TEMPORAL OPERATOR ATTENTION: RECOVERING COMMON OPERATORS IN TIME-SERIES ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world time-series applications repeatedly exhibit five common structural regimes: seasonal extrapolation, trend/anomaly separation, cross-channel demixing, regime-shift alignment, and transient recovery. Each requires a signed or zero-sum linear map over time, not a convex combination of the sequence's own values, and some further require this map to adapt to the input. Linear and MLP-based mixers, despite far fewer parameters, often match or outperform attention-based forecasters on exactly this structure, because they mix directly along the temporal dimension via a signed, position-indexed operator, rather than one derived from content-based similarity and confined to a probability simplex. We prove that softmax attention's row-stochastic kernel where nonnegative weights summing to one cannot realize any of the five recurring regimes from a single head, tracing the gap to this primitive-level mismatch rather than model scale. We propose **Temporal Operator Attention (TOA)**, which restores this signed, position-indexed operator to the attention pathway itself: learnable dense residual operators wrap the score matrix, supplying signed degrees of freedom, while a gated variant further makes the operator input-adaptive. *Stochastic Operator Regularization (SOR)* controls overfitting from this added capacity by regularizing only the learned offsets, preserving identity-based signal transport. Across Autoformer, PatchTST, iTransformer, and DUET, TOA consistently improves long- and short-term forecasting, anomaly detection, and classification, with the largest gains where signed temporal mixing matters most. Code is available at this link.

1 INTRODUCTION

Time-series measurements, including temperature, electricity load, road occupancy, and industrial sensor streams, drive decisions in weather forecasting, energy management, traffic planning, and industrial monitoring, so forecasting errors translate directly into wasted energy, congestion, and missed failures. Deep learning has been widely used for these signals, with Transformer-based architectures making attention and related mechanisms a prominent family of sequence mixers in time-series modeling (Wu et al., 2021; Nie et al., 2023; Liu et al., 2024; Qiu et al., 2025). Across these applications, the same five structural regimes recur, each requiring a signed or zero-sum linear map over time, not a convex combination of the sequence's own values: **seasonal extrapolation** (weather), extending cycles past the observed window (Oppenheim et al., 1999); **trend/anomaly separation** (energy load), isolating fast fluctuations from a slow baseline (Oppenheim et al., 1999; Hodrick & Prescott, 1997); **cross-channel demixing** (power grids), recovering independent factors from correlated sensors (Comon, 1994); **regime-shift alignment** (traffic), tracking abrupt phase shifts with an input-adaptive operator (Sakoe & Chiba, 1978); and **transient recovery** (industrial monitoring), isolating short-lived events against noise (Gabor, 1946). Figure 1 summarizes these regimes and the structural motivation for TOA.

Yet on exactly this signed-mixing structure, Transformers are routinely outperformed: linear and MLP-based models, with far fewer parameters, match or surpass them (Zeng et al., 2022; Chen et al., 2023; Wang et al., 2024a; Yue et al., 2025), a gap persisting across specialized designs (Zhou et al., 2021; Wu et al., 2021; Zhou et al., 2022; Nie et al., 2023; Liu et al., 2024; Wu et al., 2023; Qiu et al., 2025) and recent benchmarks (Wang et al., 2024b). These lightweight models share what Transformers lack: a signed operator that mixes directly along the temporal dimension, indexed by position, rather than one derived from content-based similarity and confined to a probability simplex.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

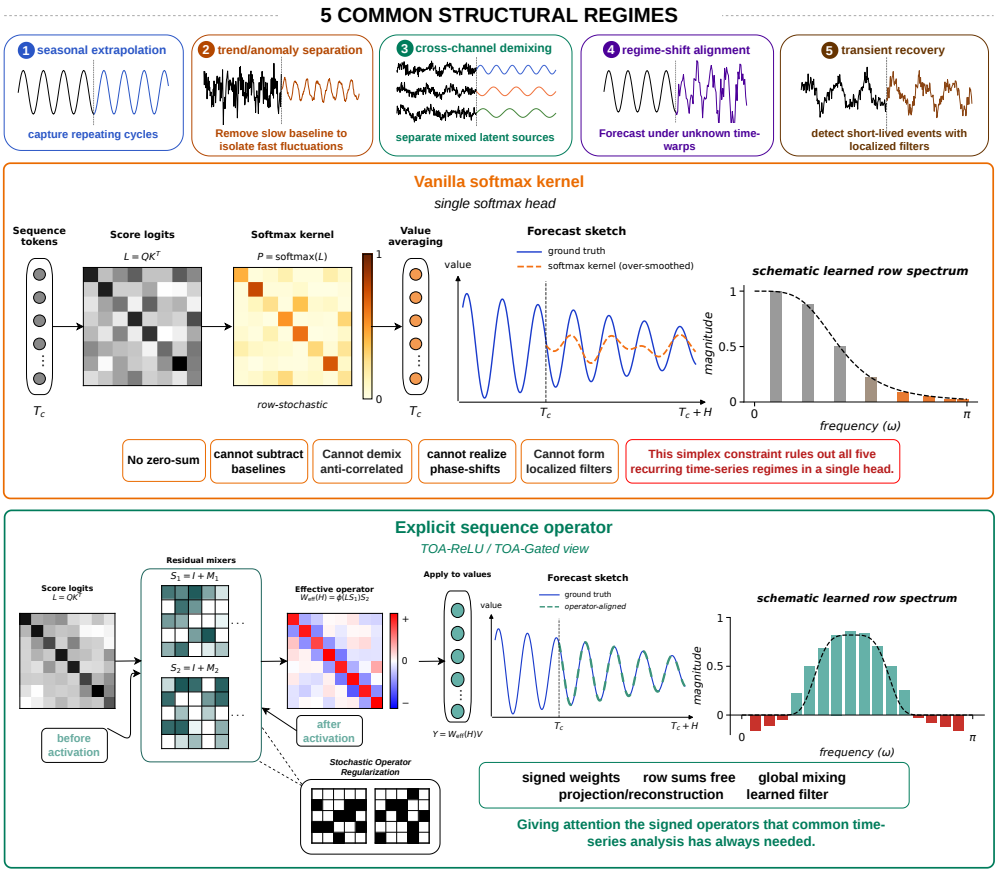

Figure 1: **Simplex-constrained mixing cannot express the operators these common regimes require. Top**: five recurring real-world regimes, namely seasonal extrapolation, trend/anomaly separation, cross-channel demixing, regime-shift alignment, and transient recovery, all need signed, non-simplex weights. **Middle**: softmax is row-stochastic by construction, so it can only average, never subtract, over-smoothing the forecast. **Bottom**: TOA wraps the score matrix in residual operators $\mathbf{S}_1 = \mathbf{I} + \mathbf{M}_1$, $\mathbf{S}_2 = \mathbf{I} + \mathbf{M}_2$, stabilized by Stochastic Operator Regularization, restoring the signed mixing time-series has always needed.

We trace this to self-attention, the core sequence-mixing primitive in the Transformer where it determines how information from other time steps is combined into each token's representation at every layer. As a sequence processor, attention itself cannot realize the five recurring regimes above. Softmax attention supplies adaptive, input-dependent routing, selecting where to look via query-key similarity, but its kernel is confined to the probability simplex, nonnegative weights summing to one. This simplex-constrained mixing limitation (Proposition 1) provably excludes all five regimes above from a single attention head (§3.3; Appendix D): none can be written as a convex combination of the sequence's own values. This does not preclude deep, multi-head Transformers, which can approximate signed operators indirectly through $\mathbf{W}_O$ and subsequent MLPs; but it identifies a primitive-level mismatch that finite data is statistically inefficient to learn around.

We propose **Temporal Operator Attention (TOA)** to let attention itself close this gap by augmenting the score-based attention pathway with explicit, learnable dense operators that add signed degrees of freedom, letting attention mix directly along the temporal dimension, indexed by position, while keeping its adaptive routing (§3.4). Because unrestricted operators risk overfitting, we pair them with **Stochastic Operator Regularization (SOR)**, regularizing the learned offsets while preserving identity-based signal transport (§3.5).

Because the fix targets the sequence-mixing primitive itself, it should transfer across architectures with different mixing mechanisms; we test this by integrating TOA into DUET (Qiu et al., 2025), Autoformer (Wu et al., 2021), PatchTST (Nie et al., 2023), and iTransformer (Liu et al., 2024)

across forecasting, anomaly detection, and classification. Letting attention model directly over time, while keeping its adaptive routing, yields consistent gains across these four architecturally distinct backbones, including decomposition-based, patch-based, inverted, and cluster-based designs.

Our contributions:

1. **Diagnosis:** We ground five common regimes in a shared requirement for signed, sometimes input-adaptive mixing over time and prove simplex-constrained attention excludes all five from a single head (Proposition 1, §3.2–§3.3).
2. **Methodology:** TOA combines adaptive score-based routing with explicit dense, signed temporal operators, realizing all five regimes in one mixing primitive.
3. **Regularization:** SOR curbs the overfitting risk of unrestricted operators by regularizing the learned offsets directly, keeping the added capacity practical.
4. **Experiments:** Across 40+ datasets, four tasks, and four backbones, TOA yields consistent gains, largest where signed mixing matters most (§4; §5).

2 Related Work

The five regimes introduced in §1 each require signed structure, and some further require it to adapt with the input. Classical solutions like Fourier projection, quadratic-penalty smoothing, independent component analysis, dynamic time warping, Gabor analysis realize this structure directly, without learning it from data; the neural sequence-mixing primitives reviewed below instead aim to learn a single mixing rule that covers all five at once. MLP-based models such as TSMixer (Chen et al., 2023), DLinear (Zeng et al., 2022), and TimeMixer (Wang et al., 2024a) parameterize free weight matrices over the sequence axis, indexed by position: this gives a direct inductive bias for signed temporal operators, but the same transformation applies regardless of input, unlike TOA's operators. Fast-weight approaches such as HyperMLP (Lu & Yang, 2026b) reintroduce input-adaptivity by generating weights from the input, but through a causal, low-rank recurrence built for autoregressive language modeling; such a structure cannot realize the acausal, arbitrary-rank operators these regimes require (§3.3), a limitation TOA does not share. Transformer-based forecasters (Zhou et al., 2021; Wu et al., 2021; Zhou et al., 2022; Li et al., 2019; Wu et al., 2023; Lu et al., 2026c), built on self-attention (Vaswani et al., 2017), take the opposite approach: query-key similarity gives input-adaptive weights, but softmax normalization confines every row to the probability simplex, while TOA keeps this routing but wraps it in operators the simplex cannot reach. TOA thus combines the input-adaptivity attention provides with the signed, arbitrary-rank operators these regimes require, realizing all five at once. Appendix B provides an extended discussion of the classical solutions and neural sequence-mixing approaches summarized above, together with additional comparisons to state-space and operator-based sequence models.

3 Methodology

3.1 Notation and Scope

Let $\mathbf{x} \in \mathbb{R}^L$ denote the scalar observed sequence and $\mathbf{X} \in \mathbb{R}^{L \times C}$ its multivariate counterpart with C channels. After any fixed tokenization, let $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_N]^\top \in \mathbb{R}^{N \times d}$ denote the token matrix. We consider three tokenization schemes: point ($N = L$), patch ($N = \lfloor \frac{L-p}{s} \rfloor + 1$ for patch length p and stride s, PatchTST style), and inverted ($N = C$, iTransformer style). We use operator for a linear map $\mathbb{R}^N \to \mathbb{R}^N$ that mixes value tokens, and write $\Delta^{N-1} = \{\mathbf{p} \in \mathbb{R}^N_{\geq 0} : \mathbf{1}^\top \mathbf{p} = 1\}$ for the probability simplex. Derivations and theoretical implications for patch and inverted tokenization are in Appendix F.

All formal limitation results are stated for a single-layer, single-head setting with scalar values $v_n = h_n^{(1)}$. This convention isolates the structural constraint at the level of the mixing primitive. The results do not assert impossibility for deep multi-head stacks, which can implement signed combinations through $\mathbf{W}_O$ and subsequent MLPs. The motivation for the dense primitive is therefore inductive bias in the finite-data regime, not pure expressivity.

3.2 The Simplex Constraint and Its Impossibility

For a single self-attention head with projections $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{d_h \times d}$ and scalar values $v_n = h_n^{(1)}$, the score matrix

$$\mathbf{A}(\mathbf{H}) = \mathbf{H}\mathbf{W}_Q^\top \mathbf{W}_K \mathbf{H}^\top \in \mathbb{R}^{N \times N}$$

is row-normalized via softmax:

$$a_{m,n}(\mathbf{H}) = \frac{\exp(\mathbf{A}_{m,n})}{\sum_{j=1}^{N} \exp(\mathbf{A}_{m,j})}, \qquad o_m(\mathbf{H}) = \sum_{n=1}^{N} a_{m,n}(\mathbf{H})\, v_n.$$

Every row of the kernel lies in the simplex: $a_{m,n} \geq 0$, $\sum_n a_{m,n} = 1$, unconditionally for all $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{H}$. This makes softmax attention an adaptive routing mechanism that selects where to look, but it cannot directly parameterize signed transformations such as the second-difference filter $[1, -2, 1, 0]$. The following result makes this precise: scaling the input toward zero forces softmax to collapse to the uniform distribution, which is incompatible with any non-simplex target.

Proposition 1 (Simplex-constrained mixing limitation)**.** *Consider a single-head self-attention layer with scalar values $v_n = h_n^{(1)}$. Let $\mathbf{T}^\star \in \mathbb{R}^{N \times N}$ satisfy $\mathbf{T}^\star_{m,:} \notin \Delta^{N-1}$ for some row m. Then for any choice of $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{d_h \times d}$, there is no parameter setting such that*

$$o_m(\mathbf{H}) = \mathbf{T}^\star_{m,:}\, \mathbf{v}(\mathbf{H}) \qquad \textit{for all } \mathbf{H} \in \mathbb{R}^{N \times d}, \tag{1}$$

where $\mathbf{v}(\mathbf{H}) = (h_1^{(1)}, \ldots, h_N^{(1)})^\top$.

Proof sketch. Suppose toward contradiction that the hypothesis holds. Substituting $\alpha\mathbf{H}$ and dividing by $\alpha > 0$, the values cancel while the softmax sees logits scaling as α^2. As $\alpha \to 0^+$ the kernel collapses to $\mathbf{1}^\top/N$, forcing $\mathbf{T}^\star_{m,:} = \mathbf{1}^\top/N \in \Delta^{N-1}$, a contradiction. Full proof in Appendix C.

Since each of the five common regimes in §3.3 requires a target operator that violates simplex membership in at least one row, Proposition 1 excludes all five from a single softmax attention head, providing the diagnosis that motivates TOA below.

3.3 Five Common Temporal Operator Families

The five common regimes motivated in §1 recur throughout time-series analysis: seasonal extrapolation, trend/anomaly separation, cross-channel demixing, regime-shift alignment, and transient recovery. Each corresponds, in this order, to Cases A–E below, whose exact solution can be expressed as a linear operator $\mathbf{T}^\star \in \mathbb{R}^{N \times N}$ over the sequence axis. In each case, the required operator structure follows from signal-processing or estimation theory rather than from a modeling choice.

Case A: Seasonal extrapolation. Extending a seasonal pattern requires weights that cancel the mean: sampled sinusoids are orthogonal to the DC component (Oppenheim et al., 1999), yielding zero-sum rows ($\mathbf{T}^\star\mathbf{1} = \mathbf{0}$), which violate the simplex's unit row sum (Lemma 1).

Case B: Trend/anomaly separation. Isolating rapid fluctuations from a slowly varying baseline requires negative off-diagonal entries and zero row sums (Oppenheim et al., 1999; Hodrick & Prescott, 1997) (Lemma 2).

Case C: Cross-channel demixing. Recovering sources from sensor mixtures requires the MMSE projection $\mathbf{P}_\mathbf{A}$ (Comon, 1994), which generically has non-unit row sums, non-binary diagonals, and negative entries (Lemma 3).

Case D: Regime-shift alignment. Phase- or warp-dependent alignment induces a family of input-dependent target operators, as in dynamic time warping (Sakoe & Chiba, 1978); in the warped harmonic setting considered here, these operators are generically non-simplex (Lemma 4).

Case E: Transient recovery. Recovering a short-lived quadrature event requires a rank-2 projection (Gabor, 1946) that vanishes under the $\mathbf{1}$-vector (Lemma 5).

Appendix D proves each case's non-realizability under softmax attention and shows how TOA's operators realize it instead, building on the single argument §3.2 already gave for all five at once; Appendix E confirms it synthetically, including learned mixing-matrix visualizations.

