Residual Clock Accumulation and Rotary Decoherence under Time Warping

Anonymous authors
Paper under double-blind review

Abstract

Rotary sequence models rely on an internal positional clock that should remain coherent with the sequence parameter being modeled. We study a failure mode that arises when observed index time is a warped version of the latent sequence clock: fixed or learned clocks can accumulate residual phase error, causing rotary bands to decohere over long offsets. We formulate this as residual-clock accumulation. If a model estimates the true local warp speed $c'(t)$ with residual $\epsilon(t) = \widehat{c}'(t) - c'(t)$, then the relevant quantity at offset u is the accumulated residual $W_u = \sum_{r=t}^{t+u-1} \epsilon(r)$. We derive a two-moment approximation to the decoherence curve from empirical $\mu(u) = \mathbb{E}[W_u]$ and $v(u) = \mathrm{Var}(W_u)$, using the empirical characteristic function only to diagnose moment-closure failure. In controlled synthetic warp generators with known clock speed, we evaluate fixed, static-oracle, MLP, Mamba-Δ-based, and frozen-Δ clocks. We also compare residual-derived band horizons with an explicit single-layer rotary-attention logit diagnostic. The residual band horizon closely tracks the attention-logit horizon, and the two-moment approximation matches attention-logit horizons with median relative error around 8% over all finite crossing pairs in both weakly supervised and forecast-only settings, though errors are larger on adaptive learned clocks alone. Under warp-distribution shift, learned clocks delay the transition from discrimination-limited to warp-coherence-limited behavior.

1 Introduction

“‘tex id="jbc8d7" Position encodings are often treated as architectural details, but in long-range sequence modeling they define an implicit clock. Since attention alone does not encode token order, Transformer models require explicit or implicit positional mechanisms to represent sequence structure (Vaswani et al., 2017; Dufter et al., 2022). This clock-like role becomes especially visible in work on length extrapolation and context-window extension, where positional mechanisms strongly affect whether models can operate beyond the range of positions seen during training or adaptation (Su et al., 2021; Press et al., 2021; Chen et al., 2023; Peng et al., 2023; Ding et al., 2024).

This paper studies a different but related question. Existing context-extension methods typically ask whether a model trained, adapted, or position-rescaled on one range of observed positions can operate at larger index-space contexts. We instead ask whether the model's internal positional clock remains coherent with the latent parameter that generates the sequence. This distinction matters when observed index time is not the natural clock of the underlying signal. In a warped sequence, the latent parameter may evolve nonuniformly relative to index time. A model may then represent long offsets in observed index space while still accumulating phase error relative to the latent clock.

We study this failure mode in rotary sequence representations. Suppose a sequence is generated from a latent signal $z(c(t))$, where $c(t)$ is a monotone warp and $c'(t)$ is the true local clock speed. A model may use a fixed clock, a learned local clock $g(x_t)$, or an input-dependent state-space step size. In our controlled setting, a Mamba-style input-dependent Δ-like quantity provides a natural candidate adaptive clock (Gu & Dao, 2024; Dao & Gu, 2024). In all cases, the relevant quantity is the residual

$$\epsilon(t) = \widehat{c}'(t) - c'(t).$$

The error that affects a rotary offset u is its accumulated window sum

$$W_u(t) = \sum_{r=t}^{t+u-1} \epsilon(r).$$

Thus pointwise clock RMSE alone is not sufficient: a small residual can matter if it accumulates coherently, while a larger fluctuating residual may cancel over long offsets.

We turn this observation into a measurable diagnostic. For a rotary frequency θ_j, residual clock error induces a band error governed by $\sin(\theta_j W_u/2)$. A two-moment closure based on empirical

$$\mu(u) = \mathbb{E}[W_u], \qquad v(u) = \mathrm{Var}(W_u)$$

yields a predicted decoherence curve and horizon. The prediction should work when the first two residual-window statistics dominate the band error, and fail when higher-order or non-Gaussian structure in W_u becomes important. We therefore use the empirical characteristic function of W_u as a diagnostic for moment-closure failure, not as an independent validation metric.

Empirically, we construct synthetic warp generators for which $c'(t)$ is known, train data-dependent clocks under forecasting and weak auxiliary clock supervision, and measure the resulting residuals. We compare two-moment residual predictions with residual-derived rotary band horizons and with an explicit single-layer rotary-attention logit diagnostic, in which fixed random query/key projections are rotated using either the true latent clock or the learned clock. This diagnostic is not a full downstream task evaluation, but it checks that the residual-derived band horizon corresponds to an explicit rotary attention computation rather than only to a closed-form residual statistic. We also test warp-distribution shift to determine when the active bottleneck switches from ordinary rotary discrimination to warp coherence.

The results support residual-clock accumulation as a useful but imperfect diagnostic of rotary decoherence. The residual-derived band horizon closely tracks the explicit attention-logit horizon, and the two-moment approximation matches attention-logit horizons with median relative error around 8% over all finite crossing pairs in both weakly supervised and forecast-only settings. The approximation is less accurate on adaptive learned clocks alone, reflecting structured residual regimes where moment closure is harder. Under warp-distribution shift, learned clocks delay the transition from discrimination-limited to warp-coherence-limited behavior, whereas fixed clocks and frozen Mamba-Δ controls are warp-limited from the start.

Our contribution is a diagnostic account of rotary decoherence under time warping. We identify residual-clock accumulation as a measurable mechanism by which rotary sequence representations lose latent-clock coherence, validate the resulting band diagnostic against an explicit rotary-attention logit computation, and separate ordinary frequency-spectrum discrimination limits from residual-driven warp-coherence limits.

Contributions.

1. We formulate rotary decoherence as a residual-clock accumulation problem, expressing band-level error in terms of $W_u = \sum_{r=t}^{t+u-1} \epsilon(r)$.
2. We derive and evaluate a two-moment horizon approximation based on empirical $\mu(u)$ and $v(u)$, showing that it gives a useful but imperfect estimate of decoherence horizons.
3. We validate the residual-derived band diagnostic against an explicit single-layer rotary-attention logit diagnostic, reducing the concern that the evaluation is only a closed residual-statistics calculation.
4. We use the empirical characteristic function of W_u as a diagnostic for moment-closure failure, separating failures of the two-moment approximation from failures of the residual-decoherence identity.
5. We evaluate fixed, static-oracle, MLP, Mamba-Δ-based, and frozen-Δ clocks under both weakly supervised and forecast-only training.
6. We show that warp-distribution shift induces a transition from discrimination-limited to warp-coherence-limited regimes for learned clocks, while fixed and frozen clocks are warp-limited from the start.

