Causal Relational Toeplitz Attention

Peter Racioppo
pcracioppo@gmail.com

***Abstract*—We propose a deep learning framework for... To enable robust forecasting in this low-data, high-sparsity regime, we introduce the Causal Relational Toeplitz Attention (CRTA) mechanism, which enforces temporal stationarity and causality within a Transformer-based architecture. (change motivation to emphasize efficiency). The proposed CRTA-Net integrates: (1) a dual-path CNN encoder with Gated Residual Blocks for feature fidelity, (2) a CRTA temporal core that constructs a single, content-aware Toeplitz kernel from query-key correlations, and (3) explicit structural constraints that promote temporal consistency without redundant recurrence.**

***Index Terms*—Toeplitz Attention, Structural Regularization**

I. Introduction

A. The Challenge of Temporal Generalization

Convolutional recurrent models such as ConvLSTM and GRU enforce causal, Markovian dynamics by design, producing stable but low-capacity temporal representations. In contrast, Transformer models capture long-range dependencies through self-attention but lack built-in temporal regularity: each timestep computes an independent attention map $\mathbf{A}_t = \mathrm{softmax}(\mathbf{Q}_t\mathbf{K}^\top)$, allowing arbitrary non-stationary interactions that can easily overfit small datasets.

We therefore seek a formulation that preserves attention's non-recurrent, parallelizable structure while reintroducing the inductive bias of recurrent filtering. The proposed Causal Relational Toeplitz Attention (CRTA) mechanism achieves this by constraining attention to the space of causal Toeplitz operators—stationary, shift-invariant filters whose rows are temporally shifted copies of a single canonical kernel. This enforces temporal consistency across timesteps while retaining data adaptivity through $\mathbf{QK}$ correlations. In effect, CRTA restores a temporal stationarity prior within a fully attention-based, non-iterative framework.

II. Causal Relational Toeplitz Attention (CRTA)

A. Structural Prior: Stationary and Causal Attention

Standard self-attention computes $\mathbf{A} = \mathbf{QK}^\top$ independently for each sequence, allowing non-stationary, fully connected dependencies. CRTA instead constrains $\mathbf{A}$ to the set of causal Toeplitz matrices, where each row is a shifted copy of a canonical kernel and all future entries are masked:

$$\mathbf{A}_{ij} = \begin{cases} w[i-j], & j \le i, \\ 0, & j > i. \end{cases}$$

This constraint enforces temporal stationarity (shared correlations across lags) and strict causality, aligning the model with adaptive FIR filtering and ARMA processes.

B. Relational Query Kernel Construction

To derive a single, time-invariant filter that preserves context-awareness and causality, CRTA transforms the raw Query matrix ($\mathbf{Q}$) into a highly compressed vector, $\mathbf{Q}_{\mathrm{rel}} \in \mathbb{R}^{1\times D}$, where D is the embedding dimension. This process is necessary to enforce the Temporal Stationarity of the Toeplitz attention mechanism.

The core mechanism is a feature polishing pipeline that ensures robustness and expressive power before final aggregation:

$$\mathbf{Q}_{\mathrm{rel}} = \mathrm{Avg}_L\big(\mathrm{LayerNorm}(\mathrm{GELU}(\mathrm{Conv1D}_{\mathrm{causal}}^{G=D_H}(\mathbf{Q})))\big). \quad (1)$$

The $\mathrm{Conv1D}_{\mathrm{causal}}$ provides causal temporal smoothing, guaranteeing dependence only on past information ($t' \le t$). The depthwise grouping ($G = D_H$) ensures head independence, allowing each feature subspace to accumulate unique temporal statistics without channel intermixing. The subsequent non-linearity (GELU) and normalization (LayerNorm) are critical for capturing non-linear dynamics and maintaining kernel stability during optimization, respectively.

The final Avg_L step collapses the filtered sequence into the single, stationary vector $\mathbf{Q}_{\mathrm{rel}}$. This vector is then cross-correlated with the Key matrix ($\mathbf{K}$) to determine the canonical kernel:

$$\mathbf{W}_{\mathrm{kernel}} = \mathbf{Q}_{\mathrm{rel}}\mathbf{K}^\top, \quad \mathbf{W}_{\mathrm{kernel}} \in \mathbb{R}^{1\times L}, \quad (2)$$

requiring $\mathcal{O}(LD)$ complexity. $\mathbf{W}_{\mathrm{kernel}}$ provides the adaptable coefficients required for the subsequent causal Toeplitz operation.

C. Causal Toeplitz Assembly

The core of the CRTA mechanism is the construction of the Causal Toeplitz Attention Matrix ($\mathbf{A}$). The use of a Toeplitz matrix explicitly enforces the temporal stationarity assumption: the relationship (or kernel weight) between two time steps depends only on the lag $|i-j|$, not on their absolute position in the sequence. The matrix $\mathbf{A}$ is constructed entirely by temporal shifts of the derived kernel vector $\mathbf{W}_{\mathrm{kernel}} \in \mathbb{R}^{1\times L}$:

$$\mathbf{A}_{ij} = \mathbf{W}_{\mathrm{kernel}}[|i-j|]. \quad (3)$$

For time series forecasting, the influence must be strictly retrospective. We apply a causal mask, $\mathbf{M}_{\mathrm{causal}}$, to guarantee that information flow is limited to the past ($j \le i$). The mask's entries are set to zero for all $j \le i$ and negative infinity for all future connections ($j > i$).

