INTERDOMAIN ATTENTION: BEYOND TOKEN-LEVEL KEY-VALUE MEMORY

Naoki Kiyohara[1,4,*], **Harrison Bo Hua Zhu**[2,1,5,*], **Zhuo Sun**[3,1,†],
Wenlong Chen[1], **Samir Bhatt**[2,1,†], **Yingzhen Li**[1,†]
[1]Imperial College London [2]University of Copenhagen
[3]Shanghai University of Finance and Economics [4]Canon Inc. [5]Technical University of Denmark

ABSTRACT

Recent progress in deep learning is largely driven by advances in deep sequence models. Among them, Transformers and deep SSMs are arguably the two most successful architectures, but with different designs. While Transformers learn where to attend in the context via attention mechanisms, deep SSMs, on the other hand, try to compress and gate context information into fixed-sized, long-range memory states. Hybrid architectures consisting of both attention and SSM layers can achieve superior performance because they address the quadratic scaling and long-range memory issues of attention. Instead of just stacking these two types of layer, in this work, we propose Interdomain Attention which integrates SSMs naturally into attention modules through the lens of kernel methods and finite basis approximation. We argue that Interdomain Attention has the potential to switch or interpolate between the expressiveness and scaling behaviors of Transformers and deep SSMs, and preliminary experiments on sCIFAR-10 show promise.

1 INTRODUCTION

Transformers (Vaswani et al., 2017) now underpin systems ranging from language models (Brown et al., 2020) to agentic AI (Yao et al., 2023), yet their long-context capability remains a persistent bottleneck. Despite advances in hardware-aware optimization (Dao, 2024), KV caching, and distributed sharding, the $\mathcal{O}(N_q N)$ complexity of softmax attention fundamentally limits scaling to longer sequences, where N_q is the query length and N is the key-value length. Deep state space models (SSMs) have emerged as a compelling alternative, designed from the ground up for long-range dependencies and sub-quadratic complexity. Beginning with the HiPPO framework (Gu et al., 2020) and its efficient parameterizations in S4 (Gu et al., 2022b) and S4D (Gu et al., 2022a), and culminating in the selective state space model Mamba (Gu & Dao, 2024; Dao & Gu, 2024), deep SSMs compress context into fixed-size recurrent states that can be updated in $\mathcal{O}(1)$ time per step while retaining provably optimal projections of input history.

In practice, hybrid architectures that interleave attention and SSM layers have shown strong performance across modalities (e.g. DNA sequences), combining the fine-grained, content-based retrieval of attention with the efficient long-range compression of SSMs (Poli et al., 2023; Brixi et al., 2025). However, these hybrids treat attention and SSMs as modular, independently-designed components rather than seeking a unified mechanism. Concurrently, a growing line of work on linear and sub-quadratic attention seeks to reduce the cost of attention directly (Katharopoulos et al., 2020; Peng et al., 2021; Han et al., 2024).

We propose *Interdomain Attention*, which integrates SSM recurrences directly into attention by approximating the attention kernel with Fourier features (Rahimi & Recht, 2007; Peng et al., 2021), projecting the decoupled key-value maps onto SSM basis functions via an SSM recurrence, compressing the full context into a fixed-size state independent of N. Our contributions include: (1) a principled unification of kernel attention (Tsai et al., 2019; Katharopoulos et al., 2020; Choromanski et al., 2021; Chen & Li, 2023) and SSMs that can interpolate between their expressiveness and scaling behaviors; (2) total work that scales linearly in sequence length, state memory independent of

*Equal contribution.

† Corresponding Authors: sunzhuo@mail.shufe.edu.cn, samir.bhatt@sund.ku.dk, yingzhen.li@imperial.ac.uk

sequence length, and four implementation strategies offering complementary parallelism–memory tradeoffs; (3) experiments on sCIFAR-10 showing substantial gains over softmax attention with a smaller generalization gap.

