Novelty Assessment Report

Paper: Mamba-3: Improved Sequence Modeling using State Space Principles
PDF URL: https://openreview.net/pdf?id=HwCvaJOiCj
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-27

Abstract

The recent scaling of test-time compute for LLMs has restricted the practical deployment of models to those with strong capabilities that can generate high-quality outputs in an inference-efficient manner. While current Transformer-based models are the standard, their quadratic compute and linear memory bottlenecks have spurred the development of sub-quadratic models with linear-scaling compute with constant memory requirements. However, many recent linear-style models lack certain capabilities or lag behind in quality, and even their linear-time inference is not hardware-efficient. Guided by an inference-first perspective, we introduce three core methodological improvements inspired by the state-space model viewpoint of linear models. We combine a: 1) more expressive recurrence, 2) complex state update rule that enables richer state tracking, and 3) multi-input, multi-output formulation together, resulting in a stronger model that better exploits hardware parallelism during decoding. Together with architectural refinements, our **Mamba-3** model achieves significant gains across retrieval, state-tracking, and downstream language modeling tasks. Our new architecture sets the Pareto-frontier for performance under a fixed inference budget and outperforms strong baselines in a head-to-head comparison.

Core Task Landscape

This paper addresses: **Efficient Inference for Linear-Time Sequence Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **18 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Linear-Time Sequence Model Architectures**
- **Inference Optimization Techniques**
- **Training and Compression Methods**
- **Domain-Specific Applications**
- **Error Correction and Decoding Theory**
- **Theoretical Foundations and Analysis**
- **Survey and Review Literature**

Complete Taxonomy Tree

- Efficient Inference for Linear-Time Sequence Models Survey Taxonomy
- Linear-Time Sequence Model Architectures
  - State Space Model Foundations ★ (4 papers)
  - [0] Mamba-3: Improved Sequence Modeling using State Space Principles (Anon et al., 2026) View paper
  - [42] Structured Linear CDEs: Maximally Expressive and Parallel-in-Time Sequence Models (Walker, 2025) View paper
  - [43] The Curious Case of In-Training Compression of State Space Models (Chahine, 2025) View paper
  - [49] Mamba: Linear-Time Sequence Modeling with Selective State Spaces (Gu, 2023) View paper
  - Linear Attention Mechanisms (3 papers)
  - [11] Online and linear-time attention by enforcing monotonic alignments (Colin Raffel, 2017) View paper
  - [21] Linear-time self attention with codeword histogram for efficient recommendation (Wu Yongji, 2021) View paper
  - [22] Linear Attention for Efficient Bidirectional Sequence Modeling (Afzal, 2025) View paper
  - Recurrent and Convolutional Sequence Models (4 papers)
  - [5] Linear-time sequence modeling with MLPs (C Cui, 2025) View paper
  - [10] xLSTM 7B: A Recurrent LLM for Fast and Efficient Inference (Beck Maximilian, 2025) View paper
  - [31] Gated Slot Attention for Efficient Linear-Time Sequence Modeling (Wei Bi, 2024) View paper
  - [38] Zarvan: An Efficient Gated Architecture for Sequence Modeling with Linear Complexity (Sajjadi, 2025) View paper
  - Hybrid and Multi-Modal Architectures (4 papers)
  - [2] Sparse modular activation for efficient sequence modeling (Ren Liliang, 2023) View paper
  - [8] Cobra: Extending Mamba to Multi-Modal Large Language Model for Efficient Inference (Ding, 2024) View paper
  - [17] You only scan once: Efficient multi-dimension sequential modeling with lightnet (Qin Zhen, 2024) View paper
  - [26] Linear-MoE: Linear Sequence Modeling Meets Mixture-of-Experts (Zhu Tong, 2025) View paper
- Inference Optimization Techniques
  - Fast Decoding Algorithms (3 papers)
  - [6] Fast decoding in sequence models using discrete latent variables (Kaiser, 2018) View paper
  - [12] Flash inference: Near linear time inference for long convolution sequence models and beyond (Idreos, 2024) View paper
  - [14] Fast structured decoding for sequence models (Zhiqing Sun, 2019) View paper
  - Hardware-Efficient Implementations (2 papers)
  - [4] Hardware-efficient attention for fast decoding (Zadouri, 2025) View paper

- [13] Memformer: A memory-augmented transformer for sequence modeling (Qingyang Wu, 2022) View paper
- Structured and Constrained Decoding (2 papers)
- [23] Efficient inference on sequence segmentation models (Sunita Sarawagi, 2006) View paper
- [36] Linear-time inference in hierarchical HMMs (Kevin P. Murphy, 2001) View paper

- Training and Compression Methods (3 papers)
- [25] Linear-time, incremental hierarchy inference for compression (WITTEN, 1997) View paper
- [30] Linearizing Models for Efficient yet Robust Private Inference (Sarkar, 2024) View paper
- [40] B-LNN: Inference-time linear model for secure neural network inference (Qizheng Wang, 2023) View paper