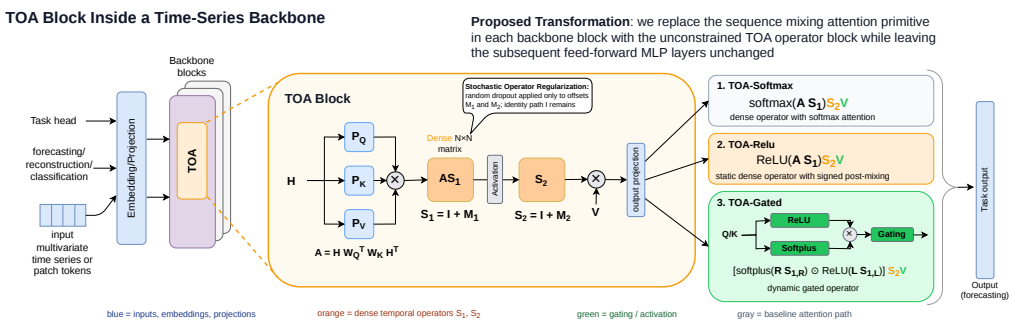

Figure 2: **The TOA Block.** TOA replaces the standard sequence-mixing layer while leaving the MLP layers unchanged. The internal block shows dense, unconstrained sequence operators $(\mathbf{S}_1, \mathbf{S}_2)$ that provide signed temporal mixing, stabilized via Stochastic Operator Regularization on the learned offsets.

3.4 Temporal Operator Attention (TOA)

TOA augments the attention pathway with explicit, learnable dense operators over the temporal axis, supplying the signed degrees of freedom that simplex normalization removes (Figure 2). Let $\mathbf{M}_1^{(h)}, \mathbf{M}_2^{(h)} \in \mathbb{R}^{N\times N}$ be learnable matrices for head h, and define the residual operators

$$\mathbf{S}_1^{(h)} = \mathbf{I}_N + \mathbf{M}_1^{(h)}, \qquad \mathbf{S}_2^{(h)} = \mathbf{I}_N + \mathbf{M}_2^{(h)},$$

with entries of $\mathbf{M}_i^{(h)}$ initialized i.i.d. from $\mathcal{N}(0, \sigma_M^2)$ at a small scale ($\sigma_M = 10^{-3}$). At initialization $\mathbf{S}_i^{(h)} \approx \mathbf{I}_N$, so the model begins as a near-standard attention layer and departs smoothly as gradients shape the offsets toward task-specific temporal filters. In practice, $\mathbf{M}_i^{(h)}$ is allocated at the maximum sequence length $N_{\max}$ and sliced to $\mathbf{M}_i^{(h)}[:N, :N]$ at runtime.

The full mixing operator is:

$$\mathbf{W}_{\text{mix}}^{\text{TOA}}(\mathbf{H}) = \sigma\big(\mathbf{H}\mathbf{W}_Q^\top \mathbf{W}_K \mathbf{H}^\top \mathbf{S}_1\big)\,\mathbf{S}_2,$$

with σ a non-saturating, non-normalizing activation. $\mathbf{S}_1$ reshapes each query's similarity profile before activation; $\mathbf{S}_2$ applies a signed linear map before value aggregation. Together they supply the signed degrees of freedom needed to leave the nonnegative orthant.

Because $\mathbf{S}_2 = \mathbf{I}_N + \mathbf{M}_2$ spans all of $\mathbb{R}^{N\times N}$, it can represent any single target operator from §3.3 directly (set $\mathbf{M}_2 = \mathbf{T}^\star - \mathbf{I}_N$); where the optimal operator itself varies with the input, realizing the full family additionally requires the input-dependent routing of TOA-Gated, described below.

Structural Variants and Activation Transfer. The choice between softmax, ReLU, and gated activations is inspired by HyperMLP, which showed that replacing softmax with ReLU or gating increases a sequence mixer's expressive capacity. HyperMLP realizes this through a causal, low-rank recurrence built for autoregressive language modeling, which cannot realize the acausal, arbitrary-rank operators of §3.3. We instead apply the activation choice to dense $\mathbf{S}_1, \mathbf{S}_2$, yielding three variants (TOA-Softmax, TOA-ReLU, TOA-Gated) that isolate its contribution. Let d_v denote the projected value dimension per head, and define the pre-activation score matrix $\mathbf{A}^{(h)} = \mathbf{H}\mathbf{W}_Q^{(h)\top}\mathbf{W}_K^{(h)}\mathbf{H}^\top \in \mathbb{R}^{N\times N}$.

TOA-Softmax: signed operators around a normalized kernel. Retain row-normalizing softmax, with $\mathbf{S}_1$ inside and $\mathbf{S}_2$ outside it:

$$\mathbf{O}^{(h)} = \text{softmax}\big(\mathbf{A}^{(h)}\,\mathbf{S}_1^{(h)}\big)\,\mathbf{S}_2^{(h)}\,\mathbf{H}\mathbf{W}_V^{(h)\top}.$$

Because $\mathbf{S}_2$ acts after the softmax, the overall map is signed even though the kernel is row-normalized. This isolates the effect of row normalization: TOA-Softmax has the same added operators as the other variants but keeps the softmax kernel.

TOA-ReLU: unconstrained signed mixing. Replace softmax with ReLU:

$$\mathbf{O}^{(h)} = \text{ReLU}\big(\mathbf{A}^{(h)}\,\mathbf{S}_1^{(h)}\big)\,\mathbf{S}_2^{(h)}\,\mathbf{H}\,\mathbf{W}_V^{(h)\top}.$$

ReLU removes the kernel's normalization while $\mathbf{S}_2$ still supplies the signs, testing whether dropping normalization adds to the operators' gains.

TOA-Gated: input-dependent adaptation. TOA-ReLU's operator pathway is static per head. TOA-Gated restores input adaptivity by doubling the QK head count to $2H$ and partitioning them into left and right groups. For head h, define

$$\mathbf{L}^{(h)} = \mathbf{H}\mathbf{W}_{Q,L}^{(h)\top}\mathbf{W}_{K,L}^{(h)}\mathbf{H}^{\top}, \qquad \mathbf{R}^{(h)} = \mathbf{H}\mathbf{W}_{Q,R}^{(h)\top}\mathbf{W}_{K,R}^{(h)}\mathbf{H}^{\top},$$

with independent QK projections per group and separate pre-activation operators $\mathbf{S}_1^{(h,L)}, \mathbf{S}_1^{(h,R)}$. The per-head output is:

$$\mathbf{O}^{(h)} = \Big(\mathrm{softplus}\big(\mathbf{R}^{(h)}\,\mathbf{S}_1^{(h,R)}\big) \odot \mathrm{ReLU}\big(\mathbf{L}^{(h)}\,\mathbf{S}_1^{(h,L)}\big)\Big)\,\mathbf{S}_2^{(h)}\,\mathbf{H}\,\mathbf{W}_V^{(h)\top},$$

The left branch provides a nonnegative similarity profile via ReLU; the right branch provides a strictly positive, sample-dependent gate via softplus. Their elementwise product forms a dynamically gated mixing kernel. This tests whether the operator itself benefits from input-dependent structural adaptation. All heads are concatenated and projected: $\mathbf{O} = \mathrm{Concat}(\mathbf{O}^{(1)}, \ldots, \mathbf{O}^{(H)})\mathbf{W}_O$, where $\mathbf{W}_O \in \mathbb{R}^{Hd_v \times d}$.

3.5 Stochastic Operator Regularization

Unrestricted operators provide the signed degrees of freedom that the simplex removes. But those same degrees of freedom create overfitting risk on finite time-series data. The residual decomposition $\mathbf{S}_i = \mathbf{I}_N + \mathbf{M}_i$ provides a natural solution: regularize the new capacity ($\mathbf{M}_i$) while preserving the baseline transport ($\mathbf{I}_N$).

At each forward pass, we sample a drop rate $p \sim U[0, 1)$ and an independent Bernoulli mask $\mathbf{B} \in \{0, 1\}^{N\times N}$ with $B_{jk} \overset{\text{iid}}{\sim} \mathrm{Bern}(1 - p)$, replacing the offset with its inverted-dropout image:

$$\tilde{\mathbf{M}}_i^{(h)} = \frac{1}{1-p}\big(\mathbf{M}_i^{(h)} \odot \mathbf{B}\big), \qquad \tilde{\mathbf{S}}_i^{(h)} = \mathbf{I}_N + \tilde{\mathbf{M}}_i^{(h)}.$$

The offset is preserved in expectation ($\mathbb{E}[\tilde{\mathbf{M}}_i^{(h)}] = \mathbf{M}_i^{(h)}$), but the per-entry variance $\frac{p}{1-p}(M_{i,jk}^{(h)})^2$ diverges as $p \to 1^-$, acting as an implicit regularizer that discourages over-reliance on the learned offsets. When the offset collapses ($p \to 1^-$), the operator reduces to $\tilde{\mathbf{S}}_i^{(h)} \to \mathbf{I}_N$, recovering identity transport. The model therefore cannot rely on $\mathbf{M}_i$ for stability and must stay robust to its removal.

4 Structural Limitation in Synthetic Tasks

The theory in §3 predicts that simplex-constrained mixing cannot represent the signed operators the five common regimes of §3.3 require. Before testing this on real benchmarks (§5), we isolate the mixing primitive directly: this section covers two regimes in the main text, with the full five visualizations and analysis in Appendix E.

We test two related signal variants per case from Appendix D: Case A (seasonal extrapolation, where a static signed operator suffices) and Case D (regime-shift alignment, which requires input-adaptive routing). Specifically, we use a magnitude-shift variant, for the learned query-key mixing matrix, and a signed-polarity variant, for the forecast comparison, which flips sign rather than merely rescaling and so more directly tests whether the operator can leave the nonnegative orthant. All models use PatchTST with $N = 11$ patches (patch length 16, stride 8, $d_{\text{model}} = 512$) under a severe 2-layer, single-head bottleneck (Figure 3).

Case A: Seasonal extrapolation. The signed-polarity variant is $x_t = \sum_{k=1}^{3} a_k(t)\sin(2\pi f_k t + \phi_k) + \varepsilon_t$ ($f_k \in \{1/80, 1/35, 1/20\}$, $\phi_k \sim U[0, 2\pi)$, $\varepsilon_t \sim \mathcal{N}(0, 0.01^2)$), where two of the three amplitudes flip sign at the sequence's 70% boundary, creating a discontinuity that no nonnegative combination of past values can reproduce. On the magnitude-shift variant's mixing-matrix task, softmax collapses onto a narrow nonnegative column band, consistent with Lemma 1's zero-sum requirement being unreachable. On the signed-polarity variant's forecasting task, softmax loses phase and amplitude outright (MSE 0.606); TOA-ReLU and TOA-Gated (0.021, 0.022) track ground truth.

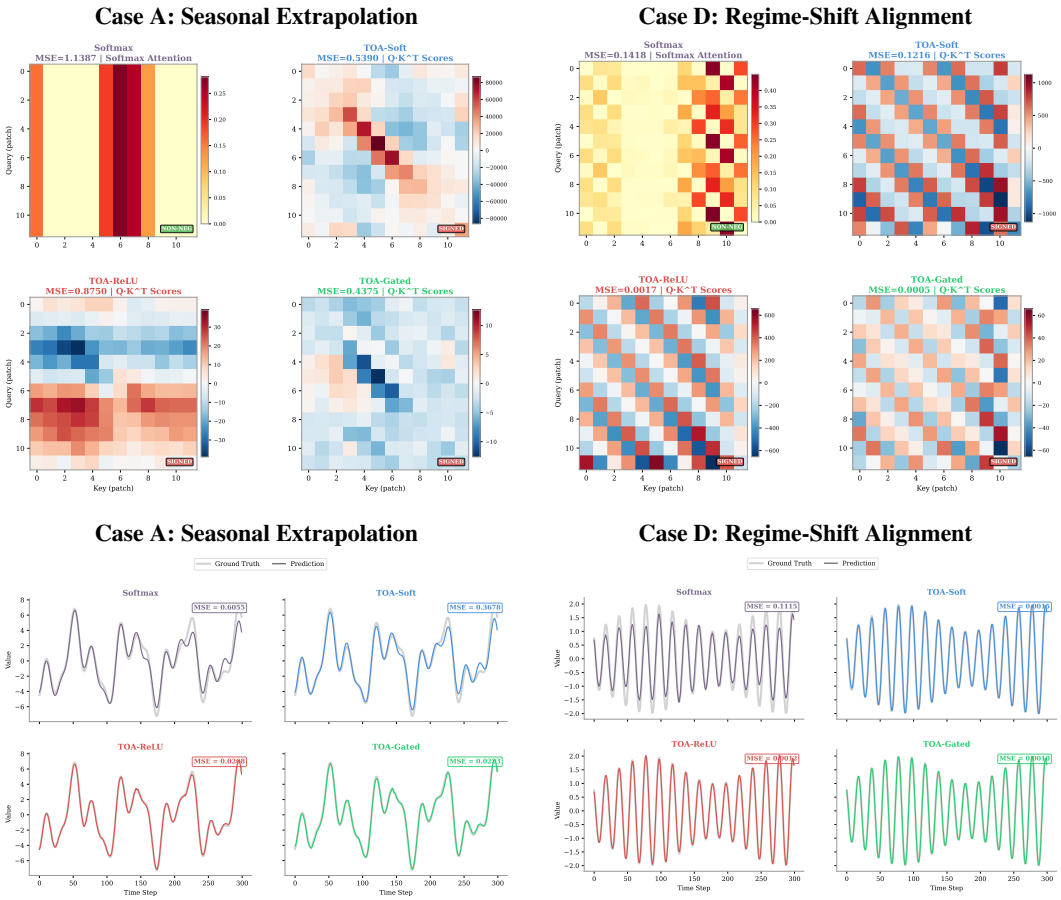

Figure 3: **Visualization for Cases A and D.** Top: learned query-key mixing matrices. Softmax stays nonnegative (NON-NEG), while every TOA variant learns signed weights (SIGNED). Bottom: the resulting forecasts against ground truth; softmax's forecast MSE is $\sim$27$\times$ (Case A) and $\sim$112$\times$ (Case D) higher than TOA-Gated's.

Case D: Regime-shift alignment. The signed-polarity variant is $x_t = a(t)\sin(\theta_t) + \varepsilon_t$ with $\theta_t = \theta_{t-1} + 2\pi f(t)$, $f(t)$ shifting from $1/50$ to $1/20$ at the sequence midpoint, $a(t) = 1.5 + 0.5\sin(2\pi t/200)$, and $\varepsilon_t \sim \mathcal{N}(0, 0.01^2)$. On the magnitude-shift variant's mixing-matrix task, softmax collapses onto a fixed column band, unable to track the phase-dependent target family (Lemma 4). On the signed-polarity variant's forecasting task, softmax drifts out of phase entirely (MSE 0.112), while TOA-Softmax, TOA-ReLU, and TOA-Gated track the ground truth (0.0015, 0.0012, 0.0010).

The same pattern holds for the remaining three regimes (Appendix E), and carries over to real benchmarks next (§5).

5 Improvements in Empirical Experiments

We integrate TOA into PatchTST, iTransformer, DUET, and Autoformer across four tasks, testing whether the advantage predicted by the common structural regime diagnosis of §3.3 survives on real data.

5.1 Improvements in Long- and Short-Term Forecasting Tasks

We integrate TOA into DUET, PatchTST, iTransformer, and Autoformer for long-term forecasting across nine datasets, and into PatchTST, iTransformer, and Autoformer for short-term forecasting on the PEMS traffic benchmarks. Detailed setup is in Appendix H.

Tables 1 and 2 show that adding the signed $\mathbf{S}_1, \mathbf{S}_2$ operators to attention yields gains across backbones and horizons, with the best variant depending on the backbone. Since TOA-Softmax differs from the baseline only by the added operators, its gains already trace to the signed-operator mechanism.

The most important observation is cross-backbone consistency: some TOA variant improves over softmax on every one of these four architecturally distinct backbones, including decomposition-based Autoformer, patch-based PatchTST, inverted iTransformer, and cluster-based DUET, even though which variant wins differs by backbone. This indicates the gains stem from the mixing primitive itself, not from backbone-specific interactions, and is the strongest evidence that the simplex constraint is the limitation. Full tables are in Appendix G. Multi-seed robustness for a representative subset of these datasets is in Appendix I.

Table 1: **PEMS Results** with forecasting horizons $L_p \in \{12, 24, 48, 96\}$. Average test MSE. Best in **bold** (r), second underlined (b). See Table 8 for full results.

	PatchTST Backbone				iTransformer Backbone				Autoformer Backbone			
Model	**Softmax**	**Gated**	**ReLU**	**TOA-S**	**Softmax**	**Gated**	**ReLU**	**TOA-S**	**Softmax**	**Gated**	**ReLU**	**TOA-S**
PEMS03	0.091	**0.087**	0.090	**0.087**	0.150	**0.135**	**0.135**	0.150	0.596	**0.582**	0.583	0.588
PEMS04	0.094	**0.090**	**0.090**	**0.090**	0.115	**0.106**	**0.106**	0.109	**0.757**	0.759	0.771	0.759
PEMS07	0.067	0.069	**0.064**	**0.064**	0.158	0.093	**0.092**	0.138	0.525	**0.520**	0.558	0.551
PEMS08	0.118	**0.106**	0.108	0.107	0.182	0.156	**0.154**	0.183	0.852	0.830	**0.810**	0.834
AvgRank	3.75	1.75	2.00	**1.25**	3.50	1.50	**1.00**	3.25	2.75	**1.50**	2.75	2.75
#Top1	0	**3**	2	**3**	0	2	**4**	0	1	**2**	1	0

Table 2: Summary of Multivariate TSF Results with forecasting horizons $L_p \in \{96, 192, 336, 720\}$. Averaged test set MSE are reported. Best results are in **bold** (r) and second best are underlined (b) (compared within each backbone group). See Tables 6 and 7 for full results.

