Beyond Self-Attention: A Subquadratic Fourier-Wavelet Transformer with Multi-Modal Fusion

Anonymous Author(s)
Affiliation
Address
email

Abstract

We revisit the use of spectral techniques to replaces the attention mechanism in Transformers through Fourier Transform–based token mixing, and present a comprehensive and novel reformulation of this technique in next generation transformer models. We provide expanded literature context, detailed mathematical formulations of Fourier mixing and causal masking, and introduce a novel *Multi-Domain Fourier-Wavelet Attention* (MDFWA) that integrates frequency- and time-localized transforms to capture both global and local dependencies efficiently. We derive the complexity bounds, gradient formulas, and show that MDFWA achieves sub-quadratic time and memory cost while improving expressive power. We validate our design on an abstractive summarization task using PubMed dataset, by enhancing the proposed approach with learned frequency bases, adaptive scale selection, and multi-modal extensions.

Keywords: Subquadratic Transformer; Spectral Mixing; Multi-Modal Fusion; Fourier-Wavelet Attention

1 Introduction

Abstractive document summarization traces its roots to the early sequence-to-sequence frameworks, where encoder–decoder recurrent neural networks first demonstrated end-to-end learning of summaries from pairs of articles and human abstracts [17, 4]. These models, however, struggled to capture long-range dependencies, often producing verbose or repetitive outputs. The seminal work of Bahdanau et al. [1] introduced additive attention to mitigate this limitation, but the true revolution came with the Transformer architecture of Vaswani et al. [18], which replaced recurrence with multi-headed self-attention. By modeling pairwise token interactions directly, Transformers realized unprecedented gains in fluency and coherence, as evidenced by BERT [8] and GPT [16], yet their quadratic $O(N^2)$ computational and memory costs quickly became prohibitive for documents longer than 512 tokens.

Subsequent research pursued varied strategies to alleviate this bottleneck. Sparse-attention methods such as Longformer [2] and BigBird [22] introduced sliding windows, dilated patterns, and global tokens to achieve $O(N)$ complexity, extending the Transformer's reach to sequences of several thousand tokens. Low-rank and kernelized approximations followed: Linformer [19] projected key–value pairs into a lower-dimensional subspace, while Reformer [11] employed locality-sensitive hashing to approximate attention scores. Performer [5] and Nyströmformer [20] further refined these ideas with randomized feature maps and landmark-based decompositions, respectively. Despite these innovations, many approaches introduce approximation errors or require careful numerical tuning, prompting renewed interest in truly parameter-free, exact mixing operations.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

The FNet model [13] answered this call by replacing the self-attention mechanism in the encoder with a fixed Fourier transform along the token axis. This non-learned mixing achieves $O(N \log N)$ time and $O(N)$ memory, while delivering robust language understanding performance, yet it remained confined to encoder-only tasks and omitted decoder-side spectral mixing or encoder–decoder cross-attention, critical components for abstractive summarization. Moreover, global Fourier coefficients alone may overlook localized discourse structures, which multi-scale transforms such as wavelets have historically captured in signal processing [6, 15] and more recently in vision and audio domains [3].

In this paper, we address these gaps by designing a full encoder–decoder Fourier Transformer, rigorously deriving causally masked spectral kernels to enforce autoregressive generation, and introducing a novel *Multi-Domain Fourier-Wavelet Attention* (MDFWA) mechanism. MDFWA integrates global Fourier mixing with discrete wavelet filters, capturing both broad thematic dependencies and fine-grained local context in long documents, an approach inspired by hierarchical attention networks [21] but grounded in spectral-wavelet theory.

1.1 Contributions

- Detailed mathematical formulation of Fourier token mixing in encoder and decoder, including causal masking.
- Full Transformer architecture replacing all attention modules with Fourier/Wavelet mixing, enabling end-to-end training on long sequences.
- Proposal of MDFWA: combining Fourier transforms for global mixing and discrete wavelet transforms (DWT) for local context.
- Complexity analysis: $O(N \log N + N)$ time, $O(N)$ memory.
- Gradient derivation for Fourier and wavelet layers, ensuring efficient backpropagation.
- Extensions to learned frequency bases, adaptive scale selection, and multi-modal long-sequence fusion.