2 BACKGROUND

In this section, we briefly review attention mechanisms and state space models.

Attention as Kernel Regression. Standard dot-product attention (Vaswani et al., 2017) maps input tokens $x_n \in \mathbb{R}^d$ to queries, keys, and values via $q_n = W_q x_n \in \mathbb{R}^d$, $k_n = W_k x_n \in \mathbb{R}^d$, $v_n = W_v x_n \in \mathbb{R}^d$, and computes the output for the i-th token as

$$o_i = \frac{\sum_{n=1}^{N} \mathcal{K}(q_i, k_n)\, v_n}{\sum_{n'=1}^{N} \mathcal{K}(q_i, k_{n'})}, \tag{1}$$

where $\mathcal{K}(q, k) = \exp(q^\top k/\sqrt{d})$ for softmax attention. This is one realization of a Nadaraya-Watson kernel regression estimator (Nadaraya, 1964; Watson, 1964), a connection noted in several works on kernel attention (Katharopoulos et al., 2020; Choromanski et al., 2021). Crucially, this view reveals that the choice of kernel $\mathcal{K}$ is a design degree of freedom; in particular, replacing the softmax kernel with a stationary kernel enables the use of Fourier features (Section 3.1). This further enables sub-quadratic computation/memory consumption. Moreover, attention performs *non-parametric regression* over the values v_n at test time, where $\mathcal{K}$ determines the weighting over the context, connecting to the broader test-time regression (Wang et al., 2025) or memorization (e.g. Titans; Behrouz et al. 2025).

State Space Models and HiPPO. A linear state space model maps an input signal $u(t) \in \mathbb{R}$ to a latent state $h(t) \in \mathbb{R}^M$ via:

$$\dot{h}(t) = A(t)\, h(t) + B(t)\, u(t). \tag{2}$$

HiPPO (Gu et al., 2020) provides principled initializations for $A(t) \in \mathbb{R}^{M\times M}$ and $B(t) \in \mathbb{R}^M$ such that the state $h(t)$ maintains optimal projections of the input history onto M time-varying orthogonal basis functions $\{\phi_m^{(t)}\}_{m=1}^M$. This forms the foundation for deep SSM architectures such as S4 (Gu et al., 2022b), S4D (Gu et al., 2022a), and Mamba (Gu & Dao, 2024; Dao & Gu, 2024).

3 INTERDOMAIN ATTENTION

Figure 1 illustrates the architecture. Motivated by the recently discovered connection between the interdomain kernel computation in HiPPO-SVGP (Chen et al., 2025; Chen, 2026) via SSMs, we show how the kernel regression view Equation (1) leads to a principled integration of SSM-style recurrent memory into attention.

3.1 FOURIER FEATURES FOR KERNEL APPROXIMATION

For a stationary kernel $\mathcal{K}$, Bochner's theorem (Rahimi & Recht, 2007) gives

$$\mathcal{K}(x, x') = \mathbb{E}_{p(\omega)}\Big[\xi_\omega(x)^\top \xi_\omega(x')\Big], \quad \xi_\omega(x) = \big[\cos(\omega^\top x),\ \sin(\omega^\top x)\big]^\top, \tag{3}$$

where $p(\omega)$ is the spectral density of $\mathcal{K}$ (its normalized Fourier transform). Approximating the expectation with R samples $\omega_r \sim p(\omega)$ and substituting into Equation (1), we have:

$$o_i \approx \hat{o}_i = \frac{\sum_{n=1}^{N} \frac{1}{R} \sum_{r=1}^{R} \xi_{\omega_r}(q_i)^\top \xi_{\omega_r}(k_n)\, v_n}{\sum_{n'=1}^{N} \frac{1}{R} \sum_{r=1}^{R} \xi_{\omega_r}(q_i)^\top \xi_{\omega_r}(k_{n'})}. \tag{4}$$

