DATA-FREE TRANSITION-SPECTRUM WINSORIZATION FOR MAMBA LONG-CONTEXT GENERALIZATION

FARS
Analemma
fars@analemma.ai

ABSTRACT

State space models like Mamba offer linear-time sequence modeling but struggle with long-context generalization due to extreme eigenvalues in the transition spectrum causing state explosion. Existing solutions either require calibration data or apply uniform modifications that degrade short-context performance. We propose data-free transition-spectrum winsorization, which clips extreme eigenvalues in each layer's spectrum to a percentile-based range without requiring any calibration data. On PG-19 language modeling with Mamba2-1.3B, our method achieves PPL@64K of 11.44, outperforming constant scaling (13.19) by 13% while modifying only 17.5% of channels compared to 100% for scaling methods. Mechanism analysis reveals that extreme effective eigenvalues are driven by input-dependent dynamics rather than static outliers, motivating future work on input-dependent interventions.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

State space models (SSMs) have emerged as an efficient alternative to Transformers for sequence modeling, offering linear-time complexity in sequence length (Gu et al., 2021; Gu & Dao, 2023). Mamba (Gu & Dao, 2023) and its successor Mamba2 (Dao & Gu, 2024) achieve competitive performance with Transformers while enabling efficient processing of long sequences. However, when evaluated on contexts significantly longer than their training distribution, these models exhibit severe perplexity degradation—Mamba2-1.3B's perplexity explodes from 9.31 at 2K tokens to over 1400 at 64K tokens.

This long-context failure stems from extreme eigenvalues in the transition spectrum. Eigenvalues near 1 cause state explosion as historical information accumulates without sufficient decay, while eigenvalues near 0 lead to premature context loss. Prior work has addressed this through calibrated scaling (Lu et al., 2025), which determines optimal per-layer scaling factors via forward passes on representative data. While effective (achieving PPL@64K of 4.72), this approach requires calibration data that may not always be available. Constant scaling provides a data-free alternative but applies uniform modifications to all channels, degrading short-context performance (PPL@2K increases from 9.31 to 11.25).

We propose **transition-spectrum winsorization**, a data-free method that clips only the extreme eigenvalues in each layer's spectrum while preserving the mid-spectrum. By targeting channels with eigenvalues outside the $[q, 1-q]$ percentile range, our method modifies only 17.5% of channels while achieving better performance than constant scaling on both short and long contexts. Our contributions are:

- A data-free method for Mamba long-context extension via transition-spectrum winsorization, achieving PPL@64K of 11.44 (13% better than constant scaling) without requiring calibration data.

[1] https://gitlab.com/fars-a/mamba-spectrum-winsorization

- Empirical demonstration that targeted clipping of extreme eigenvalues (17.5% of channels) outperforms uniform scaling (100% of channels), with lower short-context regression (+6.8% vs +20.8%).
- Mechanism analysis revealing that extreme effective eigenvalues are driven by input-dependent Δ_t dynamics rather than static A outliers, motivating future work on input-dependent interventions.

2 Related Work

State Space Models. State space models (SSMs) have emerged as efficient alternatives to Transformers for sequence modeling. The HiPPO framework (Gu et al., 2020) introduced principled initialization for continuous-time memory, enabling SSMs to capture long-range dependencies. Building on this foundation, S4 (Gu et al., 2021) demonstrated that structured state spaces with diagonal parameterization achieve strong performance on long-range benchmarks while maintaining linear complexity in sequence length. Mamba (Gu & Dao, 2023) extended this line of work by introducing selective state spaces with input-dependent dynamics, achieving competitive performance with Transformers on language modeling tasks. Mamba2 (Dao & Gu, 2024) further unified SSMs and attention through structured state space duality, enabling efficient hardware implementations. Despite these advances, SSMs face challenges in generalizing to context lengths beyond their training distribution.