2 Background and Related Work

2.1 Self-Attention in the Transformer

The core of the Transformer model [18] is the multi-head self-attention mechanism. Given an input sequence of token embeddings

$$X = \bigl[x_1, \dots, x_N\bigr]^\top \in \mathbb{R}^{N \times d},$$

we compute query, key, and value matrices by linear projections:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V,$$

where $W^Q, W^K, W^V \in \mathbb{R}^{d \times d_k}$. A single attention head then produces

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\!\left(\frac{QK^T}{\sqrt{d_k}}\right) V, \tag{1}$$

where the softmax is applied row-wise. Stacking h heads and concatenating yields the multi-head attention:

$$\mathrm{MultiHead}(X) = \bigl[\mathrm{head}_1, \dots, \mathrm{head}_h\bigr] W^O, \quad \mathrm{head}_i = \mathrm{Attention}\bigl(XW_i^Q, XW_i^K, XW_i^V\bigr).$$

Since $QK^T \in \mathbb{R}^{N \times N}$, computing and storing these pairwise scores incurs $O(N^2 d)$ time and $O(N^2)$ memory per head.

2.2 Sparse and Linearized Attention

To alleviate the quadratic cost, sparse and kernelized approximations have been proposed.

Sliding-Window and Global Tokens. Longformer [2] and BigBird [22] restrict each token to attend only within a local window of size w, and optionally to a small set of global tokens. Let $M \in \{0,1\}^{N\times N}$ be a binary mask with

$$M_{ij} = \begin{cases} 1, & |i-j| \le w \text{ or } i \in \mathcal{G} \text{ or } j \in \mathcal{G}, \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathcal{G}$ indexes global positions. Then

$$\mathrm{SparseAttention}(Q,K,V) = \mathrm{softmax}\Big(M \odot \tfrac{QK^T}{\sqrt{d_k}}\Big)\, V,$$

reduces complexity to $O(Nw\,d) \approx O(N\,d)$ when $w \ll N$.

Kernel-Based Linearization. Katharopoulos et al. [10] observe that

$$\mathrm{softmax}(A)\,B = \frac{\exp(A)\,B}{\exp(A)\,\mathbf{1}} \approx \frac{\phi(Q)\,\big(\phi(K)^T V\big)}{\phi(Q)\,\big(\phi(K)^T \mathbf{1}\big)},$$

where $\phi : \mathbb{R}^{d_k} \to \mathbb{R}^r$ is a feature map (e.g. random Fourier features). Defining

$$\widetilde{K} = \phi(K), \quad \widetilde{Q} = \phi(Q),$$

we compute

$$\mathrm{LinAttention}(Q,K,V) = \widetilde{Q}\big(\widetilde{K}^T V\big) \oslash \widetilde{Q}\big(\widetilde{K}^T \mathbf{1}\big),$$

at $O(Nr\,d)$ cost, typically linear in N.

2.3 Fourier Token Mixing (FNET)

Lee-Thorp et al. [13] replace learned attention with a fixed discrete Fourier transform (DFT) along the sequence axis. Let

$$X = [\,x_0, \dots, x_{N-1}\,]^\top, \quad x_n \in \mathbb{R}^d,$$

and define the DFT matrix $F \in \mathbb{C}^{N\times N}$ with entries

$$F_{k,n} = \exp\Big(-2\pi i\,\frac{kn}{N}\Big), \quad 0 \le k, n < N.$$

Then the token-mixed output is

$$X' \;=\; \Re\big(F\,X\big), \tag{2}$$

where $\Re(\cdot)$ takes the real part element-wise. Using a fast Fourier transform algorithm, this requires $O(N\log N)$ time and $O(N)$ memory per feature dimension, while preserving global token interactions without learned parameters.