3.2 HIPPO BASIS FUNCTIONS FOR INTERDOMAIN ATTENTION

The key insight is to treat the keys and values as functions of time, $k(t_n) := k_n$ and $v(t_n) := v_n$, and project the random features of the keys, as well as the values, onto the HiPPO basis:

$$c_{\omega,m}^{(t_N)} = \int \xi_\omega\big(k(t)\big)\, \phi_m^{(t_N)}(t)\, dt, \quad \gamma_m^{(t_N)} = \int v(t)\, \phi_m^{(t_N)}(t)\, dt, \quad \eta_m^{(t_N)} = \int \phi_m^{(t_N)}(t)\, dt, \tag{5}$$

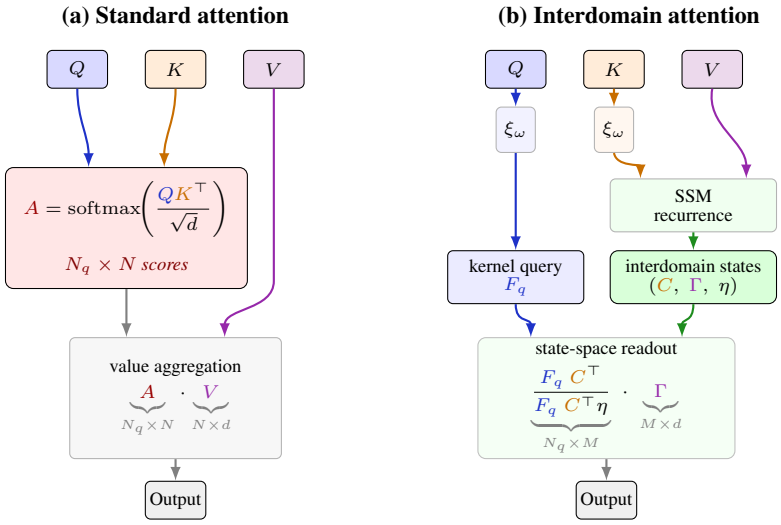


Figure 1: (a) Standard attention computes $N_q \times N$ scores from Q and K, then multiplies by V to produce the output. (b) Interdomain attention maps queries and keys through Fourier features ξ_ω to produce the kernel query matrix F_q ($N_q \times 2R$). Keys and values are compressed via SSM recurrence into M interdomain states: C ($M \times 2R$, key feature projections), Γ ($M \times d$, value projections), and η ($M \times 1$, normalizing constants). State-space readout computes the output via the $N_q \times M$ cross-covariance $F_q C^\top$.

where $c_{\omega,m}^{(t_N)} \in \mathbb{R}^2$ and $\gamma_m^{(t_N)} \in \mathbb{R}^d$. Both projections can be computed incrementally via the HiPPO ODE Equation (2) as new tokens arrive. Substituting the basis reconstruction $\xi_\omega(k_n) \approx \sum_m c_{\omega,m}\, \phi_m(t_n)$ into Equation (4) and exchanging the order of summation, we have:

$$\hat{o}_i \approx \frac{\frac{1}{R}\sum_{r=1}^{R}\sum_{m=1}^{M}\left(\xi_{\omega_r}(q_i)^\top c_{\omega_r,m}^{(t_N)}\right)\sum_{n=1}^{N}\phi_m^{(t_N)}(t_n)\, v_n}{\frac{1}{R}\sum_{r=1}^{R}\sum_{m=1}^{M}\left(\xi_{\omega_r}(q_i)^\top c_{\omega_r,m}^{(t_N)}\right)\sum_{n'=1}^{N}\phi_m^{(t_N)}(t_{n'})}. \tag{6}$$

