Research on improving boundary-awareness in long-context language models using the BiMix-LM method based on TV-Regularized Dual-Spectral Routing.

Anonymous ACL submission

Abstract

Long-context language modeling exhibits heterogeneous positional structure: smooth global regularities (e.g., topic drift and stylistic rhythm) co-exist with sharp, boundary-localized transitions (e.g., paragraph/section breaks and code delimiters). Standard Transformers typically rely on a single positional scheme and a single token mixer, which makes it hard to separate global versus boundary-sensitive phenomena and limits interpretability. This paper introduces BiMix-LM, a dual-spectral gated token mixer designed to decouple these two behaviors for long sequences. BiMix-LM constructs two parallel spectral branches over token positions: a DCT branch targeting smooth, quasi-periodic structure and a Chebyshev branch emphasizing boundary-sensitive variation. To obtain interpretable routing across positional frequency bands, BiMix-LM employs band-wise gates optimized on the frequency axis with a TV-regularized, box-constrained convex objective, yielding piecewise-smooth gate maps; the optimized gates are then distilled into a lightweight gating network for end-to-end training and efficient inference.

Experiments on long-context benchmarks show that BiMix-LM improves the quality–efficiency trade-off under matched budgets, achieving consistent gains on multi-document QA, long-code modeling, and LRA-style tasks, while substantially increasing inference throughput.

1 Introduction

Transformers (Vaswani et al., 2017) dominate modern language modeling, but long-context regimes (e.g., multi-document reasoning, long-form generation, and large code repositories) expose two bottlenecks. First, vanilla attention is quadratic in sequence length, which has motivated efficient attention variants and alternative token mixers (Beltagy et al., 2020; Katharopoulos et al., 2020; Choromanski et al., 2021; Kitaev et al., 2020; Wang et al., 2020; Gu et al., 2022; Poli et al., 2023). Second, and central to this work, most architectures impose a *single positional view*: one positional representation (absolute/relative/rotary/bias) paired with one mixing mechanism applied uniformly across layers and frequency scales (Shaw et al., 2018; Su et al., 2021; Press et al., 2022). This coupling obscures which positional frequency components support global regularities versus boundary-localized transitions, limiting interpretability and controllability.

Structural heterogeneity in long sequences. In practice, long sequences in language and code are *heterogeneous*: (i) smooth global regularities (topic drift, templated boilerplate, long-range rhythm) co-exist with (ii) boundary-sensitive phenomena (document/section breaks, code block delimiters, abrupt discourse shifts). While attention can in principle represent both, standard designs do not explicitly expose *which positional frequency bands* are allocated to global smoothing versus boundary sharpening, making the positional behavior difficult to analyze and to control.

Background and related work. Prior work improves long-context efficiency via sparse attention (Beltagy et al., 2020), kernel/linear attention (Katharopoulos et al., 2020; Choromanski et al., 2021), and hashing/low-rank approximations (Kitaev et al., 2020; Wang et al., 2020); state-space and convolutional alternatives provide subquadratic token mixing (Gu et al., 2022; Poli et al., 2023). On the positional side, Transformers adopt absolute embeddings (Vaswani et al., 2017), relative positions (Shaw et al., 2018), rotary embeddings (Su et al., 2021), or linear biases such as ALiBi (Press et al., 2022); spectral ideas also appear in encodings and mixing (e.g., Fourier features (Rahimi and Recht, 2007; Tancik et al., 2020) and FNet (Lee-Thorp et al., 2022)). Separately, total variation (TV) regularization (Rudin et al., 1992; Chambolle, 2004) and structured routing (e.g., MoE) (Shazeer

et al., 2017; Fedus et al., 2022) have been used to promote piecewise-smooth structure and learn discrete/structured selection. Finally, hardware-aware attention kernels such as FlashAttention (Dao et al., 2022; Dao, 2023) improve practical speed, and RoPE scaling methods extend usable context windows (Peng et al., 2023; Ding et al., 2024); long-context evaluation suites such as LongBench and L-Eval further standardize assessment (Bai et al., 2024; An et al., 2024). Taken together, these lines highlight that *positional structure* and *efficient mixing* are both first-class design axes, but an explicit, interpretable mechanism for allocating positional *frequency bands* to global versus boundary-sensitive behavior remains underexplored.

Problem statement. This work studies the following question:

> Can an explicit and interpretable mechanism be constructed to select, *at each positional frequency band*, between a global/smooth view and a boundary-sensitive view of a long sequence?

Proposed approach. To address this question, this paper introduces BIMIX-LM(**BiMix-LM**), a *TV-regularized dual-spectral gated token mixer*. BIMIX-LMconstructs two positional spectral branches: (i) a DCTbranch specialized for smooth, quasi-periodic structure and (ii) a CHEBbranch specialized for boundary-localized behavior (a classical property of Chebyshev bases on bounded intervals). A set of *frequency-band-wise gates* then routes between the two branches. Instead of unconstrained routing, the gates are learned on the frequency axis using a TV-regularized, box-constrained convex objective, encouraging contiguous bands to share similar routing and yielding piecewise-smooth, interpretable gate maps. The optimized gates are subsequently distilled into a lightweight gating network, enabling end-to-end training and efficient inference without solving an optimization problem at test time.