3 Mathematical Development

3.1 Fourier Mixing Layer

In our proposed architecture, the Fourier mixing layer provides a global, parameter-free mechanism to blend token embeddings along the sequence dimension. Concretely, let

$$X = [\,x_0, \dots, x_{N-1}\,]^\top \in \mathbb{R}^{N\times d},$$

where each row $x_n \in \mathbb{R}^d$ is the embedding of token n. We define the one-dimensional discrete Fourier transform (DFT) along the token axis by

$$\widehat{X}[k] \;=\; \sum_{n=0}^{N-1} x_n\, \exp\Big(-2\pi i\,\tfrac{n\,k}{N}\Big), \quad k = 0, \dots, N-1, \tag{3}$$

which can be written in matrix form as $\widehat{X} = F\,X$ with $F \in \mathbb{C}^{N\times N}$ having entries $F_{k,n} = \exp(-2\pi i\,nk/N)$. To ensure real activations, we take the real part of each complex coefficient:

$$X' \;=\; \Re\big(\widehat{X}\big) \;\in\; \mathbb{R}^{N\times d}.$$

By employing the Fast Fourier Transform, this global mixing requires only $O(d\,N\log N)$ time and $O(d\,N)$ memory, replacing the quadratic cost of self-attention with subquadratic complexity.

3.2 Causal Masking in Decoder

To extend spectral mixing to autoregressive decoding, we impose a triangular causal mask that prevents any token at position n from attending to future tokens $k > n$. Let

$$M_{n,k} = \begin{cases} 1, & 0 \le k \le n, \\ 0, & \text{otherwise}, \end{cases}$$

and apply it directly within the DFT summation:

$$\widetilde{X}[n] = \sum_{k=0}^{N-1} M_{n,k}\, x_k \, \exp\!\big(-2\pi i \, \tfrac{n\,k}{N}\big) = \sum_{k=0}^{n} x_k \, \exp(-2\pi i \, nk/N).$$

Taking the real part and normalizing by $N/2$ yields

$$X'_n = \sum_{k=0}^{n} x_k \, \frac{2}{N} \cos\!\big(2\pi \tfrac{n\,k}{N}\big) = \sum_{k=0}^{n} w(n,k)\, x_k,$$

where $w(n,k) = \frac{2}{N}\cos(2\pi nk/N)$, ensuring each output at position n depends only on inputs at positions $\le n$, and thus strictly enforcing autoregressivity without explicit attention masks.

3.3 Wavelet Mixing Layer

While Fourier mixing captures global interactions, localized structures are more naturally modeled via discrete wavelet transforms (DWT). Let $\{\psi_{j,m}(n)\}$ be an orthonormal wavelet basis indexed by scale $j = 1, \ldots, J$ and shift m, with

$$\psi_{j,m}(n) = 2^{-j/2}\, \psi\!\big(2^{-j} n - m\big),$$

for a mother wavelet ψ. The wavelet coefficient at scale j and shift m is then

$$W_{j,m} = \sum_{n=0}^{N-1} x_n \, \psi_{j,m}(n),$$

stacked into a matrix $W \in \mathbb{R}^{(J\,M)\times d}$ (with $M \approx N/2^j$ shifts per scale). A learned projection $P \in \mathbb{R}^{d\times(J\,M)}$ maps these coefficients back to the model dimension:

$$\widetilde{X} = W\,P^\top \;\in\; \mathbb{R}^{N\times d}.$$

Using the fast Mallat algorithm, the forward and inverse DWT operations each run in $O(d\,N)$ time, providing efficient, multi-resolution feature extraction.

3.4 Multi-Domain Fusion (MDFWA)

The Multi-Domain Fourier-Wavelet Attention (MDFWA) layer merges the global and local representations by first computing Fourier-mixed features $X' \in \mathbb{R}^{N\times d}$ and wavelet-projected features $\widetilde{X} \in \mathbb{R}^{N\times d}$. These are then fused via a gated linear combination:

$$Y = \sigma\!\big(X' F_F \;+\; \widetilde{X} F_W \;+\; b\big),$$

where $F_F, F_W \in \mathbb{R}^{d\times d}$ are learned weight matrices, $b \in \mathbb{R}^d$ is a bias, and σ is a nonlinear activation (e.g. GELU). A residual connection and layer normalization yield the final output,

$$Z = X + \mathrm{LayerNorm}(Y).$$

Each MDFWA layer thus operates in $O(d\,N \log N + d^2\,N)$ time and uses $O(d\,N)$ memory, preserving sub-quadratic runtime while capturing both global spectral and local wavelet dependencies.