Model	DUET				PatchTST				iTransformer				Autoformer			
	Softmax-Attention	TOA-Gated	TOA-ReLU	TOA-Softmax	Softmax-Attention	TOA-Gated	TOA-ReLU	TOA-Softmax	Softmax-Attention	TOA-Gated	TOA-ReLU	TOA-Softmax	Softmax-Attention	TOA-Gated	TOA-ReLU	TOA-Softmax
Weather	0.250	0.243	**0.239**	0.247	0.230	0.230	0.230	**0.229**	0.259	0.255	0.254	**0.252**	0.296	0.300	0.287	**0.276**
Solar	0.231	0.231	**0.217**	0.230	0.198	0.200	0.190	**0.187**	**0.234**	0.242	0.240	0.238	0.716	**0.701**	0.763	0.722
ECL	0.177	**0.170**	**0.170**	0.176	0.163	0.162	0.163	**0.160**	0.180	0.176	0.175	**0.167**	0.456	0.453	0.455	**0.436**
ETT	0.399	0.400	0.396	**0.393**	0.345	0.345	0.345	**0.342**	0.385	0.380	**0.379**	0.382	0.429	0.431	0.436	**0.422**
Exchange	0.266	0.264	**0.258**	0.262	0.673	0.567	**0.548**	0.669	0.363	**0.360**	0.361	0.362	0.676	0.686	0.681	**0.675**
Traffic	0.466	0.464	0.470	**0.461**	0.400	**0.397**	0.400	0.399	0.422	0.423	0.422	**0.420**	1.070	1.078	1.073	**1.062**
AvgRank	3.50	2.50	**1.67**	2.00	2.83	2.17	2.17	**1.50**	3.17	2.83	2.00	**1.83**	2.50	3.00	3.17	**1.33**
#Top1	0	1	**4**	2	0	1	1	**4**	1	1	1	**3**	0	1	0	**5**

5.2 Improvements in Classification Tasks

Classification is the most generic task in our evaluation: unlike forecasting, there is no explicit temporal reconstruction target. If signed operators help here, the advantage is task-general. We evaluate on 28 UEA multivariate datasets (setup in Appendix H).

Table 3: **UEA Results** on 28 datasets for PatchTST and Autoformer. Acc. denotes mean accuracy, Rank denotes mean rank within each backbone, and W/T/L is measured against Softmax. See Table 9 for full results.

Backbone	**Model**	**Acc.**	**Avg Rank↓**	**W/T/L**
PatchTST	Softmax (Base)	65.62%	3.11	—
	TOA-S (Abl.)	64.80%	3.36	9/2/17
	TOA-ReLU	69.13%	**1.54**	**26/1/1**
	TOA-Gated	**69.16%**	1.64	24/1/3
Autoformer	Softmax (Base)	55.67%	2.79	—
	TOA-S (Abl.)	56.36%	2.50	16/1/11
	TOA-ReLU	57.52%	**2.27**	**17/1/10**
	TOA-Gated	**57.91%**	2.45	15/2/11

TOA-Gated and TOA-ReLU substantially outperform softmax on both backbones (Table 3), with a more one-sided win rate on PatchTST than on Autoformer, though positive on both, which indicates a systematic structural advantage, not dataset-specific tuning. The TOA-Softmax ablation directly supports the variant design from §3.4: it trails both backbones' full TOA variants by a wide margin, even underperforming baseline on PatchTST, showing that added operator capacity ($\mathbf{S}_1, \mathbf{S}_2$) alone is insufficient; the larger gains instead come from also removing row normalization. Full per-dataset results are in Appendix G (Table 9).

5.3 Improvements in Anomaly Detection Tasks

Anomaly detection provides a direct test of this diagnosis. Detection depends on the reconstruction error between observed and predicted normal dynamics: a mixing layer that can represent sharper temporal operators should produce tighter reconstructions of normal behavior, making true anomalies stand out more clearly. We evaluate on five standard benchmarks: PSM, MSL, SMAP, SMD, and SWAT (setup in Appendix H).

Table 4 shows consistent F1-score improvements when TOA replaces the standard mixing layer across all three backbones, with the largest gains on SWAT (PatchTST) and SMAP (Autoformer) where both are driven by the operator's ability to model the signed, multi-scale structure that distinguishes normal from anomalous windows, precisely the trend/anomaly separation regime of Case B (§3.3), which a mixing layer confined to convex combinations smooths away.

Table 4: Summary of Anomaly Detection Results. F1-scores are reported. Best results are in **bold** (r) and second best are underlined (b) (compared within each backbone group).

Dataset	iTransformer				PatchTST				Autoformer			
	Softmax	TOA-Gated	TOA-ReLU	TOA-Softmax	Softmax	TOA-Gated	TOA-ReLU	TOA-Softmax	Softmax	TOA-Gated	TOA-ReLU	TOA-Softmax
PSM	0.948	**0.949**	**0.949**	0.948	0.962	0.962	**0.968**	0.962	0.882	0.883	**0.900**	**0.900**
MSL	0.725	0.726	0.725	**0.744**	0.788	**0.794**	0.790	0.788	0.820	**0.834**	0.827	0.820
SMAP	0.668	**0.670**	0.669	0.668	0.689	0.688	**0.690**	0.688	0.740	**0.913**	0.836	0.776
SMD	**0.824**	**0.824**	**0.824**	**0.824**	0.845	0.846	0.846	**0.848**	**0.712**	0.711	0.711	0.711
SWAT	**0.927**	**0.927**	**0.927**	0.926	0.853	**0.897**	0.860	0.853	**0.792**	**0.792**	**0.792**	**0.792**
AvgRank	2.20	**1.20**	1.60	2.40	2.80	1.80	**1.60**	2.40	2.60	**1.60**	**1.60**	2.00
#Top1	2	**4**	3	2	0	**2**	**2**	1	2	**3**	2	2

6 Ablation Study

Unrestricted operators provide the signed degrees of freedom that the simplex removes (§3.4), but risk overfitting on finite data. SOR tests whether TOA's gains come from this structural capacity or from memorizing training-set patterns: if SOR were mere noise injection, degradation would be uniform; instead, we expect the largest gains where the operator has the most room to overfit.

Table 5: **SOR Ablation.** Test MSE on the PatchTST backbone averaged across $L_p \in \{96, 192, 336, 720\}$. Lower MSE is in **bold** (r) for each architecture variant. See full table in Table 10.

	TOA-ReLU		TOA-Gated	
Dataset	**w/o SOR**	**+ SOR**	**w/o SOR**	**+ SOR**
Weather	0.239	**0.228**	0.238	**0.230**
ETTm2	0.273	**0.261**	0.271	**0.259**
ETTm1	**0.353**	0.355	**0.352**	0.355
ETTh2	0.381	**0.354**	0.377	**0.360**
ETTh1	**0.421**	**0.421**	0.423	**0.421**

Table 5 supports this: SOR yields substantial gains on complex, high-frequency datasets, where the dense operator is especially prone to memorizing complex correlations. For example, ETTh2 drops from 0.381 to 0.354 with TOA-ReLU, while ETTm2 drops from 0.271 to 0.259 with TOA-Gated. The small regression on ETTm1 fits another pattern: its smooth dynamics are already well-served by the identity-dominated operator, so SOR occasionally interferes rather than helps.

Upon more ablation studies, Appendix K tests whether the dense operators SOR regularizes can be factorized at low rank without losing accuracy on the datasets tested, shrinking the footprint from $O(N^2)$ to $O(Nr)$. Yet only the dense family provably realizes every recurring regime's target operator exactly (§3.4), so we report it as default, at a training cost comparable to Transformer-based forecasters (Appendix J).

7 Conclusion

Time-series analysis relies on common regimes (seasonal extrapolation, trend/anomaly separation, cross-channel demixing, regime-shift alignment, and transient recovery) that require signed, position-indexed transformations, not merely adaptive ones. Standard attention supplies adaptive, content-based routing confined to a probability simplex; explicit mixers supply the missing signed structure but sacrifice adaptivity. We proved the simplex constraint excludes all five regimes from a single attention head, providing a primitive-level explanation for why signed linear and MLP mixers can be advantageous in forecasting settings dominated by such structure. TOA resolves this with dense signed operators, parameterized as residual offsets from the identity and stabilized by SOR: synthetic error falls by up to two orders of magnitude once the target operator must change sign, and the gains hold across four backbones and 40+ datasets spanning forecasting, anomaly detection, and classification, indicating that the gain comes from the operator rather than the backbone.

AI USE STATEMENT

We used large language models (GPT, Claude) for LaTeX formatting assistance, prose editing, and debugging experimental scripts. All theoretical results, algorithm design, and experimental protocols were developed by the authors. LLM-generated text was reviewed and rewritten where necessary. LLM-generated code was tested and verified for correctness by at least two authors before inclusion. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies the mathematical structure of sequence-mixing primitives and proposes a general-purpose architectural modification. The societal impact is indirect and application-dependent. As with any advance in time-series modeling, improved predictive performance could benefit domains such as healthcare monitoring, climate science, and infrastructure planning, but could also be applied in sensitive contexts. We view TOA as a methodological contribution whose real-world impact depends on the transparency and oversight of downstream systems.

REPRODUCIBILITY STATEMENT

All experiments are fully reproducible. The complete codebase, including training scripts, evaluation pipelines, and hyperparameter configurations, is available at the anonymous repository linked in the abstract. Section 3 provides full architectural specifications; Appendix H details all hyperparameters, dataset splits, and training protocols. Appendices C, D, and F contain the complete proofs and theoretical extensions. Synthetic data generation code is included in the repository.

REFERENCES

Zeyuan Allen-Zhu. Physics of language models: Part 4.1, architecture design and the magic of canon layers, 2025. URL https://arxiv.org/abs/2512.17351.

Si-An Chen, Chun-Liang Li, Nate Yoder, Sercan O. Arik, and Tomas Pfister. Tsmixer: An all-mlp architecture for time series forecasting, 2023. URL https://arxiv.org/abs/2303.06053.

Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994. doi: 10.1016/0165-1684(94)90029-9.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth, 2023. URL https://arxiv.org/abs/2103.03404.

Dennis Gabor. Theory of communication. part 1: The analysis of information. *Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering*, 93(26): 429–457, 1946.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024. URL https://arxiv.org/abs/2312.00752.

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Re. Hippo: Recurrent memory with optimal polynomial projections, 2020. URL https://arxiv.org/abs/2008.07669.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces, 2022. URL https://arxiv.org/abs/2111.00396.

Ankit Gupta, Albert Gu, and Jonathan Berant. Diagonal state spaces are as effective as structured state spaces, 2022. URL https://arxiv.org/abs/2203.14343.

Robert J Hodrick and Edward C Prescott. Postwar u.s. business cycles: an empirical investigation. *Journal of Money, Credit and Banking*, 29(1):1–16, 1997.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks, 2018. URL https://arxiv.org/abs/1703.07015.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018a. ISBN 978-1-57735-800-8.

Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Advances in neural information processing systems*, 32, 2019.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting, 2018b.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024.

Jiecheng Lu and Shihao Yang. Free energy mixer, 2026a. URL https://arxiv.org/abs/2602.07160.

Jiecheng Lu and Shihao Yang. Hypermlp: An integrated perspective for sequence modeling, 2026b. URL https://arxiv.org/abs/2602.12601.

Jiecheng Lu and Shihao Yang. Linear transformers as var models: Aligning autoregressive attention mechanisms with autoregressive forecasting, 2026c. URL https://arxiv.org/abs/2502.07244.

Jiecheng Lu, Xu Han, Yan Sun, and Shihao Yang. WAVE: Weighted autoregressive varying gate for time series forecasting. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Qqn5ktBUxH.

Jiecheng Lu, Xu Han, Yan Sun, Viresh Pati, Yubin Kim, Siddhartha Somani, and Shihao Yang. Zeros: Zero-sum linear attention for efficient transformers, 2026a. URL https://arxiv.org/abs/2602.05230.

Jiecheng Lu, Xu Han, Yan Sun, and Shihao Yang. Cats: Enhancing multivariate time series forecasting by constructing auxiliary time series as exogenous variables, 2026b. URL https://arxiv.org/abs/2403.01673.

Jiecheng Lu, Xu Han, and Shihao Yang. Arm: Refining multivariate forecasting with adaptive temporal-contextual learning, 2026c. URL https://arxiv.org/abs/2310.09488.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers, 2023. URL https://arxiv.org/abs/2211.14730.

Alan V. Oppenheim, Ronald W. Schafer, and John R. Buck. *Discrete-time signal processing (2nd ed.)*. Prentice-Hall, Inc., USA, 1999. ISBN 0137549202.

Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. Duet: Dual clustering enhanced multivariate time series forecasting, 2025. URL https://arxiv.org/abs/2412.10859.

Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting, 2024a. URL https://arxiv.org/abs/2405.14616.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Mingsheng Long, and Jianmin Wang. Deep time series models: A comprehensive survey and benchmark. 2024b.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22419–22430, 2021.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training, 2024. URL https://arxiv.org/abs/2312.06635.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule, 2025. URL https://arxiv.org/abs/2412.06464.

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer, 2025. URL https://arxiv.org/abs/2410.05258.

Wenzhen Yue, Yong Liu, Haoxuan Li, Hao Wang, Xianghua Ying, Ruohao Guo, Bowei Xing, and Ji Shi. Olinear: A linear model for time series forecasting in orthogonally transformed domain, 2025. URL https://arxiv.org/abs/2505.08550.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting?, 2022. URL https://arxiv.org/abs/2205.13504.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting, 2022. URL https://arxiv.org/abs/2201.12740.

A LIMITATIONS

A.1 LIMITATIONS

While Temporal Operator Attention (TOA) resolves the simplex-constrained bottleneck at the primitive level, it introduces an $O(N^2)$ memory and compute footprint through its dense sequence operators, which limits the maximum sequence length that can be handled directly without token compression or patching (Computation Efficiency Analysis can be found for more details in Appendix J). Our low-rank ablation (Appendix K) shows that rank-16 and rank-32 factorizations match dense performance on the datasets tested, but only the dense parameterization is theoretically guaranteed to realize arbitrary instances of the five regimes exactly; whether structured operators preserve this guarantee at scale remains future work. In practice, dense signed operators are also more prone to overfitting than standard softmax attention, particularly on smaller datasets, making stable training dependent on effective regularization through our Stochastic Operator Regularization (SOR) scheme. Finally, our formal impossibility results apply specifically to the single-layer, single-head softmax primitive; characterizing the approximation efficiency and sample-complexity tradeoffs of deep, multi-head Transformer stacks relative to TOA remains an important open problem.

B EXTENDED RELATED WORKS

This section expands the comparison in §2 with further detail and additional related work.

B.1 SOLUTIONS TO THE FIVE REGIMES.

The five regimes in §3.3 have established solutions under their respective signal models: seasonal extrapolation through harmonic projection and continuation using sinusoidal least-squares regression or Fourier analysis (Oppenheim et al., 1999); trend/anomaly separation through smoothing and residualization, as in the Hodrick–Prescott filter (Hodrick & Prescott, 1997); cross-channel demixing through latent-subspace projection or, under source-independence assumptions, ICA (Comon, 1994); regime-shift alignment through warped harmonic fitting or temporal alignment methods such as DTW (Sakoe & Chiba, 1978); and transient recovery through localized time-frequency projection and reconstruction, as in Gabor/short-time Fourier analysis (Gabor, 1946). Each solution already has exactly the structure these regimes require: signed and, in the case of DTW, input-adaptive. However, each is computed directly rather than learned from data. This is precisely what motivates the sequence-mixing primitives reviewed below: rather than five regime-specific closed-form operators, they aim to learn a single mixing rule general enough to cover all five at once, at the risk of not being expressive enough to represent any of them.

B.2 MLP-BASED SEQUENCE MIXING.

TSMixer (Chen et al., 2023), DLinear (Zeng et al., 2022), and TimeMixer (Wang et al., 2024a) replace attention with explicit sequence-space mixing, achieving competitive or superior forecasting performance with signed weights indexed by position, although the transformation is fixed after training, regardless of input. CATS (Lu et al., 2026b) extends this line by learning auxiliary exogenous series as a preprocessing operator over the temporal dimension, achieving strong performance even with a simple MLP backbone. TOA retains this explicit, position-indexed operator structure while reintroducing the input-adaptivity these models lack.

B.3 DYNAMIC PARAMETERIZATION AND FAST-WEIGHT PERSPECTIVES.