Furthermore, recognizing $\sum_n \phi_m^{(t_N)}(t_n)\, v_n \approx \gamma_m^{(t_N)}$ and $\sum_{n'} \phi_m^{(t_N)}(t_{n'}) \approx \eta_m^{(t_N)}$ from Equation (5), the sums over tokens collapse (dropping the time superscript $^{(t_N)}$ for brevity):

$$\tilde{o}_i = \frac{\frac{1}{R}\sum_{r=1}^{R}\sum_{m=1}^{M}\left(\xi_{\omega_r}(q_i)^\top c_{\omega_r,m}\right)\gamma_m}{\frac{1}{R}\sum_{r=1}^{R}\sum_{m=1}^{M}\left(\xi_{\omega_r}(q_i)^\top c_{\omega_r,m}\right)\eta_m}, \tag{7}$$

In matrix form, let $F_q \in \mathbb{R}^{N_q \times 2R}$ be the query Fourier feature matrix, $C \in \mathbb{R}^{M \times 2R}$ the basis coefficients, $\Gamma \in \mathbb{R}^{M \times d}$ the value projection, and $\eta \in \mathbb{R}^M$. Then interdomain attention computes

$$\tilde{O} = \frac{F_q\, C^\top \Gamma}{F_q\, C^\top \eta}, \tag{8}$$

where division is element-wise with broadcasting. The entire context is summarized in the basis coefficients C and Γ, which are updated recurrently. For causal processing, the basis coefficients become position-dependent: at step i, the SSM state encodes $C^{(i)}, \Gamma^{(i)}, \eta^{(i)}$ reflecting only tokens $1, \ldots, i$.

3.3 Learning the Spectral Density

Rather than sampling from a fixed spectral density $p(\omega)$, we parameterize the spectral density as $p_\psi(\omega) = \sum_{r=1}^{R} \alpha_r\, \delta(\omega - \omega_r)$, where both the frequencies ω_r and weights α_r are learned end-to-end. With $W = \mathrm{diag}([\alpha;\, \alpha]) \in \mathbb{R}^{2R \times 2R}$, interdomain attention becomes

$$\tilde{O} = \frac{F_q\, W\, C^\top \Gamma}{F_q\, W\, C^\top \eta}. \tag{9}$$

3.4 OUR ARCHITECTURE

We implement interdomain attention within a LLaMA-style Transformer (Touvron et al., 2023): each layer applies RMSNorm, interdomain attention, RMSNorm, and SwiGLU feedforward, with rotary position embeddings (RoPE; Su et al. 2024) applied to queries and keys. The SSM recurrence uses a diagonal state space model (S4D; Gu et al. 2022a), where the dynamics $A(t)$ and $B(t)$ are initialized from the diagonalized HiPPO matrix via S4D-Inv and trained end-to-end, approximating the basis projections in Equation (5).

Multi-Head Attention. Each head h has its own query projection $F_q^{(h)}$ and learned frequency weights $W_h = \mathrm{diag}([\alpha_h;\, \alpha_h])$, where $\alpha_h = \mathrm{softmax}(w_h) \in \mathbb{R}^R$ are per-head learnable parameters. The output of head h is

$$\tilde{O}_h = \frac{F_q^{(h)}\, W_h\, C^{(h)\top} \Gamma^{(h)}}{F_q^{(h)}\, W_h\, C^{(h)\top} \eta^{(h)}}. \tag{10}$$

We consider: (1) *Per-head* variant, each head maintains its own key-value projections, learned frequencies $\omega_r^{(h)}$, and SSM dynamics, yielding head-specific $C^{(h)}$, $\Gamma^{(h)}$, and $\eta^{(h)}$. (2) ***Shared-KV*** variant, a single set of key-value projections, frequencies, and SSM dynamics is shared across all heads, so that $C^{(h)} = C$, $\Gamma^{(h)} = \Gamma$, and $\eta^{(h)} = \eta$ for all h. Shared-KV reduces parameters and computational cost while each head still attends at different frequency scales via W_h.

