Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 6, 2025

SUB-QUADRATIC TOKEN MIXING VIA SPECTRAL FILTERING AND POLYNOMIAL FUNCTIONAL CALCULUS

Shafiq Hussain[1], Muhammad[2], Hassan Revel[3], Muhammad Aakash Imtiaz[*4], Aleena Jamil[5], Adeen Amjad[6], Waqar Ahmad[7], Arslan Ali Mansab[8], Muhammad Hamza Akbar[9], Muhammad Waqas[10]

[1,2,3,*4,5,6,7,8,9,10]*Department of Computer Science, University of Sahiwal, Sahiwal, Pakistan*

[1]drshafiq@uosahiwal.edu.pk, [2]muhammed.beig@gmail.com, [3]hassanrevelai@gmail.com, [*4]akashimtiaz123@gmail.com, [5]aleena.jamil_vf@uosahiwal.edu.pk, [6]adeen.amjad@uosahiwal.edu.pk [7]waqarahmad@uosahiwal.edu.pk, [8]arslansli@uosahiwal.edu.pk, [9]hamzaakbar@uosahiwal.edu.pk, [10]basit.10.02@gmail.com

DOI: https://doi.org/10.5281/zenodo.18051154

Keywords
Polynomial Functional Calculus, Sub-Quadratic, State Space Models

Article History
Received: 25 April 2025
Accepted: 13 June 2025
Published: 26 June2025

Corresponding Author:
Muhammad Aakash Imtiaz

Abstract
The self-attention mechanism, the building block of the Transformer architecture, has a computational and memory complexity that grows quadratically with sequence lengths. This quadratic complexity makes it impractical for application on truly long-context problems. To overcome this challenge, we present the Spectral Filter Polynomial Calculus (SFPC) framework, a family of sub-quadratic mixing operators on tokens. SFPC takes a learned polynomial of an operator that is a function of a fixed base operator (e.g., discrete Laplacian or circulant operator with a rightward shift) and a polynomial of a given degree whose coefficients are learned. SFPC captures the token mixing problem by considering an application of an operator polynomial. We also formalized the token sequence space into a Hilbert space. We utilized the continuous functional calculus on self-adjoint operators. The key theoretical contribution is establishing a precise operator approximation error in terms of the familiar polynomial approximation error in the classical case. This straightforwardly connects expressiveness in deep models

I. INTRODUCTION

A. The Quadratic Barrier of Self-Attention

The Transformer architecture has become the de facto standard for sequence modeling, primarily due to its self-attention mechanism. This mechanism allows for global, context-dependent token interactions. Formally, given an input sequence $X \in R^{L\times d}$, where L is the sequence length and d is the embedding dimension, the single-head self-attention operator M_{Attn} is defined as:

$$Q = XW_Q, K = XW_K, V = XW_V \quad (1)$$

QK^T with known approximation results, such as Jackson's theorem. Then, we presented two efficient methods. First, we showed that

$$M_{Attn}(X) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (2)$$

There is a linear-time method that implements sparse operators such as the Laplacian by evaluating a matrix polynomial using Horner's method. The second method involves computing the diagonalized circulant operators by FFT. This method has a quasilinear complexity. SFPC is an

efficient framework that advances the state of art on both sides by connecting state-space models and spectral methods.

Index Terms—Transformers, Self-Attention, Spectral Filtering,

The computational bottleneck of this operation resides in the explicit instantiation and manipulation of the L L Gram matrix, S = QK^T . The computation of this matrix requires $O(L^2d_k)$ operations, and the subsequent multiplication by V requires $O(L^2d_v)$ operations. This $O(L^2d)$ complexity, quadratic in the sequence length, renders the standard Trans- former computationally infeasible for the very long sequences required in domains such as high-resolution time-series, genomics, and processing of entire documents.