The final attention weights are then computed:

$$\mathbf{A}^{\mathrm{final}} = \mathrm{Softmax}(\mathbf{A} + \mathbf{M}_{\mathrm{causal}})\,\mathbf{V}. \quad (4)$$

By combining the Toeplitz structure with the causal mask, CRTA achieves a temporal filtering mechanism that is stationary, context-aware, and strictly causal.

D. Efficient Convolutional Implementation

The multiplication of a Toeplitz matrix with a vector can be exactly recast as a circular convolution operation. This principle allows the entire CRTA attention mechanism to be implemented using standard convolution algorithms, replacing the inherently costly $\mathcal{O}(L^2)$ matrix-vector product. Specifically, the attention output $\mathbf{Y} = \mathbf{A}^{\text{final}}\mathbf{V}$ is computed as a causal convolution of the Value matrix ($\mathbf{V}$) with a normalized version of the canonical kernel ($\mathbf{W}_{\text{kernel}}$):

$$\mathbf{Y} = \tilde{\mathbf{w}} * \mathbf{V}, \quad \tilde{\mathbf{w}} = \frac{\mathbf{W}_{\text{kernel}}}{\|\mathbf{W}_{\text{kernel}}\|_1 + \varepsilon}. \tag{5}$$

Here, $\tilde{\mathbf{w}}$ is the ℓ_1-normalized kernel, which functionally replaces the Softmax normalization of the attention matrix. This computation achieves $\mathcal{O}(LK)$ complexity for a kernel band-limited to size $K \ll L$, or $\mathcal{O}(L \log L)$ via the Fast Fourier Transform when the full sequence length L is used as the kernel. This method unifies adaptive, $\mathbf{Q}_{\text{rel}}\mathbf{K}^\top$-derived attention with efficient, stationarity-enforcing causal filtering.

E. Application to Spatiotemporal Data

Forecasting progression from longitudinal retinal images requires modeling the full spatiotemporal domain ($\mathbf{X} \in \mathbb{R}^{B\times T\times H\times W\times C}$), where B is the batch size, T the sequence length, H image height, W image width, and C the channel dimension. Applying global attention over the full flattened sequence yields an intractable time complexity of $\mathcal{O}((THW)^2)$. To maintain the computational efficiency of CRTA, we employ Axial Factorization. This technique decomposes the full 4D problem into two sequential, lower-complexity operations: Time-Width (TW) and Time-Height (TH) attention passes. The resulting complexity is significantly reduced, as it is proportional to $\mathcal{O}((TW)^2 + (TH)^2)$.

In the most rigorous implementation, the Block-Causal Axial Design ($\mathcal{O}(L^2)$), the sequence length L is set to the product of the time dimension (T) and one spatial dimension (e.g., W). By restructuring the input as $(B \cdot H, T \cdot W, C)$, the attention layer operates over the entire spatio-temporal plane $T \cdot W$. The batch dimension B is grouped with the un-attended spatial dimension H to form the effective batch size. This configuration allows a token at time t to communicate with all spatial tokens (both at the current time t and in the past $t' < t$). This communication is controlled by a specific Block Causal Mask ($\mathbf{M}_{\text{block}}$) that permits unrestricted mixing within the current time block t, while strictly enforcing temporal causality between blocks. For the computationally efficient convolutional variant, leveraging the maximal $\mathcal{O}(L \log L)$ speedup necessitates a design compromise, resulting in a Decoupled Time-Causal approach detailed in the subsequent section.

To leverage the maximal $\mathcal{O}(L \log L)$ speedup offered by the convolutional implementation, the stringent spatial requirements of the Block-Causal design must be relaxed. We transition to the Decoupled Time-Causal approach. This variant factors the spatiotemporal input into $H \cdot W$ independent temporal sequences of length T:

$$(B, T, H, W, C) \to (B \cdot H \cdot W, T, C).$$

In this simplified form, the CRTA mechanism acts as a set of parallel, purely temporal causal filters, operating independently on the time sequence of every spatial location. This decoupling minimizes memory and computation by reducing the effective sequence length to T. Spatial coherence, which is lost in this step, is subsequently recovered by relying on localized spatial mixing layers, applied between the axial passes. This design provides a practical and scalable solution for high-resolution spatiotemporal forecasting.

III. Related Work

a) Linearized Attention.: Linear Transformers [1] and Performer [2] approximate $\mathrm{softmax}(\mathbf{Q}\mathbf{K}^\top)$ via kernel factorization, reducing cost but remaining non-stationary and non-causal.

b) Stationary Long Convolutions.: X and Hyena [3] enforce stationarity through parameterized long convolutions. These kernels are efficient and causal but fixed or parametric, not derived from $\mathbf{Q}, \mathbf{K}$ correlations.