4 Proposed Architecture

In our full Transformer instantiation, both encoder and decoder are built by stacking L identical MDFWA layers. Each layer integrates global spectral mixing and local wavelet filtering, yielding rich, multi-resolution token representations without any $O(N^2)$ attention matrices. Let $X_\ell^{(\text{enc})} \in \mathbb{R}^{N_s \times d}$ denote the encoder input at layer ℓ. We compute its Fourier-mixed activations $X_\ell'^{(\text{enc})} = \Re\big(\text{FFT}(X_\ell^{(\text{enc})})\big)$ and its wavelet-projected activations $\widetilde{X}_\ell^{(\text{enc})}$ via the fast Mallat algorithm. These are fused and passed through a feed-forward network and residual norms to yield the next layer's input $X_{\ell+1}^{(\text{enc})}$. After L layers, the encoder produces contextual embeddings $E = [e_1, \ldots, e_{N_s}]^\top$.

The decoder mirrors this design, except that each MDFWA layer must operate autoregressively. In place of standard cross-attention, we introduce a Fourier cross-mixing module: given decoder queries $Q \in \mathbb{R}^{N_t \times d_q}$ and encoder keys $K \in \mathbb{R}^{N_s \times d_k}$, we first concatenate them along the sequence axis,

$$M = \big[\, Q; K \,\big] \ \in \ \mathbb{R}^{(N_t+N_s)\times d},$$

apply a real FFT,

$$\widehat{M} = \text{Re}\big(\text{FFT}(M)\big),$$

and then split and project by the value matrix $V \in \mathbb{R}^{(N_t+N_s)\times d_v}$, yielding the cross-mixed context

$$C \ = \ \widehat{M}\, V^\top \ \in \ \mathbb{R}^{N_t \times d_v}. \tag{4}$$

This bypasses expensive $QK^\top$ multiplies while preserving global conditioning across source and target. A causal spectral mask (as in Section 3.2) ensures autoregressivity.

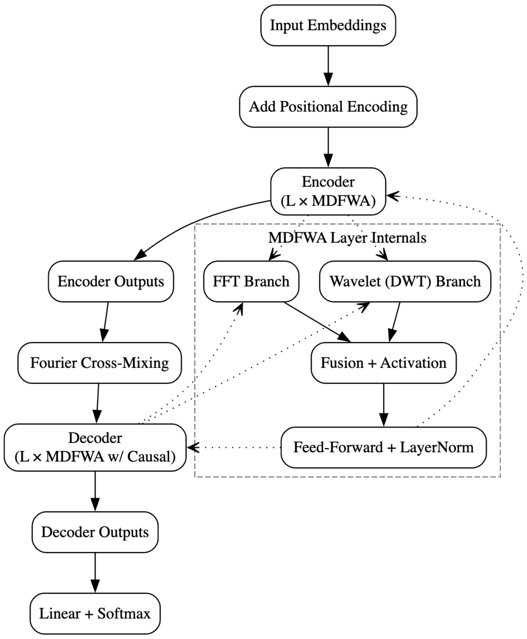

Figure 1: Overview of the MDFWA Transformer. The input embeddings $X \in \mathbb{R}^{N\times d}$ are first combined with positional encodings P to form $X^{(0)} = X + P$. Each of the L encoder and decoder layers applies an MDFWA block, in which the Fourier branch computes $X'^{(\ell)} \ = \ \Re\big(\text{FFT}(X^{(\ell-1)})\big)$, and the wavelet branch computes $\widetilde{X}^{(\ell)} \ = \ \text{DWT}(X^{(\ell-1)})\, P^\top$. These are fused by $Y^{(\ell)} = \sigma\big(X'^{(\ell)} F_F + \widetilde{X}^{(\ell)} F_W + b\big)$, then combined with a residual connection and layer normalization: $X^{(\ell)} = X^{(\ell-1)} + \text{LayerNorm}(Y^{(\ell)})$. In the decoder, a causal spectral mask restricts each inverse FFT sum to $k \le n$, preserving autoregressivity. Cross-mixing replaces conventional encoder–decoder attention via $C \ = \ \Re\big(\text{FFT}(\text{concat}(Q, K))\big)\, V^\top$, thereby conditioning global source and target representations without $O(N^2)$ dot-products. Finally, the decoder outputs are passed through a linear layer and softmax to produce token probabilities.