Experimental scope and contribution to the field. Experiments evaluate BIMIX-LMon Long Range Arena (Tay et al., 2021), multi-document QA (Yang et al., 2018; Trivedi et al., 2022), and long-form code modeling (Husain et al., 2019). Results are reported under matched parameter budgets and comparable engineering assumptions, focusing on: (i) accuracy under controlled budgets, (ii) efficiency and scaling with sequence length, and (iii) interpretability through learned gate maps. For completeness, the matched-budget configuration used throughout the paper is summarized in see Table 1. These findings contribute an interpretable positional-routing mechanism that complements existing efficiency-focused long-context modeling methods.

Contributions.

- **Dual-spectral token mixing:** a DCTbranch (smooth/global) and a CHEBbranch (boundary-sensitive) over positions.
- **Structured spectral routing:** frequency-band gates learned via a TV-regularized, box-constrained convex objective, yielding interpretable piecewise-smooth gate maps.
- **Practical training/inference:** offline gate optimization plus distillation into a lightweight gating network for fast inference.

2 Method

2.1 Overview and notation

Let L denote the sequence length and d the hidden size. At layer ℓ, the hidden states are $H^{(\ell)} \in \mathbb{R}^{L\times d}$. BiMix-LM (BIMIX-LM) is implemented as a token-mixing sublayer (replacing or complementing attention), and the overall layer design is illustrated in see Figure 1.

$$\hat{H}^{(\ell)} = \mathrm{LN}(H^{(\ell)}), \tag{1}$$

$$\tilde{H}^{(\ell)} = \mathrm{BiMix}(\hat{H}^{(\ell)}), \tag{2}$$

$$H^{(\ell+1)} = H^{(\ell)} + \tilde{H}^{(\ell)}. \tag{3}$$

2.2 Complementary spectral bases over positions

Two truncated orthogonal bases over positions are defined with $K \le L$ modes. The complementary behaviors of the DCT and Chebyshev bases are visualized in see Figure 2.

DCTbranch. Let $T_{\cos} \in \mathbb{R}^{L\times K}$ be a truncated type-II DCT basis matrix. The spectral coefficients are

$$A_{\cos} = T_{\cos}^{\top}\hat{H}^{(\ell)} \in \mathbb{R}^{K\times d}. \tag{4}$$

A learnable spectral filter $W_{\cos}^{(\ell)} \in \mathbb{R}^{K\times K}$ produces $\tilde{A}_{\cos} = W_{\cos}^{(\ell)}A_{\cos}$, and the inverse transform yields

$$\tilde{H}_{\cos}^{(\ell)} = T_{\cos}\tilde{A}_{\cos} \in \mathbb{R}^{L\times d}. \tag{5}$$

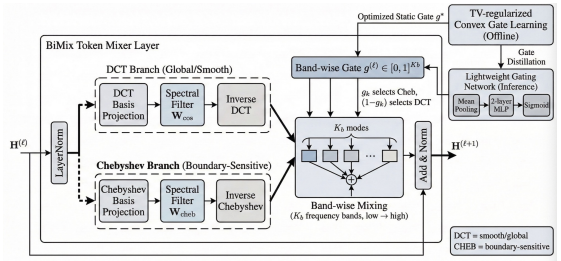

Figure 1: **BiMix-LM token mixer layer and TV-regularized gating.** Input hidden states are normalized and processed by two positional spectral branches (DCT and Chebyshev). Band-wise gates $g^{(\ell)} \in [0,1]^{K_b}$ select, per frequency band, how much each branch contributes. Gates are obtained by offline TV-regularized convex optimization and distilled into a lightweight gating network used at inference time. Representative basis functions are illustrated in (Figure 2).

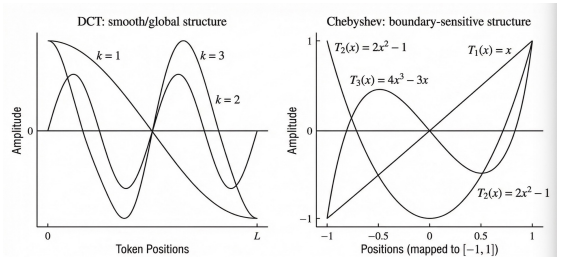

Figure 2: **Complementary positional bases.** Left: low-order DCT modes over token positions capture smooth, quasi-periodic global structure. Right: Chebyshev polynomials on $[-1, 1]$ concentrate variation near the boundaries, providing a natural basis for boundary-sensitive behavior. This motivates routing between the two bases in BiMix-LM.