Toeplitz neural networks [4] On the Integration of Self-Attention and Convolution [5]

c) Dynamic Convolution and Meta-Kernels.: Dynamic Conv [**?**] and CondConv [**?**] generate input-conditioned spatial filters, typically small and non-stationary. These methods produce a unique kernel per position, maximizing spatial flexibility but lacking temporal regularization. In contrast, CRTA generates a single kernel shared across all time steps, enforcing temporal stationarity and providing a global, content-aware representation. Dynamic Conv aims for spatial adaptivity, whereas CRTA enforces temporal coherence—offering an explicit bias toward consistent dynamics.

d) Causal State-Space Models.: Recent selective-state models (S4?) [**?**] (e.g., RetNet [6], Mamba [7]) approximate attention through recurrent exponential decays, achieving implicit stationarity but without explicit Toeplitz constraints. CRTA provides an explicit analytical form linking attention and filtering within that causal regime.

e) Summary.: CRTA unifies the efficiency of long convolutional models with the adaptivity of attention. Unlike prior dynamic or Toeplitz approaches, it derives a single stationary kernel directly from causal running statistics of $\mathbf{Q}$ and cross-correlates it with $\mathbf{K}$, forming a data-dependent yet stable causal operator.

Set Transformers [**?**] and Perceiver architectures [**? ?**] employ a small set of learned or latent *global queries* to aggregate sequence or image representations efficiently. However, these global queries serve as fixed latent tokens for summarization, not as dynamic, content-adaptive convolutional kernels. Efficient attention variants such as Longformer [**?**], Linformer [**?**], and Nyströmformer [**?**] reduce the quadratic cost via low-rank or sparse projections but preserve position-specific,

non-stationary attention patterns. FFT-based formulations exploiting the Toeplitz structure of relative-position encodings have also been explored for long-sequence efficiency [**? ?**], yet these rely on fixed or learned positional kernels rather than adaptive $\mathbf{QK}$-derived ones. CRTA differs by constructing a *per-sequence, stationary, and causal* kernel directly from $\mathbf{QK}$ correlations, combining the adaptivity of attention with the efficiency and regularity of convolutional filtering.

CRTA's novelty resides in its unique combination of adaptivity, stationarity, and efficiency. While Dynamic Convolution schemes (e.g., CondConv) generate context-aware kernels, they typically focus on spatial flexibility by producing small, unique filters per position, thus lacking the global temporal stationarity required for robust time-series forecasting. In contrast, Stationary Long Convolution models (e.g., Hyena) enforce stationarity and causality efficiently ($\mathcal{O}(L \log L)$), but their filters are often parameterized or fixed, preventing them from being fully data-adaptive to the global $\mathbf{QK}$ correlation statistics. CRTA explicitly bridges this gap: it uses the core $\mathbf{Q}_{\text{rel}}\mathbf{K}^\top$ correlation to dynamically derive a single, long kernel $\mathbf{W}_{\text{kernel}}$. This kernel is then rigidly constrained by the Toeplitz structure and causal convolution implementation to guarantee both global temporal consistency and maximum computational efficiency. CRTA therefore provides an explicit, content-aware temporal filter bank that is structurally superior for learning stable, non-Markovian dynamics.

TABLE I: Complexity Comparison of Attention Mechanisms

Mechanism	**Computation**	**Complexity**
Standard Attention	$\mathbf{QK}^\top$	$\mathcal{O}(L^2 D)$
Linear / Performer	$\Phi(\mathbf{Q})\Phi(\mathbf{K})^\top$	$\mathcal{O}(LD)$
Toeplitz / Hyena	Parameterized kernel (FFT)	$\mathcal{O}(L \log L)$
CRTA (Ours)	$\mathbf{Q}_{\text{rel}}\mathbf{K}^\top$ + causal conv	$\mathcal{O}(LD) + \mathcal{O}(L \log L)$

IV. Results

V. Conclusion

We introduced the Causal Relational Toeplitz Attention (CRTA) mechanism and integrated it into a hybrid CNN–Transformer framework... CRTA enforces temporal stationarity and causality through a $\mathbf{QK}$-derived Toeplitz kernel interpreted as a causal filter, providing a principled link between adaptive attention and temporal regularization.

References

[1] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention, 2020.

[2] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers, 2022.

[3] Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models, 2023.

[4] Zhen Qin, Xiaodong Han, Weixuan Sun, Bowen He, Dong Li, Dongxu Li, Yuchao Dai, Lingpeng Kong, and Yiran Zhong. Toeplitz neural network for sequence modeling, 2023.

[5] Xuran Pan, Chunjiang Ge, Rui Lu, Shiji Song, Guanfu Chen, Zeyi Huang, and Gao Huang. On the integration of self-attention and convolution, 2022.

[6] Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models, 2023.

[7] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024.