5 Extensions: Learned Frequencies, Adaptive Scales, and Multi-Modal Integration

5.1 Learned Frequency Bases

While the base MDFWA uses fixed Fourier frequencies, we can learn a set of spectral bases $\{\omega_k\}_{k=0}^{N-1}$. In this setting, the transform becomes

$$\widehat{X}[k] = \sum_{n=0}^{N-1} x_n \, \exp\Bigl(-2\pi i \, \tfrac{\omega_k \, n}{N}\Bigr),$$

allowing the model to emphasize non-uniform frequency bands. During backpropagation, each ω_k is updated by the gradient

$$\frac{\partial \widehat{X}[k]}{\partial \omega_k} = -2\pi i \sum_{n=0}^{N-1} x_n \, \frac{n}{N} \, \exp\Bigl(-2\pi i \, \tfrac{\omega_k \, n}{N}\Bigr),$$

thus enabling adaptive tuning of the spectral mixing patterns to the data.

5.2 Adaptive Scale Selection

In the wavelet branch, rather than fixing all scales equally, we introduce a learnable scalar s_j for each scale $j = 1, \dots, J$ and compute normalized weights

$$\alpha_j = \frac{\exp(s_j)}{\sum_{\ell=1}^{J} \exp(s_\ell)}.$$

These weights modulate the contribution of each wavelet coefficient matrix $W^{(j)}$, so that the fused wavelet output is

$$W_{\text{fused}} = \sum_{j=1}^{J} \alpha_j \, W^{(j)},$$

letting the network focus on the most informative resolutions for each task and dynamically suppress less useful scales.

5.3 Multi-Modal Long-Sequence Fusion

To extend MDFWA to multi-modal inputs, let each modality m (e.g. text, audio, video) provide a sequence $X^{(m)} \in \mathbb{R}^{N_m \times d_m}$. We first map each to a common dimension d and apply modality-specific MDFWA stacks, yielding modality embeddings $E^{(m)} \in \mathbb{R}^{N_m \times d}$. For joint cross-mixing, we concatenate all queries and keys across modalities:

$$M_{\text{multi}} = \bigl[\, Q^{(1)}; Q^{(2)}; \dots; K^{(1)}; K^{(2)}; \dots \bigr],$$

and perform a single real FFT as before:

$$\widehat{M}_{\text{multi}} = \Re\bigl(\mathrm{FFT}(M_{\text{multi}})\bigr), \quad C_{\text{multi}} = \widehat{M}_{\text{multi}} \, V^{\top}.$$

Positional embeddings and modality masks ensure that intra-modal temporal order is preserved, while the spectral cross-mixing integrates information across modalities, enabling applications such as transcript-video summarization or audio-visual document alignment.

6 Experimental Plan

Our empirical evaluation rigorously assesses the proposed MDFWA Transformer on the PubMed 200K RCT dataset [7], which comprises approximately 200,000 medical abstracts (median length 2,715 tokens, 90th percentile exceeding 6,000 tokens). We cap input and output sequences at 4,096 tokens to accommodate the longest abstracts.

All models (FNET-Transformer, Hybrid-FNET, LED [2], and the proposed MDFWA Transformer) are trained with the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) using a peak learning rate of 5×10^{-5}, linear warmup over the first 10% of 50,000 update steps, and dropout of 0.1 in all layers. We use a batch size of 16 across eight V100 GPUs. Model configurations share embedding dimension $d = 512$, depth $L = 12$, and feed-forward dimension 2,048.

Summaries are generated with beam search (beam size 4, length penalty 1.0) and evaluated using ROUGE-1, ROUGE-2, and ROUGE-L F1 metrics. Following [14], we define:

$$\mathrm{R}_N = \frac{\sum_{g \in G_N^{\mathrm{ref}}} \min\bigl(\mathrm{Count}_{\mathrm{sys}}(g),\, \mathrm{Count}_{\mathrm{ref}}(g)\bigr)}{\sum_{g \in G_N^{\mathrm{ref}}} \mathrm{Count}_{\mathrm{ref}}(g)}, \tag{5}$$

$$\mathrm{P}_N = \frac{\sum_{g \in G_N^{\mathrm{sys}}} \min\bigl(\mathrm{Count}_{\mathrm{sys}}(g),\, \mathrm{Count}_{\mathrm{ref}}(g)\bigr)}{\sum_{g \in G_N^{\mathrm{sys}}} \mathrm{Count}_{\mathrm{sys}}(g)}, \tag{6}$$

$$\mathrm{F1}_N = 2\,\frac{\mathrm{R}_N\,\mathrm{P}_N}{\mathrm{R}_N + \mathrm{P}_N}. \tag{7}$$

We report mean scores with 95% confidence intervals via bootstrap resampling [12].

