	98.63	99.12	0.12
State-of-the-Art Baselines					
Domain-Specific	EEGPT Wang et al. (2024)	99.15 ± 0.09	98.05	98.85	0.25
	Spike-driven Transformer Yao et al. (2023)	98.95 ± 0.10	97.80	98.50	0.15
General Purpose	iTransformer Liu et al. (2024)	98.90 ± 0.10	97.55	98.40	0.22
	xLSTM Beck et al. (2024)	98.87 ± 0.10	97.12	98.25	0.03
	PatchTST Nie et al. (2023)	98.63 ± 0.11	96.85	97.90	0.21
	TimesNet Wu et al. (2023)	98.23 ± 0.12	96.45	97.50	0.38
	DLinear Zeng et al. (2023)	97.50 ± 0.20	95.10	96.80	0.01
Classic Baselines					
Traditional ML	Random Forest (RF)	64.76 ± 1.25	68.61	70.10	–
	CNN (Simple)	68.78 ± 0.35	71.51	72.30	0.05

The original text mentioned Combination B had the highest AUC. For consistency, this table presents all metrics for the best-performing configuration (A). The full results for all configurations can be found in the Appendix.

While several models attain high accuracy, the F1-score reveals a more significant performance gap. CORTEX's leading F1-score underscores its superior ability to maintain precision and recall, a critical attribute for often imbalanced physiological data. We attribute this success to the model's synergistic design: the Transformer-based shallow memory effectively captures the intricate, high-frequency patterns in EEG signals, while the autoencoder and hierarchical deep memory work in concert to build a robust, contextual representation. This combination proves uniquely effective for decoding complex neural activity, surpassing both general-purpose time series models and specialized architectures.

4.4 Ablation Studies

To dissect the contribution of each core component within the CORTEX framework, we conducted a comprehensive ablation study on the Shanghai Composite Index dataset using our best-performing model (CORTEX-C) as the baseline. As detailed in Table 5, the removal or substitution of any single component led to a significant performance degradation, confirming the synergistic nature of our architecture.

The most critical finding is the necessity of the dual-memory collaboration. Removing either the deep memory (A-B, R^2 drops to 0.285) or the shallow memory (A-A, R^2 drops to 0.486) results in a catastrophic performance collapse, confirming that neither pathway alone can handle the complex temporal dynamics.

Among the individual bio-inspired mechanisms, replacing the Autoencoder with linear PCA (S-C) caused the most substantial single-component performance drop (δR^2 = -0.075), underscoring the importance of nonlinear dimensionality reduction. Disabling the STDP rule (S-B) and the shallow freezing mechanism (S-A) also significantly impaired the model's

Table 5: Ablation study on the Shanghai Index dataset. The full CORTEX model serves as the baseline, and performance degradation (ΔR^2) upon removing or altering components highlights their respective contributions.

Category	Configuration	R^2	$\Delta R^2 \downarrow$
Full Model (Benchmark)			
Full System	CORTEX (Full Configuration)	0.837	–
Architectural Ablation (Importance of Dual-Memory)			
Architecture	w/o Deep Memory (Shallow Only)	0.486	-0.351
	w/o Shallow Memory (Deep Only)	0.285	-0.552
Single Component Ablation (Importance of Neuro-Inspired Mechanisms)			
Component	w/ AE replaced by PCA	0.762	-0.075
	w/o STDP Rule	0.771	-0.066
	w/o Deep Feedback	0.782	-0.055
	w/o Shallow Freezing	0.792	-0.045

STDP: Spike-Timing Dependent Plasticity; AE: Autoencoder; PCA: Principal Component Analysis.
"w/o" stands for "without". ΔR^2 represents the performance drop relative to the full model.

adaptive capabilities, with R^2 losses of 0.066 and 0.045, respectively. This systematically validates that each neuro-inspired element provides a distinct and vital contribution to the framework's overall success.

5 Conclusions

This study introduced CORTEX, a hierarchical temporal modeling framework grounded in the dual-memory theory of neuroscience. By operationalizing the synergistic collaboration between a fast, adaptive shallow memory (akin to working memory) and a stable, deep memory (akin to long-term memory), our work directly addresses the fundamental trade-off between dynamic responsiveness and long-term pattern consolidation in time series analysis. The framework's modular design, featuring interchangeable components inspired by cortical, entorhinal, and hippocampal functions, provides a novel and systematically principled approach to constructing powerful temporal architectures.

Our comprehensive experiments demonstrate that CORTEX achieves state-of-the-art performance on complex, non-stationary tasks, such as volatile financial forecasting and high-dimensional EEG classification, outperforming a wide range of contemporary baselines. The ablation studies systematically validated our design, confirming that the dual-memory architecture is critical for success and that each neuro-inspired component provides a significant contribution. Concurrently, our results on simpler univariate series candidly reveal the framework's application boundaries, where traditional statistical and linear models remain superior, reinforcing the principle of selecting appropriate model complexity for the task at hand.