2 Related Work

2.1 Position encodings and long-context extrapolation

A central question in long-context sequence modeling is how positional information should be represented. Early Transformer models used absolute sinusoidal position embeddings (Vaswani et al., 2017), while later work introduced relative position representations directly into self-attention (Shaw et al., 2018). Transformer-XL combined segment-level recurrence with a relative positional encoding scheme to model dependencies beyond a fixed-length context (Dai et al., 2019). RoPE represented positions through rotations, making relative-position dependence explicit in the self-attention formulation (Su et al., 2021).

Recent long-context methods further show that positional mechanisms are central to extrapolation behavior. ALiBi replaces learned position embeddings with distance-dependent attention biases and enables input-length extrapolation (Press et al., 2021). Position Interpolation extends RoPE-based pretrained LLMs by down-scaling position indices to the original context range (Chen et al., 2023). YaRN improves the efficiency of RoPE-based context extension (Peng et al., 2023), while LongRoPE uses non-uniform interpolation and progressive extension to reach much longer context windows (Ding et al., 2024). XPos and recent surveys likewise identify position modeling as a key factor in length extrapolation (Sun et al., 2023; Zhao et al., 2023).

Closest to our discrimination-horizon analysis, Xu et al. (2024) study the relationship between RoPE base and attainable context length through the long-term decay behavior of RoPE. The spectrum-level kernel $K_B(u)$ used in our analysis should be viewed in this line: it captures an ordinary frequency-spectrum discrimination limit. Our contribution is orthogonal to this base-controlled limit. We introduce a separate warp-coherence horizon driven by residual clock error and study which of the two limits becomes active.

2.2 Warped positional encodings and adaptive clocks

Most directly related to our setting is StretchTime (Kim et al., 2026), which formalizes observed-index-versus-latent-clock misalignment and shows that RoPE cannot represent non-affine temporal warping. StretchTime addresses this limitation by proposing Symplectic Positional Embeddings, which generalize RoPE's rotation group and use an input-dependent adaptive warp module for multivariate time-series forecasting.

Our goals are complementary. We do not propose a new positional encoding, and we do not rely on a representability impossibility result. Instead, we take the residual clock error of an arbitrary clock as the object of study. Given a fixed, oracle, MLP, or Mamba-Δ-based clock head, we ask how its accumulated residual W_u accounts for rotary decoherence and when the resulting warp-coherence horizon becomes the active bottleneck. In this sense, StretchTime motivates why adaptive clocks may be needed, while our work quantifies how much decoherence a given clock still incurs.

2.3 State-space models and input-dependent clocks

State-space sequence models provide an alternative to attention for long-sequence modeling. S4 showed that structured state-space models can efficiently model long sequences (Gu et al., 2022). Subsequent work studied how state-space and convolutional alternatives can close the gap with attention in language modeling and long-context tasks (Fu et al., 2023; Poli et al., 2023). Mamba builds on this line by making state-space parameters input-dependent, allowing the model to selectively propagate or forget information depending on the current token while retaining linear scaling in sequence length through a hardware-aware implementation (Gu & Dao, 2024). Mamba-2 and the structured state-space duality framework further connect selective state-space models with attention-like structured matrix views (Dao & Gu, 2024).

The interpretation of input-dependent temporal modulation as an adaptive clock has a precedent in recurrent models. Tallec & Ollivier (2018) show that learnable gates in recurrent networks can provide quasi-invariance to general time transformations of the input. This motivates our use of a Mamba-derived input-dependent Δ-like representation as a candidate adaptive clock in a controlled setting. Our residual-clock diagnostic is complementary: it measures the residual warp error left by such input-dependent adaptation and relates it to

rotary decoherence. The frozen-Δ control separates the marginal scale of the learned step-size representation from its input dependence.

2.4 Time warping, reparameterization, and irregular time

Time warping has long been used to compare sequences whose meaningful temporal coordinates are not aligned with observed index time. Dynamic time warping provides a classical alignment-based formulation for time-normalized sequence comparison (Sakoe & Chiba, 1978). Differentiable variants such as soft-DTW make such alignment objectives usable as learning losses and remain robust to shifts or dilatations along the time dimension (Cuturi & Blondel, 2017).

A related classical line in functional and shape data analysis treats the distinction between latent parameterization and observed sampling as a reparameterization problem. Velocity-based representations such as the square-root velocity function yield reparameterization-invariant metrics and phase–amplitude separation (Srivastava et al., 2011; Srivastava & Klassen, 2016). This line concerns how to represent or align signals invariantly to warping. We instead study the downstream effect of imperfectly estimating the local warp speed on rotary phase coherence: the ideal clock is the latent speed $c'(t)$, and the residual $\epsilon(t) = \widehat{c}'(t) - c'(t)$ measures deviation from it.

A separate line of work models temporal dynamics in continuous time, including neural ordinary differential equations and neural controlled differential equations for irregularly sampled time series (Chen et al., 2018; Kidger et al., 2020). Our use of warping differs from both alignment-based and continuous-time forecasting approaches. We do not learn an alignment objective, and we do not propose a general continuous-time sequence model. Instead, we use controlled synthetic warps with known $c'(t)$ so that residual clock error can be measured directly and related to rotary decoherence.

2.5 Mechanistic diagnostics for learned sequence models

Mechanistic interpretability aims to explain learned model behavior by identifying internal computations rather than only measuring input-output performance. Work on transformer circuits and induction heads studies concrete mechanisms that can explain observable behavior such as in-context learning (Olsson et al., 2022). Recent surveys characterize mechanistic interpretability as an approach that seeks granular, causal accounts of neural network computation (Bereska & Gavves, 2024; Rai et al., 2024).

Our diagnostic is narrower than circuit-level mechanistic interpretability. We do not reverse-engineer a full language model or identify circuits in a pretrained Transformer. Instead, we construct a controlled setting in which the latent clock $c(t)$ and speed $c'(t)$ are known, allowing us to measure residual clock error and its consequences directly. The contribution is a controlled diagnostic for a specific failure mode: residual-clock decoherence under time warping.

3 Residual-Clock Decoherence

3.1 Warped sequence model

Consider a latent periodic signal $z(s) \in \mathbb{R}^d$ observed through a monotone warp $c(t)$:

$$x_t = \phi(z(c(t))) + \eta_t,$$

where t is observed index time, $s = c(t)$ is latent clock time, ϕ is an observation map, and η_t is noise. The true local clock speed is $c'(t)$. A clocked model estimates this local speed by $\widehat{c}'(t)$, producing the residual

$$\epsilon(t) = \widehat{c}'(t) - c'(t).$$

For an offset u, the residual accumulated over the window is

$$W_u(t) = \sum_{r=t}^{t+u-1} \epsilon(r).$$

The distribution of W_u determines whether local clock errors cancel, diffuse, or accumulate coherently. A persistent residual bias can make W_u grow approximately linearly in u, while nearly white residuals lead to approximately diffusive growth. Structured or anticorrelated residuals can produce sublinear or non-Gaussian accumulation.