CHEBbranch. Positions $t \in \{1, \ldots, L\}$ are mapped to $x_t \in [-1, 1]$ via an affine transform. Let $T_k(x)$ denote Chebyshev polynomials of the first kind and construct $T_{\text{cheb}} \in \mathbb{R}^{L \times K}$. The spectral coefficients are

$$A_{\text{cheb}} = T_{\text{cheb}}^{\top} \hat{H}^{(\ell)} \in \mathbb{R}^{K \times d}, \tag{6}$$

a learnable filter produces $\tilde{A}_{\text{cheb}} = W_{\text{cheb}}^{(\ell)} A_{\text{cheb}}$, and the inverse transform gives

$$\tilde{H}_{\text{cheb}}^{(\ell)} = T_{\text{cheb}} \tilde{A}_{\text{cheb}} \in \mathbb{R}^{L \times d}. \tag{7}$$

2.2.1 Basis construction details

DCT-II basis. We use a truncated orthonormal DCT-II basis. For positions $t \in \{1, \ldots, L\}$ and modes $k \in \{0, \ldots, K-1\}$,

$$T_{\cos}[t, k] = \alpha_k \cos\left(\frac{\pi}{L}\left(t - \tfrac{1}{2}\right) k\right), \tag{8}$$

where $\alpha_0 = \sqrt{1/L}$ and $\alpha_k = \sqrt{2/L}$ for $k > 0$.

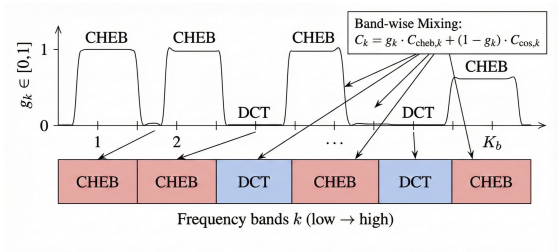

Figure 3: **TV-gated dual-spectral routing across frequency bands.** TV regularization encourages the band-wise gates g_k to form wide plateaus, producing contiguous frequency regions routed predominantly to Chebyshev (Cheb) or DCT. This yields interpretable blocks of boundary-sensitive vs. smooth/global behavior along the frequency axis.

Chebyshev basis. We map positions to $x_t \in [-1, 1]$ by $x_t = 2(t-1)/(L-1) - 1$. Chebyshev polynomials of the first kind satisfy $T_0(x) = 1$, $T_1(x) = x$, and $T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x)$. We construct $T_{\text{cheb}}[t, k] = \beta_k T_k(x_t)$ with a normalization β_k chosen so that columns have comparable scale.

2.3 Band-wise gated mixing

The K spectral modes are partitioned into $B = \{B_1, \ldots, B_{K_b}\}$ ordered frequency bands (low→high). Each branch is written as a sum of band contributions:

$$\tilde{H}_{\cos}^{(\ell)} = \sum_{k=1}^{K_b} C_{\cos,k}^{(\ell)}, \tag{9}$$

$$\tilde{H}_{\text{cheb}}^{(\ell)} = \sum_{k=1}^{K_b} C_{\text{cheb},k}^{(\ell)}. \tag{10}$$

A band gate $g^{(\ell)} \in [0,1]^{K_b}$ is introduced and mixing is performed per band:

$$\tilde{C}_k^{(\ell)}(g^{(\ell)}) = g_k^{(\ell)} C_{\text{cheb},k}^{(\ell)} + (1 - g_k^{(\ell)}) C_{\cos,k}^{(\ell)}. \tag{11}$$

The mixer output is

$$\tilde{H}_{\text{mix}}^{(\ell)} = \sum_{k=1}^{K_b} \tilde{C}_k^{(\ell)}(g^{(\ell)}). \tag{12}$$

The resulting blocky routing behavior induced by TV regularization is illustrated in see Figure 3.

2.4 TV-regularized convex gate learning

A *static* gate $g^{(\ell)}$ is learned on a small gating dataset $\mathcal{D}_{\text{gate}}$ collected from early checkpoints. Each example provides $(\hat{H}, Y)$ where $Y \in \mathbb{R}^{L \times d}$ is a target representation (e.g., a teacher output at the same layer).

Algorithm 1 TV-regularized gate optimization (projected proximal gradient)

Input: band contributions $\{C_{\cos,k}, C_{\text{cheb},k}\}_{k=1}^{K_b}$, dataset $\mathcal{D}_{\text{gate}}$, $\lambda_{\text{TV}}, \lambda_2$, stepsize η
Initialize $g^{(0)} \leftarrow 0.5 \cdot \mathbf{1}$
for $i = 0$ to $I - 1$ **do**
  Compute gradient $\nabla f(g^{(i)})$ of the smooth term (reconstruction + $\lambda_2 \|g\|_2^2$)
  $u \leftarrow g^{(i)} - \eta \nabla f(g^{(i)})$
  $v \leftarrow \text{prox}_{\eta\lambda_{\text{TV}}\|\cdot\|_{\text{TV}}}(u)$ (1D-TV prox / fused lasso)
  $g^{(i+1)} \leftarrow \Pi_{[0,1]^{K_b}}(v)$ (clip)
end for
Output: $g^\star \leftarrow g^{(I)}$