Self-attention can be viewed as dynamic weight generation, where input-dependent similarity scores define a sequence-mixing operator, a perspective shared with fast-weight programming and formulations such as HyperMLP (Lu & Yang, 2026b) and Canon-style layers (Allen-Zhu, 2025). A broader line of work increases attention's expressivity beyond softmax: Gated Linear Attention and Gated DeltaNet add input-dependent gates to fast-weight recurrences (Yang et al., 2024; 2025); Differential Transformer differences attention maps to induce signed interactions (Ye et al., 2025); and ZeroS (Lu et al., 2026a) and Free Energy Mixer (Lu & Yang, 2026a) introduce a zero-sum reparameterization and a value-aware free-energy posterior, respectively. These approaches remain causal, low-rank,

or recurrent by construction. TOA instead introduces shared, dense, position-indexed operators that model global temporal structure directly, supplying signed mixing capacity without tying it to any particular rank or causal structure, thereby enabling precise signal cancellation in bidirectional time-series settings.

B.4 State-space and operator-based sequence models.

Recent architectures replace explicit attention with structured dynamical operators derived from state-space systems, including S4 (Gu et al., 2022), DSS (Gupta et al., 2022), and Mamba (Gu & Dao, 2024), achieving strong long-range performance through continuous-time dynamics, convolutional kernels, or input-dependent recurrence. These share TOA's broader goal of replacing similarity-based token aggregation with richer parameterized operators, but impose strong structural priors for efficient long-context propagation, including HiPPO-based dynamics (Gu et al., 2020), convolutional recurrence, and causal selective scans. TOA instead wraps content-based attention in signed, position-indexed operators, preserving input-adaptive routing while removing the simplex constraint.

A complementary line of work shows that linear attention admits an explicit autoregressive interpretation: single-layer linear attention can be reformulated as a dynamic VAR system, and deeper linear Transformers aligned with autoregressive forecasting objectives through reparameterization (Lu et al., 2025; Lu & Yang, 2026c). Unlike TOA, these approaches still operate within a constrained linear-attention family, preserving similarity-driven aggregation rather than introducing signed temporal operators.

B.5 Transformer variants for time series.

Transformer variants such as Informer (Zhou et al., 2021), Autoformer (Wu et al., 2021), and FEDformer (Zhou et al., 2022) improve efficiency or incorporate frequency-domain structure, while LogTrans (Li et al., 2019) and ARM (Lu et al., 2026c) bias attention toward local or band-limited interactions. None alter the simplex-confined kernel itself, so attention remains a row-stochastic averaging operator, a limitation theoretically supported by rank collapse in deep pure-attention stacks (Dong et al., 2023) and analogous over-smoothing in graph neural networks (Li et al., 2018a). TOA instead targets the mixing primitive directly: by adding signed, position-indexed operators over the temporal axis, it enables the global, signed interactions standard attention cannot reach, addressing a different part of the design space than efficiency or locality.

C Full Proofs for Impossibility and Realization

This section provides the complete mathematical proofs for the structural limits of standard softmax attention and the expressive capacity of the dense residual operators, expanding on the proof sketches provided in Section 3.

C.1 Proof of Proposition 1 (The Simplex-Constrained Mixing Limitation)

Proposition 1. *Consider a single-head self-attention layer with scalar values $v_n = h_n^{(1)}$. Let $\mathbf{T}^\star \in \mathbb{R}^{N\times N}$ satisfy $\mathbf{T}^\star_{m,:} \notin \Delta^{N-1}$ for some row m. Then there exist no parameters $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{d_h\times d}$ such that $o_m(\mathbf{H}) = \mathbf{T}^\star_{m,:}\,\mathbf{v}(\mathbf{H})$ for all $\mathbf{H} \in \mathbb{R}^{N\times d}$, where $\mathbf{v}(\mathbf{H}) = (h_1^{(1)}, \dots, h_N^{(1)})^\top$.*

Proof. Suppose toward contradiction that the hypothesis holds for all $\mathbf{H} \in \mathbb{R}^{N\times d}$.

Fix $\mathbf{H} \in \mathbb{R}^{N\times d}$ and consider the scaling family $\{\alpha\mathbf{H} : \alpha > 0\}$. The value vector scales linearly, $\mathbf{v}(\alpha\mathbf{H}) = \alpha\,\mathbf{v}(\mathbf{H})$, and the output scales as

$$o_m(\alpha\mathbf{H}) \;=\; \sum_{n=1}^{N} a_{m,n}(\alpha\mathbf{H})\,\alpha v_n \;=\; \alpha \sum_{n=1}^{N} a_{m,n}(\alpha\mathbf{H})\,v_n, \tag{2}$$

where $v_n = h_n^{(1)}$. Applying the hypothesis at input $\alpha\mathbf{H}$ and dividing both sides by $\alpha > 0$ yields

$$\sum_{n=1}^{N} a_{m,n}(\alpha\mathbf{H})\, v_n \;=\; \mathbf{T}^{\star}_{m,:}\,\mathbf{v}(\mathbf{H}) \qquad \forall\, \alpha > 0, \tag{3}$$

with $\mathbf{v} = \mathbf{v}(\mathbf{H})$ on both sides. Now the (m, n)-entry of the score matrix at input $\alpha\mathbf{H}$ satisfies

$$\mathbf{A}_{m,n}(\alpha\mathbf{H}) \;=\; \alpha^2\,(\mathbf{W}_Q\mathbf{h}_m)^{\top}\mathbf{W}_K\mathbf{h}_n \;=\; \alpha^2\,\mathbf{A}_{m,n}(\mathbf{H}), \tag{4}$$

so every logit scales as α^2. Since softmax is continuous and $\mathrm{softmax}(\mathbf{0})_n = 1/N$, we have $a_{m,n}(\alpha\mathbf{H}) \to 1/N$ as $\alpha \to 0^+$ for every n. Taking the limit of equation 3 gives

$$\tfrac{1}{N}\mathbf{1}^{\top}\,\mathbf{v}(\mathbf{H}) \;=\; \mathbf{T}^{\star}_{m,:}\,\mathbf{v}(\mathbf{H}). \tag{5}$$

Equation 5 must hold for every $\mathbf{H} \in \mathbb{R}^{N\times d}$. The map $\mathbf{H} \mapsto \mathbf{v}(\mathbf{H}) = (h_1^{(1)}, \ldots, h_N^{(1)})^{\top}$ is surjective onto $\mathbb{R}^N$ (its components are N independent free coordinates of $\mathbf{H}$), so the linear identity $\frac{1}{N}\mathbf{1}^{\top}\mathbf{v} = \mathbf{T}^{\star}_{m,:}\,\mathbf{v}$ holds for every $\mathbf{v} \in \mathbb{R}^N$. By uniqueness of representation of linear functionals, $\mathbf{T}^{\star}_{m,:} = \frac{1}{N}\mathbf{1}^{\top}$. But $\frac{1}{N}\mathbf{1}^{\top} \in \Delta^{N-1}$, contradicting the premise that $\mathbf{T}^{\star}_{m,:} \notin \Delta^{N-1}$. □

D Five Common Temporal Operator Families

This section provides the complete mathematical formulations and proofs for the five recurring time-series regimes discussed in Section 3.3, demonstrating why they cannot be realized by standard softmax attention and how the TOA framework parameterizes them. Each Realization proposition is stated at the level of the post-mixer family $\{\mathbf{S}_2 = \mathbf{I}_N + \mathbf{M}_2\}$, which, as a simple affine shift, spans all of $\mathbb{R}^{N\times N}$.

D.1 Case A: Seasonal Extrapolation

Setup. Let $\mathbf{t} = (0, 1, \ldots, L-1)^{\top}$. For K candidate periods $\{P_k\}$ with angular frequencies $\omega_k = 2\pi/P_k$, define the quadrature feature $\mathbf{\Phi}_k = [\cos(\omega_k\mathbf{t}), \sin(\omega_k\mathbf{t})] \in \mathbb{R}^{L\times 2}$ and the full matrix $\mathbf{\Phi} = [\mathbf{\Phi}_1 \mid \cdots \mid \mathbf{\Phi}_K] \in \mathbb{R}^{L\times 2K}$. The latent signal is $y_t = \sum_{k=1}^{K}\big(a_k\cos(\omega_k t) + b_k\sin(\omega_k t)\big)$, observed as $x_t = y_t + \varepsilon_t$ where $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$. Evaluating $\mathbf{\Phi}$ at indices $t = 0, \ldots, L-1$ yields $\mathbf{\Phi}_{\mathrm{obs}}$, and at indices $t = L, \ldots, L+T-1$ yields $\mathbf{\Phi}_{\mathrm{fcast}}$. In the high-SNR limit $\sigma \to 0$, the exact MMSE forecasting operator fits Fourier coefficients by least-squares: $\mathbf{T}^{\star} = \mathbf{\Phi}_{\mathrm{fcast}}\,\mathbf{\Phi}_{\mathrm{obs}}^{+} \in \mathbb{R}^{T\times L}$, where $\mathbf{\Phi}_{\mathrm{obs}}^{+} = (\mathbf{\Phi}_{\mathrm{obs}}^{\top}\mathbf{\Phi}_{\mathrm{obs}})^{-1}\mathbf{\Phi}_{\mathrm{obs}}^{\top}$. Proposition 1 extends verbatim from self-attention to the cross-attention setting $T \neq L$, since the small-scale scaling argument depends only on the row-stochastic property of softmax.

Lemma 1 (Failure of Softmax Attention for Seasonal Extrapolation)**.** *Suppose (a) L is an integer multiple of P_k for every k, (b) all frequencies are non-trivial ($\omega_k \neq 0 \bmod 2\pi$), and (c) $L \geq 2K$ such that $\mathbf{\Phi}_{\mathrm{obs}}$ has full column rank. Then every row of $\mathbf{T}^{\star}$ has strictly zero sum: $\mathbf{T}^{\star}\mathbf{1} = \mathbf{0}$. Consequently, no parameter setting for single-head softmax attention (in its cross-attention extension) can realize the seasonal extrapolation operator.*

Proof. Each column of $\mathbf{\Phi}_{\mathrm{obs}}$ represents a sampled sinusoid over an integer number of periods. The discrete sum over the sequence is computed via the geometric series: $\sum_{t=0}^{L-1} e^{i\omega_k t} = \frac{1-e^{i\omega_k L}}{1-e^{i\omega_k}} = 0$, as $e^{i\omega_k L} = 1$ and $e^{i\omega_k} \neq 1$. Thus, every column of $\mathbf{\Phi}_{\mathrm{obs}}$ sums to zero: $\mathbf{1}^{\top}\mathbf{\Phi}_{\mathrm{obs}} = \mathbf{0}^{\top}$. Right-multiplying by the components of the Moore–Penrose inverse gives $\mathbf{\Phi}_{\mathrm{obs}}^{+}\mathbf{1} = (\mathbf{\Phi}_{\mathrm{obs}}^{\top}\mathbf{\Phi}_{\mathrm{obs}})^{-1}(\mathbf{\Phi}_{\mathrm{obs}}^{\top}\mathbf{1}) = \mathbf{0}$. It follows that $\mathbf{T}^{\star}\mathbf{1} = \mathbf{\Phi}_{\mathrm{fcast}}(\mathbf{\Phi}_{\mathrm{obs}}^{+}\mathbf{1}) = \mathbf{0}$. Since every row of $\mathbf{T}^{\star}$ sums to zero, every row fails simplex membership ($\sum_n T^{\star}_{m,n} = 0 \neq 1$). By Proposition 1, softmax attention cannot realize this operator. □

Proposition 2 (Realization by TOA)**.** *The post-mixer family $\{\mathbf{S}_2 = \mathbf{I}_N + \mathbf{M}_2 : \mathbf{M}_2 \in \mathbb{R}^{N\times N}\}$ contains the seasonal extrapolation operator $\mathbf{T}^{\star}$ exactly. In particular, the parameter choice $\mathbf{M}_2^{\star} = \mathbf{T}^{\star} - \mathbf{I}_N$ yields $\mathbf{S}_2 = \mathbf{T}^{\star}$.*

Proof. Because the map $\mathbf{M}_2 \mapsto \mathbf{I}_N + \mathbf{M}_2$ is a trivial affine bijection on $\mathbb{R}^{N\times N}$, the post-mixer can represent any arbitrary matrix. Setting the learnable weights to $\mathbf{M}_2^\star = \mathbf{T}^\star - \mathbf{I}_N$ immediately yields $\mathbf{S}_2 = \mathbf{T}^\star$. □

Remark 1. *Proposition 2 establishes that $\mathbf{T}^\star$ lies in the span of the post-mixer alone. In the full TOA-ReLU pipeline $\mathrm{ReLU}(\mathbf{A}\mathbf{S}_1)\mathbf{S}_2\mathbf{H}\mathbf{W}_V^\top$ the left factor is constrained to be entry-wise nonnegative, so realizing $\mathbf{T}^\star$ requires the left factor to be driven to a positive multiple of the identity (or another invertible nonnegative pattern) so that $\mathbf{S}_2$ supplies the signed degrees of freedom.*

D.2 Case B: Trend/Anomaly Separation

Setup. Anomaly detection models typically identify deviations by scoring the residual sequence $\mathbf{r} = (\mathbf{I}_N - \mathbf{T}^\star_{\mathrm{normal}})\mathbf{x}$, where $\mathbf{T}^\star_{\mathrm{normal}}$ represents a smoothing or reconstruction operator for "normal" temporal patterns. We analyze the high-pass anomaly separation operator $\mathbf{T}^\star = \mathbf{I}_N - \mathbf{T}^\star_{\mathrm{normal}}$, which generalizes first-order differences, moving-average residuals, and Laplacian-style filtering.

Lemma 2 (Failure of Softmax Attention for Trend/Anomaly Separation)**.** *Let $\mathbf{T}^\star = \mathbf{I}_N - \mathbf{G}$ be a trend/anomaly separation operator where $\mathbf{G} \in \mathbb{R}^{N\times N}$ is any valid smoothing kernel satisfying $\mathbf{G}\mathbf{1} = \mathbf{1}$ and $\mathbf{G} \geq \mathbf{0}$. Then:*

(i) Every row of $\mathbf{T}^\star$ has zero sum: $\mathbf{T}^\star\mathbf{1} = \mathbf{0}$.

(ii) For every row t with $\mathbf{G}_{t,:} \neq \mathbf{e}_t^\top$, the row $\mathbf{T}^\star_{t,:}$ contains at least one strictly negative entry.

Consequently, if $\mathbf{G} \neq \mathbf{I}_N$, at least one row $\mathbf{T}^\star_{t,:} \notin \Delta^{N-1}$, and by Proposition 1 single-head softmax attention cannot realize $\mathbf{T}^\star$.

Proof. (i) Using $\mathbf{G}\mathbf{1} = \mathbf{1}$: $\mathbf{T}^\star\mathbf{1} = (\mathbf{I}_N - \mathbf{G})\mathbf{1} = \mathbf{1} - \mathbf{1} = \mathbf{0}$.

(ii) For any row t with $\mathbf{G}_{t,:} \neq \mathbf{e}_t^\top$, the constraint $\sum_j G_{t,j} = 1$ together with $G_{t,t} \neq 1$ and $G_{t,j} \geq 0$ for all j forces some off-diagonal entry $G_{t,j_0} > 0$ for $j_0 \neq t$. Then $T^\star_{t,j_0} = -G_{t,j_0} < 0$.

If $\mathbf{G} \neq \mathbf{I}_N$, at least one row t_0 satisfies $\mathbf{G}_{t_0,:} \neq \mathbf{e}_{t_0}^\top$. By (i)–(ii), the row $\mathbf{T}^\star_{t_0,:}$ has zero sum and a strictly negative entry, so $\mathbf{T}^\star_{t_0,:} \notin \Delta^{N-1}$. Proposition 1 then forbids realization by single-head softmax attention. □

Proposition 3 (Realization by TOA)**.** *The post-mixer family $\{\mathbf{S}_2 = \mathbf{I}_N + \mathbf{M}_2 : \mathbf{M}_2 \in \mathbb{R}^{N\times N}\}$ contains every trend/anomaly separation operator $\mathbf{T}^\star = \mathbf{I}_N - \mathbf{G}$ exactly. The parameter choice $\mathbf{M}_2^\star = -\mathbf{G}$ yields $\mathbf{S}_2 = \mathbf{T}^\star$.*

Proof. Because the affine map $\mathbf{M}_2 \mapsto \mathbf{I}_N + \mathbf{M}_2$ spans all of $\mathbb{R}^{N\times N}$, we can directly set $\mathbf{M}_2^\star = -\mathbf{G}$, yielding $\mathbf{S}_2 = \mathbf{I}_N - \mathbf{G} = \mathbf{T}^\star$. □

Remark 2. *The post-mixer alone supplies the unconstrained negative off-diagonal weights and zero row sums required to compute residuals; in the full pipeline, the left ReLU factor must be driven to a nonnegative pattern under which $\mathbf{S}_2$ acts cleanly.*

D.3 Case C: Cross-Channel Demixing

Setup. Let $\mathbf{z}_t \in \mathbb{R}^r$ be a latent factor vector ($r < C$) and $\mathbf{A} \in \mathbb{R}^{C\times r}$ a feature mixing matrix such that $\mathbf{x}_t = \mathbf{A}\mathbf{z}_t + \boldsymbol{\varepsilon}_t$. Under iTransformer-style tokenization, the sequence dimension is the channels ($N = C$). The MMSE clean-signal estimator is the orthogonal projection matrix $\mathbf{P}_\mathbf{A} = \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top$.