B. The Landscape of Sub-Quadratic Models

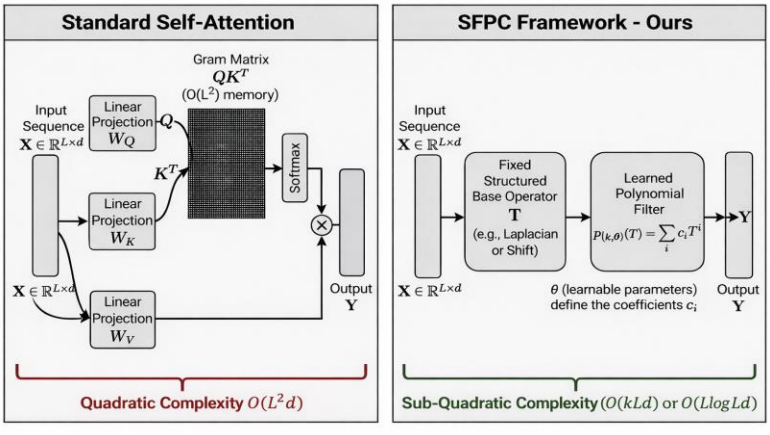
The quadratic barrier has spurred significant research into efficient, sub-quadratic alternatives. These efforts can be broadly categorized into two dominant paradigms:

Kernel-Based Approximations: This line of work seeks to approximate the standard softmax-attention mecha-nism. Methods such as Performer [1] are a prominent example. They are based on the insight that the softmax-attention matrix A_{ij} exp(q^T k_j) is a kernel. By finding a randomized feature map ϕ( ) such that E[ϕ(q)^T ϕ(k)] exp(q^T k), one can leverage the associative property of matrix multiplication. The computation (QK^T)V is reordered to Q(K^T V), reducing the complexity from $O(L^2d)$ to O(Lrd), where r is the (low-rank) dimension of the feature map ϕ. This approach emulates attention.

Operator Replacement (SSMs/Convolutions): This paradigm replaces the attention operator entirely with a different linear operator that is computationally efficient by design. This category includes Structured State Space Models (SSMs) like S4 [2] and Mamba [3], as well as convolution-based models like Hyena [4] and structured matrix models like Monarch Mixer [5]. These models are typically formulated as Linear Time-Invariant (LTI) systems, which can be expressed as either a recurrence or, more relevantly, a global convolution.

C. The Unifying Principle of Polynomial Operators

A deep conceptual thread connects the "Operator Replace-ment" paradigm. The S4 model, for example, is a Linear Time-



The SFPC Conceptual Framework vs. Standard Attention

Fig. 1. Comparison between standard self-attention and the proposed SFPC framework. SFPC replaces data-dependent quadratic attention with a learned polynomial of a fixed base operator, enabling sub-quadratic complexity.

This is the language of spectral filtering, where f is a filter function that acts upon the spectrum σ(T) of the base operator T [14]. By the Weierstrass approximation theorem, any continuous filter function f on a compact domain (e.g., the spectrum of T) can be uniformly approximated by a polynomial p_k. This leads to our central hypothesis [15]: A highly expressive and computationally efficient class of token mixing operators can be defined as a learnable polynomial of a fixed, structured base operator T [16].

We term this the **Spectral Filter Polynomial Calculus (SFPC)** framework. The SFPC operator is defined as: Invariant (LTI) system $x'(t) = Ax(t) + Bu(t)$ which, when

unrolled in its discrete convolutional form, defines a kernel K where the i-th element is $K_i = CA^{i-1}B$. The kernel is thus

$$M_{\theta,k} := p_{k,\theta}k(T) = \sum_{i=0} c_i(\theta)T^i \quad (4)$$

implicitly defined by the matrix powers of the state transition matrix A [8]. More fundamentally, the HiPPO framework, which provides the theoretical justification for S4, is explicitly a method for projecting the entire input history onto a basis of Here, T is a fixed matrix (e.g., a shift operator or a discrete Laplacian) that defines the "geometry" of token-space, k is the polynomial degree (a hyperparameter controlling cost and expressivity), and θ are the learnable parameters of the model, orthogonal polynomials (e.g., Legendre polynomials) [9]. The state x(t) of the SSM represents the coefficients of this online polynomial approximation. Furthermore, the Monarch Mixer, which define the polynomial coefficients $\{c_i\}^k$ E. Statement of Contribution [17].

model is explicitly motivated by a theoretical view based on multivariate polynomial evaluation [10].