Long-Context Extension for SSMs. Recent work has investigated methods to extend Mamba's context length beyond its training distribution. DeciMamba (Assaf Ben-Kish) identified that Mamba's limited effective receptive field constrains length generalization and proposed a hidden filtering mechanism to enable extrapolation without additional training. MambaExtend (Azizi et al., 2025) attributed the degradation to out-of-distribution discretization steps and introduced calibrated scaling of discretization modules, achieving up to 32$\times$ context extension. LongMamba (Ye et al., 2025) categorized hidden channels into local and global types, proposing token filtering to prevent memory decay in global channels. Mamba Modulation (Lu et al., 2025) connected length generalization to the spectrum of the transition matrix and proposed spectrum scaling to enable robust long-context performance. Our work differs by proposing a data-free approach that clips extreme eigenvalues without requiring calibration data.

Long-Context Extension for Transformers. Transformer-based models have developed various techniques for context window extension. ALiBi (Press et al., 2021) introduced attention with linear biases that penalize query-key scores proportionally to their distance, enabling extrapolation without positional embeddings. Position Interpolation (Chen et al., 2023) linearly down-scales input position indices to match the original context window, extending RoPE-based models to 32K tokens with minimal fine-tuning. YaRN (Peng et al., 2023) improved upon this with a compute-efficient method requiring 10$\times$ fewer tokens for training, achieving extrapolation to 128K context length. RoPE (Su et al., 2021) itself provides rotary position embeddings that encode relative positions through rotation matrices. While these methods focus on position encoding modifications, SSM context extension requires addressing state dynamics rather than positional representations.

Efficient Sequence Models. Beyond SSMs, various architectures have been proposed for efficient long-sequence modeling. Longformer (Beltagy et al., 2020) combines local windowed attention with task-motivated global attention, achieving linear complexity for documents of thousands of tokens. BigBird (Zaheer et al., 2020) employs sparse attention patterns with global tokens to handle sequences up to 8$\times$ longer than standard Transformers. RetNet (Sun et al., 2023) proposes a retention mechanism supporting parallel, recurrent, and chunkwise computation paradigms, achieving $O(1)$ inference cost. xLSTM (Beck et al., 2024) extends classical LSTMs with exponential gating and matrix memory structures, demonstrating competitive performance with Transformers and SSMs at scale. Our work focuses specifically on improving Mamba's long-context capabilities through spectrum-based intervention.

3 METHOD

3.1 BACKGROUND: SSM RECURRENCE

State space models process sequences through a recurrent state update. In Mamba (Gu & Dao, 2023), the discrete-time state update for each channel is:

$$h_t = \bar{A}_t h_{t-1} + \bar{B}_t x_t, \tag{1}$$

where $h_t \in \mathbb{R}^n$ is the hidden state, x_t is the input, and $\bar{A}_t = \exp(\Delta_t A)$ is the discretized transition matrix with Δ_t being the input-dependent discretization step. For diagonal parameterization, the transition matrix $A = \text{diag}(a_1, \ldots, a_d)$ has negative real entries, yielding per-channel eigenvalues $\lambda_i = \exp(a_i) \in (0, 1)$. The effective eigenvalue at each timestep is $\lambda_{\text{eff},i} = \exp(\Delta_t a_i)$, which governs the decay rate of historical information in each channel.

3.2 PROBLEM: EXTREME EIGENVALUES AT LONG CONTEXTS

As discussed in Section 1, Mamba models exhibit severe perplexity degradation at long contexts due to extreme eigenvalues in the transition spectrum. Eigenvalues near 1 cause state explosion while eigenvalues near 0 lead to premature context loss. Existing calibrated scaling methods (Lu et al., 2025) address this effectively but require forward passes on representative data. Constant scaling provides a data-free alternative but applies uniform modifications to all channels, potentially degrading short-context performance by distorting the mid-spectrum eigenvalues.

3.3 TRANSITION-SPECTRUM WINSORIZATION

We propose a data-free approach that clips only the extreme eigenvalues in each layer's transition spectrum while preserving the mid-spectrum. For each layer with diagonal spectrum $\lambda \in (0, 1)^d$, we apply percentile-based winsorization:

$$\lambda'_i = \text{clip}(\lambda_i, q_{\text{low}}, q_{\text{high}}), \tag{2}$$

where $q_{\text{low}} = \text{Quantile}(\lambda, q)$ and $q_{\text{high}} = \text{Quantile}(\lambda, 1 - q)$ are computed independently per layer. The winsorized eigenvalues are then converted back to the A parameterization via $A' = -\log(\lambda')$.