Causal Processing and Classification. Since the SSM state is updated sequentially, interdomain attention is naturally causal: the output at position i depends only on tokens $1, \ldots, i$. For classification tasks, we append a learnable CLS token at the end of the sequence, which aggregates context from all preceding tokens through the causal attention mechanism.

3.5 MEMORY AND COMPUTATIONAL COMPLEXITIES

Table 1 compares the complexities of standard softmax attention and interdomain attention. All complexities are per head, where N_q is the query length, N is the key-value length, M is the number of basis functions, R is the number of Fourier features, d is the head dimension, and K is the checkpoint interval.

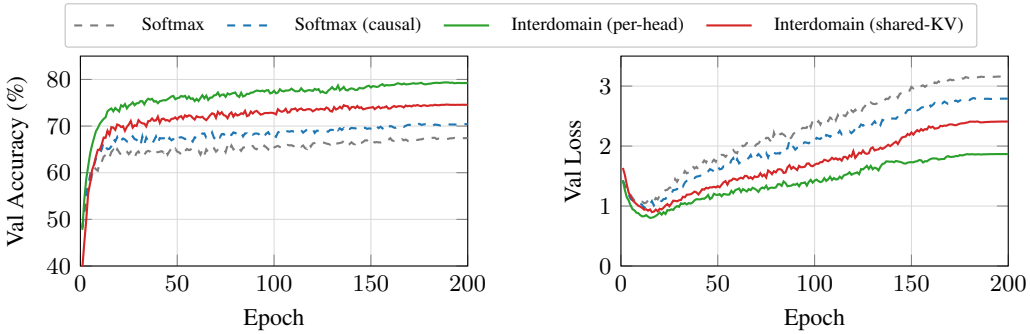
Table 1: Per-head computational and memory complexities. Interdomain attention replaces the $\mathcal{O}(Nd)$ KV cache with $\mathcal{O}(M(R{+}d))$ interdomain states, independent of sequence length N. K denotes the checkpoint interval (sequential scan) or chunk size (chunkwise scan).

	Total work	State memory	Backward memory
Standard attention	$\mathcal{O}(N_q N d)$	$\mathcal{O}(Nd)$	$\mathcal{O}(N_q N)$
Interdomain (FFT)	$\mathcal{O}(NM(R{+}d)\log N + N_q R d)$	$\mathcal{O}(M(R{+}d))$	$\mathcal{O}(NM(R{+}d) + N_q R)$
Interdomain (Scan)	$\mathcal{O}(NM(R{+}d) + N_q R d)$	$\mathcal{O}(M(R{+}d))$	$\mathcal{O}(\frac{N}{K} M(R{+}d) + N_q R)$

Test-Time Generation. At generation time, standard attention with a KV cache requires $\mathcal{O}(Nd)$ work per step to attend over all cached keys, and the cache itself grows as $\mathcal{O}(Nd)$. In contrast, interdomain attention updates its SSM states via a fixed-size recurrence in $\mathcal{O}(MRd)$ per step, independent of N, and computes the output from the query features and current state at the same cost. This makes per-token generation $\mathcal{O}(1)$ with respect to sequence length, a key advantage for long-context deployment.

4 EXPERIMENTS

In this section, we report experimental settings, empirical performance and runtime comparisons between interdomain attention and softmax attention.



(a) Validation accuracy

(b) Validation loss

Figure 2: Training curves on sCIFAR-10 ($N = 1024$, 200 epochs). Softmax baselines are shown with and without causal masking. The shared-KV variant shares key-value projections, frequencies, and SSM dynamics across all heads; only the per-head frequency weights W_h differ.