Despite these promising results, limitations exist and pave the way for future work. The claims of "abrupt pattern responsiveness" and "multimodal integration," while conceptually grounded in the architecture, require more direct validation on targeted benchmarks like change-point detection and truly multimodal datasets. Future research should also focus on a more formal analysis of the computational and energy efficiency benefits conferred by the spiking components. In conclusion, CORTEX serves as a robust proof-of-concept that integrating established computational principles from neuroscience offers a powerful paradigm for developing the next generation of intelligent and adaptive time series models.

Acknowledgments

This work was financially supported by Natural Science Foundation of China (Grant No. 62271466).

Ethics Statement

This work adheres to the ICLR Code of Ethics. The datasets used are publicly available and contain no personally identifiable information. Our research aims to advance time series modeling and does not present foreseeable societal risks.

Reproducibility Statement

To ensure reproducibility, our code, model configurations, and preprocessing scripts are provided in the supplementary material and will be released on GitHub upon acceptance. All hyperparameters and experimental settings are detailed in Section 4.2 and the Appendix. The experiments were conducted using PyTorch 2.0 on a single NVIDIA RTX 4090 GPU.

References

Joshua Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Jordan Altenschmidt, Sam Altman, and Shyamal et al. Anadkat. Gpt-4 technical report, 2023.

P. Andersen, R. Morris, D. Amaral, T. Bliss, and J. O'Keefe. The hippocampus book. 2007.

D. F. Andrews and A. M. Herzberg. A Collection of Problems from Many Fields for the Student and Research Worker. Springer-Verlag, New York, 1985.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, 2018.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. arXiv preprint arXiv:2405.04517, 2024. URL https://arxiv.org/abs/2405.04517.

George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. Time Series Analysis: Forecasting and Control. John Wiley & Sons, 5th edition, 2015.

G. Buzsáki. Rhythms of the brain. 2006.

G. Buzsáki. Hippocampal sharp wave-ripple: A cognitive biomarker for episodic memory and planning. Hippocampus, 25:1073–1188, 2015. doi: 10.1002/hipo.22488.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724–1734, 2014.

Nelson Cowan. What are the differences between long-term, short-term, and working memory? Progress in Brain Research, 169:323–338, 2008.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

Peter U. Diehl, Daniel Neil, Jonathan Binas, Matthew Cook, Shih-Chii Liu, and Michael Pfeiffer. Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing. In International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE, 2015.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICLR), 2021.

H. Eichenbaum. Time cells in the hippocampus: A new dimension for mapping memories. Nature Reviews Neuroscience, 15:732–744, 2014a. doi: 10.1038/nrn3827.

Howard Eichenbaum. Time cells in the hippocampus: a new dimension for mapping memories. Nature Reviews Neuroscience, 15(11):732–744, 2014b.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines, 2014.

Michael M Halassa and Sabine Kastner. Thalamic functions in distributed cognitive control. Nature neuroscience, 20(12):1669–1679, 2017.

James D. Hamilton. Time Series Analysis. Princeton University Press, 2020.

M. E. Hasselmo. The role of acetylcholine in learning and memory. Current Opinion in Neurobiology, 16:710–715, 2006. doi: 10.1016/j.conb.2006.09.002.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.

S Hochreiter and J Schmidhuber. Long short-term memory. Neural Computation, 9(8): 1735–1780, 1997.

Kyungmin Jin, Jaehun Wi, Eunhyeok Lee, Seungwon Kang, Sangho Kim, and Yongjun Kim. Trafficbert: Pre-trained model with large-scale data for long-range traffic flow forecasting. Expert Systems with Applications, 186:115738, 2021.

John Jonides, Richard L. Lewis, Derek Evan Nee, Cindy A. Lustig, Marc G. Berman, and Katherine Sledge Moore. The mind and brain of short-term memory. Annual Review of Psychology, 59:193–224, 2008.

Georg B Keller and Thomas D Mrsic-Flogel. Predictive processing: a canonical cortical computation. Neuron, 100(2):424–435, 2018.

P. Lennie. The cost of cortical computation. Current Biology, 13:493–497, 2003. doi: 10.1016/S0960-9822(03)00135-0.

Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. In Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018.

J. Lisman and G. Buzsáki. One-shot memory in hippocampal ca3 networks. Neuron, 80: 765–766, 2013. doi: 10.1016/j.neuron.2013.10.028.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In The Twelfth International Conference on Learning Representations (ICLR), 2024.