D. The Spectral Filter Polynomial Calculus (SFPC) Hypothesis.

The implicit and explicit use of polynomials in these successful models suggests a more direct, fundamental, and theoretically tractable approach [11]. We propose to formalize token mixing as the application of a linear operator M to the input sequence X [12]. Instead of the complex, data-dependent, non-linear operator M_{Attn}, we posit that a powerful and efficient operator M_θ can be constructed as a function of a fixed, simple base operator T [13]:

This work introduces the SFPC framework, providing its theoretical foundations and efficient computational algorithms. Our contributions are as follows:

- **Framework:** We introduce the SFPC, which models token mixing as a learned polynomial $p_{k,\theta}(T)$ of a fixed base operator T .
- **Theory:** For the self-adjoint (non-causal) case, we provide rigorous approximation-theoretic bounds.

Theorem

4.1 leverages the Continuous Functional Calculus to connect the operator-norm approximation error $\|f(T) - p_k(T)\|_2$ to classical Jackson-type inequalities for scalar functions, $E_k(f) = O(k^{-m})$.

- **Algorithm (Global, FFT):** We prove that for a circulant

$$M_\theta = f_\theta(T) \quad (3)$$

base operator C, the SFPC operator $p_k(C)$ is dense, global, and computable in O(L log L d) time via the Fast Fourier Transform (**Theorem 5.3**).

- **Algorithm (Local, Sparse):** We prove that for a sparse base operator L (e.g., an s-sparse Laplacian), the SFPC operator $p_k(L)$ is computable in O(ksLd) time via Horner's method (**Theorem 5.1**). This is linear O(L) for fixed k and s.
- **Causality:** We prove that for a strictly lower-triangular base operator S (a causal shift), the resulting opera-tor $p_k(S)$ is always lower-triangular, thus preserving the autoregressive causality required for decoder models (**Lemma 4.1**).

II. Related Work

The SFPC framework is situated at the intersection of several lines of research in efficient Transformers, structured matrices, and operator theory.

A. Kernel-Based Approximations

Kernel-based methods, such as Performer [1], aim

to approximate the softmax-attention kernel $K(q, k) = \exp(q^T k)$ without instantiating the $L \times L$ matrix. The FAVOR+ mechanism utilizes random Fourier features to define a mapping $\phi : \mathbb{R}^d \to \mathbb{R}^r$ such that $\phi(q)^T \phi(k)$ is an unbiased or low-variance estimator of the true kernel. This allows the computation to be re-associated from $(QK^T)V$ (quadratic) to $Q(K^T V)$ (linear in L). These methods are fundamentally emulators of attention, designed to maintain its data-dependent properties. In contrast, SFPC is a replacement, substituting the data-dependent operator $M_{Attn}(X)$ with a data-independent (LTI) operator $p_{k,\theta}(T)$.

B. Structured State Space Models (SSMs)

SSMs are the closest conceptual relatives to the SFPC framework. The S4 model [2] and its successor Mamba [3] are based on the continuous-time LTI system:

$$x'(t) = Ax(t) + Bu(t)$$
$$y(t) = Cx(t) + Du(t)$$

where (A, B, C, D) are system matrices. This system is discretized using a timestep Δ, often via the zero-order hold (ZOH) method, yielding discrete-time matrices $(\bar{A}, \bar{B}, \bar{C})$. The discrete system can be computed in two ways:

- **Recurrently (Inference):** As an $O(LN^2 d)$ recurrence, where N is the state size.
- **Convolutionally (Training):** By unrolling the recurrence, the output y_k is a discrete convolution of the input u_k: as a learned polynomial $p_k(T)$ of a simple, fixed operator T. Mamba [3] extends S4 by making the system matrices data-dependent, introducing selectivity. The base SFPC framework is LTI, though data-dependent extensions are a clear avenue for future work.