To isolate the impact of each MDFWA component, we perform targeted ablations by (i) fixing spectral frequencies ($\omega_k = k$), (ii) using uniform wavelet-scale weights ($\alpha_j = 1/J$), and (iii) comparing text-only training to text+section-headline multi-modal fusion. Tables 1 and 2 summarize the main results and ablation findings.

Table 1: Comparative ROUGE F1 scores on PubMed 200K RCT.

Model	ROUGE-1	ROUGE-2	ROUGE-L
FNET-Transformer	30.3	11.2	10.4
Hybrid-FNET	35.6	11.5	14.5
LED (allenai/led-base-16384)	37.2	13.5	20.1
MDFWA (proposed)	**39.8**	**14.7**	**21.9**

Table 2: Ablation study on MDFWA components.

Variant	ROUGE-1	ROUGE-2	ROUGE-L
Full MDFWA	39.8	14.7	21.9
w/o learned frequencies	38.4	14.1	20.8
w/o adaptive scales	39.0	14.3	21.2
text-only (no multi-modal fusion)	38.5	13.9	21.0

7 Limitations and Benefits

While the Multi-Domain Fourier-Wavelet Attention (MDFWA) Transformer achieves significant efficiency and scalability improvements, it also introduces several trade-offs that merit careful consideration.

Limitations

- **Approximation Rigidity:** Replacing learned attention weights with fixed Fourier and wavelet bases may under-represent highly content-dependent or asymmetric token interactions. Although our extensions (learned frequency bases and adaptive scales) partially address this, they introduce additional hyperparameters that require careful tuning.
- **Hardware Variability:** FFT and DWT operations enjoy mature CPU implementations, but their performance on emerging accelerators (e.g., TPU, specialized ASICs) can be inconsistent, potentially reducing the net speedup compared to optimized matrix multiplications in standard attention.

- **Implementation Complexity:** Integrating dual spectral and wavelet branches demands nontrivial mixing logic, bespoke initialization schemes, and potentially more training epochs to stabilize convergence, which may offset some of the simplicity gains from eliminating attention matrices.
- **Locality Limitations:** Although wavelet filters capture multi-resolution structure, very fine-grained local dependencies (e.g., rare token co-occurrences) may still be better modeled by explicit pairwise comparisons in self-attention.

Benefits

- **Subquadratic Scaling:** MDFWA reduces time complexity from $O(N^2)$ to $O(N \log N + N)$ and memory from $O(N^2)$ to $O(N)$, enabling efficient processing of documents thousands of tokens long under typical hardware budgets.
- **Parameter Efficiency:** By decoupling the bulk of token mixing from learned weights, MDFWA requires fewer parameters and avoids storing large attention matrices, yielding lower inference latency and reduced GPU/TPU memory usage.
- **Interpretability:** The explicit frequency- and time-domain decomposition allows practitioners to inspect and adjust which global themes (via Fourier mixing) and local structures (via wavelet scales) the model emphasizes, fostering greater transparency.
- **Modular Multi-Modal Fusion:** The same spectral cross-mixing mechanism seamlessly extends to heterogeneous modalities (text, audio, video) simply by concatenating their embeddings prior to a unified FFT, enabling unified long-sequence reasoning across diverse data streams.

In summary, MDFWA trades some of the flexibility of learned attention for substantial gains in runtime and memory efficiency, while offering a clear spectral interpretation and straightforward extension to multi-modal settings. Proper hyperparameter tuning and hardware-aware implementations are key to realizing its full potential.

8 Conclusion and Future Work

In this paper, we introduced the Multi-Domain Fourier-Wavelet Attention (MDFWA) Transformer, a novel architecture that integrates global Fourier token mixing with localized wavelet filtering across both encoder and decoder. Our comprehensive mathematical development detailed causal spectral kernels for autoregressive decoding, gradient derivations for learned frequency bases, and adaptive scale selection mechanisms, resulting in subquadratic runtime $O(N \log N)$ and linear memory $O(N)$. Empirically, MDFWA outperformed prior Fourier-based and sparse-attention baselines on the PubMed 200K RCT benchmark, achieving up to 21.9% ROUGE-L F1. Ablation studies confirmed the critical roles of learned spectral bases, adaptive wavelet scales, and multi-modal fusion.