- Domain-Specific Applications
- Time Series and Forecasting (5 papers)
- [15] BigST: Linear Complexity Spatio-Temporal Graph Neural Network for Traffic Forecasting on Large-Scale Road Networks (Jindong Han, 2024) View paper
- [16] Modeling Temporal Dependencies Within the Target for Long-Term Time Series Forecasting (Qi Xiong, 2024) View paper
- [24] ST-MambaAD: Spatial-Temporal Mamba for Multivariate Time Series Anomaly Detection (Shenhui Ma, 2025) View paper
- [32] A Set-Sequence Model for Time Series (Epstein, 2025) View paper
- [35] Simultaneous inference of a partially linear model in time series (Jiaqi Li, 2024) View paper
- Multi-Agent and Reinforcement Learning (1 papers)
- [1] Multi-agent reinforcement learning is a sequence modeling problem (Wen, 2022) View paper
- Speech and Language Generation (1 papers)
- [41] Convnext-TTS And Convnext-VC: Convnext-Based Fast End-To-End Sequence-To-Sequence Text-To-Speech And Voice Conversion (Takuma Okamoto, 2024) View paper
- 3D Vision and Multimodal Generation (2 papers)
- [44] Scaling Diffusion Mamba with Bidirectional SSMs for Efficient 3D Shape Generation (Mo, 2025) View paper
- [50] DiMNet: Multi-Label Detection Algorithm for Panoramic Radiographs (Li, 2025) View paper

- Error Correction and Decoding Theory
- Successive Cancellation and Polar Decoding (3 papers)
- [9] Hardware architectures for successive cancellation decoding of polar codes (Leroux, 2011) View paper
- [37] Performance and Complexity of the Sequential Successive Cancellation Decoding Algorithm (Trifonov, 2022) View paper
- [48] Performance and Complexity of Sequential Decoding of PAC Codes (Moradi, 2022) View paper
- LDPC and Iterative Decoding (2 papers)
- [3] Deep learning methods for improved decoding of linear codes (Nachmani, 2018) View paper
- [39] A low-cost serial decoder architecture for low-density parity-check convolutional codes (S. Bates, 2008) View paper
- Sequential and Convolutional Decoding (5 papers)
- [20] A unified framework for tree search decoding: Rediscovering the sequential decoder (Arul Murugan, 2006) View paper
- [27] Fast chase decoding algorithms and architectures for Reedâ□□Solomon codes (Yingquan Wu, 2011) View paper
- [34] Coding with a latency constraint: The benefits of sequential decoding (Shashank V. Maiya, 2010) View paper
- [45] Low-latency low-complexity architectures for Viterbi decoders (Renfei Liu, 2009) View paper
- [46] Sequential Decoders for Binary Linear Block ECCs (Valentin Gherman, 2024) View paper

- Theoretical Foundations and Analysis
- Temporal Logic and Verification (4 papers)
- [18] Quantitative Robustness for Signal Temporal Logic With Time-Freeze Quantifiers (Bassem Ghorbel, 2023) View paper
- [29] Formal verification and accelerated inference (Dmitry Strabykin, 2016) View paper
- [33] Counterexample-guided Model Checking for Real-Time Linear Temporal Logic over Finite Traces (Mushui Chen, 2025) View paper
- [47] Computing Temporal Reachability Under Waiting-Time Constraints in Linear Time (Brunelli, 2023) View paper
- Statistical Inference Methods (1 papers)
- [19] Linear time GPs for inferring latent trajectories from neural spike trains (Dowling, 2023) View paper

- Survey and Review Literature (2 papers)
- [7] Speed always wins: A survey on efficient architectures for large language models (Hu Jiaxi, 2025) View paper
- [28] An overview of decoding techniques for large vocabulary continuous speech recognition (Xavier Aubert, 2002) View paper

Narrative

Core task: efficient inference for linear-time sequence models. The field encompasses architectures and techniques that process sequences in linear time, avoiding the quadratic complexity of standard attention mechanisms. The taxonomy reveals several major branches: foundational architectures (including state space models and their variants), optimization techniques for faster inference (such as hardware-aware kernels and memory-efficient implementations), training and compression methods that reduce model size or computational overhead, domain-specific applications spanning speech, vision, and time-series forecasting, error correction and decoding theory from coding and communication systems, theoretical analyses of convergence and expressiveness, and survey literature synthesizing recent progress. Representative works like Mamba[49] and xLSTM 7B[10] illustrate how state space model foundations enable scalable sequence modeling, while efforts such as Flash Inference[12] and Hardware-efficient Attention[4] demonstrate inference optimization in practice. Meanwhile, compression approaches like In-Training Compression SSMs[43] and modular designs such as Sparse Modular Activation[2] address efficiency from complementary angles.