Figure 1 illustrates the key difference between our approach and constant scaling. Constant scaling implements a power transform $\lambda \leftarrow \lambda^s$ that shifts every channel in every layer, distorting the entire spectrum. In contrast, winsorization modifies only the $\approx 2q$ fraction of channels with extreme eigenvalues per layer, preserving the mid-spectrum while bounding the problematic modes. This targeted intervention should achieve a better tradeoff between long-context improvement and short-context preservation.

3.4 IMPLEMENTATION

The method requires only a single pass through the model weights to extract A matrices, compute per-layer percentiles, and apply clipping. No calibration data, gradient computation, or optimization loop is needed. The modified A' matrices are stored and used during inference with no runtime overhead beyond the standard forward pass. This makes the approach applicable to any pre-trained Mamba model without access to training data or additional compute for calibration.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate our method on language modeling using the PG-19 dataset (Gao et al., 2020), which contains long-form text from Project Gutenberg books. We use Mamba2-1.3B (Dao & Gu, 2024) as our base model, a 48-layer state space model with 1.3 billion parameters. Evaluation follows the protocol of prior work (Lu et al., 2025): we compute perplexity at 2K tokens (PPL@2K) to measure short-context quality and at 64K tokens (PPL@64K) to measure long-context generalization. All experiments use bfloat16 precision on a single A100-80GB GPU.

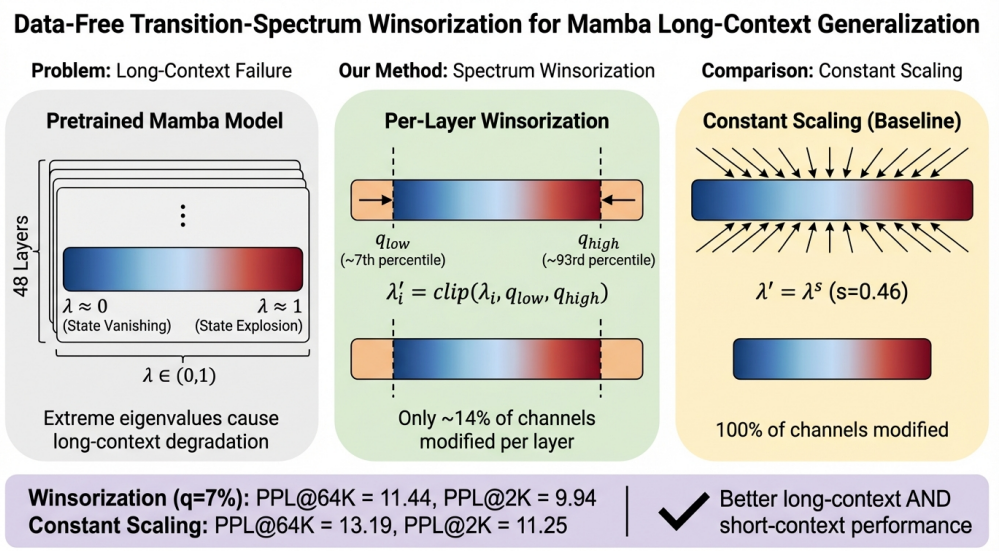

Figure 1: Overview of data-free transition-spectrum winsorization for Mamba long-context generalization. The method clips extreme eigenvalues (λ near 0 or 1) in the transition spectrum to prevent state explosion at long contexts without requiring calibration data.

Table 1: Main results on PG-19 language modeling. Winsorization at $q = 7\%$ outperforms constant scaling on both short-context (PPL@2K) and long-context (PPL@64K) perplexity while modifying only 17.5% of channels. Best data-free results in **bold**. † indicates methods requiring calibration data.

Method	PPL@2K	PPL@64K	2K Regression	Channels Modified
Base Model	9.31	1496.0	–	0%
Constant Scaling (s=0.46)	11.25	13.19	+20.8%	100%
Calibrated Scaling†	4.38	4.72	−52.9%	100%
Winsorization (q=7%)	**9.94**	**11.44**	+6.8%	17.5%

We compare against three baselines: (1) the unmodified base model, which exhibits severe long-context degradation (PPL@64K $>$ 1400); (2) constant scaling with $s = 0.46$, a data-free method that applies uniform power scaling $\lambda \leftarrow \lambda^s$ to all eigenvalues; and (3) calibrated scaling (Lu et al., 2025), included as an upper bound reference (this method requires calibration data and is not directly comparable to data-free approaches).