3.2 Band error from accumulated residual

Let θ_j denote a rotary frequency. At offset u, accumulated clock residual $W_u(t)$ induces a phase error $\theta_j W_u(t)$. For unit-amplitude rotary components, the normalized discrepancy between the correct and residual-shifted phases is proportional to

$$\left|\sin\left(\frac{\theta_j W_u(t)}{2}\right)\right|.$$

We therefore define the normalized RMS band error

$$E_j(u) = \sqrt{\mathbb{E}\left[\sin^2\left(\frac{\theta_j W_u}{2}\right)\right]}.$$

Using $\sin^2(a/2) = (1-\cos a)/2$,

$$E_j(u) = \sqrt{\frac{1-\mathbb{E}[\cos(\theta_j W_u)]}{2}}.$$

The observed decoherence horizon for frequency θ_j and threshold δ is

$$u_j^{\mathrm{obs}}(\delta) = \inf\{u : E_j(u) \geq \delta\}.$$

3.3 Two-moment closure

Let

$$\mu(u) = \mathbb{E}[W_u], \qquad v(u) = \mathrm{Var}(W_u).$$

If W_u is approximately Gaussian, then

$$\mathbb{E}[e^{i\theta_j W_u}] = \exp\left(i\theta_j \mu(u) - \frac{1}{2}\theta_j^2 v(u)\right),$$

and hence

$$\mathbb{E}[\cos(\theta_j W_u)] \approx \exp\left(-\frac{1}{2}\theta_j^2 v(u)\right)\cos(\theta_j \mu(u)).$$

This yields the two-moment prediction

$$\widehat{E}_{\mu v,j}(u) = \sqrt{\frac{1-\exp\left(-\frac{1}{2}\theta_j^2 \widehat{v}(u)\right)\cos(\theta_j \widehat{\mu}(u))}{2}}.$$

The corresponding predicted horizon is

$$\widehat{u}_{\mu v,j}(\delta) = \inf\{u : \widehat{E}_{\mu v,j}(u) \geq \delta\}.$$

This prediction is a moment-closure approximation rather than an identity. It can fail when residual window sums are non-Gaussian, multimodal, skewed, or strongly correlated.

3.4 Characteristic-function diagnostic

To distinguish failures of the moment closure from failures of the residual-decoherence identity, we also compute the empirical characteristic coherence

$$\widehat{C}_j(u) = \left|\frac{1}{N_u}\sum_t \exp(i\theta_j W_u(t))\right|,$$

and the corresponding diagnostic curve

$$\widehat{E}_{\mathrm{char},j}(u) = \sqrt{\frac{1 - \widehat{C}_j(u)}{2}}.$$

This curve is not an independent validation metric: it is computed from the same residual window sums W_u as the observed band error. Its role is diagnostic. When the two-moment curve fails but the characteristic-function curve agrees with the observed band error, the discrepancy is localized to the Gaussian closure rather than to the residual-clock formulation.

3.5 Discrimination horizon and warp-coherence horizon

Rotary representations also have an ordinary discrimination horizon determined by their frequency spectrum. For a RoPE base B and dimension d, the frequencies are

$$\theta_j = B^{-2j/d}, \qquad j = 0, \ldots, d/2 - 1.$$

We define a spectrum-level discrimination kernel

$$K_B(u) = \frac{2}{d} \sum_{j=0}^{d/2-1} \cos(\theta_j u),$$

and a discrimination horizon

$$u_{\mathrm{disc}}(B; \tau) = \inf\{u : K_B(u) \leq \tau\}.$$

This horizon depends only on the RoPE spectrum and the chosen threshold.

The warp-coherence horizon is instead driven by residual clock error and is measured at a probe frequency θ:

$$u_{\mathrm{warp}} = \inf\{u : E_\theta(u) \geq \delta\}.$$

The usable horizon is the smaller of the two:

$$u_{\mathrm{use}} = \min(u_{\mathrm{disc}}, u_{\mathrm{warp}}).$$

This separation is central to our interpretation. RoPE base changes the ordinary discrimination horizon, while clock quality and warp shift determine the warp-coherence horizon. Our experiments study which of these two limits becomes active.

4 Experimental Setup

4.1 Synthetic warp generators

We use controlled synthetic sequences because the true local clock speed $c'(t)$ must be known in order to measure residual clock error. Each sequence instantiates the warped model above by sampling a monotone clock $c(t)$, evaluating a multi-frequency latent signal $z(c(t))$, and adding observation noise. The latent signal contains multiple sinusoidal components so that phase coherence can be evaluated at different rotary frequencies.

We use three warp families.

- **Affine**: sequence-specific constant-speed warps. These are absorbed by a per-sequence static clock and serve as a sanity check.
- **Orbit-in**: smooth non-affine warps whose speed profiles remain within the training support. These require local adaptation but do not strongly shift the speed distribution.

- **Outside**: non-affine warps with stronger local speed variation. These create structured residuals and test whether learned clocks remain coherent outside the easiest regime.

The affine family verifies that the metric behaves correctly when the true speed can be represented by a constant clock. The non-affine families test whether data-dependent clocks reduce local residuals and whether the remaining residuals predict decoherence horizons. Full generator parameters, including amplitude ranges, frequency ranges, and shift schedules, are reported in Appendix B.1.

4.2 Clock models

We evaluate five clock families.

Fixed clock. The fixed clock uses a constant unit speed and performs no input-dependent adaptation. It is the baseline for measuring residual clock mismatch.

Static oracle. The static oracle uses the average true speed of each sequence. It is not a deployable model, but it is an important control: it should solve affine warps while failing to represent non-affine local speed variation.

MLP clocks. The `mlp_low` and `mlp_mid` clocks map local input features to positive speed estimates. They test whether a direct feed-forward local predictor can reduce residual clock error without a state-space mechanism.

Mamba-Δ clock. We use the official Mamba implementation and extract an input-dependent Δ-like representation from the selective state-space block. A lightweight clock head maps this representation to a positive local speed estimate. This construction is used as a controlled probe of input-dependent clocking, not as a claim that Mamba was designed to estimate physical or latent time.

Frozen-Δ control. The frozen-Δ control replaces the input-dependent Mamba-Δ speed with its train-set constant mean. This preserves the marginal scale of the learned step size while removing input dependence. If the benefit of Mamba-Δ comes from adaptive clocking rather than scale alone, the frozen control should behave similarly to a fixed clock.