Lemma 3 (Failure of Softmax Attention for Cross-Channel Demixing)**.** *For $\mathbf{A}$ outside a Lebesgue-measure-zero subset of $\mathbb{R}^{C\times r}$ (i.e., generically), each row $c \in \{1,\dots,C\}$ of $\mathbf{P}_\mathbf{A}$ fails simplex membership: (i) the diagonal entry $(\mathbf{P}_\mathbf{A})_{c,c} \in (0,1)$; (ii) some off-diagonal entry in each row is strictly negative on a Lebesgue-positive open subset of parameterizations; (iii) the row sum $(\mathbf{P}_\mathbf{A}\mathbf{1})_c \neq 1$. Consequently, by Proposition 1, softmax attention cannot realize $\mathbf{P}_\mathbf{A}$.*

Proof. Concretely: for two anti-correlated sensors ($C = 2, r = 1, \mathbf{A} = (1, -1)^\top$), $\mathbf{P_A} = \frac{1}{2}\left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right)$ has a negative off-diagonal entry and each row sums to 0, not 1. This is generic, not a special case:

(i) $\mathbf{P_A}$ is idempotent and symmetric, so $(\mathbf{P_A})_{c,c} = \|\mathbf{P_A}\mathbf{e}_c\|_2^2 \in [0, 1]$, with equality only if $\mathbf{e}_c \in \mathrm{col}(\mathbf{A})$, which is a codimension-≥ 1 condition on $\mathbf{A}$ and hence has measure zero.

(ii) The off-diagonal entries are rational functions of $\mathbf{A}$; since the example above gives a strictly negative value, continuity propagates this sign to an open neighborhood in $\mathbb{R}^{C\times r}$.

(iii) The row sum equals $(\mathbf{P_A}\mathbf{1})_c$, which equals 1 only if $\mathbf{1} \in \mathrm{col}(\mathbf{A})$, again a codimension-$(C - r)$ condition of measure zero.

Outside this measure-zero set, every row of $\mathbf{P_A}$ has a non-unit row sum, a diagonal strictly between 0 and 1, and a negative entry, so $(\mathbf{P_A})_{c,:} \notin \Delta^{C-1}$ for every c, and Proposition 1 forbids realization by single-head softmax attention. □

Proposition 4 (Realization by TOA)**.** *The post-mixer family contains the orthogonal projection $\mathbf{T}^\star = \mathbf{P_A}$ exactly. The parameter choice $\mathbf{M}_2^\star = \mathbf{P_A} - \mathbf{I}_C$ yields $\mathbf{S}_2 = \mathbf{T}^\star$.*

Proof. Because the affine map $\mathbf{M}_2 \mapsto \mathbf{I}_C + \mathbf{M}_2$ spans all of $\mathbb{R}^{C\times C}$, we can directly set $\mathbf{M}_2^\star = \mathbf{P_A} - \mathbf{I}_C$, yielding $\mathbf{S}_2 = \mathbf{P_A} = \mathbf{T}^\star$. □

Remark 3. *Under iTransformer-style tokenization ($N = C$), the post-mixer parameterizes channel-wise projections directly, including the asymmetric and signed entries required by $\mathbf{P_A}$ for generic $\mathbf{A}$.*

D.4 Case D: Regime-Shift Alignment

Setup. Let $\phi \sim \mathrm{Uniform}[0, 2\pi)$ be a per-sample phase drawn i.i.d., and let $\tau : \mathbb{Z} \to \mathbb{R}$ be a smooth monotone time-warping function (e.g., periodic vibrato or quadratic chirp). The observed signal is $x_t = \sum_{k=1}^{K} a_k \cos(\omega_k \tau(t) + \phi) + \varepsilon_t$. The target $y_{t'}$ is the clean continuation under the same phase and warping regime. In this non-stationary setting, the MMSE predictor is a phase-dependent operator $\mathbf{T}^\star(\phi)$, where $\mathbf{T}^\star(\phi)$ fits the warped quadrature basis $\mathbf{\Phi}(\phi)$ with columns $\cos(\omega_k\tau(t) + \phi)$ and $\sin(\omega_k\tau(t) + \phi)$.

Lemma 4 (Failure of Softmax Attention for Regime-Shift Alignment)**.** *Let $\{\mathbf{T}^\star(\phi)\}_{\phi\in[0,2\pi)}$ be the family of MMSE predictors under per-sample phase ϕ and a fixed smooth monotone warping τ. Assume that for some fixed ϕ_0, the operator $\mathbf{T}^\star(\phi_0)$ has at least one row outside Δ^{N-1}. Then no parameter setting for single-head softmax attention can realize the family $\{\mathbf{T}^\star(\phi)\}_\phi$.*

Proof. By Proposition 1, no choice of $\mathbf{W}_Q, \mathbf{W}_K$ can realize the single operator $\mathbf{T}^\star(\phi_0)$ as the input-output map of softmax attention. A fortiori, no fixed parameter setting can realize the parameterized family $\{\mathbf{T}^\star(\phi)\}_\phi$, since realizing the family requires in particular realizing the member $\mathbf{T}^\star(\phi_0)$.

The hypothesis that some $\mathbf{T}^\star(\phi_0)$ has a row outside Δ^{N-1} is generic for the warped quadrature setting: each $\mathbf{T}^\star(\phi)$ is a least-squares projector onto the span of the warped basis $\{\cos(\omega_k\tau(t) + \phi), \sin(\omega_k\tau(t) + \phi)\}_k$, and the row-sum and sign analysis of Lemma 1 extends to this basis whenever the warped sample set is approximately balanced over the unit circle, a condition satisfied for non-degenerate τ and sufficiently long L. □

Proposition 5 (Realization by TOA-Gated)**.** *For each fixed phase $\phi \in [0, 2\pi)$, some head's post-mixer $\mathbf{S}_2^{(h)}$ can equal $\mathbf{T}^\star(\phi)$ exactly, as in Proposition 4. Across the family $\{\mathbf{T}^\star(\phi)\}_\phi$, different heads specialize to different phases, and the input-dependent gate selects which head dominates for a given sample.*

Proof. Fix ϕ: setting $\mathbf{M}_2^\star(\phi) = \mathbf{T}^\star(\phi) - \mathbf{I}_N$ for one head's post-mixer gives $\mathbf{S}_2 = \mathbf{T}^\star(\phi)$ for that head, by the same affine-span argument as Proposition 4. Because $\mathbf{M}_2$ is a fixed, learned parameter, no single head can realize every member of $\{\mathbf{T}^\star(\phi)\}_\phi$ at once. TOA-Gated resolves this across heads instead: different heads' post-mixers specialize to different phases, and the gate

$$\mathrm{softplus}(\mathbf{R}\mathbf{S}_1^R) \odot \mathrm{ReLU}(\mathbf{L}\mathbf{S}_1^L),$$

which depends on the input through $\mathbf{L}, \mathbf{R}$, up-weights the head whose post-mixer best matches the sample's phase before the heads are combined by $\mathbf{W}_O$. The family is thus realized by input-dependent selection among per-head static operators, not by a single operator varying continuously with ϕ. □

Remark 4. *Unlike Cases A–C, the static post-mixer alone cannot realize the entire family because the input dependence of the operator is supplied by the gating product.*

D.5 Case E: Transient Recovery

Setup. Let local phase atoms be defined as $\psi_{r,\phi,c}(t) = w_r(t-c)\cos(\omega_r(t-c)+\phi)$, where w_r is a strictly positive, compactly supported window of scale r, c is the event center, and $\phi \in [0, 2\pi)$ is the local phase. The observed signal is $x_t = \sum_j \alpha_j \psi_{r_j,\phi_j,c_j}(t) + \varepsilon_t$. This models transient phenomena such as ECG QRS-complexes or mechanical vibration transients. For recovering a single atom at known (r, c) with unknown phase ϕ, the MMSE denoiser is the rank-2 projection $\mathbf{T}^\star_{\phi,c}$ onto the 2D quadrature subspace at (r, c): the sum of the individual rank-1 projectors onto each quadrature component[1],

$$\mathbf{T}^\star_{\phi,c} = \frac{\boldsymbol{\psi}_{\phi,c}\,\boldsymbol{\psi}_{\phi,c}^\top}{\|\boldsymbol{\psi}_{\phi,c}\|^2} + \frac{\boldsymbol{\psi}_{\phi+\pi/2,c}\,\boldsymbol{\psi}_{\phi+\pi/2,c}^\top}{\|\boldsymbol{\psi}_{\phi+\pi/2,c}\|^2}. \tag{6}$$

Lemma 5 (Failure of Softmax Attention for Transient Recovery)**.** *Suppose the local atom $\boldsymbol{\psi}_{\phi,c} \in \mathbb{R}^N$ contains entries of both signs and is zero-mean ($\mathbf{1}^\top \boldsymbol{\psi}_{\phi,c} = 0$); the same is assumed for the quadrature companion $\boldsymbol{\psi}_{\phi+\pi/2,c}$. Then for every (ϕ, c), single-head softmax attention cannot realize the optimal rank-2 quadrature projector $\mathbf{T}^\star_{\phi,c}$.*

Proof. Write $\mathbf{T}^\star_{\phi,c} = \boldsymbol{\psi}\boldsymbol{\psi}^\top/\|\boldsymbol{\psi}\|^2 + \boldsymbol{\psi}'\boldsymbol{\psi}'^\top/\|\boldsymbol{\psi}'\|^2$ where $\boldsymbol{\psi} = \boldsymbol{\psi}_{\phi,c}$ and $\boldsymbol{\psi}' = \boldsymbol{\psi}_{\phi+\pi/2,c}$.

Zero row sum. For each row t, $(\boldsymbol{\psi}\boldsymbol{\psi}^\top)_{t,:}\mathbf{1} = \psi(t)(\boldsymbol{\psi}^\top\mathbf{1}) = 0$ by the zero-mean assumption, and similarly for $\boldsymbol{\psi}'$. Hence $\mathbf{T}^\star_{\phi,c}\mathbf{1} = \mathbf{0}$.

Sign violation. Pick any row t with $\psi(t) \neq 0$ or $\psi'(t) \neq 0$. The row of $\mathbf{T}^\star_{\phi,c}$ at t equals $\psi(t)\boldsymbol{\psi}^\top/\|\boldsymbol{\psi}\|^2 + \psi'(t)\boldsymbol{\psi}'^\top/\|\boldsymbol{\psi}'\|^2$, a linear combination of two mixed-sign vectors. Generically this row contains at least one strictly negative entry.

Combined, these give $\mathbf{T}^\star_{\phi,c}\mathbf{1} = \mathbf{0} \neq \mathbf{1}$ and at least one strictly negative entry per non-zero row. Therefore some row of $\mathbf{T}^\star_{\phi,c}$ lies outside Δ^{N-1}, and by Proposition 1, single-head softmax attention cannot realize $\mathbf{T}^\star_{\phi,c}$. □

Proposition 6 (Realization by TOA-Gated)**.** *The TOA-Gated forward pass realizes the location-and phase-adaptive operator family $\{\mathbf{T}^\star_{\phi,c}\}_{\phi,c}$ in the same sense as Proposition 5: for each fixed (ϕ, c), the post-mixer family contains $\mathbf{T}^\star_{\phi,c}$, and the input-dependent gating product supplies the sample-dependent routing required to select among them.*

Proof. The argument parallels Proposition 5 with the additional dependence on the temporal center c. The gating product $\mathrm{softplus}(\mathbf{R}\mathbf{S}_1^R) \odot \mathrm{ReLU}(\mathbf{L}\mathbf{S}_1^L)$ acts as a localized feature detector over (t, c) pairs: the QK structure of $\mathbf{L}, \mathbf{R}$ encodes which (scale, center) regime the input occupies, and the post-mixer $\mathbf{S}_2$ supplies the signed quadrature projection within each regime. □

Remark 5. *The combined location-and-phase adaptivity is the principal mechanism by which TOA-Gated outperforms TOA-ReLU on transient-rich signals.*

E Synthetic Operator Recovery Across All Five Cases

§4 reports Cases A and D in the main text for space; this section gives the full synthetic results for all five cases of Appendix D, evaluated on the query-key mixing-matrix recovery task. For each case, we generate a signal that instantiates the case's failure mode via an abrupt magnitude shift partway

[1] Assuming the quadrature pair is orthogonal; for non-orthogonal atoms, $\mathbf{T}^\star$ involves the inverse of the local Gram matrix.

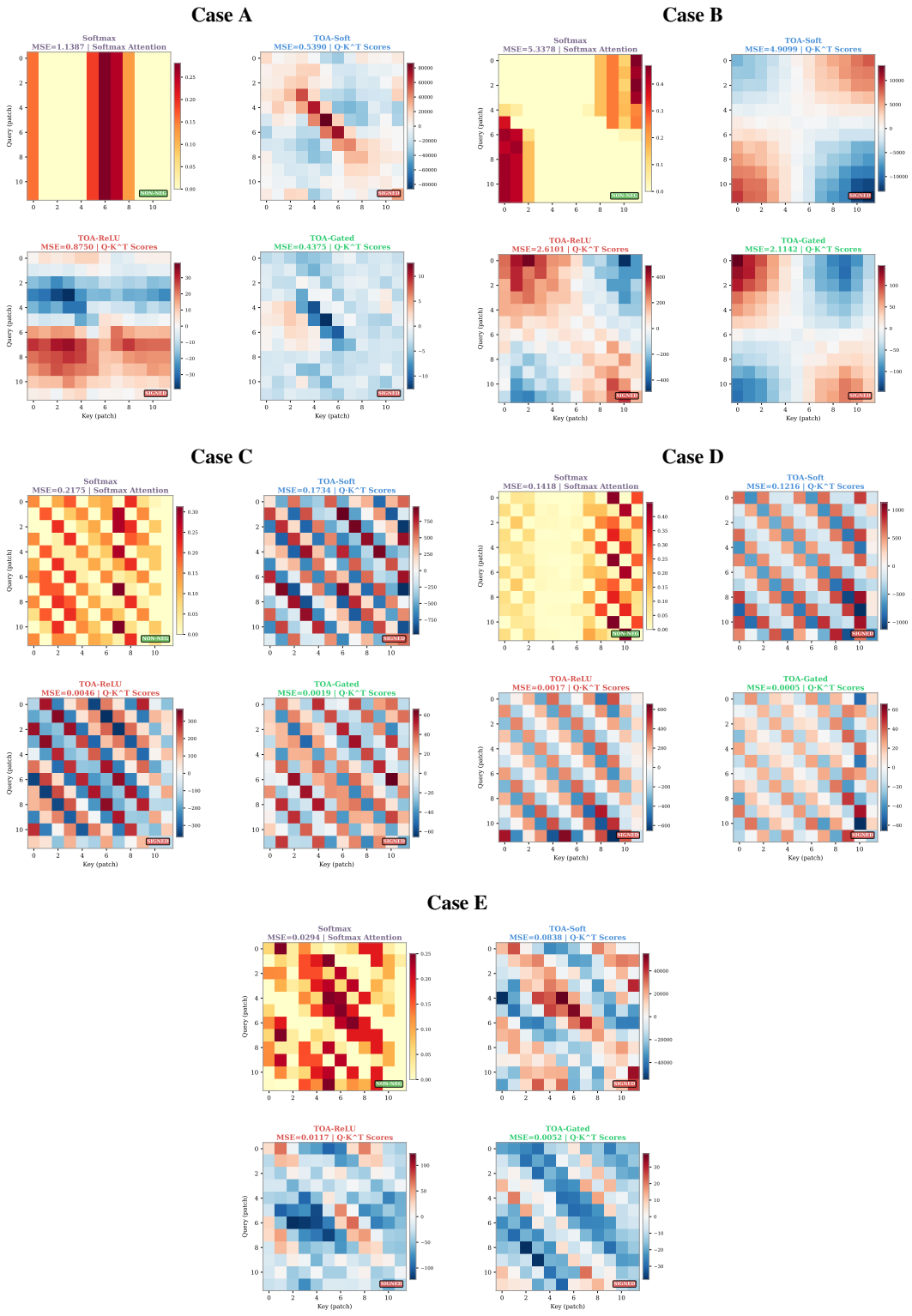

Figure 4: **Learned query-key mixing matrices for all five cases.** Softmax (NON-NEG) is visibly confined to a nonnegative pattern in every case; all three TOA variants (SIGNED) are not.

through the sequence, train four mixing-primitive variants (Softmax, TOA-Softmax, TOA-ReLU, and TOA-Gated) and visualize each variant's learned query-key matrix directly. All models use PatchTST with $N = 11$ patches (patch length 16, stride 8, $d_{\text{model}} = 512$) under a 2-layer, single-head limitation. Visualizations for all cases are provided above in Figure 4.