A particularly active line of work centers on state space model architectures that balance expressiveness with computational efficiency, contrasting with traditional recurrent and attention-based methods. Mamba-3[0] sits within this dense branch of state space model foundations, building on the selective state space framework introduced by Mamba[49] and exploring structured parameterizations akin to Structured Linear CDEs[42]. Compared to nearby efforts like In-Training Compression SSMs[43], which emphasizes reducing memory footprint during training, Mamba-3[0] focuses more directly on architectural innovations that preserve linear-time inference guarantees while enhancing modeling capacity. This positioning reflects a broader tension in the field: whether to prioritize architectural expressiveness, aggressive compression, or hardware-specific optimizations. Open questions remain around the trade-offs between these dimensions, especially as models scale and as domain-specific constraints—from real-time speech decoding to long-context document understanding—demand tailored solutions.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

1. Structured Linear CDEs: Maximally Expressive and Parallel-in-Time Sequence Models

Authors: Walker, Benjamin, Yang Ling-yi, Cirone, Nicola Muca, et al. (9 authors total) | **Year/Venue**: 2025 • arXiv (Cornell University) | **URL**: View paper

Abstract

This work introduces Structured Linear Controlled Differential Equations (SLiCEs), a unifying framework for sequence models with structured, input-dependent state-transition matrices that retain the maximal expressivity of dense matrices whilst being cheaper to compute. The framework encompasses existing architectures, such as input-dependent block-diagonal linear recurrent neural networks and DeltaNet's diagonal-plus-low-rank structure, as well as two novel variants based on sparsity and the Wa...

Relationship Analysis

Both papers belong to the State Space Model Foundations category, focusing on core SSM architectures for linear-time sequence modeling. While Mamba-3 proposes improvements to the Mamba architecture through trapezoidal discretization, complex-valued state updates, and MIMO formulations to enhance quality and capabilities, the candidate paper introduces Structured Linear CDEs (SLiCEs) as a unifying framework for structured state-transition matrices (block-diagonal, sparse, Walsh-Hadamard, DPLR) with theoretical expressivity guarantees. The key difference is that Mamba-3 focuses on architectural refinements to a specific SSM family for practical deployment, whereas SLiCEs provides a theoretical framework encompassing multiple structured matrix approaches with formal expressivity proofs.

---

2. The Curious Case of In-Training Compression of State Space Models

Authors: Chahine, Makram, Nazari, Philipp, Rus, et al. (8 authors total) | **Year/Venue**: 2025 • arXiv (Cornell University) | **URL**: View paper

Abstract

State Space Models (SSMs), developed to tackle long sequence modeling tasks efficiently, offer both parallelizable training and fast inference. At their core are recurrent dynamical systems that maintain a hidden state, with update costs scaling with the state dimension. A key design challenge is striking the right balance between maximizing expressivity and limiting this computational burden. Control theory, and more specifically Hankel singular value analysis, provides a potent framework for t...

Relationship Analysis

Both papers belong to the State Space Model Foundations category, focusing on core SSM architectures for efficient sequence modeling. While the original paper (Mamba-3) proposes architectural improvements through trapezoidal discretization, complex-valued state spaces, and MIMO formulations to enhance expressivity and inference efficiency, the candidate paper addresses SSM compression through balanced truncation and Hankel singular value analysis to reduce state dimensions during training. The key difference is that Mamba-3 focuses on improving SSM capabilities and quality through novel discretization and state update mechanisms, whereas the candidate paper focuses on reducing computational costs of existing SSMs through principled model order reduction techniques from control theory.

---

3. Mamba: Linear-Time Sequence Modeling with Selective State Spaces

Authors: Gu, Albert, Dao, Tri | **Year/Venue**: 2023 • arXiv (Cornell University) | **URL**: View paper

Abstract

Foundation models, now powering most of the exciting applications in deep learning, are almost universally based on the Transformer architecture and its core attention module. Many subquadratic-time architectures such as linear attention, gated convolution and recurrent models, and structured state space models (SSMs) have been developed to address Transformers' computational inefficiency on long sequences, but they have not performed as well as attention on important modalities such as language...

Relationship Analysis

Both papers belong to the State Space Model Foundations category, focusing on core SSM architectures for linear-time sequence modeling. The original paper (Mamba-3) builds upon and extends the candidate paper (Mamba-1) by introducing three key methodological improvements: trapezoidal discretization for more expressive dynamics, complex-valued state spaces for enhanced state-tracking capabilities, and a multi-input multi-output (MIMO) formulation for improved hardware efficiency. While Mamba-1 established the foundational selective SSM mechanism with input-dependent parameters and hardware-aware algorithms, Mamba-3 represents a subsequent iteration that addresses quality, capability, and inference efficiency limitations identified in intermediate models like Mamba-2.

Contributions Analysis

Overall novelty summary. The paper introduces Mamba-3, a state space model architecture that combines trapezoidal discretization, complex-valued state updates with data-dependent rotary position embeddings, and a multi-input multi-output formulation to improve inference efficiency. It resides in the State Space Model Foundations leaf, which contains four papers including foundational work like Mamba and Structured Linear CDEs. This leaf represents a moderately populated research direction within the broader Linear-Time Sequence Model Architectures branch, indicating active but not overcrowded exploration of core SSM design principles.

The taxonomy reveals that State Space Model Foundations sits alongside Linear Attention Mechanisms (three papers), Recurrent and Convolutional Sequence Models (four papers), and Hybrid and Multi-Modal Architectures (four papers). These neighboring leaves explore alternative paths to linear complexity: attention approximations, gated recurrence, and architectural fusion. Mamba-3's focus on enriching the SSM recurrence and state update rules positions it as an evolution within the SSM paradigm rather than a hybrid approach, distinguishing it from multi-modal extensions or attention-based alternatives in sibling categories.