4.3 Training objectives

The full learned-clock setting uses

$$\mathcal{L} = \mathcal{L}_{\text{forecast}} + \lambda_{\text{clock}} \text{MSE}(\widehat{c}', c') + \lambda_{\text{smooth}} \mathcal{L}_{\text{smooth}}.$$

We use weak clock supervision with $\lambda_{\text{clock}} = 0.01$. This setting is designed to produce learned residuals whose accumulation can be tested against decoherence horizons; it is not intended as a claim of fully unsupervised clock discovery. We also run a forecast-only ablation with

$$\lambda_{\text{clock}} = 0, \qquad \lambda_{\text{smooth}} = 0,$$

which tests whether the residual-to-horizon relationship persists without direct supervision of $c'(t)$.

4.4 Evaluation metrics

For each model, warp family, and frequency θ, we compute:

- clock residual RMSE;
- residual accumulation statistics $\widehat{\mu}(u)$ and $\widehat{v}(u)$;
- directly observed band-error horizon u^{obs};

- two-moment predicted horizon $\widehat{u}_{\mu v}$;
- characteristic-function diagnostic horizon $\widehat{u}_{\mathrm{char}}$;
- phase-forecast proxy MSE.

For horizon prediction, the primary metric is agreement between $\widehat{u}_{\mu v}$ and u^{obs}. We report R^2 against the identity line $y = x$ and median relative error:

$$\mathrm{MedRelErr} = \mathrm{median}\frac{|\widehat{u}_{\mu v} - u^{\mathrm{obs}}|}{\max(u^{\mathrm{obs}}, 10^{-9})}.$$

Characteristic-function statistics are reported separately as diagnostics.

4.5 Implementation and reproducibility

Experiments are run with PyTorch, CUDA, the official `mamba-ssm` package, and `causal-conv1d`. The reported runs use a fixed seed. Because fused CUDA kernels may not guarantee bitwise determinism, we treat the reported numbers as fixed-seed estimates and focus on patterns that are stable across controls, ablations, and shift criteria.

5 Experiments

We evaluate the residual-clock account in two stages. First, we test whether residual window statistics predict directly observed rotary decoherence horizons in learned clocks. This stage evaluates the two-moment prediction, the forecast-only ablation, and controls that separate fixed, oracle, feed-forward, Mamba-Δ, and frozen-Δ clocks. Second, we test whether warp-distribution shift changes the active bottleneck from ordinary rotary discrimination to warp coherence. Throughout, the two-moment prediction is the primary validation metric, and characteristic-function curves are used only as diagnostics for moment-closure failure.

5.1 Residual horizons in learned clocks

Table 1 summarizes the horizon prediction results. The primary prediction is the two-moment residual prediction based on empirical $\widehat{\mu}(u)$ and $\widehat{v}(u)$. We evaluate it against two targets. The first is the residual-derived rotary band horizon from Section **??**. The second is an explicit single-layer rotary-attention logit horizon: fixed random query/key projections are rotated using either the true latent clock or the learned clock, and we measure offset-wise logit error.

The attention-logit diagnostic is not a full downstream task evaluation, but it is not computed from the closed-form band-error identity. It therefore checks whether the residual-derived band horizon corresponds to an explicit rotary attention computation. In both weakly supervised and forecast-only settings, the residual-derived band horizon closely matches the attention-logit horizon, with median relative error around 1.1–1.3%. The two-moment prediction then matches the attention-logit horizon with median relative error around 8% over all finite crossing pairs. On adaptive learned clocks alone, the error is higher, around 15–17%, reflecting structured residual regimes where moment closure is less accurate.

The near-equivalence between residual-derived band horizons and attention-logit horizons is also strong quantitatively. The residual-band horizon matches the attention-logit horizon with $R^2 = 0.9997$ and median relative error 1.28% in the weakly supervised setting, and $R^2 = 0.9995$ with median relative error 1.14% in the forecast-only setting. This supports using the residual band horizon as a compact diagnostic for rotary logit decoherence, while preserving the distinction between this diagnostic and full downstream performance.

5.2 Controls and residual structure

The controls behave as expected. The static oracle absorbs affine speed almost exactly, while fixed and frozen clocks remain inaccurate. Mamba-Δ substantially reduces residual error relative to fixed clocks in affine and orbit-in regimes, and remains competitive in the harder outside-warp regime.

Table 1: Horizon prediction summary. The attention-logit horizon is measured by an explicit single-layer rotary-attention diagnostic using true-clock and learned-clock rotary phases. Characteristic-function curves are diagnostic only and are reported separately.

Setting	Target / subset	Crossing pairs	R^2 vs. $y = x$	Median rel. error
Weak clock supervision	residual band, all pairs	59	0.916	7.86%
Weak clock supervision	attention logit, all pairs	59	0.919	7.96%
Weak clock supervision	attention logit, adaptive clocks	19	0.791	16.94%
Forecast-only	residual band, all pairs	55	0.871	7.89%
Forecast-only	attention logit, all pairs	55	0.872	8.00%
Forecast-only	attention logit, adaptive clocks	15	0.644	14.68%

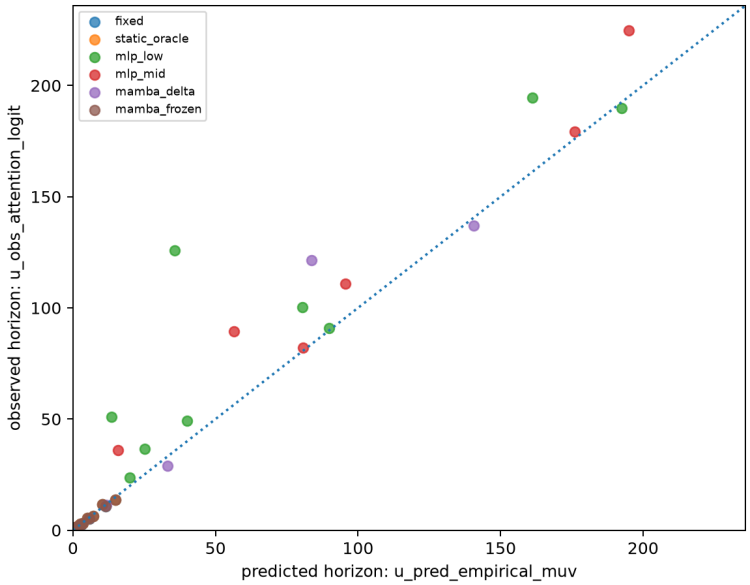


Figure 1: Two-moment predicted horizon versus explicit rotary-attention logit horizon. The prediction uses empirical $\widehat{\mu}(u)$ and $\widehat{v}(u)$, while the target is measured from a single-layer rotary-attention logit diagnostic.