Case A: Seasonal extrapolation. $x_t = \sum_{k=1}^{3} a_k(t)\sin(2\pi f_k t + \phi_k) + \varepsilon_t$, $f_k \in \{1/47, 1/19, 1/11\}$, amplitudes jumping from $(1.0, 0.7, 0.4)$ to $(3.0, 0.2, 1.8)$ at the sequence midpoint, $\varepsilon_t \sim \mathcal{N}(0, 0.02^2)$. Softmax's kernel collapses into a narrow nonnegative column band, effectively forming a low-pass filter that cannot express the zero-sum structure Lemma 1 requires (MSE 1.139). TOA-Gated learns a signed kernel that tracks the amplitude shift (MSE 0.438).

Case B: Trend/anomaly separation. $x_t = \text{slow}(t) + a(t)\,\text{fast}(t) + \varepsilon_t$, $\text{slow}(t) = 2.0\sin(2\pi t/200) + 0.8\cos(2\pi t/130)$, $\text{fast}(t) = \sin(2\pi t/8) + 0.5\cos(2\pi t/11)$, with the residual amplitude $a(t)$ jumping sixfold, $0.3 \to 1.8$, at the midpoint ($\varepsilon_t \sim \mathcal{N}(0, 0.02^2)$). Softmax's kernel is confined to a blockwise nonnegative pattern that cannot subtract the slow component, consistent with Lemma 2's negative-entry requirement (MSE 5.338). TOA-Gated learns signed weights that isolate the residual (MSE 2.114).

Case C: Cross-channel demixing. $x_t = a_1(t)s_1(t) + a_2(t)s_2(t) + \varepsilon_t$, $s_1(t) = \sin(2\pi t/30)$, $s_2(t) = \sin(2\pi t/13 + 1.0)$, with a smooth dominance swap $a_1 : 2.0 \to 0.3$, $a_2 : 0.3 \to 2.0$ across the midpoint ($\varepsilon_t \sim \mathcal{N}(0, 0.02^2)$). Softmax's kernel forms a nonnegative checkerboard, unable to realize the negative off-diagonal entries the MMSE demixing projection $\mathbf{P_A}$ requires (Lemma 3, MSE 0.218). TOA-Gated's signed kernel separates the two sources cleanly (MSE 0.0019), a $> 100\times$ reduction.

Case D: Regime-shift alignment. $x_t = a(t)\big[\sin(\theta_t) + 0.3\sin(2\theta_t)\big] + \varepsilon_t$, $\theta_t = \theta_{t-1} + 2\pi f(t)$ with $f(t)$ tripling from $1/30$ to $1/10$ and $a(t)$ doubling from 1.0 to 2.0 at the midpoint ($\varepsilon_t \sim \mathcal{N}(0, 0.02^2)$). Softmax's kernel is a fixed column band that cannot adapt to the frequency jump, consistent with Lemma 4 (MSE 0.142). TOA-Gated's kernel shifts weight to track the new frequency (MSE 0.0005).

Case E: Transient recovery. $x_t = a(t)\,\text{env}(t)\cos(2\pi t/8) + \varepsilon_t$, $\text{env}(t) = \sin(2\pi t/200) + 0.4\sin(2\pi t/70)$, with the envelope amplitude $a(t)$ jumping from 1.0 to 2.5 at the midpoint ($\varepsilon_t \sim \mathcal{N}(0, 0.02^2)$). Softmax's kernel forms a diagonally localized nonnegative band, unable to realize the signed, zero-sum quadrature projection Lemma 5 requires (MSE 0.029). Uniquely among the five cases, TOA-Softmax (MSE 0.084) is worse than plain softmax: the added operators do not help isolate a signed envelope while the kernel stays row-normalized. Only once normalization is removed does error fall sharply: TOA-ReLU (0.012) and TOA-Gated (0.005) both learn signed, localized kernels that track the envelope polarity.

F Theoretical Implications for Advanced Tokenization Schemes

The foundational limits established in Proposition 1 and their resolution via Temporal Operator Attention (TOA) apply to generic sequence mixers $\mathbb{R}^N \to \mathbb{R}^N$. In modern time-series architectures, the semantic meaning of the sequence length N depends heavily on the chosen tokenization scheme. In this section, we explicitly map our theoretical findings to the two dominant advanced tokenization paradigms: patch tokenization (e.g., PatchTST) and inverted tokenization (e.g., iTransformer).

We demonstrate that the structural restrictions of softmax attention severely constrain these state-of-the-art backbones, and prove that integrating TOA mathematically unlocks their intended representational capacities.

F.1 Patch Tokenization

Setup. In patch tokenization, the raw scalar sequence $\mathbf{x} \in \mathbb{R}^L$ is converted into patches of length p with stride s. Without padding, the sequence dimension becomes $N = \left\lfloor \frac{L-p}{s} \right\rfloor + 1$, and each token $\mathbf{h}_n \in \mathbb{R}^d$ encodes the local dynamics of the n-th patch. At this macroscopic level, optimal sequence mixing often requires modeling autoregressive transitions, such as discrete derivatives to capture trend changes ($\mathbf{o}_n = \mathbf{h}_n - \mathbf{h}_{n-1}$) or seasonal differencing to isolate cyclic anomalies ($\mathbf{o}_n = \mathbf{h}_n - \mathbf{h}_{n-S}$), where S is the macroscopic seasonal lag). Let $\mathbf{T}^\star \in \mathbb{R}^{N\times N}$ denote a generalized patch-differencing operator.

Lemma 6 (Failure of Softmax Attention for Inter-Patch Differencing)**.** *Let $\mathbf{T}^\star$ be a patch-differencing operator such that for some target patch m and reference patch k, the optimal operation requires subtraction: $\mathbf{T}^\star_{m,m} > 0$ and $\mathbf{T}^\star_{m,k} < 0$. Then $\mathbf{T}^\star_{m,:} \notin \Delta^{N-1}$. Consequently, standard softmax attention over patches cannot compute discrete derivatives or seasonal differences between patch embeddings.*

Proof. By definition, the simplex constraint Δ^{N-1} strictly requires all kernel weights $a_{m,n} \geq 0$. An operation requiring $\mathbf{T}^\star_{m,k} < 0$ forces the target operator outside the non-negative orthant. When standard attention attempts to relate patch m to reference patches, it computes a convex combination $\mathbf{o}_m = \sum_{i=1}^{N} a_{m,i}\mathbf{h}_i$ subject to $\sum_i a_{m,i} = 1$. Because $a_{m,i} \geq 0$, the model is structurally forced to smooth (average) the patches rather than differentiate them, mathematically annihilating inter-patch rate-of-change information. By Proposition 1, softmax attention cannot realize $\mathbf{T}^\star$. □

Proposition 7 (Realization by TOA Patch-Mixing)**.** *When TOA is applied over patched tokens, the post-activation unconstrained operator $\mathbf{S}_2 \in \mathbb{R}^{N\times N}$ provides the signed degrees of freedom required to compute exact discrete derivatives across patches.*

Proof. Unlike standard attention, the Temporal Operator Attention (TOA) module parameterizes the sequence mixing operation via an unconstrained weight matrix $\mathbf{M}_2$, establishing the operator $\mathbf{S}_2 = \mathbf{I}_N + \mathbf{M}_2$. Because this formulation does not enforce a probability simplex Δ^{N-1}, the affine transformation $\mathbf{S}_2$ spans the entirety of $\mathbb{R}^{N\times N}$. Consequently, the residual mixer can be directly optimized such that $(\mathbf{S}_2)_{m,m} = 1$, $(\mathbf{S}_2)_{m,k} = -1$, and zeros elsewhere. This perfectly recovers the inter-patch differencing operator $\mathbf{o}_m = \mathbf{h}_m - \mathbf{h}_k$, moving beyond simplex-constrained mixing and enabling exact macroscopic trend and seasonality analysis. □

F.2 Inverted Tokenization

Setup. In inverted tokenization (iTransformer style), the tokenization axis is flipped. Given a multivariate series $\mathbf{X} \in \mathbb{R}^{L\times C}$, the tokens are formed over the full temporal history of each variate, yielding a sequence length of $N = C$. Each token $\mathbf{h}_n \in \mathbb{R}^d$ represents the entire timeline of variable n. In this regime, the sequence mixer acts as a cross-channel interaction operator. If the observed variables are entangled linear combinations of true independent physical mechanisms (e.g., Independent Component Analysis), the optimal operator $\mathbf{T}^\star \in \mathbb{R}^{N\times N}$ is a channel-demixing matrix containing both positive and negative correlations.

Lemma 7 (Failure of Softmax Attention for Cross-Channel Demixing)**.** *Let $\mathbf{T}^\star$ be a cross-channel demixing operator such that for some target channel m and reference channel k, the optimal operation requires subtraction: $\mathbf{T}^\star_{m,m} > 0$ and $\mathbf{T}^\star_{m,k} < 0$. Then $\mathbf{T}^\star_{m,:} \notin \Delta^{N-1}$. Consequently, standard softmax attention over channels cannot compute this cross-variate projection.*

Proof. By definition, the simplex constraint Δ^{N-1} strictly requires all kernel weights to be non-negative. The attention mechanism calculates the output as $\mathbf{o}_m = \sum_{i=1}^{N} a_{m,i}\mathbf{h}_i$ subject to $\sum_{i=1}^{N} a_{m,i} = 1$ and $a_{m,i} \geq 0$. An operation requiring $\mathbf{T}^\star_{m,k} < 0$ forces the operator outside the non-negative orthant. As a result, the attention mechanism is structurally forbidden from producing negative coefficients, defaulting instead to a positive correlation bias. □

Proposition 8 (Realization by TOA Channel-Mixing)**.** *When TOA is applied under inverted tokenization, the post-activation unconstrained operator $\mathbf{S}_2 \in \mathbb{R}^{N\times N}$ (where $N = C$) provides the signed degrees of freedom required to compute exact cross-channel demixing.*

Proof. The TOA sequence operator is inherently parameterized as an affine shift $\mathbf{S}_2 = \mathbf{I}_N + \mathbf{M}_2$, which, free from simplex constraints, spans the entirety of $\mathbb{R}^{N\times N}$. Because $\mathbf{M}_2$ is unconstrained, we can directly set the residual mixer such that $(\mathbf{S}_2)_{m,m} = 1$, $(\mathbf{S}_2)_{m,k} = -1$, and zeros elsewhere. This directly realizes the cross-channel subtraction operator $\mathbf{o}_m = \mathbf{h}_m - \mathbf{h}_k$, circumventing the softmax bottleneck and providing the signed interactions required for cross-channel demixing. □

G FULL MULTIVARIATE TIME SERIES ANALYSIS RESULTS TABLE

Table 6: Full Multivariate TSF Results (1/2: DUET and PatchTST). We report MSE and MAE. Best results are in **bold** (b) and second best are underlined (r) (compared within each backbone group). AvgRank and #Top1 are calculated based on MSE across all horizons.

Model		DUET								PatchTST							
		Softmax-Attention		TOA-Gated		TOA-ReLU		TOA-Softmax		Softmax-Attention		TOA-Gated		TOA-ReLU		TOA-Softmax	
Dataset	Len	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	96	0.173	0.219	0.156	0.207	**0.154**	**0.203**	0.168	0.217	0.152	0.200	0.150	0.198	0.150	0.198	**0.149**	**0.196**
	192	0.218	0.261	0.211	0.258	**0.204**	**0.252**	0.216	0.257	0.195	0.241	0.195	**0.239**	0.194	0.240	**0.193**	0.241
	336	0.269	0.300	**0.258**	**0.295**	0.261	0.299	0.266	0.297	0.250	0.284	0.252	0.284	0.249	**0.283**	**0.248**	**0.283**
	720	**0.338**	0.349	0.348	0.351	**0.338**	0.353	**0.338**	**0.346**	0.325	0.336	**0.323**	**0.333**	0.325	**0.333**	0.326	0.336
Solar	96	0.199	0.250	0.212	0.236	**0.180**	**0.227**	0.197	0.237	0.176	**0.228**	0.195	0.282	**0.163**	0.238	0.166	0.237
	192	0.234	0.276	0.246	0.263	**0.217**	**0.256**	0.233	0.276	0.188	0.257	**0.185**	**0.252**	0.187	0.256	**0.185**	0.257
	336	0.247	0.283	**0.230**	0.278	0.239	0.271	0.236	**0.267**	0.194	0.256	0.205	0.256	0.202	0.257	**0.188**	**0.252**
	720	0.243	**0.274**	0.237	0.278	**0.232**	0.277	0.255	0.284	0.233	0.269	0.213	0.272	0.209	0.266	**0.208**	**0.262**
ECL	96	0.145	0.239	0.147	0.244	**0.141**	**0.237**	0.147	0.241	0.131	0.223	0.129	0.223	0.130	0.223	**0.128**	**0.221**
	192	0.160	0.252	**0.156**	**0.251**	0.157	0.255	0.161	0.253	0.150	0.241	0.149	0.240	0.149	0.241	**0.147**	**0.239**
	336	0.178	0.272	0.172	0.274	**0.171**	**0.270**	0.179	0.273	**0.165**	0.259	0.166	0.259	0.169	0.264	**0.165**	**0.258**
	720	0.224	0.312	**0.204**	**0.302**	0.211	0.307	0.219	0.310	0.205	0.294	0.206	0.294	0.203	**0.291**	**0.202**	0.293
ETTh1	96	0.399	0.415	0.400	**0.410**	**0.396**	0.412	0.398	0.413	**0.370**	**0.394**	0.373	0.397	0.372	0.396	0.371	**0.394**
	192	0.452	0.442	0.447	0.444	**0.440**	**0.436**	0.441	0.442	**0.407**	**0.416**	0.411	0.418	0.409	0.417	0.409	0.417
	336	0.497	0.475	**0.473**	**0.456**	0.481	0.461	0.492	0.467	0.434	0.434	0.437	0.435	0.436	**0.433**	**0.433**	**0.433**
	720	0.512	0.498	**0.496**	**0.496**	0.500	0.501	0.512	0.503	0.447	**0.462**	0.450	0.463	0.450	0.463	**0.446**	**0.462**
ETTh2	96	0.317	0.361	0.306	0.355	0.307	0.356	**0.303**	**0.353**	**0.274**	**0.335**	0.277	0.338	0.294	0.348	0.275	0.336
	192	0.416	0.419	0.397	0.408	0.394	0.405	**0.378**	**0.399**	0.380	0.401	0.373	0.397	**0.339**	**0.377**	0.342	0.379
	336	**0.423**	**0.432**	0.438	0.440	0.439	0.437	0.448	0.446	0.390	0.413	0.389	0.413	0.383	0.411	**0.377**	**0.406**
	720	0.464	0.464	**0.437**	**0.447**	0.440	0.448	0.440	0.452	**0.391**	**0.429**	0.393	0.430	0.404	0.440	0.396	0.432
ETTm1	96	0.354	0.382	0.371	0.394	0.358	0.387	**0.342**	**0.376**	0.291	0.343	0.288	0.342	**0.283**	**0.339**	0.288	0.340
	192	**0.381**	**0.398**	0.403	0.411	0.404	0.412	0.382	**0.398**	0.331	0.368	**0.326**	**0.366**	0.332	0.372	0.329	0.368
	336	**0.414**	**0.418**	0.432	0.429	0.441	0.434	0.422	0.425	0.370	0.394	0.366	**0.387**	0.368	**0.387**	**0.365**	0.391
	720	0.501	0.472	0.497	0.467	0.489	0.464	**0.480**	**0.461**	0.415	0.422	**0.412**	**0.417**	0.418	0.422	0.413	0.421
ETTm2	96	0.182	0.262	0.180	0.263	0.180	0.266	**0.177**	**0.258**	**0.163**	**0.253**	0.166	0.255	0.167	0.256	0.165	**0.253**
	192	**0.239**	0.305	0.244	**0.304**	0.261	0.325	0.247	0.309	0.222	0.295	0.222	**0.293**	0.223	0.295	**0.221**	**0.293**
	336	**0.299**	**0.342**	0.350	0.379	0.363	0.388	0.338	0.369	**0.273**	**0.327**	0.277	0.330	0.278	0.331	0.274	**0.327**
	720	0.527	0.472	0.533	0.480	**0.440**	**0.430**	0.488	0.449	0.368	0.386	**0.360**	**0.383**	0.361	0.384	0.367	**0.383**
Exchange	96	0.088	0.206	**0.080**	**0.199**	0.081	0.200	0.084	0.204	0.367	0.367	0.122	0.250	**0.120**	**0.248**	0.149	0.270
	192	0.173	0.302	**0.168**	**0.297**	0.172	0.302	0.176	0.307	0.499	0.458	0.322	0.424	**0.306**	**0.411**	0.460	0.458
	336	**0.277**	**0.389**	0.289	0.398	0.280	0.392	0.279	0.391	0.575	0.557	0.537	0.542	**0.444**	**0.509**	0.774	0.625
	720	0.524	0.556	0.518	0.551	**0.499**	**0.542**	0.508	0.546	**1.250**	**0.821**	1.289	0.833	1.322	0.863	1.294	0.849
Traffic	96	**0.422**	**0.274**	0.434	**0.274**	0.448	0.278	0.431	**0.274**	0.366	**0.250**	0.368	0.254	0.370	0.257	**0.365**	**0.250**
	192	0.447	0.288	0.453	0.291	0.449	0.292	**0.442**	**0.279**	0.385	0.258	0.386	0.260	0.387	0.262	**0.381**	**0.256**
	336	0.487	0.300	0.474	0.302	0.484	0.307	**0.461**	**0.286**	0.398	0.266	0.400	0.267	0.408	0.277	**0.396**	**0.265**
	720	0.509	0.308	**0.497**	0.315	0.498	0.317	0.510	**0.307**	0.453	0.308	**0.434**	**0.286**	0.436	0.290	0.453	0.308
AvgRank		2.89		2.42		**2.14**		2.36		2.67		2.56		2.69		**1.78**	
#Top1		8		10		**12**		8		8		6		6		**18**	

Table 7: Full Multivariate TSF Results (2/2: iTransformer and Autoformer); continuation of Table 6, with the same conventions.