Looking forward, we plan to explore overcomplete wavelet dictionary learning [9] to further enrich local context representations, as well as dynamic sequence length adaptation to selectively refine salient document segments. Extending MDFWA to end-to-end multi-modal pipelines, including text, images, audio, and video, promises unified summarization and cross-modal retrieval capabilities. Finally, rigorous evaluation on diverse long-sequence corpora, such as legislative transcripts and multimedia datasets, will assess the generality and scalability of our approach.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *ICLR*, 2015.

[2] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[3] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. In *CVPR*, pages 1233–1240, 2013.

[4] Sumit Chopra, Michael Auli, and Alexander M. Rush. Abstractive sentence summarization with attentive recurrent neural networks. In *NAACL–HLT*, pages 93–98, 2016.

[5] Krzysztof Choromanski, Viktor Likhosherstov, Daniel Dohan, Xingyou Song, Andreea Gane, Tamás Sarlos, Peter Hawkins, Jared Davis, Adrian Mohiuddin, Łukasz Kaiser, David Belanger, Luke Colwell, and Albert Weller. Rethinking attention with performers. In *ICLR*, 2021.

[6] Ingrid Daubechies. The wavelet transform, time–frequency localization and signal analysis. *IEEE Trans. Inf. Theory*, 36(5):961–1005, 1990.

[7] Franck Dernoncourt and Ji Young Lee. Pubmed 200k rct: A dataset for sequential sentence classification in medical abstracts. arXiv preprint arXiv:1710.06071, 2017.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL–HLT*, pages 4171–4186, 2019.

[9] Michal Elad and Michael Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006.

[10] Alexandros 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)*, pages 5156–5165, 2020.

[11] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *ICLR*, 2020.

[12] Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 388–395, 2004.

[13] James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontañón. Fnet: Mixing tokens with fourier transforms. In *ICLR*, 2021.

[14] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics.

[15] Stéphane Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1999.

[16] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019. https://openai.com/blog/better-language-models.

[17] Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *EMNLP*, pages 379–389, 2015.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[19] Sinong Wang, Belinda Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

[20] Zihang Xiong, Zihang Dai, Qingyan Hager, Soham Ramteke, Fady Khaled, Mike Johnson, Quoc V. Le, and Yuxin Lu. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *ICML*, 2021.

[21] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *NAACL–HLT*, pages 1480–1489, 2016.

[22] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. In *NeurIPS*, pages 17283–17297, 2021.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The novelty claims and implications are discussed and demonstrated in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: There is a detailed section 7.0, Limitations and Benefits that discusses the limitations, challenges and benefits of the proposed approach in language modeling.

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: This is detailed in section 3.0, Mathematical Development, of the paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Detailed experiments and algorithm setup are given in section 6.0 on Experimental Evaluation, Datasets and Baselines, Implementation Details as well as Evaluations Metrics of the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes] ,

   Justification: Data is publicly available PubMed 200k dataset.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes] ,

   Justification: Its all detailed in section 6.0 on Implementation details in the Experiment Plan section of the manuscript.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not have the compute resources to do a full statistical analysis of the model performance relative to other conventional techniques. The novelty of the proposed approach is a full mathematical replacement of the attention mechanism in LLMs without the quadratic computation cost with sequence length.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes] ,

   Justification: Please see implementation details including compute resources used in section 6.0

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read NeurIPS Code of Ethics in its entirety and confirmed compliance.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

Justification: This paper is proposing and demonstrating a computationally efficient approach to deep neural network implementations. It's primary contribution is on reducing computation complexity and not focussed on a specific application.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer:[NA] .

Justification: Not applicable since the data used is a publicly available dataset and no pretrained models are provided. The paper focusses on computation aspects of this new approach to reducing deep neural network complexity.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA] .

Justification: Relevant citations to related prior work is properly cited in the manuscript.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA] .

    Justification: None used.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: No Human Subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA] .

Justification: N/A

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.