The two-moment prediction fails most clearly in structured residual regimes. A representative case is the outside-warp Mamba-Δ model at $\theta = 4$:

$$\widehat{u}_{\mu v} \approx 236, \qquad \widehat{u}_{\mathrm{char}} \approx 32, \qquad u^{\mathrm{obs}} \approx 32.$$

The characteristic-function diagnostic agrees with the observed band error, while the two-moment prediction overestimates the horizon. This localizes the discrepancy to the Gaussian moment closure rather than to the residual-decoherence identity. Additional residual-structure diagnostics, including variance-growth slopes and the full characteristic-function scatter, are reported in Appendix C.

5.3 Warp shift and active bottlenecks

We next test when warp coherence becomes the active bottleneck. Clocks are trained on a mild outside-warp distribution and evaluated under increasing warp-distribution shift. The shift parameter $s \in [0, 1]$ increases warp amplitude and changes the speed profile; we measure shift strength using the Wasserstein distance between train and test distributions of $c'(t)$.

Table 2: Representative full-setting clock errors and phase-forecast proxy errors. Values are reported from the fixed-seed full run. The affine rows verify that the static oracle absorbs constant-speed warps, while the non-affine rows show the benefit of data-dependent clocks.

Warp family	Model	Speed RMSE	Phase proxy MSE
Affine	fixed	0.20531	8.29×10^{-3}
Affine	static oracle	1.23×10^{-7}	1.29×10^{-11}
Affine	mamba_delta	0.00246	1.20×10^{-6}
Affine	mamba_frozen	0.20513	8.28×10^{-3}
Orbit-in	mlp_mid	0.02375	1.02×10^{-4}
Orbit-in	mamba_delta	0.00553	5.29×10^{-6}
Outside	mlp_mid	0.01247	6.21×10^{-5}
Outside	mamba_delta	0.01451	1.01×10^{-5}

Table 3: First warp-binding shift $s^\dagger$ using the observed direct band-error criterion. Fixed and frozen clocks are warp-limited from the start, while learned clocks transition as shift increases.

Model	Base 10^3	Base 10^4	Base 10^5	Base 10^6
fixed	0.0	0.0	0.0	0.0
mamba_frozen	0.0	0.0	0.0	0.0
mlp_low	0.5	0.3	0.1	0.0
mlp_mid	0.8	0.6	0.6	0.4
mamba_delta	0.7	0.4	0.4	0.4

For each shift level and RoPE base, we compute

$$u_{\text{use}} = \min(u_{\text{disc}}, u_{\text{warp}}).$$

We report three criteria for u_{warp}: the two-moment prediction, the characteristic-function diagnostic, and the directly observed band-error horizon. The observed criterion is the primary bottleneck measure.

Table 3 reports the first shift $s^\dagger$ at which the observed warp horizon becomes smaller than the discrimination horizon. Fixed and frozen clocks are warp-limited from the start. Learned clocks remain discrimination-limited near the training distribution and become warp-limited only after sufficient shift.

The transition is stable across criteria. For Mamba-Δ, the μ, v, characteristic, and observed criteria give similar transition points:

criterion	10^3	10^4	10^5	10^6
μ, v	0.7	0.5	0.4	0.4
char	0.7	0.4	0.4	0.4
obs	0.7	0.4	0.4	0.4

Although characteristic-function horizons are diagnostic rather than independent validation metrics, their agreement with observed transition points supports the qualitative conclusion that the bottleneck switch is not an artifact of the two-moment approximation. Full μ, v and characteristic-function transition tables are reported in Appendix D.

For a representative Mamba-Δ base-10^6 trace, the observed transition occurs between $s = 0.3$ and $s = 0.4$, where the warp horizon drops below the discrimination horizon. Detailed traces, residual-RMSE curves, and downstream phase-proxy errors are reported in Appendix D.

This experiment clarifies the role of RoPE base. In our setup, base changes the discrimination horizon u_{disc}, while the warp-coherence horizon is measured at a fixed probe frequency. The result should therefore not be read as evidence that increasing base directly increases the warp horizon. Instead, base controls discrimination, while clock quality and warp shift control warp coherence.

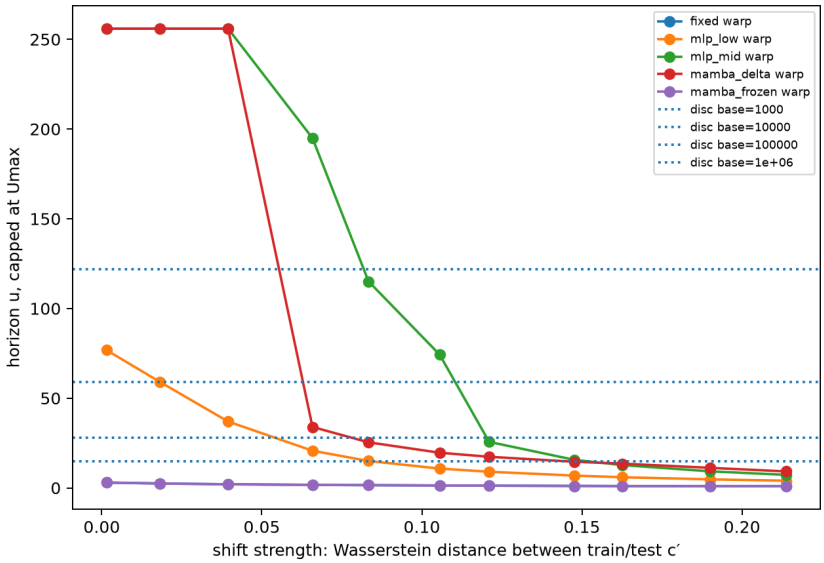


Figure 2: Observed warp horizon versus discrimination horizons under increasing warp shift. Fixed and frozen clocks are warp-limited from the start. Learned clocks transition from discrimination-limited to warp-limited regimes as shift increases.

6 Discussion

6.1 Residual-clock accumulation as a mechanism

The results support a specific mechanism claim: rotary decoherence under time warping is governed by accumulated residual clock error. The key object is not the pointwise residual $\epsilon(t)$ alone, but its window sum $W_u = \sum_{r=t}^{t+u-1} \epsilon(r)$. This explains why clock errors with similar RMSE can lead to different decoherence horizons: biased or correlated residuals can accumulate coherently, while fluctuating residuals may cancel over longer offsets.

The two-moment approximation provides a useful summary of this mechanism, but it is not a complete description. It matches explicit attention-logit horizons well over all finite crossing pairs, but is less accurate on adaptive learned clocks alone. When residual-window distributions are structured, non-Gaussian, or dominated by higher-order effects, $\mu(u)$ and $v(u)$ are insufficient. In such cases, the characteristic-function diagnostic localizes the failure to the moment closure rather than to the residual-decoherence identity itself.