Model		iTransformer								Autoformer							
		Softmax-Attention		TOA-Gated		TOA-ReLU		TOA-Softmax		Softmax-Attention		TOA-Gated		TOA-ReLU		TOA-Softmax	
Dataset	Len	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	96	0.175	0.215	0.170	0.211	**0.169**	**0.210**	0.171	0.211	0.247	0.305	0.247	0.328	0.225	0.303	**0.207**	**0.278**
	192	0.223	0.256	0.220	0.257	0.217	**0.254**	**0.215**	0.255	0.276	0.326	0.268	0.334	0.273	0.346	**0.251**	**0.310**
	336	0.282	0.299	0.275	0.297	0.275	0.296	**0.272**	**0.295**	0.309	0.344	0.303	0.358	0.299	0.356	**0.293**	**0.334**
	720	0.358	0.350	0.355	0.350	0.356	0.350	**0.348**	**0.346**	**0.351**	0.374	0.383	0.424	0.352	0.389	0.353	**0.373**
Solar	96	0.206	0.233	0.202	**0.222**	0.201	**0.222**	**0.198**	0.224	**0.547**	**0.559**	0.548	0.565	0.760	0.673	0.550	0.562
	192	**0.231**	0.257	0.245	0.257	0.237	**0.251**	0.241	0.255	0.768	0.677	**0.724**	**0.645**	0.764	0.679	0.784	0.677
	336	**0.248**	0.272	0.256	0.272	0.255	**0.269**	0.252	0.270	0.777	0.682	**0.758**	**0.673**	0.764	0.679	0.778	0.690
	720	**0.251**	**0.274**	0.265	0.283	0.265	0.281	0.260	0.277	0.771	0.696	0.776	0.700	**0.766**	**0.693**	0.776	0.703
ECL	96	0.147	0.239	0.147	0.239	0.147	0.239	**0.141**	**0.235**	0.423	0.500	0.412	0.493	0.418	0.497	**0.401**	**0.483**
	192	0.163	0.255	0.166	0.257	0.167	0.258	**0.161**	**0.253**	0.445	0.517	0.433	0.508	0.433	0.507	**0.415**	**0.493**
	336	0.177	0.270	0.183	0.276	0.181	0.274	**0.173**	**0.267**	0.451	0.519	0.454	0.522	0.453	0.521	**0.429**	**0.502**
	720	0.231	0.314	0.206	0.295	0.206	0.294	**0.194**	**0.287**	0.505	0.552	0.513	0.557	0.518	0.561	**0.500**	**0.547**
ETTh1	96	0.385	0.403	0.386	0.402	**0.384**	**0.401**	0.385	0.403	0.470	0.481	**0.457**	**0.466**	0.499	0.483	0.467	0.477
	192	0.440	0.435	0.440	0.433	**0.437**	**0.430**	0.442	0.436	0.487	0.490	**0.469**	**0.470**	0.501	0.487	0.478	0.483
	336	0.488	0.458	0.488	0.458	0.488	0.457	**0.484**	**0.456**	0.490	0.492	**0.474**	**0.476**	0.497	0.488	0.486	0.489
	720	0.522	0.502	**0.499**	**0.486**	0.501	0.487	0.508	0.494	0.502	0.513	**0.476**	**0.492**	0.518	0.518	0.498	0.511
ETTh2	96	0.299	0.350	**0.295**	**0.346**	**0.295**	0.347	0.299	0.350	0.370	0.413	0.370	0.412	0.370	0.411	**0.358**	**0.406**
	192	0.378	0.398	**0.376**	**0.396**	**0.376**	**0.396**	0.381	0.399	0.386	0.428	0.386	**0.423**	0.387	0.425	**0.384**	0.426
	336	0.424	0.433	0.419	**0.430**	**0.418**	**0.430**	0.424	0.434	0.386	0.431	**0.381**	**0.424**	**0.381**	0.425	0.388	0.433
	720	0.428	0.447	**0.425**	**0.444**	**0.425**	**0.444**	0.428	0.447	0.426	0.457	0.427	0.457	0.428	0.461	**0.418**	**0.452**
ETTm1	96	0.347	0.379	0.343	**0.373**	**0.333**	**0.373**	0.335	**0.373**	0.502	0.484	0.502	0.485	**0.470**	0.484	0.492	**0.479**
	192	0.380	0.394	0.377	0.392	**0.373**	**0.390**	0.374	0.392	0.490	0.481	0.516	0.495	0.491	0.483	**0.486**	**0.480**
	336	0.418	0.418	0.408	**0.411**	0.409	0.413	**0.407**	0.415	0.513	0.494	0.568	0.509	0.554	0.511	**0.501**	**0.489**
	720	0.488	0.457	0.478	0.450	**0.472**	**0.447**	0.481	0.454	**0.541**	**0.509**	0.577	0.529	0.589	0.538	0.546	0.512
ETTm2	96	0.184	0.269	**0.178**	**0.264**	0.180	0.266	0.184	0.267	0.263	0.334	0.261	0.335	0.256	0.331	**0.230**	**0.316**
	192	0.253	0.313	**0.245**	**0.307**	0.247	0.308	0.250	0.310	0.294	0.352	0.295	0.354	0.292	0.351	**0.279**	**0.346**
	336	0.313	0.351	**0.311**	**0.348**	**0.311**	**0.348**	0.315	0.352	0.333	0.373	0.330	0.372	0.327	**0.370**	**0.325**	0.373
	720	0.412	0.405	0.411	0.404	**0.410**	**0.403**	0.411	0.404	0.415	0.423	0.414	0.421	**0.409**	**0.417**	0.413	0.423
Exchange	96	0.087	0.207	**0.085**	**0.204**	**0.085**	0.205	0.087	0.207	0.340	0.437	0.339	0.438	0.343	0.439	**0.332**	**0.432**
	192	0.178	0.301	**0.177**	**0.299**	0.178	**0.299**	0.178	0.301	0.465	0.514	0.461	0.511	0.463	0.513	**0.456**	**0.510**
	336	0.333	0.419	**0.331**	**0.418**	0.333	0.419	0.332	**0.418**	**0.634**	0.612	0.640	0.614	0.636	0.612	**0.634**	**0.611**
	720	0.853	0.696	**0.846**	**0.693**	0.847	0.694	0.853	0.696	**1.264**	**0.864**	1.303	0.877	1.282	0.870	1.277	0.867
Traffic	96	0.401	0.272	0.400	0.271	0.399	0.270	**0.393**	**0.267**	**1.036**	**0.619**	1.055	0.627	1.046	0.623	1.050	0.626
	192	0.415	0.278	0.417	0.279	0.416	0.280	**0.412**	**0.274**	1.063	0.630	1.067	0.632	1.083	0.638	**1.047**	**0.621**
	336	0.426	**0.287**	0.428	0.291	0.427	0.289	**0.423**	0.289	1.074	0.630	1.073	0.631	1.058	0.623	**1.056**	**0.621**
	720	0.447	0.296	0.447	0.305	**0.446**	**0.295**	0.450	**0.295**	1.107	0.643	1.114	0.647	1.105	0.642	**1.093**	**0.635**
AvgRank		2.94		2.22		**1.89**		2.17		2.67		2.64		2.72		**1.69**	
#Top1		3		11		**14**		13		6		7		4		**21**	

Table 8: Full PEMS Results. We report MSE and MAE. Best results are in **bold** (b) and second best are underlined (r) (compared within each backbone group). AvgRank and #Top1 are calculated based on MSE across all horizons.

Model		PatchTST								iTransformer								Autoformer							
		Softmax-Attention		TOA-Gated		TOA-ReLU		TOA-Softmax		Softmax-Attention		TOA-Gated		TOA-ReLU		TOA-Softmax		Softmax-Attention		TOA-Gated		TOA-ReLU		TOA-Softmax	
Dataset	Len	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
PEMS03	12	0.058	**0.157**	**0.057**	**0.157**	**0.057**	**0.157**	0.058	0.159	0.066	**0.171**	0.067	0.172	0.067	**0.171**	**0.064**	0.172	0.299	0.418	0.316	0.435	0.318	0.439	**0.285**	**0.404**
	24	0.073	0.176	0.072	**0.174**	0.072	0.175	**0.071**	0.175	0.094	0.204	**0.092**	**0.200**	0.093	0.202	0.096	0.207	0.368	0.473	0.373	0.477	0.355	0.464	**0.332**	**0.440**
	48	0.101	0.203	**0.094**	**0.197**	0.098	0.201	0.096	0.200	0.158	0.268	0.148	**0.257**	**0.147**	**0.257**	0.163	0.273	0.762	0.717	0.740	0.710	**0.718**	**0.700**	0.824	0.746
	96	0.131	0.232	0.125	**0.222**	0.131	0.227	**0.123**	0.223	0.281	0.372	**0.234**	**0.330**	0.236	0.333	0.275	0.364	0.956	0.826	**0.899**	0.803	0.942	0.821	0.911	**0.801**
PEMS04	12	**0.071**	**0.169**	**0.071**	0.172	**0.071**	0.172	0.072	0.172	0.077	0.182	**0.076**	**0.180**	**0.076**	0.181	0.077	0.182	0.358	0.469	0.359	0.466	0.342	0.452	**0.338**	**0.441**
	24	0.083	0.184	0.082	0.183	**0.081**	**0.181**	0.082	0.185	0.095	0.204	**0.092**	**0.200**	**0.092**	**0.200**	0.093	0.202	0.698	0.692	**0.680**	**0.683**	0.766	0.739	0.709	0.698
	48	0.100	0.202	0.096	**0.195**	0.095	**0.195**	**0.094**	0.196	0.129	0.242	**0.115**	**0.226**	0.116	**0.226**	0.119	0.232	**0.903**	**0.809**	0.929	0.818	0.916	0.816	0.922	0.817
	96	0.124	0.227	0.111	**0.211**	0.111	**0.211**	**0.110**	**0.211**	0.160	0.271	0.141	0.251	**0.138**	**0.249**	0.146	0.257	1.068	0.891	1.067	0.891	**1.058**	**0.887**	1.068	0.889
PEMS07	12	**0.048**	**0.138**	0.055	0.153	0.050	0.144	0.050	0.145	0.062	**0.158**	0.062	0.159	**0.061**	**0.158**	0.064	0.159	0.261	0.389	0.293	0.419	0.243	0.372	**0.214**	**0.344**
	24	**0.057**	0.150	**0.057**	**0.147**	**0.057**	0.153	**0.057**	0.151	0.082	0.183	0.080	**0.181**	**0.079**	0.182	0.085	0.186	0.296	0.420	0.305	0.432	0.400	0.507	**0.279**	**0.404**
	48	0.071	0.168	0.072	0.170	0.069	0.166	**0.068**	**0.164**	0.123	0.230	0.104	0.206	**0.102**	**0.205**	0.275	0.355	0.608	0.639	**0.569**	**0.609**	0.658	0.661	0.791	0.742
	96	0.091	0.193	0.092	0.195	0.082	**0.177**	**0.080**	0.178	0.366	0.432	0.127	**0.230**	**0.125**	**0.230**	0.127	0.234	0.935	0.801	**0.912**	**0.799**	0.931	0.803	0.918	0.803
PEMS08	12	0.067	0.161	**0.064**	**0.157**	0.065	0.161	0.066	0.163	0.084	0.188	0.083	0.186	0.083	0.186	**0.082**	**0.184**	0.494	0.516	0.514	0.533	0.499	0.523	**0.479**	**0.504**
	24	0.088	0.180	0.081	**0.172**	**0.080**	0.174	0.083	0.177	0.130	0.235	**0.124**	**0.228**	0.125	0.229	0.125	0.229	0.756	0.678	0.641	0.618	**0.599**	**0.591**	0.786	0.690
	48	0.125	0.201	**0.117**	0.190	0.122	**0.189**	0.123	0.193	0.235	0.279	0.188	**0.230**	**0.185**	0.231	0.267	0.308	0.958	0.787	0.974	0.790	0.967	0.788	**0.864**	**0.740**
	96	0.191	0.224	0.162	**0.205**	0.164	0.207	**0.158**	**0.205**	0.280	0.313	0.231	0.267	**0.223**	**0.264**	0.258	0.296	1.201	0.898	1.189	0.892	**1.176**	**0.883**	1.206	0.900
AvgRank		3.19		2.06		**1.88**		**1.88**		3.38		1.69		**1.44**		3.00		2.62		2.62		2.44		**2.25**	
#Top1		3		6		5		**8**		0		6		**10**		2		1		4		4		**7**	

Table 9: Summary of Classification Results. Accuracies are reported. Best results are in **bold** (b) and second best are underlined (r).

Dataset	PatchTST				Autoformer			
	Softmax	TOA-ReLU	TOA-Gated	TOA-Softmax	Softmax	TOA-ReLU	TOA-Gated	TOA-Softmax
ArticularyWordRecognition	**0.9800**	**0.9800**	**0.9800**	**0.9800**	0.5322	0.5300	0.5300	**0.5511**
AtrialFibrillation	0.4000	0.5333	**0.6000**	0.4000	0.2889	0.2667	**0.4000**	0.3556
BasicMotions	0.6167	0.7000	**0.7250**	0.6083	0.5500	**0.6667**	0.6333	0.5333
CharacterTrajectories	0.9654	0.9708	**0.9721**	0.9652	0.8285	**0.9350**	0.9308	**0.9350**
Cricket	0.9167	**0.9306**	0.9028	0.8981	0.2407	**0.2500**	**0.2500**	0.2083
DuckDuckGeese	0.2000	0.2400	**0.2600**	0.2133	**0.2333**	0.1933	0.1933	0.2067
ERing	0.9457	**0.9519**	**0.9519**	0.8778	**0.6605**	0.6309	0.6210	0.5025
Epilepsy	**0.9589**	0.9275	0.9565	0.9565	0.8213	0.7464	0.8140	**0.8527**
EthanolConcentration	0.2421	0.2776	**0.2814**	0.2307	0.2421	0.2294	**0.2649**	0.2459
FaceDetection	0.6642	**0.6725**	0.6552	0.6557	0.5816	**0.6169**	0.5757	0.5743
FingerMovements	0.5533	**0.6200**	0.6100	0.5333	0.4767	0.5067	**0.5167**	0.4933
HandMovementDirection	0.4820	**0.5405**	**0.5405**	0.4369	**0.3198**	0.3108	**0.3198**	0.3153
Handwriting	0.2322	**0.2671**	0.2553	0.2267	**0.1596**	0.1439	0.1310	0.1525
Heartbeat	0.6488	**0.7171**	0.6829	0.6894	0.7138	0.7252	**0.7333**	0.7171
InsectWingbeat	0.5756	0.6024	**0.6034**	0.5718	**0.6150**	**0.6150**	0.5987	0.5737
JapaneseVowels	0.9441	**0.9541**	0.9514	0.9468	0.9550	0.9685	0.9685	**0.9712**
LSST	0.4784	**0.5008**	0.4862	0.4757	**0.4184**	0.3643	0.3850	0.3882
Libras	0.7407	**0.7778**	0.7722	0.7444	0.7667	**0.8074**	0.7944	0.7926
NATOPS	0.7481	0.7556	**0.7611**	0.7593	0.7907	**0.8778**	0.8704	0.8759
PEMS-SF	0.8170	0.8382	**0.8728**	0.7765	0.7322	**0.7592**	0.7322	0.7399
PenDigits	0.9584	**0.9711**	0.9694	0.9623	0.9737	**0.9763**	0.9725	0.9672
PhonemeSpectra	0.1179	0.1208	**0.1265**	0.1145	0.0741	0.0856	**0.0967**	0.0959
RacketSports	0.7237	**0.7763**	0.7697	0.7434	0.7763	0.8114	0.7917	**0.8136**
SelfRegulationSCP1	0.7873	0.8396	**0.8498**	0.8020	0.5438	**0.7827**	0.7782	0.5961
SelfRegulationSCP2	0.4796	0.5278	**0.5389**	0.4648	**0.5037**	0.4833	0.5019	0.4963
SpokenArabicDigits	0.9650	**0.9832**	0.9827	0.9674	0.9730	**0.9782**	0.9744	0.9771
StandWalkJump	0.4000	**0.5333**	0.4667	0.3111	0.3556	**0.3778**	**0.3778**	0.3556
UWaveGestureLibrary	0.8323	**0.8469**	0.8406	0.8302	0.4615	0.4656	0.4583	**0.4938**
AvgRank	3.11	**1.54**	1.64	3.36	2.79	**2.27**	2.45	2.50
#Top1	2	**16**	14	1	7	**12**	8	6

Table 10: Full TSF Ablation Results on Stochastic Operator Regularization. We report MSE. Best results (lower MSE) are in **bold**. AvgRank and #Top1 are calculated based on MSE across all horizons within each group.