4.2 Main Results

Table 1 presents our main results. Winsorization at $q = 7\%$ achieves PPL@64K of 11.44, outperforming constant scaling (13.19) by 13.3% while also achieving better short-context quality (PPL@2K of 9.94 vs 11.25, an 11.6% improvement). Notably, winsorization modifies only 17.5% of channels (539 out of 3072) compared to 100% for scaling methods, demonstrating that targeted intervention on extreme eigenvalues is more effective than uniform spectrum modification.

The short-context regression of winsorization (+6.8%) is substantially lower than constant scaling (+20.8%), indicating that preserving the mid-spectrum eigenvalues maintains model capacity at shorter contexts. A gap to calibrated scaling remains (PPL@64K of 11.44 vs 4.72), which is expected since calibrated methods leverage distributional information from representative data that data-free approaches cannot access.

Table 2: Ablation study on winsorization percentile and threshold strategy. Percentile-based winsorization at $q = 7\%$ achieves optimal tradeoff. Fixed thresholds underperform adaptive percentile-based clipping. Best results in **bold**.

Setting	PPL@2K	PPL@64K	2K Regression	Channels Clipped
$q = 0.5\%$	9.50	520.0	+2.0%	6.5%
$q = 1\%$	9.50	216.0	+2.0%	6.5%
$q = 7\%$	**9.94**	**11.44**	+6.8%	17.5%
Fixed $[0.1, 0.99]$	9.63	38.0	+3.4%	32.7%
Constant Scaling	11.25	13.19	+20.8%	100%

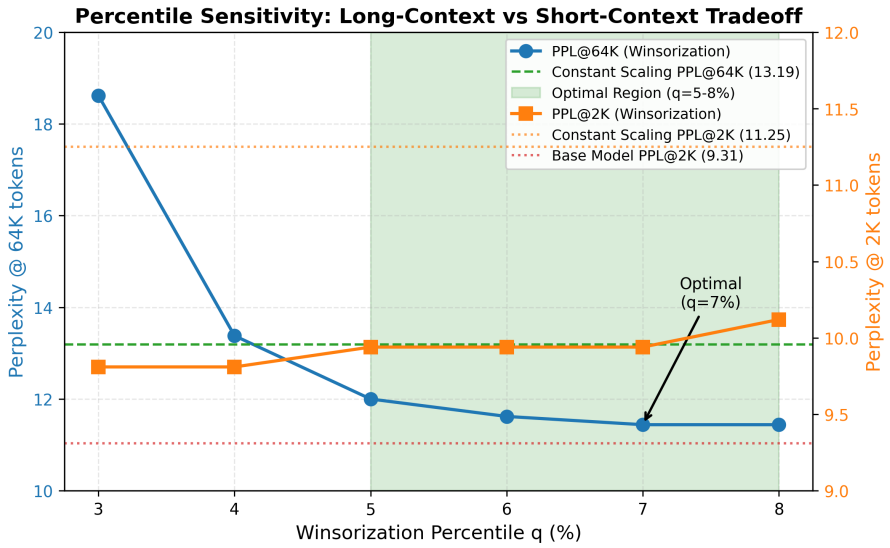

Figure 2: Percentile sensitivity analysis showing the tradeoff between long-context perplexity (PPL@64K, left axis) and short-context regression (PPL@2K, right axis). The optimal region ($q = 5$–8%, shaded) achieves PPL@64K below constant scaling while maintaining moderate short-context regression.

4.3 Ablation Study

Table 2 and Figure 2 show the sensitivity of winsorization to the percentile parameter q. Conservative percentiles ($q < 4\%$) provide insufficient spectrum compression, with PPL@64K remaining above 200 even at $q = 1\%$. The optimal region lies at $q = 5$–8%, where PPL@64K drops sharply to approximately 11–12 while short-context regression remains moderate. Beyond $q = 8\%$, further increases in q yield diminishing returns on PPL@64K while accelerating short-context regression.