6.2 When adaptive clocks matter

The shift experiment clarifies when adaptive clocking is useful. Near the training warp distribution, learned clocks can remain limited by the ordinary rotary discrimination horizon. Under stronger warp shift, residual clock error grows and the warp-coherence horizon becomes the active bottleneck. Fixed clocks and frozen Mamba-Δ controls are warp-limited from the start, while learned clocks delay this transition.

This distinction also clarifies the role of RoPE base. In our setup, changing the base changes the discrimination horizon, not the underlying residual-clock process. Warp coherence is controlled by the learned clock and by the severity of the warp shift. The usable horizon is therefore determined by whichever of these two limits becomes active first.

6.3 Scope and limitations

The paper does not claim a new state-of-the-art sequence model, nor does it show improved real-language long-context performance. The contribution is a controlled diagnostic: it identifies a measurable failure mode and tests whether its predicted consequences appear in learned-clock models.

The main limitation is that the experiments are synthetic. This design is necessary because the true local speed $c'(t)$ must be known in order to measure residual clock error directly. It also limits the extent to which the results can be transferred to natural language or real-world time-series tasks without additional validation. A natural next step is to construct semi-synthetic or real-data settings where an interpretable latent clock can be estimated or externally specified.

A second limitation is numerical reproducibility. The reported runs use fixed seeds and record environment metadata, but official Mamba and fused CUDA kernels may not guarantee bitwise determinism. Future work should report multi-seed aggregates for all headline values and deterministic configurations where supported. Finally, the two-moment prediction is only a closure approximation. Higher-cumulant or nonparametric residual-window approximations may improve horizon prediction in structured residual regimes.

7 Conclusion

We studied rotary decoherence under time warping as a residual-clock accumulation problem. When a model estimates the true local speed $c'(t)$ with residual $\epsilon(t) = \widehat{c}'(t) - c'(t)$, the accumulated residual W_u determines the phase error seen by a rotary band at offset u. This yields a compact diagnostic: decoherence horizons can be estimated from the statistics of residual window sums, subject to the accuracy of the chosen moment closure.

Controlled experiments support this diagnostic view. A two-moment approximation based on empirical $\mu(u)$ and $v(u)$ gives useful but imperfect horizon estimates, and the residual-derived band horizon closely tracks an explicit single-layer rotary-attention logit horizon. Forecast-only ablations show that the relationship is not solely an artifact of direct clock supervision. Under warp-distribution shift, learned clocks delay the transition from discrimination-limited to warp-coherence-limited behavior, while fixed and frozen clocks are warp-limited from the start.

These results identify residual-clock accumulation as a concrete mechanism by which rotary sequence representations can fail under time warping. More broadly, they suggest that long-context behavior should be evaluated not only by index-space extrapolation, but also by whether the model's internal clock remains coherent with the latent parameterization of the sequence.

A Additional Mathematical Details

A.1 Direct Band Error Identity

For a rotary band with frequency θ, accumulated residual W_u induces phase error θW_u. The normalized RMS band error is

$$E(u) = \sqrt{\mathbb{E}\left[\sin^2\left(\frac{\theta W_u}{2}\right)\right]}.$$

Using

$$\sin^2(a/2) = \frac{1-\cos a}{2},$$

we obtain

$$E(u) = \sqrt{\frac{1-\mathbb{E}[\cos(\theta W_u)]}{2}}.$$

A.2 Gaussian Moment Closure

If $W_u \sim \mathcal{N}(\mu(u), v(u))$, then

$$\mathbb{E}[e^{i\theta W_u}] = e^{i\theta\mu(u)-\theta^2 v(u)/2}.$$

Taking the real part gives

$$\mathbb{E}[\cos(\theta W_u)] = e^{-\theta^2 v(u)/2}\cos(\theta\mu(u)).$$

This yields the two-moment prediction used in the main text.

A.3 Why Characteristic Curves Are Not Independent Validation

The empirical characteristic function is

$$\widehat{\varphi}(\theta; u) = \frac{1}{N_u}\sum_t e^{i\theta W_u(t)}.$$

Its modulus is

$$|\widehat{\varphi}(\theta; u)| = \sqrt{\left(\frac{1}{N_u}\sum_t \cos(\theta W_u(t))\right)^2 + \left(\frac{1}{N_u}\sum_t \sin(\theta W_u(t))\right)^2}.$$

The observed band error depends on

$$\frac{1}{N_u}\sum_t \cos(\theta W_u(t)).$$

When the sine average is near zero and the cosine average is nonnegative, the two quantities are close. Therefore, the characteristic-function curve is expected to closely track the observed band-error curve in symmetric residual regimes. This is why we use it only as a diagnostic for moment-closure failure.

B Additional Experimental Details

B.1 Synthetic Generator Details

Each sequence is generated on an index grid $t = 0, \ldots, T-1$ with $T = 512$. We first sample a positive local speed sequence $c'(t)$, normalize it to have mean one when appropriate, and define the latent clock by cumulative summation,

$$c(t) = \sum_{r=0}^{t} c'(r).$$

The latent signal is a six-dimensional multi-frequency sinusoidal signal

$$\begin{aligned} z(c)_1 &= \sin(c+\psi), & z(c)_2 &= \cos(c+\psi), \\ z(c)_3 &= 0.35\sin(0.5c+0.7\psi), & z(c)_4 &= 0.35\cos(0.5c+0.7\psi), \\ z(c)_5 &= 0.20\sin(2c-0.3\psi), & z(c)_6 &= 0.20\cos(2c-0.3\psi), \end{aligned}$$

where the sequence phase $\psi \sim \mathrm{Unif}(0, 2\pi)$. The observed signal z_{obs} is obtained by adding Gaussian observation noise. Unless otherwise stated, the noise scale is 0.006; for affine Exp A sequences, we cap the noise at 0.002 to keep the oracle sanity check sharp.

The model input feature vector concatenates the noisy signal, first differences, second differences, and local difference norms:

$$x_t = \left[z_{\mathrm{obs}}(t), \Delta z_{\mathrm{obs}}(t), \Delta^2 z_{\mathrm{obs}}(t), \|\Delta z_{\mathrm{obs}}(t)\|_2, \|\Delta z_{\mathrm{obs}}(t)\|_2^2, \|\Delta^2 z_{\mathrm{obs}}(t)\|_2\right],$$

giving 21 input dimensions.