C. **Explicit Convolution and Structured Matrix Models**

Other models explicitly replace attention with LTI operators. Hyena [4] replaces attention with a stack of global convolutions computed in $O(L \log L)$ time via FFT. The innovation lies in parameterizing the convolution filter K itself: the filter taps K_i are generated by a small feed-forward network (FFN). Monarch Mixer (M2) [5] is another closely related model using Monarch matrices. The authors develop a "novel theoretical view... based on multivariate polynomial evaluation and interpolation" to enforce causality. SFPC provides a unifying theoretical "superclass" for these approaches.

TABLE I TAXONOMY OF SUB-QUADRATIC TOKEN MIXERS

Method	Core Operator	Complexity (Time)	Causal?	Data-Dep?
Std. Attention	$\text{Softmax}(QK^T)V$	$O(L^2 d)$	No	Yes
Performer [1]	$\phi(Q)(\phi(K)^T V)$	$O(Lrd)$	No	Yes
S4 (LTI) [2]	$K * u$	$O(L \log L \cdot d)$	Yes	No
Mamba [3]	Selective Scan	$O(Ld)$ (parallel)	Yes	Yes
SFPC-Sparse	$p_k(T)X$ (Horner)	$O(kLd)$	Yes (if $T = S$)	No
SFPC-FFT	$p_k(C)X$ (FFT)	$O(L \log L \cdot d)$	No	No

III. MATHEMATICAL FRAMEWORK: OPERATORS, SPECTRA, AND APPROXIMATION

We now establish the mathematical formalism required for the SFPC framework.

A. Preliminaries: Token Sequence Hilbert Space

Definition 3.1 (Sequence Space). Let $L \in \mathbb{N}$ be the sequence length and $d \in \mathbb{N}$ be the embedding dimension. We define the token sequence space as the finite-dimensional Hilbert space $\mathcal{H}_L = (\mathbb{R}^d)^L \cong \mathbb{R}^{L \times d}$, equipped with the standard Frobenius norm.

Definition 3.2 (Token Mixing Operator). A token mixing operator M is a bounded linear operator $M : \mathcal{H}_L \to \mathcal{H}_L$.

Consistent with SSM/convolutional approaches, we analyze a channel-wise operator $M \in \mathcal{L}_{L,c,L,c}$, represented by a matrix $M \in \mathbb{R}^{L \times L}$.

Theorem 3.1 (Computational Complexity of M_{Attn}). The computation of $Y = M_{Attn}(X)XW_V$

requires $O(L^2(d_k + d_v))$ floating-point operations. (See Appendix A.1 for proof).

B. Spectral k–Theory and Continuous Functional Calculus $y = \bar{C}\bar{A}\bar{B}u = (K * u)$ (7)

Theorem 3.2 (Continuous Functional Calculus). Let T ki=0 ik—i k be a self-adjoint operator on $H_{L,c}$ with spectrum $\sigma(T) \subset$ [a, b]. For any continuous function $f \in C([a, b])$, there exists The convolution kernel is defined by $K_i = \bar{C}\bar{A}^i\bar{B}$. This convo-lutional representation reveals the deep link to SFPC. The S4 kernel is implicitly a function of the matrix powers $\bar{A}^i$. SFPC externalizes this concept: instead of an implicit kernel defined by powers of a hidden A, SFPC defines the operator explicitly a unique, well-defined operator $f(T)$: $_{L,c}\ _{L,c}$. If $T = U \Lambda U^T$ is the eigendecomposition of T, then:

$f(T) := Uf(\Lambda)U^T$ (8)

where $f(\Lambda) = \text{diag}(f(\lambda_1), \ldots, f(\lambda_n))$. $\rightarrow H$