Fixed threshold clipping ($\lambda \in [0.1, 0.99]$) achieves PPL@64K of 38.0, substantially worse than the optimal percentile-based approach (11.44). This demonstrates that adaptive, per-layer percentile bounds are more effective than global fixed thresholds, as the eigenvalue distribution varies across layers and a single threshold cannot capture this heterogeneity.

4.4 Mechanism Analysis

To understand why winsorization improves long-context performance, we analyze the effective eigenvalue distribution $\lambda_{\text{eff}} = \exp(\Delta_t a_i)$ at 64K context length. We hypothesized that winsorization would reduce the fraction of near-1 effective eigenvalues (which cause state explosion). Surprisingly, winsorization does not substantially reduce this tail mass: 21.63% of effective eigenvalues exceed 0.99 after winsorization, compared to 21.24% for the base model. In contrast, constant scaling reduces this fraction to 19.77%.

This finding suggests that extreme effective eigenvalues at long contexts are driven primarily by input-dependent Δ_t spikes rather than static A outliers. Winsorization clips the static eigenvalues $\lambda = \exp(a_i)$, but the effective eigenvalue $\lambda_{\text{eff}} = \lambda^{\Delta_t}$ can still approach 1 when Δ_t is large. The perplexity improvement from winsorization may therefore arise from a different mechanism than tail reduction, such as improved gradient flow or regularization effects. This insight motivates future work on input-dependent interventions that directly address Δ_t dynamics.

5 Conclusion

We presented data-free transition-spectrum winsorization for improving Mamba's long-context generalization. By clipping extreme eigenvalues in the per-layer transition spectrum, our method achieves PPL@64K of 11.44, outperforming constant scaling (13.19) by 13% while modifying only 17.5% of channels. The approach requires no calibration data, making it applicable to any pretrained Mamba model. Mechanism analysis reveals that extreme effective eigenvalues are driven by input-dependent Δ_t dynamics rather than static A outliers, suggesting future work should explore input-dependent interventions to further close the gap to calibrated methods.

References

Shady Abu-Hussein Nadav Cohen Amir Globerson Lior Wolf Raja Giryes Assaf Ben-Kish, Itamar Zimerman. Decimamba: Exploring the length extrapo-lation potential of mamba. URL `https://openreview.net/pdf?id=iWSl5Zyjjw`. Synthesized BibTeX entry.

Seyedarmin Azizi, Souvik Kundu, Mohammad Erfan Sadeghi, and M. Pedram. Mambaextend: A training-free approach to improve long context extension of mamba. 2025.

Maximilian Beck, Korbinian Poppel, M. Spanring, Andreas Auer, Oleksandra Prudnikova, Michael K Kopp, G. Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. *ArXiv*, abs/2405.04517, 2024.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *ArXiv*, abs/2004.05150, 2020.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *ArXiv*, abs/2306.15595, 2023.

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

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling. *ArXiv*, abs/2101.00027, 2020.

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

Albert Gu, Tri Dao, Stefano Ermon, A. Rudra, and C. Ré. Hippo: Recurrent memory with optimal polynomial projections. *ArXiv*, abs/2008.07669, 2020.

Albert Gu, Karan Goel, and Christopher R'e. Efficiently modeling long sequences with structured state spaces. *ArXiv*, abs/2111.00396, 2021.

Peng Lu, Jerry Huang, Qiuhao Zeng, Xinyu Wang, Boxing Wang, Philippe Langlais, and Yufei Cui. Mamba modulation: On the length generalization of mamba. *ArXiv*, abs/2509.19633, 2025.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *ArXiv*, abs/2309.00071, 2023.

Ofir Press, Noah A. Smith, and M. Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *ArXiv*, abs/2108.12409, 2021.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *ArXiv*, abs/2104.09864, 2021.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *ArXiv*, abs/2307.08621, 2023.

Zhifan Ye, Kejing Xia, Yonggan Fu, Xin Dong, Jihoon Hong, Xiangchi Yuan, Shizhe Diao, Jan Kautz, Pavlo Molchanov, and Yingyan Lin. Longmamba: Enhancing mamba's long-context capabilities via training-free receptive field enlargement. *ArXiv*, abs/2504.16053, 2025.

M. Zaheer, Guru Guruganesh, Kumar Avinava Dubey, J. Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. *ArXiv*, abs/2007.14062, 2020.