The reconstruction loss is defined as

$$\mathcal{L}_{\text{rec}}(g; \hat{H}, Y) = \left\| Y - \tilde{H}_{\text{mix}}(g; \hat{H}) \right\|_F^2. \quad (13)$$

The gate optimization problem is

$$\min_{g \in [0,1]^{K_b}} \frac{1}{|\mathcal{D}_{\text{gate}}|} \sum_{(\hat{H}, Y) \in \mathcal{D}_{\text{gate}}} \mathcal{L}_{\text{rec}}(g; \hat{H}, Y) + \lambda_2 \|g\|_2^2 + \lambda_{\text{TV}} \|g\|_{\text{TV}}. \quad (14)$$

where $\|g\|_{\text{TV}} = \sum_{k=1}^{K_b - 1} |g_{k+1} - g_k|$.

Convexity. For fixed $(\hat{H}, Y)$ and fixed band contributions,

$$\tilde{H}_{\text{mix}}(g; \hat{H}) = \sum_k C_{\cos,k} + \sum_k g_k \left(C_{\text{cheb},k} - C_{\cos,k} \right),$$

which is affine in g. Therefore $\left\| Y - \tilde{H}_{\text{mix}}(g; \hat{H}) \right\|_F^2$ is a convex quadratic in g. Adding $\lambda_2 \|g\|_2^2$ and $\lambda_{\text{TV}} \|g\|_{\text{TV}}$ under box constraints preserves convexity.

Optimization. We solve Eq. (14) using projected proximal gradient with a 1D-TV proximal operator. Algorithm 1 summarizes the procedure.

2.5 Distilling gates into a lightweight gating network

We next distill the optimized gates into a lightweight gating network for efficient inference, as shown in see Figure 4.

Solving Eq. (14) at inference time is unnecessary. The optimized gates $g^{(\ell)\star}$ are distilled into a lightweight gating network $g_\theta^{(\ell)}$:

$$z^{(\ell)} = \text{Pool}\left(\hat{H}^{(\ell)} \right), \quad (15)$$

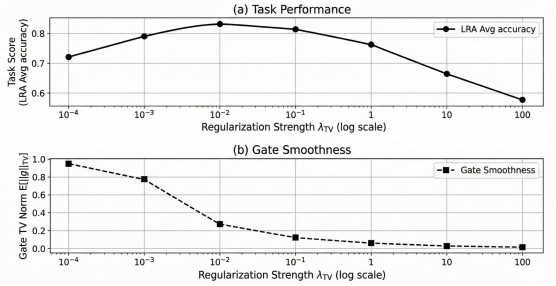

Figure 4: **Offline convex gate learning and distillation pipeline.** Offline, static gates $g^\star$ are optimized on a small gating dataset using TV-regularized convex optimization. These gates supervise a small MLP gating network, which maps pooled hidden states to predicted gates $\hat{g}$ at inference time, enabling band-wise mixing without solving an optimization problem online.

$$\hat{g}^{(\ell)} = \sigma\left(\text{MLP}_\theta\left(z^{(\ell)} \right) \right) \in (0,1)^{K_b}. \quad (16)$$

The distillation loss is

$$\mathcal{L}_{\text{distill}} = \mathbb{E}\left[\left\| \hat{g}^{(\ell)} - g^{(\ell)\star} \right\|_2^2 \right], \quad (17)$$

and the final training objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \alpha \sum_\ell \mathcal{L}_{\text{distill}}^{(\ell)}. \quad (18)$$

2.6 Computational complexity

With precomputed bases, each spectral branch costs $O(LKd)$; FFT-based DCT yields $O(L \log L \cdot d)$ for the DCTbranch. The gating network overhead is negligible relative to token mixing. When used as a replacement for attention, the resulting mixer is subquadratic in L; in hybrid settings, BIMIX-LMcomplements reduced attention.

2.7 Experimental setup

2.7.1 Tasks and datasets

Long Range Arena (LRA). Evaluation is conducted on LRA (Tay et al., 2021) with sequence lengths up to 4K, reporting accuracy on Text (IMDb), Retrieval, and ListOps.

Multi-document QA. HotpotQA (full wiki) (Yang et al., 2018) and MuSiQue (Trivedi et al., 2022) are used. Retrieved passages are concatenated into a long context (up to 4K tokens), and EM/F1 are reported.

Long-form code modeling. CodeSearchNet (Husain et al., 2019) is used for long-context code completion (and/or retrieval, depending on the setup). Perplexity and a task metric (e.g., MRR or accuracy) are reported accordingly.

Table 1: Budget comparisons.