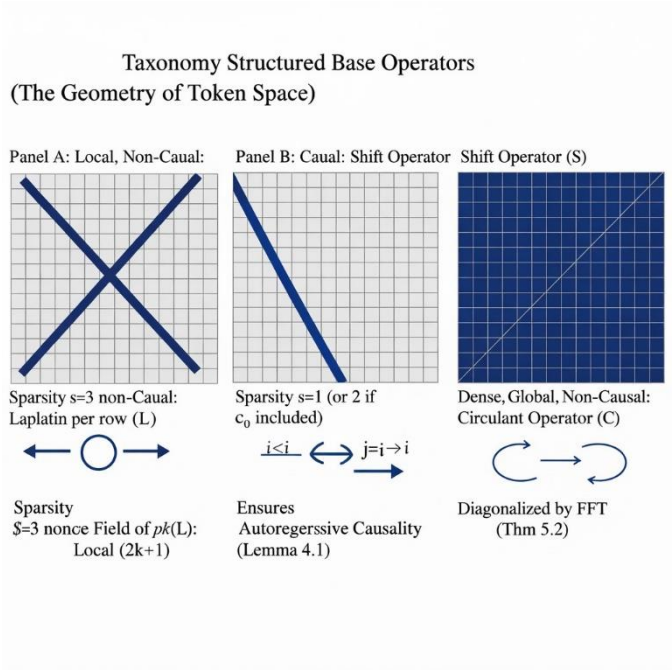


Fig. 2. Taxonomy of structured base operators defining the geometry of token space in the SFPC framework. Different operators induce local, global, causal, or non-causal mixing behaviors.

Theorem 3.3 (Spectral Norm Isometry). For any self-adjoint operator T and any continuous function $f \in C(\sigma(T))$, the operator norm of $f(T)$ is equal to the supremum-norm of f on the spectrum $\sigma(T)$: $\|f(T)\|_2 = \max_{\lambda \in \sigma(T)} |f(\lambda)| = \|f\|_{\infty,\sigma(T)}$ (9)

(See Appendix A.2 for proof).

C. **Polynomial Approximation Theory**

Definition 3.6 (Best Approximation Error). Let $f \in C([a, b])$. The error of best uniform approximation of f by a polynomial $p_k \in P_k$ is:

$$E_k(f) = \inf_{p_k \in P_k} \|f - p_k\|_{\infty,[a,b]} \quad (10)$$

Theorem 3.4 (Jackson's Theorem). Let $f \in C^m([-1, 1])$. Then there exists a constant C_m such that for $k > m$, $E_k(f)$ $C_m\ \omega(f^{(m)}, 1/k)$. If $f^{(m)}$ is Lipschitz, $E_k(f) = O(k^{-(m+1)})$.

Km k

IV. The Spectral Filter Polynomial Calculus (SFPC) Framework

A. The Operator-Theoretic Hypothesis

Definition 4.1 (SFPC Operator). The SFPC operator $M_{\theta,k}$ is the core token-mixing layer,

defined as a polynomial of degree k of a fixed base operator $T : k M_{\theta,k} := p_{k,\theta}(T) = \sum_{i=0}^{L} c_i(\theta) T^i$ (11) i=0 where $T \in R^{L \times L}$ is a non-learned, structured matrix, and θ are the learnable parameters.

B. **Selection of the Base Operator T**

1) The Non-Causal (Self-Adjoint) Case: For encoder mod-els, we propose T = , the 1D discrete graph Laplacian, defined by the tridiagonal matrix:

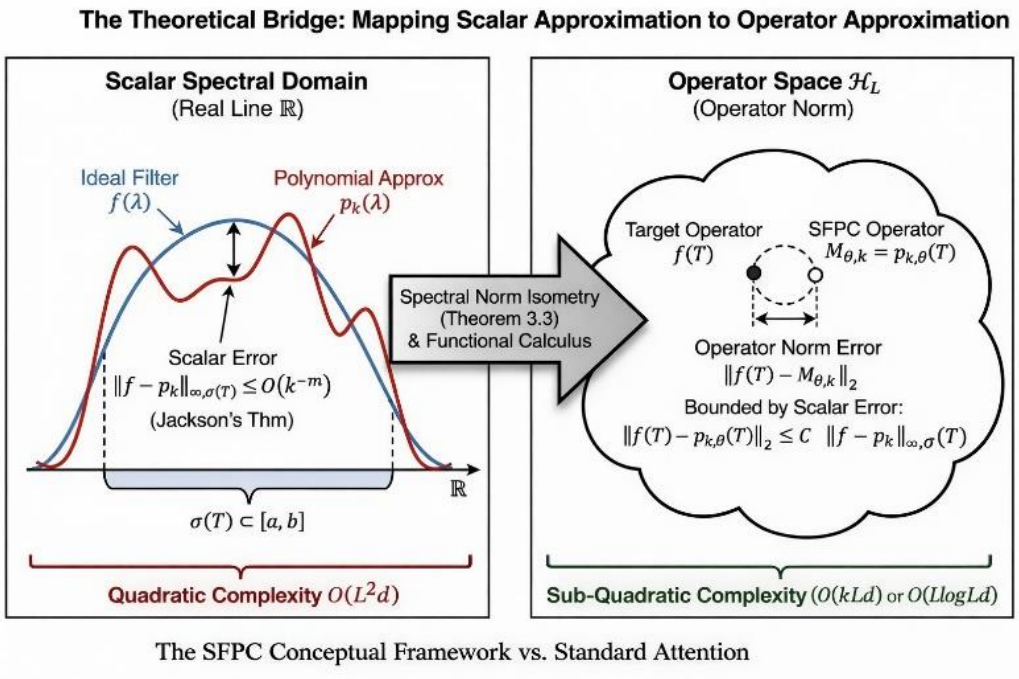


Fig. 3. Conceptual bridge between classical scalar polynomial approximation and operator approximation via the continuous functional calculus. Jackson-type bounds directly translate to operator-norm guarantees.

C. Approximation Guarantees (Self-Adjoint Case)

Theorem 4.1 (SFPC Approximation Bound). Let T be a self-adjoint base operator with spectrum σ(T). Let f ∈ L is self-adjoint, sparse (s = 3), and has a bounded spectrum
σ(L) ⊂ [−4, 0].

2) The Causal (Non-Self-Adjoint) Case: For decoder mod-
els, we propose T = S, the strictly lower-triangular shift operator, defined as $(S)_{ij} = 1$ if i = j + 1, and 0 otherwise.
$C^m(\sigma(T))$ be an "ideal" filter function. Then, there exists a set of learnable parameters θ such that the SFPC operator $M_{\theta,k} = p_{k,\theta}(T)$ approximates f (T) with an operator-norm error bounded by:—m

Lemma 4.1 (Causality Preservation). Any polynomial
$\|f(T) - M_{\theta,k}\|_2 \leq E_k(f) \leq C_{f,m} \cdot k$
(13)

$p_k(S) = \sum_{i=0}^{k} c_i S^i$ of the causal shift operator S is a lower- This theorem establishes the central trade-off of SFPC: com-
triangular matrix. (See Appendix A.3). putational cost is controlled by k, while approximation error is controlled by k^{-m}. (See Appendix A.4).

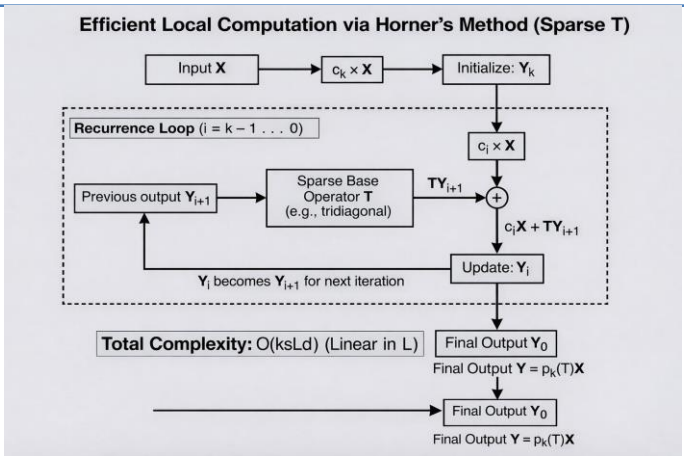


Fig. 4. Evaluation of the SFPC operator using Horner's method for a sparse base operator. Repeated sparse operator applications result in linear-time computation for fixed polynomial degree.