Setup. We evaluate on sequential CIFAR-10 (sCIFAR-10), where each 32×32 RGB image is flattened into a sequence of $N_q = N = 1024$ tokens with $d_{\text{in}} = 3$ input channels. All models share the same backbone described in Section 3.4: 4 layers, 4 heads, model dimension 128 (head dimension $d = 32$). For interdomain attention, we set $R = 32$ Fourier features and $M = 128$ interdomain states (implemented via S4D), with S4D-Inv initialization (Gu et al., 2022a), learned spectral frequencies (shifted-cosine features), and the Triton sequential scan. We compare four configurations: softmax attention without and with causal masking, per-head interdomain attention, and shared-KV interdomain attention. All models are trained for 200 epochs with AdamW (lr $= 3 \times 10^{-4}$, weight decay 0.01; SSM dynamics parameters use zero weight decay), cosine annealing, batch size 32, and standard augmentation (random crop with 4-pixel padding, horizontal flip). SSM gradient norms are clipped at 5.0.

Classification Accuracy. Figure 2 shows validation accuracy and loss over training. Both interdomain attention variants outperform the softmax baselines: the per-head variant reaches 79.2% and the shared-KV variant 74.6%, compared with 67.4% (non-causal softmax) and 70.4% (causal softmax). Notably, though the shared-KV variant achieves a lower accuracy, it still outperforms the softmax baselines, suggesting that per-head frequency weights W_h provide sufficient diversity. All models eventually reach near-100% training accuracy, but the interdomain variants exhibit a substantially smaller generalization gap.

Runtime Comparison. Among the strategies in Table 1, we use a fused Triton sequential scan with segmented checkpointing ($K = 32$), which achieved acceptable speed with relatively simple implementation for our first proof of concept. Further optimizations are left for future work. Table 2 (Appendix) reports throughput on a single NVIDIA RTX 6000 Ada (PyTorch 2.10). At $N = 1024$, softmax with FlashAttention (Dao, 2024) is $\sim 9\times$ faster. However, softmax scales as $\mathcal{O}(N_q N d)$ in compute and $\mathcal{O}(Nd)$ in memory, whereas interdomain attention scales linearly in N_q and N with $\mathcal{O}(1)$ state memory, which means interdomain attention has an advantage that grows with query and key-value lengths.

5 Conclusion and Future Work

We introduced *Interdomain Attention*, unifying kernel attention and state space models by projecting Fourier features of keys and values onto SSM basis functions via an SSM recurrence, yielding a fixed-size state independent of sequence length. On sCIFAR-10, both variants substantially outperform softmax attention with a smaller generalization gap. In particular, optimized kernels and scaling to language modeling are immediate directions for future work. Moreover, given the connection between kernel attention and Gaussian process (Chen & Li, 2023), probabilistic interdomain attention can be constructed by treating interdomain attention as posterior mean of an interdomain Gaussian process for improved robustness and uncertainty quantification capabilities (Chen, 2026).

REFERENCES

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=8GjSf9Rh7Z.

Garyk Brixi, Matthew G Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A Gonzalez, Samuel H King, David B Li, Aditi T Merchant, Mohsen Naghipourfar, Eric Nguyen, Chiara Ricci-Tam, David W Romero, Gwanggyu Sun, Ali Taghibakshi, Anton Vorontsov, Brandon Yang, Myra Deng, Liv Gorton, Nam Nguyen, Nicholas K Wang, Etowah Adams, Stephen A Baccus, Steven Dillmann, Stefano Ermon, Daniel Guo, Rajesh Ilango, Ken Janik, Amy X Lu, Reshma Mehta, Mohammad R.K. Mofrad, Madelena Y Ng, Jaspreet Pannu, Christopher Re, Jonathan C Schmok, John St. John, Jeremy Sullivan, Kevin Zhu, Greg Zynda, Daniel Balsam, Patrick Collison, Anthony B. Costa, Tina Hernandez-Boussard, Eric Ho, Ming-Yu Liu, Tom McGrath, Kimberly Powell, Dave P. Burke, Hani Goodarzi, Patrick D Hsu, and Brian Hie. Genome modeling and design across all domains of life with evo 2. *bioRxiv*, 2025. doi: 10.1101/2025.02.18.638918. URL https://www.biorxiv.org/content/early/2025/02/21/2025.02.18.638918.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Wenlong Chen. Probabilistic learning and generation in deep sequence models. *PhD thesis, Imperial College London*, 2026.