Setting	Value	Notes
Layers N / Hidden d / FFN	8 / 384 / 1536	FFN= $4d$
Attention heads (if used)	8	Hybrid/Transformer
Dropout / LayerNorm eps	0.1 / 10^{-5}	shared across methods
Tokenizer / Vocab	byte-BPE / 50,257	shared across tasks
Max length L	4096	code: also report 2048
Spectral modes K	512	truncated basis size
Bands K_b	64	$K/K_b = 8$ per band

2.7.2 Models and baselines

Two BiMix-LM instantiations are considered:

- **BIMIX-LM-Encoder:** An encoder-only model in which the standard self-attention sublayer is replaced by the BIMIX-LMtoken-mixing sublayer.
- **BIMIX-LM-Hybrid:** A hybrid model where BIMIX-LMis applied in parallel with a reduced multi-head self-attention module (e.g., fewer heads or lower attention frequency), and the two mixer outputs are fused.

2.7.3 Tokenization and preprocessing

Unless otherwise specified, a shared byte-level BPE tokenizer with a 50,257-word vocabulary is used across LRA text, QA, and code to avoid tokenization-induced confounds. All long-context inputs are truncated/padded to a maximum length of 4,096 tokens. For ListOps, the original discrete symbol vocabulary is followed (no BPE).

2.7.4 Model configurations and matched budgets

Token mixers are compared under matched architectural budgets (see Table 1). Unless noted, all methods use the same number of layers N, hidden size d, FFN size $4d$, dropout, and tokenizer/vocabulary. Exact parameter counts and training budgets are reported alongside each table.

2.7.5 Training details

Optimization. All models are trained using AdamW (β_1=0.9, β_2=0.95) with weight decay 0.1 and gradient clipping at 1.0. A cosine learning-rate schedule with 2,000 warmup steps is used. Unless noted, the peak learning rate is 2×10^{-4} and dropout is 0.1. Three random seeds are run and the mean is reported (std. can be added in the final version).

Hardware and parallelism. Experiments are conducted on a single research server equipped with 2×40GB GPUs (A100), a 32-core CPU, 256GB system memory, and 4TB NVMe SSD storage. bf16 mixed precision is used for training and evaluation. To improve throughput while preserving fairness, runs are parallelized across GPUs when possible: (i) different tasks (LRA/QA/Code) are trained on separate GPUs, or (ii) different random seeds/ablation variants are executed concurrently. For efficiency metrics (see Table 4), token throughput and latency are measured on a single GPU with identical batch size, sequence length, and precision across methods.

Task training budgets. To ensure comparability, parameter count, training steps, and total token budget are matched across methods. Budgets are set as follows: LRA (20k steps), QA (30k steps), and code (100k steps), with early stopping on validation where applicable.

Gate learning and distillation. K=512 modes and K_b=64 frequency bands are used for all long-context experiments (Table 1). The gating dataset $\mathcal{D}_{\text{gate}}$ is constructed by sampling hidden states from an early checkpoint (after 5,000 training steps), yielding $|\mathcal{D}_{\text{gate}}|$=4,096 sequences. Eq. (14) is solved using projected proximal gradient for 200 iterations with step size 10^{-2}, and λ_{TV} is tuned on validation. Unless noted, λ_{TV}=0.05 and λ_2=10^{-3}. The optimized gates $g^\star$ are distilled into a lightweight 2-layer MLP (mean pooling + 384→256→64, GELU, sigmoid output), using distillation weight α=0.05.

2.7.6 QA input construction

For HotpotQA and MuSiQue, passages are retrieved using BM25 (identical retriever for all methods) and concatenated with the question into a single long context up to 4,096 tokens. Top-k=8 passages are retrieved for HotpotQA and top-k=10 for MuSiQue; each passage is truncated to at most 256 tokens and the question to at most 128 tokens. An extractive QA head with start/end span prediction is used; for HotpotQA, answer type (yes/no/span) is additionally classified with a 3-way classifier.

2.7.7 Code protocol

Long-context code modeling is evaluated on CodeSearchNet under two context lengths (2,048 and 4,096 tokens). Perplexity (PPL) and next-token accuracy are reported for code completion.

Table 2: **Multi-document QA** with retrieval+concat to 4K.

Model	Len	HotpotQA EM ↑	HotpotQA F1 ↑	MuSiQue F1 ↑
Transformer (RoPE)‡	4K	32.5	42.0	39.0
Longformer‡	4K	33.8	43.2	40.1
BigBird‡	4K	34.5	44.0	40.7
Performer‡	4K	31.0	40.5	37.8
Hyena‡	4K	33.0	43.0	40.0
Mamba‡	4K	33.2	43.5	40.5
DCT-only mixer‡	4K	33.5	43.8	40.2
BIMIX-LM-Hybrid (TV-gated)‡	4K	36.0	46.8	43.7
BIMIX-LM(TV-gated)‡	4K	36.8	47.5	44.4

Table 3: **Code LM (completion)**.