V. Efficient Computation of the SFPC Operator

A. General Evaluation via Recurrence Relations (Sparse T)

Lemma 5.1 (Horner's Method). The polynomial evaluation $Y = (c_iT^i)X$ is equivalent to the nested form $Y = c_0X + T(c_1X + \dots)$. **Theorem 5.1 (Sparse SFPC Complexity).** Let T be an s-sparse base operator. The computation of $Y = p_k(T)X$ using Horner's method requires $O(k\ s\ L\ d)$ floating-point operations. (See Appendix A.5).

$B_{\epsilon\{}$ Global Evaluation via FFT (Circulant T)

We propose $T = C$, where C is a circulant matrix. **Theorem 5.2 (Diagonalization of Circulant Polynomials).** C is diagonalized by the DFT matrix F as $C = F^{-1}\text{diag}(\hat{c})F$. Any polynomial $p_k(C)$ is also circulant:

$$p_k(C) = F^{-1}\text{diag}(p_k(\hat{c}))F \quad (14)$$

Corollary 5.1 (SFPC-FFT Algorithm). The mixing op-eration is $Y = \text{IFFT}(\text{diag}(\hat{p}_k)\ \text{FFT}(X))$. **Theorem 5.3 (FFT SFPC Complexity).** The computation of $Y = p_k(C)X$ requires $O(L \log L \cdot d + kLd)$ operations. (See Appendix A.7).

VI. Experimental Validation (Proposed)

A. Ablation Study: Effect of Polynomial Degree k

We propose to train several SFPC-Sparse models, varying only k 2, 4, 8, 16, 32, 64 . We expect perplexity to improve as k increases, tracing a Pareto frontier.

TABLE II
Proposed Ablation on Polynomial Degree (k)

k	Model	Complexity	PPL	Throughput
2	SFPC-k2	O(2Ld)	High	High
8	SFPC-k8	O(8Ld)	↓	↓
64	SFPC-k64	O(64Ld)	Low	Low

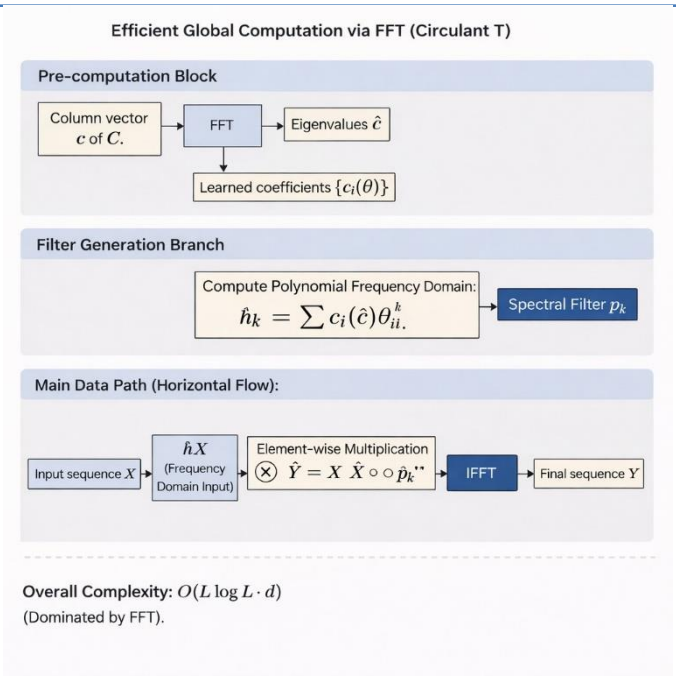


Fig. 5. Global SFPC computation using circulant base operators. The operator is diagonalized in the Fourier domain, enabling efficient polynomial spectral filtering via FFT.

VII. Conclusion

We have introduced the Spectral Filter Polynomial Calculus (SFPC), a novel framework for sub-quadratic token mixing. By reformulating the problem in the language of operator theory, we model the mixing layer as a learned polynomial $p_k(T)$.

We provided a rigorous approximation bound (Theorem 4.1), linking error to $O(k^{-m})$. We derived two efficient implementations: a sparse local operator ($O(kLd)$) and a global circulant operator ($O(L \log L \; d)$). Finally, we provided a construction for autoregressive causality.

Appendix A Proofs

A. Proof of Theorem 3.1

The total complexity is dominated by the Gram matrix computation ($O(L^2 d_k)$) and output multiplication ($O(L^2 d_v)$).