We use three warp families. The affine family samples a sequence-specific constant speed

$$c'(t) = a, \qquad a \sim \mathrm{Unif}(0.65, 1.35).$$

The orbit-in family samples

$$\tilde{c}'(t) = (t+b)^{\alpha-1}, \qquad \alpha \sim \mathrm{Unif}(0.45, 1.75), \qquad b \sim \mathrm{Unif}(20, 80),$$

clips speeds below 0.03, and normalizes the resulting speed profile to mean one. The outside family samples

$$\tilde{c}'(t) = 1 + A\sin(2\pi t/P + \varphi) + 0.15A\sin(4\pi t/P + 0.5\varphi),$$

where $\varphi \sim \mathrm{Unif}(0, 2\pi)$, clips speeds below 0.03, and normalizes to mean one. In Exp A, the outside family uses

$$A \sim \mathrm{Unif}(0.10, 0.35), \qquad P \sim \mathrm{Unif}(48, 128).$$

Exp A uses $n_{\mathrm{train}} = 1024$, $n_{\mathrm{val}} = 128$, $n_{\mathrm{test}} = 256$, $U_{\max} = 256$, and probe frequencies

$$\theta \in \{0.5, 1, 2, 4, 8\}.$$

Exp B trains on a mild outside-warp distribution with

$$A \sim \mathrm{Unif}(0.05, 0.20), \qquad P \sim \mathrm{Unif}(64, 128),$$

using $n_{\mathrm{train}} = 1536$, $n_{\mathrm{val}} = 192$, and $n_{\mathrm{test}} = 256$. At test time, the shift parameter $s \in \{0, 0.1, \ldots, 1.0\}$ changes the outside-warp distribution according to

$$A \sim \mathrm{Unif}(0.05 + 0.30s,\ 0.20 + 0.35s),$$

$$P \sim \mathrm{Unif}(64 - 40s,\ 128 - 80s),$$

and the observation-noise scale becomes $0.006 + 0.010s$. We quantify the shift strength using the Wasserstein distance between empirical train and test distributions of $c'(t)$. Exp B uses binding frequency $\theta = 4$, $U_{\max} = 256$, and RoPE bases $10^3, 10^4, 10^5, 10^6$.

B.2 Mamba-Δ Clock Extraction

The Mamba-Δ clock uses the official `mamba-ssm` implementation. We extract the internal input-dependent Δ-like representation from the selective state-space block and pass it through a lightweight clock head to obtain a positive local speed estimate $\hat{c}'(t)$. The extraction is used only for analysis and controlled comparison. It is not intended to modify the Mamba architecture or claim that Mamba directly learns physical time.

The frozen-Δ control replaces the input-dependent speed sequence with a constant train-set mean. This control tests whether performance depends on input-dependent adaptation rather than the global scale of the extracted step-size representation.

B.3 Training Details

The full learned-clock runs use:

$$\lambda_{\text{clock}} = 0.01, \qquad \lambda_{\text{smooth}} = 10^{-4}.$$

The forecast-only ablation uses:

$$\lambda_{\text{clock}} = 0, \qquad \lambda_{\text{smooth}} = 0.$$

The full Exp A setting uses $T = 512$, $n_{\text{train}} = 1024$, $n_{\text{val}} = 128$, $n_{\text{test}} = 256$, and 120 epochs. The forecast-only ablation uses the same data scale with 160 epochs. Exp B uses $n_{\text{train}} = 1536$, $n_{\text{val}} = 192$, $n_{\text{test}} = 256$, and 140 epochs.

C Additional Exp A Diagnostics

C.1 Residual accumulation regimes

Figure 3 reports residual accumulation slopes across models and warp families. These slopes are computed from the relationship between $\log \mathrm{Var}(W_u)$ and $\log u$. The wide range of slopes shows that residuals are not governed by a single accumulation law: some regimes are closer to bias-like accumulation, some are closer to diffusive accumulation, and others are sublinear or poorly fit by a simple power law. This motivates measuring $\widehat{\mu}(u)$ and $\widehat{v}(u)$ directly rather than assuming a fixed residual process.

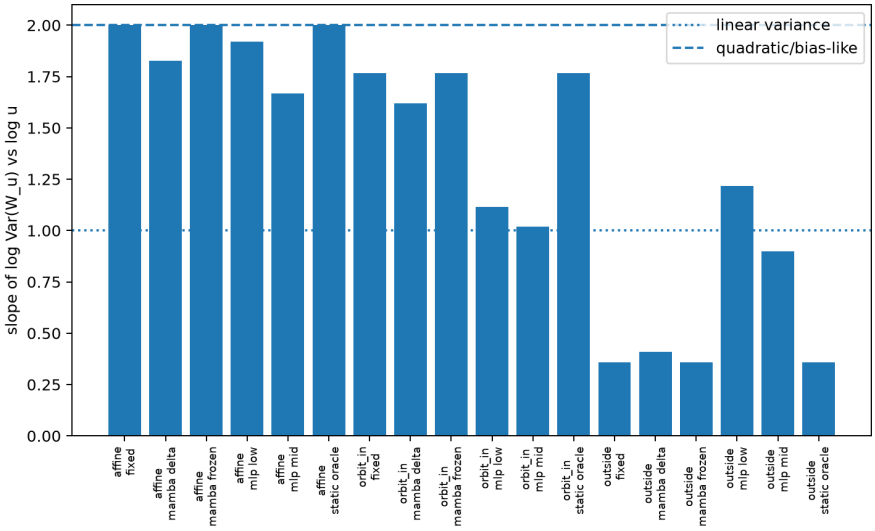


Figure 3: Residual accumulation regimes. The slope of $\log \mathrm{Var}(W_u)$ versus $\log u$ varies substantially across models and warp families.

C.2 Characteristic-function diagnostic

Figure 4 shows the characteristic-function diagnostic against the observed band-error horizon. This plot is not an independent validation result because both quantities are computed from the same residual window sums W_u. Its purpose is to diagnose whether failures of the two-moment prediction arise from moment closure rather than from the residual-decoherence identity.

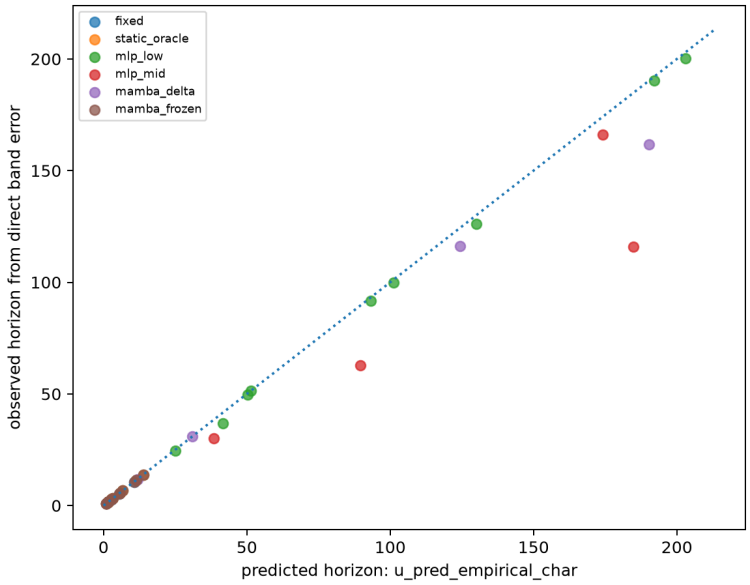


Figure 4: Characteristic-function diagnostic versus observed band-error horizon. The curve is used only to localize moment-closure failures.