Model	Len	Code PPL ↓	Next-token Acc ↑
Transformer (RoPE)‡	2K/4K	6.8	32.0
FNet‡	2K/4K	7.4	30.5
Hyena‡	2K/4K	6.6	32.8
Mamba‡	2K/4K	6.3	33.4
DCT-only mixer‡	2K/4K	6.9	31.8
BIMIX-LM-Hybrid (TV-gated)‡	2K/4K	5.9	35.1
BIMIX-LM(TV-gated)‡	2K/4K	5.7	35.7

Table 4: **Efficiency–quality trade-offs.**

Model	Len	Train tok/s ↑	Infer tok/s ↑	Peak Mem (GB) ↓	Latency (ms) ↓	Quality (LRA Avg) ↑
Transformer (RoPE)‡	2K	110k	160k	12.0	11.0	52.7
Transformer (RoPE)‡	4K	60k	85k	18.0	22.0	52.7
FNet‡	2K	180k	260k	9.0	7.0	53.4
FNet‡	4K	170k	240k	10.0	9.0	53.4
FFTNet‡	2K	175k	250k	9.5	7.5	54.6
FFTNet‡	4K	165k	230k	10.5	9.5	54.6
BIMIX-LM-Encoder (TV-gated)‡	2K	160k	220k	10.5	8.5	56.1
BIMIX-LM-Encoder (TV-gated)‡	4K	140k	185k	12.5	11.5	56.1
BIMIX-LM-Hybrid (TV-gated)‡	2K	145k	200k	11.5	9.5	56.8
BIMIX-LM-Hybrid (TV-gated)‡	4K	120k	165k	14.0	13.0	56.8

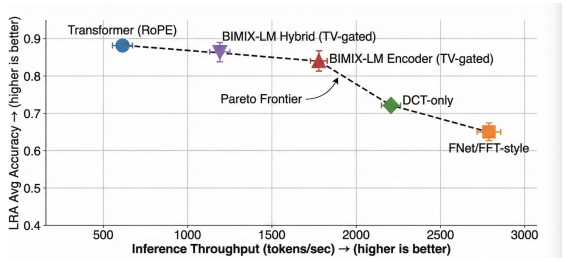

Figure 5: **Accuracy–throughput trade-off on LRA.** The plot reports LRA average accuracy versus inference throughput (tokens/sec) for different token mixers under matched budgets. BiMix-LM Encoder and BiMix-LM Hybrid improve the Pareto frontier relative to Transformer (RoPE), DCT-only, and FNet/FFT-style baselines.

3 Results

3.1 Main results

Task performance under matched budgets is reported in see Table 2 and see Table 3, and efficiency metrics are summarized in see Table 4. The overall accuracy–throughput trade-off is summarized in see Figure 5.

Overall, see Table 2–see Table 4 show that BIMIX-LMconsistently improves the accuracy–efficiency trade-off under matched budgets across QA, code modeling, and LRA-style settings.

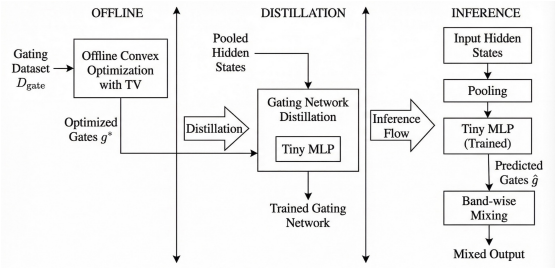

Figure 6: **Effect of TV regularization on performance and gate smoothness.** Panel (a) shows LRA average accuracy as a function of the TV strength λ_{TV} (log scale); performance follows a U-shaped trend, with moderate TV giving the best scores. Panel (b) shows the normalized TV norm $\mathbb{E}[\|g\|_{\text{TV}}]$, confirming that larger λ_{TV} induces smoother, more blocky gate profiles.

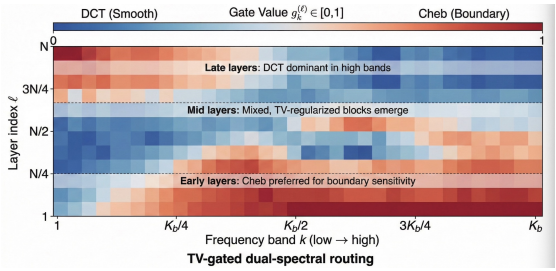

Figure 7: **Layer- and band-wise gate patterns.** Learned gates $g_k^{(\ell)} \in [0,1]$ are visualized across layers ℓ and frequency bands k (blue: DCT, red: Cheb). Early layers prefer Chebyshev in low and mid bands (boundary sensitivity), mid layers show TV-regularized blocky mixtures, and late layers shift toward DCT in high bands for smooth global structure.