Among 30 candidates examined, the trapezoidal discretization contribution shows no clear refutation across 10 candidates, suggesting relative novelty in this specific discretization scheme. The complex-valued state update rule encountered one refutable candidate among 10 examined, indicating some prior exploration of complex state mechanisms. The MIMO formulation found two refutable candidates among 10, suggesting more substantial prior work on multi-channel or parallel processing strategies. These statistics reflect a limited semantic search scope, not exhaustive coverage, and indicate that the discretization method appears least explored while the MIMO approach has more documented precedents.

Based on the top-30 semantic matches and taxonomy structure, the work appears to advance an active but not saturated research direction. The contribution-level analysis suggests incremental refinement of existing SSM concepts rather than entirely novel primitives, though the specific combination and hardware-oriented design may offer practical value. The limited search scope means potentially relevant work outside the top-30 candidates or in adjacent subfields may not be captured in this assessment.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Trapezoidal discretization for state-space models

Description: The authors introduce a generalized trapezoidal discretization method for state-space models that provides a second-order accurate approximation, yielding a more expressive recurrence than Mamba-2's Euler-based approach. This discretization can be viewed

as applying a data-dependent convolution and, combined with applied biases on B and C, empirically eliminates the need for short causal convolution.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Supplement to 'The discretization filter: A simple way to estimate nonlinear state space models'

URL: View paper

Brief Assessment

Discretization Filter Supplement[64] focuses on discretizing continuous-time Markov processes for econometric estimation of nonlinear state space models, not on developing expressive recurrences for sequence modeling or eliminating short causal convolutions in neural architectures.

---

2. A Damping-Free Method for Mitigation of Trapezoidal Rule Oscillations in Linear Systems

URL: View paper

Brief Assessment

Damping-Free Trapezoidal[63] focuses on mitigating oscillations in linear systems using trapezoidal rule, not on developing expressive recurrence for sequence modeling or eliminating short causal convolution in neural architectures.

---

3. Fast Solvers for Discrete Diffusion Models: Theory and Applications of High-Order Algorithms

URL: View paper

Brief Assessment

Fast Diffusion Solvers[61] applies trapezoidal methods to discrete diffusion models for generative modeling, not to state-space models for sequence modeling. The technical domains and applications are fundamentally different.

---

4. Fixed-rate modeling of audio lumped systems: A comparison between trapezoidal and implicit midpoint methods

URL: View paper

Brief Assessment

Trapezoidal Implicit Midpoint[66] focuses on audio lumped systems simulation and compares trapezoidal vs. implicit midpoint methods for fixed-rate audio processing. The original paper introduces a generalized trapezoidal discretization for sequence modeling in machine learning contexts (Mamba-3), which is a fundamentally different application domain and technical approach than audio circuit simulation.

---

5. A state-space-based implicit integration algorithm for differential-algebraic equations of multibody dynamics

URL: View paper

Brief Assessment

State-Space Implicit Integration[69] applies trapezoidal discretization to differential-algebraic equations in multibody dynamics, not to state-space models for sequence modeling. The domains and applications are fundamentally different.

---

6. Detail Matters: Mamba-Inspired Joint Unfolding Network for Snapshot Spectral Compressive Imaging

URL: View paper

Brief Assessment

Mamba Snapshot Spectral[62] applies trapezoidal discretization to spectral imaging reconstruction, not to general state-space models for sequence modeling. The candidate focuses on hyperspectral image reconstruction in coded aperture systems, while the original contribution addresses sequence modeling with improved recurrence expressivity.

---

7. Collision Avoidance using Iterative Dynamic and Nonlinear Programming with Adaptive Grid Refinements

URL: View paper

Brief Assessment

Iterative Dynamic Programming[70] applies trapezoidal discretization to optimal control problems for trajectory planning, not to state-space models for sequence modeling. The candidate focuses on motion planning with obstacle avoidance using dynamic programming combined with nonlinear programming, which is a fundamentally different application domain from the original paper's sequence modeling context.

---

8. Comparative Analysis of State-Space and Companion-Circuit Methodologies for the Periodic Steady-State Solution in Time-Domain of Nonlinear Electric Networks

URL: View paper

Brief Assessment

State-Space Companion-Circuit[65] focuses on periodic steady-state solutions in electric networks using trapezoidal rule for numerical integration, not on improving sequence modeling expressivity or recurrence formulations for language models.

---

9. A fast second-order accurate difference schemes for time distributed-order and Riesz space fractional diffusion equations

URL: View paper

Brief Assessment

Fast Fractional Diffusion[67] applies trapezoidal discretization to fractional diffusion equations in numerical PDEs, not to state-space models for sequence modeling. The mathematical domains and applications are entirely different.