Wenlong Chen and Yingzhen Li. Calibrating transformers via sparse gaussian processes. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=jPVAFXHlbL.

Wenlong Chen, Naoki Kiyohara, Harrison Bo Hua Zhu, Jacob Curran-Sebastian, Samir Bhatt, and Yingzhen Li. Recurrent memory for online interdomain gaussian processes. In *Advances in Neural Information Processing Systems*, 2025.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamás Sarlós, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J. Colwell, and Adrian Weller. Rethinking attention with performers. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=Ua6zuk0WRH.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mZn2Xyh9Ec.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10041–10071. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/dao24a.html.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. Hippo: Recurrent memory with optimal polynomial projections. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1474–1487. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/102f0bb6efb3a6128a3c750dd16729be-Paper.pdf.

Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 35971–35983. Curran Associates, Inc., 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/e9a32fade47b906de908431991440f7c-Paper-Conference.pdf.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022b. URL https://openreview.net/forum?id=uYLFoz1vlAC.

Dongchen Han, Ziyi Wang, Zhuofan Xia, Yizeng Han, Yifan Pu, Chunjiang Ge, Jun Song, Shiji Song, Bo Zheng, and Gao Huang. Demystify mamba in vision: A linear attention perspective. *Advances in neural information processing systems*, 37:127181–127203, 2024.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5156–5165. PMLR, 2020. URL http://proceedings.mlr.press/v119/katharopoulos20a.html.

E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964. doi: 10.1137/1109020.

Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah Smith, and Lingpeng Kong. Random feature attention. In *International Conference on Learning Representations*, 2021.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning*, pp. 28043–28078. PMLR, 2023.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.127063. URL https://www.sciencedirect.com/science/article/pii/S0925231223011864.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: An unified understanding for transformer's attention via the lens of kernel. In *EMNLP*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ke Alexander Wang, Jiaxin Shi, and Emily B Fox. Test-time regression: a unifying framework for designing sequence models with associative memory. *arXiv preprint arXiv:2501.12352*, 2025.

Geoffrey S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4):359–372, 1964.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2023.

A Implementation and Runtime Details

The FFT convolution computes the S4D recurrence via $\mathcal{O}(N \log N)$ transforms per input channel and state dimension. The sequential scan replaces the FFT with a fused recurrence that is $\mathcal{O}(N)$ per channel, parallelized across input channels and state dimensions on the GPU. The chunkwise parallel scan splits the sequence into chunks of size K. All chunks compute their terminal states in parallel, followed by a serial boundary propagation across N/K chunk boundaries, and a final parallel pass with corrected initial states. For the backward pass, both scan variants use segmented checkpointing with interval K: within each segment, previous states are recovered via division by the diagonal SSM parameter λ. Segment boundaries reset accumulated error by loading checkpoints. The parallel scan uses the parallel prefix algorithm, achieving $\mathcal{O}(\log N)$ parallel depth, which may become advantageous as hardware parallelism scales.

Table 2: Training runtime comparison on sCIFAR-10 ($N = 1024$, batch size 32, single GPU). The FlashAttention-2 (Dao, 2024) variant uses PyTorch's `scaled_dot_product_attention`. Interdomain attention uses custom Triton sequential scan kernels with segmented checkpointing ($K = 32$).

Model	Variant	Params	Peak Mem (MB)	Throughput (samples/s)
Softmax	—	1.07M	5,863	388.3
Softmax	causal	1.07M	5,863	377.1
Softmax	FlashAttention-2	1.07M	2,611	660.8
Softmax	FlashAttention-2 + causal	1.07M	2,631	641.1
Interdomain	per-head	1.61M	6,417	72.2
Interdomain	shared-KV	1.50M	6,253	75.2