D Additional Warp-Shift Diagnostics

D.1 Alternative transition criteria

Tables 4 and 5 report the first warp-binding shift $s^\dagger$ using the two-moment and characteristic-function criteria. The observed criterion is used as the main result in the paper; these tables show that the qualitative transition pattern is stable across criteria.

Table 4: First warp-binding shift $s^\dagger$ using the μ, v two-moment criterion.

Model	Base 10^3	Base 10^4	Base 10^5	Base 10^6
fixed	0.0	0.0	0.0	0.0
mamba_frozen	0.0	0.0	0.0	0.0
mlp_low	0.4	0.3	0.1	0.0
mlp_mid	0.7	0.6	0.6	0.4
mamba_delta	0.7	0.5	0.4	0.4

Table 5: First warp-binding shift $s^\dagger$ using the characteristic-function diagnostic criterion. This table is diagnostic and should not be interpreted as an independent validation result.

Model	Base 10^3	Base 10^4	Base 10^5	Base 10^6
fixed	0.0	0.0	0.0	0.0
mamba_frozen	0.0	0.0	0.0	0.0
mlp_low	0.7	0.5	0.2	0.0
mlp_mid	0.8	0.6	0.6	0.5
mamba_delta	0.7	0.4	0.4	0.4

D.2 Representative Mamba-Δ trace

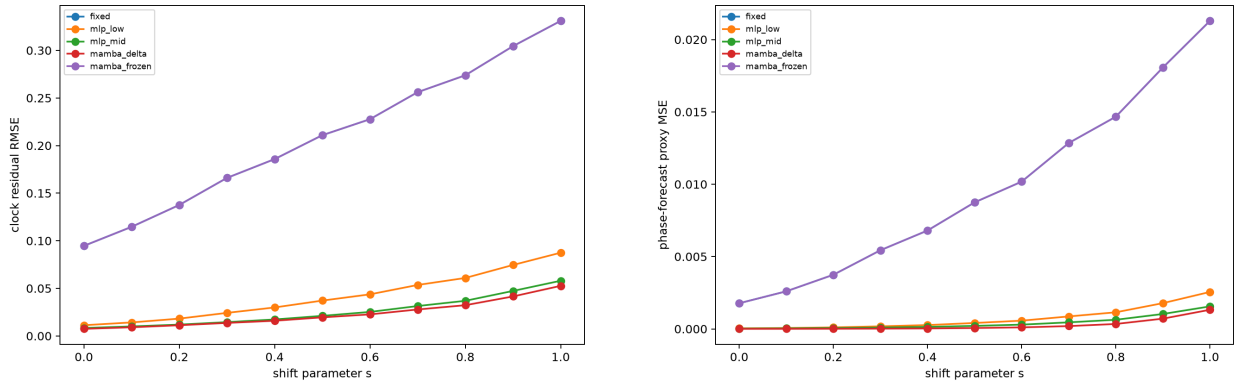
Table 6 reports a representative Mamba-Δ trace at base 10^6. The active bottleneck switches from discrimination to warp coherence between $s = 0.3$ and $s = 0.4$.

Table 6: Mamba-Δ behavior at base 10^6. The active bottleneck switches from discrimination to warp coherence as shift increases.

s	u_{disc}	$\widehat{u}_{\mu v}$	$\widehat{u}_{\text{char}}$	u^{obs}	Active obs.	Speed RMSE
0.0	121.9	193.0	256.0	193.8	disc	0.00725
0.3	121.9	125.3	129.6	125.6	disc	0.01338
0.4	121.9	30.3	26.3	26.3	warp	0.01571
0.7	121.9	14.1	14.5	14.3	warp	0.02823
1.0	121.9	7.0	9.1	8.9	warp	0.05355

D.3 Residual error and downstream proxy

Figure 5 shows residual RMSE and phase-forecast proxy MSE under increasing warp-distribution shift. The proxy degrades as shift increases, consistent with the interpretation that growing residual clock error reduces warp coherence.



(a) Residual RMSE

(b) Phase proxy MSE

Figure 5: Residual error and downstream proxy error increase under warp-distribution shift. Learned clocks, especially Mamba-Δ, remain substantially better than fixed and frozen controls.

References

Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety—a review. *Transactions on Machine Learning Research*, 2024. URL `https://openreview.net/forum?id=ePUVetPKu6`.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation, 2023.

Marco Cuturi and Mathieu Blondel. Soft-dtw: A differentiable loss function for time-series. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 894–903, 2017.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, 2019.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality, 2024.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Philipp Dufter, Martin Schmitt, and Hinrich Schütze. Position information in transformers: An overview. *Computational Linguistics*, 48(3):733–763, 2022.

Daniel Y. Fu, Tri Dao, Khaled K. Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. In *International Conference on Learning Representations*, 2023.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling*, 2024.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *The International Conference on Learning Representations*, 2022.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Yubin Kim, Viresh Pati, Jevon Twitty, Vinh Pham, Shihao Yang, and Jiecheng Lu. StretchTime: Adaptive time series forecasting via symplectic attention. *arXiv preprint arXiv:2602.08983*, 2026.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

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

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. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation, 2021.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models, 2024.

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.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 464–468, 2018.

Anuj Srivastava and Eric P. Klassen. *Functional and Shape Data Analysis*. Springer, 2016. doi: 10.1007/978-1-4939-4020-2.

Anuj Srivastava, Eric Klassen, Shantanu H. Joshi, and Ian H. Jermyn. Shape analysis of elastic curves in Euclidean spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(7):1415–1428, 2011. doi: 10.1109/TPAMI.2010.184.

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

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. A length-extrapolatable transformer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

Corentin Tallec and Yann Ollivier. Can recurrent neural networks warp time? In *International Conference on Learning Representations*, 2018.

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

Mingyu Xu, Xin Men, Bingning Wang, Qingyu Zhang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. Base of RoPE bounds context length. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Liang Zhao, Xiachong Feng, Xiaocheng Feng, Weihong Zhong, Dongliang Xu, Qing Yang, Hongtao Liu, Bing Qin, and Ting Liu. Length extrapolation of transformers: A survey from the perspective of positional encoding, 2023.