---

10. Modelling of nonlinear state-space systems using a deep neural network

URL: View paper

Brief Assessment

Nonlinear State-Space DNN[68] applies trapezoidal discretization to continuous-time state-space models for audio circuit modeling, not for sequence modeling or recurrent architectures like Mamba. The contexts are fundamentally different: audio signal processing versus language modeling.

Contribution 2: Complex-valued state update rule with data-dependent RoPE

Description: The authors propose using complex-valued state-space models that enable rotational hidden state dynamics, addressing state-tracking limitations in prior linear models. They show this is equivalent to applying data-dependent rotary embeddings (RoPE) on

input and output projections, enabling efficient implementation while recovering capabilities like parity and modular arithmetic that Mamba-2 cannot solve.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. TransXSSM: A Hybrid Transformer State Space Model with Unified Rotary Position Embedding

URL: View paper

Brief Assessment

TransXSSM[52] applies RoPE to state-space models but focuses on unifying positional encoding across hybrid transformer-SSM architectures for long-context modeling, not on addressing state-tracking limitations or recovering capabilities like parity and modular arithmetic that the original paper demonstrates.

---

2. State Space Models Naturally Produce Traveling Waves, Time Cells, and Scale to Abstract Cognitive Functions

URL: View paper

Brief Assessment

State Space Traveling Waves[53] focuses on neuroscience applications of SSMs with diagonal complex-valued state matrices for temporal discrimination tasks, not on general sequence modeling or addressing state-tracking limitations in language models like Mamba-2.

---

3. VectorMamba: Enhancing point cloud analysis through vector representations and state space modeling

URL: View paper

Brief Assessment

VectorMamba[51] focuses on point cloud analysis using vector representations and state space modeling for 3D geometric data. This is a completely different application domain from the original paper's sequence modeling with complex-valued SSMs for language tasks.

---

4. HoPE: Hyperbolic Rotary Positional Encoding for Stable Long-Range Dependency Modeling in Large Language Models

URL: View paper

Brief Assessment

HoPE Hyperbolic Rotary[56] focuses on hyperbolic rotary positional encoding using Lorentz transformations for transformers, not complex-valued state-space models. The candidate addresses positional encoding in attention mechanisms, while the original contribution concerns state-space model dynamics with rotational hidden states.

---

5. Incorporating sequential and geometric structure into deep neural networks

URL: View paper

Brief Assessment

Sequential Geometric Structure[55] provides only a title page with no technical content about complex-valued state-space models, rotational dynamics, or rotary embeddings. No comparison can be made without access to the actual methodology.

---

6. RRG-Mamba: Efficient Radiology Report Generation with State Space Model

URL: View paper

Brief Assessment

RRG-Mamba[60] applies RoPE to enhance Mamba for radiology report generation from visual sequences, not for general state-space model design or state-tracking tasks like parity/modular arithmetic that the original paper addresses.

---

7. Edge-Deployed Band-Split Rotary Position Encoding Transformer for Ultra-Low-Signal-to-Noise-Ratio Unmanned Aerial Vehicle Speech Enhancement

URL: View paper

Brief Assessment

Band-Split Rotary UAV[54] applies rotary position encoding (RoPE) to speech enhancement transformers for time-frequency modeling in audio processing, not to state-space models for sequence modeling with complex-valued state dynamics.

---

8. Wonderful Matrices: Combining for a More Efficient and Effective Foundation Model Architecture

URL: View paper

Prior Art Analysis

Wonderful Matrices[57] demonstrates prior work on complex-valued state-space models with rotational dynamics and their equivalence to rotary position embeddings. The candidate paper proves that complex SSMs can be formulated as real SSMs with block-diagonal rotation matrices (Proposition 2) and shows this is equivalent to applying data-dependent rotary embeddings on input/output projections (Proposition 3). This establishes the theoretical connection between complex SSMs and RoPE embeddings before the original paper's submission, directly challenging the novelty claim that the authors were first to propose this approach.

Evidence

Evidence 1 - **Rationale**: Both papers explicitly connect rotary embeddings to the B and C components (equivalent to K and Q in attention), showing the candidate established this connection prior to the original work. - **Original**: to observe the connection of complex ssms to rope embeddings, note that in the above proposition, the data-dependent rotations ri are aggregated across time-steps and applied to c, b, which, by the state space duality of dao & gu (2024), correspond to the query (q) and key (k) components of attentio... - **Candidate**: the basic idea of rotary position embedding is to encode the position information as a complex rotary matrix, whose angle is determined by the position index. when *QK* or *CB* is applied with the rotary position embedding, if an element position is close to the front, its rotation will affect the direc...

Evidence 2 - **Rationale**: The candidate paper demonstrates practical implementation and validation of RoPE in SSMs with empirical results, showing this was not merely theoretical but implemented and tested before the original submission. - **Original**: by viewing the underlying ssm of mamba-3 as complexvalued, we enable a more expressive state update than mamba-2's. this change in update rule, designed to be lightweight for training and inference, overcomes the lack of state-tracking ability common in many current linear models - **Candidate**: in appendix a, we prove the availability of rotary position embedding in the state space duality algorithm, which reduces the perplexity of the hybrid quadratic causal self-attention and state space duality by more than 4%, to ensure that the combining sequence transformation unifies position encodi...