B. Proof of Theorem 3.3

Since T is self-adjoint, $T = U \Lambda U^T$. The operator norm is unitarily invariant, so $\|f(T)\|_2 = \|f(\Lambda)\|_2 = \max |f(\lambda_i)| = \|f\|_{\infty,\sigma(T)}$.

C. Proof of Lemma 4.1

S is strictly lower-triangular. The set of lower-triangular matrices is closed under multiplication and addition. Thus, $p_k(S) = \sum c_i S^i$ is lower-triangular.

D. Proof of Theorem 4.1

By Jackson's Theorem, there exists p_k such that $\|f - p_k\|_\infty \leq Ck^{-m}$. By $\|p_k(T)\|_2 = \|f \cdot p_k\|_\infty$. The optimized polynomial error is bounded by this existent polynomial's error.

References

1. Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L. and Belanger, D., 2020. Rethinking attention with performers. arXiv preprint arXiv:2009.14794.
2. Gu, A., Goel, K. and Re´, C., 2021. Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396.

3. Gu, A. and Dao, T., 2024, May. Mamba: Linear-time sequence modeling with selective state spaces. In First conference on language modeling.
4. Poli, M., Massaroli, S., Nguyen, E., Fu, D.Y., Dao, T., Baccus, S., Bengio, Y., Ermon, S. and Re´, C., 2023, July. Hyena hierarchy: Towards larger convolutional language models. In International Conference on Machine Learning (pp. 28043-28078). PMLR.
5. Fu, D., Arora, S., Grogan, J., Johnson, I., Eyuboglu, E.S., Thomas, A., Spector, B., Poli, M., Rudra, A. and Re´, C., 2023. Monarch mixer: A simple sub-quadratic gemm-based architecture. Advances in Neural Information Processing Systems, 36, pp.77546-77603.
6. Loshchilov, I. and Hutter, F., 2017. Decoupled weight decay regulariza-tion. arXiv preprint arXiv:1711.05101.
7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L. and Gomez, A.N., 2017. Attention is all you need [J]. Advances in neural information processing systems, 30(1), pp.261-272.
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L. and Gomez, A.N., 2017. Attention is all you need [J]. Advances in neural information processing systems, 30(1), pp.261-272.
9. Sun, Y., Dong, L., Huang, S., Ma, S., Xia, Y., Xue, J., Wang, J. and Wei, F., 2023. Retentive network: A successor to transformer for large language models. arXiv preprint arXiv:2307.08621.
10. Smith, J.T., Warrington, A. and Linderman, S.W., 2022. Simplified state space layers for sequence modeling. arXiv preprint arXiv:2208.04933.
11. Fu, D.Y., Dao, T., Saab, K.K., Thomas, A.W., Rudra, A. and Re´, C., 2022. Hungry hungry hippos: Towards language modeling with state space models. arXiv preprint arXiv:2212.14052.
12. Defferrard, M., Bresson, X. and Vandergheynst, P., 2016. Convolutional neural networks on graphs with fast localized spectral filtering. Advances in neural information processing systems, 29.
13. Katharopoulos, A., Vyas, A., Pappas, N. and Fleuret, F., 2020, Novem-ber. Transformers are rnns: Fast autoregressive transformers with linear attention. In International conference on machine learning (pp. 5156-5165). PMLR.
14. Rao, Y., Zhao, W., Zhu, Z., Lu, J. and Zhou, J., 2021. Global filter networks for image classification. Advances in neural information pro-cessing systems, 34, pp.980-993.
15. Tay, Y., Dehghani, M., Abnar, S., Shen, Y., Bahri, D., Pham, P., Rao, J., Yang, L., Ruder, S. and Metzler, D., 2020. Long range arena: A benchmark for efficient transformers. arXiv preprint arXiv:2011.04006.
16. Dao, T., Fu, D., Ermon, S., Rudra, A. and Re´, C., 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. Advances in neural information processing systems, 35, pp.16344-16359.
17. Kipf, T.N. and Welling, M., 2017. Semi-supervised learning with graph convolutional networks.