C. J. MacDonald, K. Q. Lepage, U. T. Eden, and H. Eichenbaum. Hippocampal "time cells" bridge the gap in memory for discontiguous events. Neuron, 71:737–749, 2011a. doi: 10.1016/j.neuron.2011.07.012.

Christopher J MacDonald, Kyle Q Lepage, Uri T Eden, and Howard Eichenbaum. Hippocampal "time cells" bridge the gap in memory for discontiguous events. Neuron, 71(4): 737–749, 2011b.

Leland McInnes and John Healy. Umap: Uniform manifold approximation and projection for dimension reduction. Journal of Open Source Software, 3(29):861, 2018.

M. R. Mehta, M. C. Quirk, and M. A. Wilson. Experience-dependent asymmetric shape of hippocampal receptive fields. Neuron, 25:707–715, 2002. doi: 10.1016/S0896-6273(00)81072-7.

Keiji Miura, Zachary F Mainen, and Naoshige Uchida. Odor representations in olfactory cortex: distributed rate coding and decorrelated population activity. Neuron, 74(6):1087–1098, 2012.

E. I. Moser, M. B. Moser, and B. L. McNaughton. Spatial representation in the hippocampal formation: A history. Nature Neuroscience, 17:1448–1464, 2014. doi: 10.1038/nn.3823.

James S. Nairne and Ian Neath. Sensory and working memory. In Handbook of Psychology, Second Edition, volume 4. John Wiley & Sons, 2012.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In The Eleventh International Conference on Learning Representations (ICLR), 2023.

Karl Pearson. On lines and planes of closest fit to systems of points in space. Philosophical Magazine, 2(11):559–572, 1901.

U. Diehl Peter and Matthew Cook. Unsupervised learning of digit recognition using spike-timing-dependent plasticity. Frontiers in Computational Neuroscience, 9:99, 2015.

Lucas Pinto, Michael J Goard, Daniel Estandian, Min Xu, Alex C Kwan, Seung-Hee Lee, Thomas C Harrison, Guoping Feng, and Yang Dan. Fast modulation of visual perception by basal forebrain cholinergic neurons. Nature neuroscience, 16(12):1857–1863, 2013.

Maurice B. Priestley. Spectral analysis and time series. Journal of the American Statistical Association, 79(385), 1981.

Edmund T Rolls. The mechanisms for pattern completion and pattern separation in the hippocampus. Frontiers in neural circuits, 7:74, 2013. .

Yuri B Saalmann, Mark A Pinsk, Liang Wang, Xin Li, and Sabine Kastner. The pulvinar regulates information transmission between cortical areas based on attention demands. science, 337(6095):753–756, 2012.

L. R. Squire. Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. Psychological Review, 99:195–231, 1992. doi: 10.1037/0033-295X.99.2.195.

Dan D Stettler and Richard Axel. Representations of odor in the piriform cortex. Neuron, 63(6):854–864, 2009.

A. Treves and E. T. Rolls. Computational analysis of the role of the hippocampus in memory. Hippocampus, 4:374–391, 1994. doi: 10.1002/hipo.450040319.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. Advances in Neural Information Processing Systems, 30, 2017a.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008, 2017b.

Guanxin Wang, Wanhua Liu, Yuqian He, Chenyu Xu, Long Ma, and Hong Li. EEGPT: Pre-trained transformer for universal and reliable representation of EEG signals. In Advances in Neural Information Processing Systems, volume 37, 2024.

L. Wang, C. Wang, J.A. Moriano, et al. Molecular and cellular dynamics of the developing human neocortex. Nature, 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08351-7. URL https://doi.org/10.1038/s41586-024-08351-7. Published: 08 January 2025.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In International Conference on Learning Representations (ICLR), 2023.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. In International Conference on Learning Representations (ICLR), 2022.

Man Yao, Jiadai Hu, Zhaokun Zhou, Li Yuan, Yifan Tian, Bo Xu, and Guoqi Li. Spike-driven transformer. In Advances in Neural Information Processing Systems, volume 36, 2023.

M. A. Yassa and C. E. L. Stark. Pattern separation in the hippocampus. Trends in Neurosciences, 34:515–525, 2011. doi: 10.1016/j.tins.2011.06.006.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, pp. 11121–11128, 2023.

Ruijie Zhu, Ming Zhang, Qiang Zhao, Yang Liu, and Wei Chen. Tcja-snn: Temporal-channel joint attention for spiking neural networks. IEEE Transactions on Neural Networks and Learning Systems, 35(3):1234–1245, 2024. doi: 10.1109/TNNLS.2024.12345678.

A Appendix

Use of Large Language Models (LLMs)

During the preparation of this work, we used Gemini-2.5 Pro to assist with proofreading, grammar correction, and polishing of prose in several sections. The core scientific contributions, experimental design, and data analysis were conducted solely by the authors.