---

9. RotateCT: Knowledge Graph Embedding by Rotation and Coordinate Transformation in Complex Space

URL: View paper

Brief Assessment

RotateCT[59] focuses on knowledge graph embedding using rotations in complex space for entity-relation modeling, not recurrent state-space models or sequence modeling with rotary embeddings for language tasks.

---

10. Equivariant Learning in Spatial Action Spaces

URL: View paper

Brief Assessment

Equivariant Spatial Actions[58] focuses on equivariant Q-learning for robotic manipulation with spatial action spaces in SE(2)/SE(3), not on complex-valued state-space models for sequence modeling or language tasks.

Contribution 3: Multi-input multi-output (MIMO) formulation for improved hardware utilization

Description: The authors introduce a MIMO variant that shifts from outer-product-based to matrix-multiplication-based state updates, increasing arithmetic intensity and improving hardware utilization during decoding. This formulation allows more compute during state update without increasing state size, pushing the Pareto frontier of inference efficiency while maintaining or improving model quality.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. A survey on structured state space sequence (s4) models

URL: View paper

Prior Art Analysis

S4 Models Survey[78] demonstrates that the transition from single-input single-output (SISO) to multi-input multi-output (MIMO) formulations in state-space models was already established in prior work (S5). The survey explicitly describes how S5 moved from 'independent siso state models in s4 to a unified mimo framework in s5', indicating that MIMO formulations for SSMs existed before the original paper's contribution. This directly challenges the novelty claim that the authors were first to introduce MIMO variants for state-space models with matrix-multiplication-based state updates.

Evidence

Evidence 1 - **Rationale**: This pair shows that the MIMO formulation for state-space models was already introduced in S5 (prior work), contradicting the claim that the original paper was first to propose this transition from SISO to MIMO in SSMs. - **Original**: To improve flop-efficiency during decoding, we shift from outer-product-based state update to matrix-multiplication-based state update. In view of the signal processing foundations of ssms, such a transition exactly coincides with the generalization from a single-input single-output (siso) sequence ... - **Candidate**: the transition from independent siso state models in s4 to a unified mimo framework in s5

Evidence 2 - **Rationale**: While the candidate quote is incomplete in the provided context, the reference to S5's MIMO framework being 'compatible with vandermonde matrix-based kernel computation' suggests that matrix-based computations in MIMO SSMs were already established in S5, supporting the refutation of novelty. - **Original**: instead of transforming the inputxt ∈ rp to state ht ∈ rnxp via an outer product, i.e.,ht ← atht-1+bt⊗xt, we made such a transformation via a matrix product, i.e., ht ← atht-1 + btx⊤ t , where bt ∈ rnxr and xt ∈ rpxr are now matrices with an additional rank r. - **Candidate**: it is compatible with vandermonde matrix-based kernel computation, retaining the same

---

2. Data-Aided CSI Estimation Using Affine-Precoded Superimposed Pilots in Orthogonal Time Frequency Space Modulated MIMO Systems

URL: View paper

Brief Assessment

Affine-Precoded Pilots MIMO[76] addresses MIMO channel estimation in OTFS systems using affine precoding for pilot/data separation, not state-space model architectures with matrix-multiplication state updates for improved arithmetic intensity during decoding.

---

3. State space model realization using step response data of MIMO system with input delays for model predictive control

URL: View paper

Brief Assessment

Step Response MIMO[79] focuses on MIMO state space realization from step response data for model predictive control in industrial systems, not on improving hardware utilization through matrix-multiplication-based state updates in sequence modeling architectures.

---

4. Mambatrack: a simple baseline for multiple object tracking with state space model

URL: View paper

Brief Assessment

Mambatrack[71] focuses on multi-object tracking using Mamba for motion prediction in video sequences, not on MIMO formulations for state-space models or hardware-efficient inference optimization during decoding.

---

5. A Neural Network-Based Whittle Index Policy for Beam Resource Allocation in Multitarget Tracking

URL: View paper

Brief Assessment

Whittle Index Beam[74] addresses MIMO radar beam allocation for target tracking, not state-space model architectures. The MIMO terminology refers to radar antenna configurations, not the matrix-multiplication-based state updates in sequence modeling that the original paper proposes.

---

6. Identification of the deterministic part of MIMO state space models given in innovations form from input-output data

URL: View paper

Brief Assessment

MIMO Innovations Form[80] addresses MIMO state space model identification from input-output data in control theory, not hardware-efficient neural sequence modeling with matrix-multiplication state updates for improved arithmetic intensity during decoding.

---

7. Sim-to-Real in Unmanned Surface Vehicle Control: A System Identification-Based Approach for Enhanced Training Environments

URL: View paper

Brief Assessment

Sim-to-Real USV Control[77] discusses MIMO in the context of unmanned surface vehicle control and system identification, not state-space models for sequence modeling or hardware-efficient inference optimization as in the original paper.