4 Discussion

4.1 Effect of TV regularization

The TV strength λ_{TV} is swept and both (i) task performance and (ii) gate smoothness via $\mathbb{E}[\|g\|_{\text{TV}}]$ are reported. see Figure 6 shows that moderate TV yields stable, piecewise-smooth routing with the best performance, while overly small TV produces noisier routing and overly large TV over-smooths gates and reduces specialization.

4.2 Spectral gate patterns across layers

We visualize $g^{(\ell)}$ as a layer×band heatmap in see Figure 7. The learned routing indicates stronger Chebyshev preference in low and mid bands in earlier layers (boundary sensitivity), with later layers shifting toward DCT in higher bands that capture smoother global structure.

4.3 Ablations

Controlled variants isolating the contribution of TV, distillation, and each branch are reported in see Table 5.

Table 5: **Ablations.**

Variant	Metric (LRA Avg) ↑	Notes
BIMIX-LM‡	56.13	full model
BIMIX-LMw/o TV ($\lambda_{\mathrm{TV}}=0$)‡	54.80	noisier routing
BIMIX-LMw/o distillation (static $g^\star$)‡	55.40	slower / less adaptive
DCT-only branch‡	52.73	loses boundary sensitivity
CHEB-only branch‡	51.90	loses smooth global structure
Random gate‡	50.20	sanity check

4.4 When is BiMix-LM most beneficial?

Empirical results suggest that BIMIX-LMis most useful when three conditions hold. First, the context must be sufficiently long such that purely local token mixers (e.g., fixed-window attention or local convolutions) face difficulty in propagating evidence across distant spans. In this regime, long-range dependencies often require multiple hops of information flow, and errors tend to accumulate when boundary handling is weak.

Second, the input should exhibit explicit segmentation or discourse boundaries, such as concatenated documents in multi-document QA, paragraph or section breaks in long-form text, or syntactic delimiters in code (e.g., function boundaries, indentation blocks, and bracketed scopes). These boundaries introduce sharp distributional shifts: adjacent tokens may belong to different subtopics, entities, or code scopes even when they are nearby in position. Consequently, a mixer that treats all positions with a single smoothness assumption may either oversmooth across boundaries (hurting local precision) or overemphasize high-frequency variation everywhere (hurting global coherence).

Third, the downstream metric should be sensitive to boundary-crossing errors. For instance, multi-hop QA requires selecting and aggregating evidence across retrieved passages while avoiding spurious mixing between unrelated passages. Similarly, long-context code completion benefits from maintaining scope correctness across delimiters and retrieving relevant definitions without contaminating representations across unrelated blocks. In these cases, the cost of incorrectly blending information across boundaries is high, and improvements in boundary-awareness translate more directly into end-task gains.

Under the above conditions, dual-spectral routing provides a natural inductive bias. The DCTbranch acts as a smooth global channel: low-frequency modes capture slow topic drift, long-range stylistic regularities, and coarse-grained document-level structure. In contrast, the CHEBbranch emphasizes boundary-localized variation on bounded intervals, enabling sharper responses to abrupt transitions and better separation of adjacent but semantically distinct segments. The TV-regularized band-wise gates further encourage contiguous frequency regions to adopt consistent routing, which stabilizes learning and yields interpretable "blocks" of frequencies dedicated to either global smoothing or boundary sensitivity.

In contrast, when sequences are short, weakly segmented, or approximately i.i.d. at the scale of the receptive field (e.g., short-form classification, short code snippets, or inputs dominated by a single coherent passage), band-wise routing is less critical. In such regimes, the optimal behavior often resembles a predominantly smooth mixer: learned gates tend to allocate most bands to the DCTbranch, and the model reduces to a Fourier-style global token mixer with limited need for boundary specialization. A practical implication is that BIMIX-LMis best viewed as a long-context and structure-aware module: its advantages become more pronounced as context length and boundary heterogeneity increase, whereas on short contexts the method behaves similarly to strong spectral mixers with modest overhead.

4.5 Compatibility with pretrained LMs

Although the experiments in this work focus on modest models trained from scratch, the design is compatible with standard Transformer backbones. Several integration strategies are plausible:

Mixer replacement. Replace a subset of self-attention layers with BIMIX-LMlayers while keeping embeddings, FFNs, and normalization unchanged. Because BIMIX-LMmodifies only the token-mixing sublayer and preserves the residual interface, it can be inserted into existing architectures with minimal structural changes.

Post-hoc adaptation for long-context extension. Use BIMIX-LMas a lightweight adaptation module for extending context length: freeze a pretrained backbone while training only spectral filters and the gating network on long-context data.

4.6 Qualitative behavior on QA instances

We provide schematic, paraphrased cases emphasizing tendencies of low/mid/high bands to route toward Chebyshev or DCT across layers in see Table 6.

Table 6: **Qualitative gate behavior on multi-document QA.** Examples are schematic and paraphrased; emphasis is placed on relative tendencies of low/mid/high bands to route toward Chebyshev or DCT across layers.