Model		TOA-Gated		TOA-ReLU	
		w/o SOR	w/ SOR	w/o SOR	w/ SOR
Dataset	Len	MSE	MSE	MSE	MSE
Weather	96	0.157	**0.154**	0.154	**0.150**
	192	0.202	**0.192**	0.198	**0.195**
	336	0.254	**0.250**	0.260	**0.247**
	720	0.337	**0.324**	0.343	**0.321**
ETTh1	96	**0.370**	0.371	0.371	**0.370**
	192	0.411	**0.408**	0.412	**0.410**
	336	**0.434**	0.446	0.438	**0.433**
	720	0.478	**0.458**	**0.462**	0.470
ETTh2	96	0.299	**0.280**	0.300	**0.285**
	192	0.396	**0.364**	0.387	**0.354**
	336	0.410	**0.407**	0.419	**0.384**
	720	0.404	**0.390**	0.416	**0.393**
ETTm1	96	0.291	**0.291**	0.295	**0.292**
	192	**0.331**	0.347	**0.327**	0.335
	336	**0.362**	0.369	**0.370**	**0.370**
	720	0.423	**0.415**	**0.423**	0.426
ETTm2	96	0.177	**0.166**	0.179	**0.170**
	192	0.245	**0.228**	0.239	**0.231**
	336	0.290	**0.281**	0.278	**0.277**
	720	0.373	**0.362**	0.397	**0.366**
AvgRank		1.80	**1.20**	1.80	**1.15**
#Top1		4	**16**	4	**17**

H Datasets and Experimental Details

H.1 Experimental Datasets

We evaluate TOA on nine widely used multivariate time series benchmark datasets for long term time series forecasting and 4 datasets for short term time series forecasting. Tables 6 and 7 summarize the statistics of TOA's performance on these datasets. We provide the full names and key characteristics for each below:

Weather Dataset (Wu et al., 2021): Comprises 21 meteorological indicators (e.g., air temperature, humidity, and wind speed) recorded at 10-minute intervals throughout 2020 at the Max Planck Biogeochemistry Institute.

Solar Dataset (Lai et al., 2018): Tracks solar power production from 137 photovoltaic plants in Alabama, with observations collected every 10 minutes throughout the year 2006.

Electricity Dataset (Wu et al., 2021): Records the hourly electricity consumption of 321 users over a three-year period (2012–2014), capturing both periodic patterns and long-term dependencies.

ETT Dataset (Zhou et al., 2021): Monitors oil temperature and load metrics from electricity transformers over two years (2016–2018) at both 15-minute (ETTm1, ETTm2) and hourly (ETTh1, ETTh2) resolutions.

Exchange Dataset (Lai et al., 2018): Contains daily exchange rate records for eight countries from 1990 to 2016, providing a longitudinal view of global financial market fluctuations.

Traffic Dataset (Wu et al., 2021): Provides hourly road occupancy rates collected from 862 freeway sensors in the San Francisco Bay Area over a continuous two-year period (2015–2016).

PEMS Dataset (Li et al., 2018b): Captures spatial-temporal traffic dynamics via 5-minute sensor samples in California; this study utilizes the PEMS03, PEMS04, PEMS07, and PEMS08 subsets as standard benchmarks.

For anomaly detection and classification, we follow the standard benchmark suites provided by the Time-Series Library (TSLib) (Wu et al., 2023). Specifically, anomaly detection is evaluated on the five widely used TSLib benchmarks PSM, MSL, SMAP, SMD, and SWAT, while classification is evaluated on the 28 multivariate datasets from the UEA archive used in the TSLib evaluation pipeline. Unless otherwise noted, we adopt the default dataset splits and preprocessing conventions provided by TSLib for these tasks.

H.2 Hyper-parameter Settings and Implementation Details

Model Architecture - Forecasting Tasks. For long- and short-term multivariate forecasting, we integrate Temporal Operator Attention (TOA) into standard time-series backbones, including PatchTST, iTransformer, DUET, and Autoformer, by replacing the backbone's sequence-mixing attention primitive while leaving the surrounding feed-forward blocks unchanged. In all TOA variants, the temporal operators are parameterized as residual dense mixers, $\mathbf{S}_1 = \mathbf{I} + \mathbf{M}_1$ and $\mathbf{S}_2 = \mathbf{I} + \mathbf{M}_2$, with the learned offsets initialized at small scale ($\sigma_M = 10^{-3}$). We evaluate three variants: TOA-Softmax, TOA-ReLU, and TOA-Gated. For the forecasting benchmarks, we retain the official or official-aligned backbone hyper-parameters used by each upstream model family. PatchTST follows the standard long-horizon configuration with patch length 16 and stride 8, with dataset-dependent widths and depths taken from the official scripts. iTransformer uses its official aligned encoder-only setup with two encoder layers, eight attention heads, GELU activations, and input normalization enabled. DUET follows its official unified or nearest-official PEMS profile, using mixture-of-experts style sequence modeling with four experts, top-$k = 2$, hidden size 256, and channel-independent processing. Prediction horizons are {96, 192, 336, 720} for long-horizon forecasting and {12, 24, 48, 96} for the PEMS short-horizon benchmarks.

Model Architecture - Classification and Anomaly Detection. For time-series classification and anomaly detection, we use the TSLib benchmark suite, with TOA inserted into PatchTST and Autoformer for classification and into iTransformer, PatchTST, and Autoformer for anomaly detection. In classification, the default model width is $d_{\text{model}} = 128$ with $d_{\text{ff}} = 256$, three encoder layers, eight attention heads, dropout 0.1, and patch length 16 where applicable. The effective input sequence

length is set dynamically to the maximum sequence length of each dataset at runtime. For anomaly detection, we use the standard reconstruction-based setup with three encoder layers, eight attention heads, dropout 0.1, and dataset-specific input dimensionalities; typical anomaly configurations use $d_{\text{model}} = 128$ and short training schedules of 3 to 10 epochs depending on the dataset. TOA is applied in the same noncausal setting as in forecasting, using identity alignment and backbone-specific maximum sequence lengths for the dense operators.

Model Architecture - Synthetic Tasks. For the synthetic harmonic demixing experiments, we use a deliberately capacity-limited setup to emphasize primitive-level differences in sequence mixing. In particular, the synthetic experiments use shallow backbones with two layers, a single head, as described in the main text, so that improvements cannot be attributed simply to scale. The same TOA parameterization is used as in the real-data experiments, with dense residual temporal operators and noncausal sequence mixing.

Data Processing Strategy. For forecasting, we follow the preprocessing conventions of the underlying backbone families and their official scripts. PatchTST uses RevIN-style normalization where specified by the original configuration; iTransformer uses its standard input normalization over the observed context window; and DUET uses channel-independent processing where applicable. All forecasting runs are performed in the multivariate setting. For classification, we use the standard UEA-format pipeline with a validation split ratio of 0.2. For anomaly detection, we use the dataset-specific TSLib preprocessing pipelines and anomaly ratios defined by the benchmark configuration.

Training and Hardware. All training tasks in this paper were conducted using a single H200. For forecasting in the TSF benchmark stack, we use the AdamW optimizer with $\beta = (0.9, 0.95)$ and MSE loss. PatchTST and iTransformer use learning rate 10^{-4}, while DUET uses learning rate 5×10^{-4}. Gradient accumulation is used for larger datasets such as Electricity, Solar, Traffic, and the PEMS benchmarks to maintain stable effective batch sizes. For classification, we use the RAdam optimizer with cross-entropy loss, learning rate 10^{-3}, batch size 16, 100 training epochs, and early stopping patience 10. For anomaly detection, we use the Adam optimizer with MSE reconstruction loss, learning rate 10^{-4}, batch size typically 128, and early stopping patience 3. In forecasting and anomaly detection, we use random seed 2024, while the classification sweeps are evaluated over three seeds {2024, 2025, 2026}.

I MULTI-SEED ROBUSTNESS

Table 2, Table 6, and Table 7 report single-seed results (seed 2024, Appendix H), following common practice at this evaluation scale. To check that these differences are not an artifact of seed choice, we rerun PatchTST across 5 random seeds on three representative datasets (ETTh2, ETTm1, and Weather), spanning both smoother (ETTm1, Weather) and higher-frequency (ETTh2) dynamics.

Table 11: **Multi-Seed Robustness.** Mean ± standard deviation over 5 random seeds, PatchTST backbone. Best (lowest) value per row is **bold** (b); second-best is underlined (r), for each metric.

Model		TOA-Gated		TOA-ReLU		Softmax-Attention		TOA-Softmax	
Dataset	Len	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh2	96	0.290±0.006	0.346±0.005	0.291±0.007	0.348±0.005	**0.282±0.005**	**0.340±0.003**	0.284±0.009	0.342±0.005
	192	0.357±0.005	0.391±0.004	0.357±0.005	0.391±0.004	0.362±0.015	0.392±0.008	**0.352±0.006**	**0.385±0.004**
	336	0.396±0.007	0.422±0.006	0.394±0.014	0.419±0.009	0.383±0.003	0.411±0.002	**0.379±0.003**	**0.408±0.002**
	720	0.412±0.015	0.442±0.011	0.423±0.018	0.452±0.014	**0.396±0.008**	**0.432±0.005**	0.403±0.015	0.436±0.009
ETTm1	96	**0.285±0.001**	**0.340±0.001**	0.287±0.004	0.342±0.002	0.290±0.002	0.342±0.001	0.291±0.004	0.341±0.002
	192	**0.327±0.002**	**0.367±0.002**	0.331±0.006	0.371±0.003	0.333±0.001	0.369±0.000	0.330±0.002	0.368±0.001
	336	**0.365±0.002**	**0.388±0.003**	0.370±0.005	0.393±0.008	0.368±0.003	0.392±0.001	0.368±0.001	0.393±0.001
	720	**0.414±0.002**	**0.418±0.002**	0.420±0.003	0.425±0.005	0.417±0.002	0.422±0.002	0.417±0.003	0.423±0.003
Weather	96	**0.149±0.001**	**0.196±0.002**	0.151±0.003	0.199±0.003	0.152±0.002	0.200±0.002	0.151±0.001	0.198±0.002
	192	**0.196±0.001**	**0.240±0.002**	**0.196±0.002**	0.242±0.002	0.197±0.001	0.242±0.001	**0.196±0.001**	0.241±0.001
	336	0.250±0.002	**0.282±0.002**	0.251±0.003	0.284±0.002	0.252±0.002	0.285±0.001	**0.248±0.002**	**0.282±0.002**
	720	**0.325±0.001**	**0.335±0.002**	0.328±0.002	0.337±0.002	**0.325±0.001**	0.336±0.001	0.326±0.003	0.337±0.001

Table 11 reports mean ± standard deviation. TOA-Gated is the most consistent winner on ETTm1 and Weather, where its margin over the softmax baseline typically exceeds one standard deviation

(e.g. ETTm1 $L_p = 96$: 0.285 ± 0.001 vs. 0.290 ± 0.002). ETTh2 is the exception: at $L_p = 96$ and 720 the softmax baseline attains the lowest MSE, with the TOA variants falling within, or just outside, one standard deviation of it. This is consistent with the thin single-seed margin already visible in Table 6 (ETTh2 rows, PatchTST columns), and suggests that ETTh2 leaves less room for signed mixing to help because the near-identity operator that softmax attention already approximates is closer to sufficient there. We report this case directly rather than restricting the table to favorable datasets.

J Computational Efficiency Analysis

While replacing row-normalized similarity transport with explicit dense sequence operators inherently introduces an additional parameter footprint, the practical computational overhead remains manageable. A natural theoretical concern with unconstrained $N \times N$ mixers is that they might significantly degrade training throughput, potentially undoing the speed advantages of modern lightweight backbones.

To contextualize this overhead, we compare the training step times of TOA-augmented models against well-established architectures. As detailed in Table 12, integrating the **TOA-ReLU** or **TOA-Gated** mechanisms into PatchTST results in 18.88 ms per training step on the ETTh1 dataset. While this is an expected increase over the base PatchTST model, it remains comparable to standard canonical models such as Informer (19.57 ms) and requires less than half the time of Autoformer (45.55 ms). Similarly, an iTransformer equipped with our dense TOA variants processes a step in 12.92 ms. These results indicate that the representational capacity gained by removing the simplex constraint can be incorporated without exceeding the typical computational budgets established by prior Transformer-based forecasting models.

Table 12: **Computational Efficiency Comparison.** Average training step time (ms) on the ETTh1 dataset. While the inclusion of explicit dense sequence operators introduces an expected overhead relative to the standard baselines, the overall step time of TOA-augmented backbones remains lower than that of earlier architectures such as Informer and Autoformer.

Architecture	Sequence Operator	ETTh1 Train Step (ms)
Autoformer	Auto-Correlation	45.55
Informer	ProbSparse Attention	19.57
PatchTST	Standard Attention (Baseline)	10.57
	TOA-Softmax (Ablation)	13.67
	TOA-ReLU / TOA-Gated	18.88
iTransformer	Standard Attention (Baseline)	6.58
	TOA-Softmax (Ablation)	9.31
	TOA-ReLU / TOA-Gated	12.92

K Low-Rank Operator Factorization

The Limitations discussion (Appendix A) notes that TOA's dense operators introduce an $O(N^2)$ parameter and compute footprint. We test whether factorizing $\mathbf{M}_1, \mathbf{M}_2$ at rank $r \in \{16, 32\}$, reducing the footprint to $O(Nr)$, preserves accuracy under our default training recipe (SOR applied throughout, as elsewhere in the paper). This is a capacity probe, not a claim of equivalent generality: only the dense family provably realizes every target operator exactly (§3.4), since a rank-r factorization confines $\mathbf{M}_2$ to matrices of rank at most r, not the full $\mathbb{R}^{N\times N}$ the dense parameterization spans, while the five recurring regimes (§3.3) generically require operators whose effective rank scales with problem size, often comparable to C (channel count) or K (frequency count) rather than a fixed small constant. We compare dense, rank-16, and rank-32 factorizations of TOA-Gated and TOA-ReLU on PatchTST and iTransformer, across ETTh2, ETTm1, and Weather at horizons 96 and 720.

Table 13 shows the factorized variants match dense performance. We however report dense as the default: it is the only configuration our realization guarantees cover for arbitrary instances of the five

regimes, while low-rank factorization is a validated, cheaper substitute for settings resembling those tested here.

Table 13: **Low-Rank Operator Factorization.** Test MSE for dense vs. rank-16 vs. rank-32 factorizations of $\mathbf{M}_1, \mathbf{M}_2$, with SOR applied throughout, as in all other experiments, so differences isolate rank. Best (lowest) value per row within each variant is **bold** (b).

			TOA-Gated			TOA-ReLU		
Backbone	Dataset	Len	Dense	Rank-16	Rank-32	Dense	Rank-16	Rank-32
PatchTST	ETTh2	96	0.287	0.279	**0.277**	0.283	**0.282**	0.287
		720	0.394	0.415	**0.393**	0.437	0.414	**0.392**
	ETTm1	96	**0.283**	**0.283**	0.285	0.290	0.286	**0.284**
		720	**0.413**	0.421	**0.413**	0.422	**0.416**	**0.416**
	Weather	96	**0.149**	**0.149**	0.151	0.151	**0.149**	0.150
		720	0.326	**0.325**	0.329	0.328	0.325	**0.324**
iTransformer	ETTh2	96	**0.291**	**0.291**	**0.291**	0.293	0.293	**0.292**
		720	**0.412**	**0.412**	**0.412**	0.413	0.413	**0.411**
	ETTm1	96	**0.299**	0.301	**0.299**	0.302	**0.301**	**0.301**
		720	0.439	**0.438**	0.439	0.441	0.435	**0.432**
	Weather	96	**0.155**	0.158	0.157	0.159	0.158	**0.155**
		720	0.325	**0.324**	0.325	**0.324**	0.327	0.330