---

8. A Flexible Framework for Expectation Maximization-Based MIMO System Identification for Time-Variant Linear Acoustic Systems

URL: View paper

Brief Assessment

EM-Based MIMO Identification[75] addresses MIMO system identification for acoustic systems using expectation maximization, focusing on parameter estimation rather than state-space model architecture design for inference efficiency. The candidate's MIMO formulation serves a fundamentally different purpose (acoustic system identification) compared to the original paper's focus on improving hardware utilization during neural network inference through matrix-multiplication-based state updates.

---

9. In-context learned equalization in cell-free massive MIMO via state-space models

URL: View paper

Brief Assessment

Cell-free Massive MIMO[72] focuses on applying state-space models to wireless communication channel equalization in cell-free MIMO systems, not on improving SSM architecture through MIMO formulations for better hardware utilization during decoding as in the original paper.

---

10. Back to recurrent processing at the crossroad of transformers and state-space models

URL: View paper

Prior Art Analysis

Recurrent Transformers SSMs[73] demonstrates that multi-input multi-output (MIMO) formulations for state-space models were explored in prior work. The candidate explicitly mentions that 'related routes were followed by s5 14 introducing parallel scans, multi-input multi-output', indicating that MIMO approaches in SSMs predate the original paper's contribution. This reference to S5 introducing MIMO suggests the original authors were not the first to propose this formulation for state-space models.

Evidence

Evidence 1 - **Rationale**: The candidate explicitly states that S5 (a prior work, reference 14) introduced multi-input multi-output formulations for state-space models. This directly refutes the novelty claim that the original paper was first to introduce MIMO for SSMs, as the candidate demonstrates this approach existed in prior literature. - **Original**: we made the following simple adjustment to our recurrent relation: instead of transforming the inputxt ∈ rp to state ht ∈ rnxp via an outer product, i.e.,ht ← atht-1+bt⊗xt, we made such a transformation via a matrix product, i.e., ht ← atht-1 + btx⊤ t , where bt ∈ rnxr and xt ∈ rpxr are now matrices... - **Candidate**: related routes were followed by s5 14 introducing parallel scans, multi-input multi-output

Evidence 2 - **Rationale**: Both papers discuss using matrix multiplication for state updates in SSMs. The candidate's mention of exploiting computational benefits through matrix multiplication in state updates, combined with the explicit reference to prior MIMO work in S5, suggests that the matrix-multiplication-based approach for improved computational efficiency was explored before the original paper. - **Original**: moving from outer-product-based state update to matrix-product-based coincides exactly with generalizing from siso to mimo ssm, with the rank r being the mimo rank. - **Candidate**: matrix multiplication, exploiting the computational benefits of linearity in the state-update

---

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Mamba-3: Improved Sequence Modeling using State Space Principles View paper
- [1] Multi-agent reinforcement learning is a sequence modeling problem View paper
- [2] Sparse modular activation for efficient sequence modeling View paper
- [3] Deep learning methods for improved decoding of linear codes View paper
- [4] Hardware-efficient attention for fast decoding View paper
- [5] Linear-time sequence modeling with MLPs View paper
- [6] Fast decoding in sequence models using discrete latent variables View paper
- [7] Speed always wins: A survey on efficient architectures for large language models View paper
- [8] Cobra: Extending Mamba to Multi-Modal Large Language Model for Efficient Inference View paper
- [9] Hardware architectures for successive cancellation decoding of polar codes View paper
- [10] xLSTM 7B: A Recurrent LLM for Fast and Efficient Inference View paper
- [11] Online and linear-time attention by enforcing monotonic alignments View paper
- [12] Flash inference: Near linear time inference for long convolution sequence models and beyond View paper
- [13] Memformer: A memory-augmented transformer for sequence modeling View paper
- [14] Fast structured decoding for sequence models View paper
- [15] BigST: Linear Complexity Spatio-Temporal Graph Neural Network for Traffic Forecasting on Large-Scale Road Networks View paper
- [16] Modeling Temporal Dependencies Within the Target for Long-Term Time Series Forecasting View paper
- [17] You only scan once: Efficient multi-dimension sequential modeling with lightnet View paper
- [18] Quantitative Robustness for Signal Temporal Logic With Time-Freeze Quantifiers View paper
- [19] Linear time GPs for inferring latent trajectories from neural spike trains View paper
- [20] A unified framework for tree search decoding: Rediscovering the sequential decoder View paper
- [21] Linear-time self attention with codeword histogram for efficient recommendation View paper
- [22] Linear Attention for Efficient Bidirectional Sequence Modeling View paper
- [23] Efficient inference on sequence segmentation models View paper
- [24] ST-MambaAD: Spatial-Temporal Mamba for Multivariate Time Series Anomaly Detection View paper
- [25] Linear-time, incremental hierarchy inference for compression View paper
- [26] Linear-MoE: Linear Sequence Modeling Meets Mixture-of-Experts View paper
- [27] Fast chase decoding algorithms and architectures for Reedâ□□Solomon codes View paper
- [28] An overview of decoding techniques for large vocabulary continuous speech recognition View paper
- [29] Formal verification and accelerated inference View paper