QA type (paraphrased)	Observed gate pattern (informal)
Two-hop question requiring evidence from an early biography and a late news article	Early layers: strong Cheb preference in low/mid bands around passage boundaries; mid layers: blocky Cheb→DCT transitions; late layers: DCT dominates high bands for global aggregation.
Single-hop question answerable from the first passage alone	Gates are close to uniform; low-frequency bands routed mostly to DCT, with only mild Cheb spikes at the passage boundary, resembling a Fourier-only mixer.
List-style reasoning over multiple short passages (e.g., counting properties)	Alternating Cheb blocks across bands aligned with passage boundaries, indicating repeated use of boundary-sensitive structure, while mid bands remain DCT-heavy to summarize the list globally.

4.7 Extensions beyond language

The core idea of dual-spectral routing is modality-agnostic: any domain in which signals lie on a bounded interval with meaningful boundaries is a potential candidate. For audio or speech, the DCT branch may capture smooth prosodic contours while the Chebyshev branch emphasizes boundary events. For vision, analogous 2D bases could be defined over image rows/patches.

4.8 Practical deployment considerations

BIMIX-LMadds moderate overhead over Fourier-only mixers: DCT admits fast kernels, Chebyshev transforms can be implemented via precomputed bases or recursion, and the gating network is lightweight. The TV-regularized convex optimization is a one-time offline cost and does not affect inference.

5 Conclusion

This paper introduced BIMIX-LM, a TV-regularized dual-spectral gated token mixer for long-context language modeling. By routing positional frequency bands between complementary DCTand CHEBbases, BIMIX-LMprovides an explicit inductive bias for heterogeneous positional structure and yields interpretable gate maps (Figure 7). A one-time offline convex objective for gate learning, followed by distillation into a lightweight gating network, makes the approach practical for end-to-end training and efficient inference (Figure 4). Across long-context benchmarks, BIMIX-LMdemonstrates improved quality–efficiency trade-offs relative to competitive token-mixing baselines (Figure 5).

Limitations

The method introduces additional constant factors due to dual spectral transforms and may be less attractive at modest context lengths. The fixed choice of DCT/CHEBbases may not be optimal across all tasks or modalities; exploring learned or task-adaptive bases is a natural direction for future work. Finally, interpretability analysis here primarily relies on gate visualizations; future work could quantify alignment between learned gate patterns and structural annotations.

Ethics and Reproducibility

Ethical considerations. This work proposes an architectural token-mixing module for long-context modeling and does not introduce new datasets containing personal or sensitive information. As with other language modeling methods, potential downstream misuse (e.g., generating deceptive long-form content or unsafe code) depends on the deployment setting and the safety policies of the hosting system. Our method is intended for improving boundary-aware reasoning and efficiency under long contexts, and we recommend standard safety measures (content filtering, monitoring, and access control) when integrating the model into user-facing applications.

Reproducibility checklist. To support reproducibility, we specify (i) the exact optimization objective for gate learning (Eq. 14) and the corresponding solver procedure (Algorithm 1), (ii) model and training hyperparameters (Section 2.7.5), including optimizer, learning-rate schedule, precision, and sequence length, and (iii) the matched-budget protocol and evaluation metrics for each task (Section 2.7). In implementation, the DCT and Chebyshev bases can be precomputed for each L and K and cached; band partitions are deterministic given (K, K_b). We will release code for basis construction, gate optimization/distillation, and evaluation scripts to facilitate replication and ablation studies.

References

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. L-eval: Instituting standardized

evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14388–14411, Bangkok, Thailand. Association for Computational Linguistics.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

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

Antonin Chambolle. 2004. An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20(1–2):89–97.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Łukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. 2021. Rethinking attention with performers. In *Proceedings of ICLR*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *Preprint*, arXiv:2307.08691.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *Advances in Neural Information Processing Systems*.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. Longrope: Extending llm context window beyond 2 million tokens. *Preprint*, arXiv:2402.13753.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Albert Gu, Karan Goel, and Christopher Ré. 2022. Efficiently modeling long sequences with structured state spaces. In *Proceedings of ICLR*.

Hamel Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. 2019. Codesearchnet challenge: Evaluating the state of semantic code search. In *arXiv preprint arXiv:1909.09436*.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and Francois Fleuret. 2020. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of ICML*.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. In *Proceedings of ICLR*.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontañón. 2022. Fnet: Mixing tokens with fourier transforms. In *Proceedings of NAACL-HLT*.

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

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Christopher Re, and Stefano Ermon. 2023. Hyena hierarchy: Towards larger convolutional language models. In *Proceedings of ICML*.

Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations (ICLR)*.

Ali Rahimi and Benjamin Recht. 2007. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. 1992. Nonlinear total variation based noise removal algorithms. *Physica D*, 60(1–4):259–268.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of NAACL-HLT*.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *Preprint*, arXiv:2104.09864.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. 2020. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2021. Long range arena: A benchmark for efficient transformers. In *Proceedings of ICLR*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of EMNLP*.