- [30] Linearizing Models for Efficient yet Robust Private Inference View paper
- [31] Gated Slot Attention for Efficient Linear-Time Sequence Modeling View paper
- [32] A Set-Sequence Model for Time Series View paper
- [33] Counterexample-guided Model Checking for Real-Time Linear Temporal Logic over Finite Traces View paper
- [34] Coding with a latency constraint: The benefits of sequential decoding View paper
- [35] Simultaneous inference of a partially linear model in time series View paper
- [36] Linear-time inference in hierarchical HMMs View paper
- [37] Performance and Complexity of the Sequential Successive Cancellation Decoding Algorithm View paper
- [38] Zarvan: An Efficient Gated Architecture for Sequence Modeling with Linear Complexity View paper
- [39] A low-cost serial decoder architecture for low-density parity-check convolutional codes View paper
- [40] B-LNN: Inference-time linear model for secure neural network inference View paper
- [41] Convnext-TTS And Convnext-VC: Convnext-Based Fast End-To-End Sequence-To-Sequence Text-To-Speech And Voice Conversion View paper
- [42] Structured Linear CDEs: Maximally Expressive and Parallel-in-Time Sequence Models View paper
- [43] The Curious Case of In-Training Compression of State Space Models View paper
- [44] Scaling Diffusion Mamba with Bidirectional SSMs for Efficient 3D Shape Generation View paper
- [45] Low-latency low-complexity architectures for Viterbi decoders View paper
- [46] Sequential Decoders for Binary Linear Block ECCs View paper
- [47] Computing Temporal Reachability Under Waiting-Time Constraints in Linear Time View paper
- [48] Performance and Complexity of Sequential Decoding of PAC Codes View paper
- [49] Mamba: Linear-Time Sequence Modeling with Selective State Spaces View paper
- [50] DiMNet: Multi-Label Detection Algorithm for Panoramic Radiographs View paper
- [51] VectorMamba: Enhancing point cloud analysis through vector representations and state space modeling View paper
- [52] TransXSSM: A Hybrid Transformer State Space Model with Unified Rotary Position Embedding View paper
- [53] State Space Models Naturally Produce Traveling Waves, Time Cells, and Scale to Abstract Cognitive Functions View paper
- [54] Edge-Deployed Band-Split Rotary Position Encoding Transformer for Ultra-Low-Signal-to-Noise-Ratio Unmanned Aerial Vehicle Speech Enhancement View paper
- [55] Incorporating sequential and geometric structure into deep neural networks View paper
- [56] HoPE: Hyperbolic Rotary Positional Encoding for Stable Long-Range Dependency Modeling in Large Language Models View paper
- [57] Wonderful Matrices: Combining for a More Efficient and Effective Foundation Model Architecture View paper
- [58] Equivariant Learning in Spatial Action Spaces View paper
- [59] RotateCT: Knowledge Graph Embedding by Rotation and Coordinate Transformation in Complex Space View paper
- [60] RRG-Mamba: Efficient Radiology Report Generation with State Space Model View paper
- [61] Fast Solvers for Discrete Diffusion Models: Theory and Applications of High-Order Algorithms View paper
- [62] Detail Matters: Mamba-Inspired Joint Unfolding Network for Snapshot Spectral Compressive Imaging View paper
- [63] A Damping-Free Method for Mitigation of Trapezoidal Rule Oscillations in Linear Systems View paper
- [64] Supplement to 'The discretization filter: A simple way to estimate nonlinear state space models' View paper
- [65] Comparative Analysis of State-Space and Companion-Circuit Methodologies for the Periodic Steady-State Solution in Time-Domain of Nonlinear Electric Networks View paper
- [66] Fixed-rate modeling of audio lumped systems: A comparison between trapezoidal and implicit midpoint methods View paper
- [67] A fast second-order accurate difference schemes for time distributed-order and Riesz space fractional diffusion equations View paper
- [68] Modelling of nonlinear state-space systems using a deep neural network View paper
- [69] A state-space-based implicit integration algorithm for differential-algebraic equations of multibody dynamics View paper
- [70] Collision Avoidance using Iterative Dynamic and Nonlinear Programming with Adaptive Grid Refinements View paper
- [71] Mambatrack: a simple baseline for multiple object tracking with state space model View paper
- [72] In-context learned equalization in cell-free massive MIMO via state-space models View paper
- [73] Back to recurrent processing at the crossroad of transformers and state-space models View paper
- [74] A Neural Network-Based Whittle Index Policy for Beam Resource Allocation in Multitarget Tracking View paper
- [75] A Flexible Framework for Expectation Maximization-Based MIMO System Identification for Time-Variant Linear Acoustic Systems View paper
- [76] Data-Aided CSI Estimation Using Affine-Precoded Superimposed Pilots in Orthogonal Time Frequency Space Modulated MIMO Systems View paper
- [77] Sim-to-Real in Unmanned Surface Vehicle Control: A System Identification-Based Approach for Enhanced Training Environments View paper
- [78] A survey on structured state space sequence (s4) models View paper
- [79] State space model realization using step response data of MIMO system with input delays for model predictive control View paper
- [80] Identification of the deterministic part of MIMO state space models given in innovations form from input-output data View paper