BEHIND ZERO-SHOT FORECASTING: HOW DO TIME-SERIES FOUNDATION MODELS USE THEIR CONTEXT?

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series foundation models (TSFMs) leverage large-scale pretraining to achieve cross-domain generalization in forecasting. Among them, the zero-shot paradigm generates predictions using only the observed context, without any task-specific training or fine-tuning. This enables immediate deployment, but leaves a fundamental question open: how do these models use their context to produce forecasts? We investigate this in-context forecasting mechanism directly. Based on an architecture-agnostic framework that characterizes context use through controlled input interventions and their effects on the forecast, we establish three findings. First, TSFM forecasts rule out context parroting, indicating that the models perform inference rather than merely interpolate among values already observed in the context. Second, context influence is governed primarily by temporal position instead of phase-space distance, while similarity-based sample selection can admit a local-approximation error bound no larger than recency-based sample selection. Third, despite its influence decay with temporal distance in the majority of models, distant context can remain functionally important by providing temporal structure beyond distributional statistics. Collectively, these results reveal both the functional strengths and intrinsic limitations of current TSFMs in context utilization, providing a mechanistic basis for diagnosing zero-shot forecasting behavior and informing the design of future models and training strategies.

1 INTRODUCTION

Time-series foundation models (TSFMs) are pretrained on large, heterogeneous collections of time series and then reused across forecasting tasks (Ansari et al., 2024; Das et al., 2024; Woo et al., 2024). Some are adapted to a target domain through post-training or fine-tuning, whereas others are designed to forecast with frozen model weights. We study the latter regime, commonly referred to as zero-shot forecasting. In this setting, a model can be applied directly to a previously unseen series, with the context window providing the only information specific to that series. Understanding how the model uses its context is therefore crucial to the forecasting interpretability.

Recent work has shown both the promise of zero-shot forecasting and an unresolved puzzle about its mechanism. On chaotic systems, zero-shot TSFMs can produce forecasts competitive with models trained directly on the target system and can preserve properties of the underlying attractor beyond the pointwise prediction horizon (Zhang & Gilpin, 2025). At the same time, a parameter-free method that searches the context for a state similar to the current one and copies its continuation can outperform several leading TSFMs on the same task (Zhang & Gilpin, 2026). These results demonstrate that the context contains useful predictive information, but how that information is used remains unclear: while forecasting accuracy compares outcomes, it does not identify the mechanism that produces them.

This leaves three successive questions about the model's in-context behavior. First, what does the model actually do with the context? Does it parrot the observed series by retrieving previous continuations and interpolating among them, or does it perform inference beyond values already available in the context? Second, how does the model use the samples in the context? Are their weights determined by similarity to the current state or by temporal position, and are these weights distributed

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.

broadly across the context or concentrated near the prediction point? Finally, if distant samples have only weak influence, are they still necessary, and if so, what information do they provide? Do they contribute only distributional statistics, or does their temporal organization also matter?

To address these questions, we intervene directly on the input and measure the resulting changes in the forecast. We formalize local context sensitivity through an effective weight profile and derive four diagnostic criteria from a common local-approximation framework in delay coordinates. Each criterion is calibrated against reference forecasters whose behavior is known by construction, allowing the measurement to be validated before it is interpreted. The results demonstrate three key findings. First, TSFM forecasts escape the convex hull of retrieved continuations, ruling out context parroting and indicating that the models perform inference rather than merely interpolate among observed values. Second, context influence is governed primarily by temporal position rather than phase-space distance and is strongly concentrated near the prediction point in most models. This context utilization is also suboptimal under the local-approximation framework. Finally, although responses to localized perturbations decay with temporal distance in most models, distant context can remain necessary. Removing it degrades forecasting in some models, while disrupting its temporal order fails to preserve its contribution. Thus, the far context provides temporal structure beyond what can be captured by distributional statistics alone.

Our contributions are as follows:

(1) We introduce an architecture-agnostic criterion that distinguishes *context parroting* from *inference* through the geometry of the forecast. This provides a mechanistic characterization of the extrapolation capability of zero-shot TSFMs.

(2) We reveal how TSFMs allocate influence across context samples and derive a corresponding dominance result. We further prove that, at matched neighborhood size and fit order, similarity-based sample selection can admit a local-approximation error bound no larger than recency-based sample selection. This connects the observed weighting mechanism to a concrete and theoretically grounded opportunity for improving context utilization.

(3) We explain why distant context with weak influence can remain functionally important. By testing whether distributional information alone is sufficient, we identify temporal structure as an essential component of long-context use, with implications for synthetic pretraining and data augmentation beyond matching marginal statistics alone.

2 Related work

Pretraining and zero-shot forecasting. Pretrained forecasting models differ in how they are deployed. Some are adapted to the target domain before forecasting through post-training or fine-tuning, whereas others are applied with their model parameters fixed and evaluated on series absent from their pretraining corpus (Ansari et al., 2024; Das et al., 2024; Woo et al., 2024). We focus on the latter regime; Appendix A.1 discusses this distinction in detail. Most released zero-shot TSFMs use Transformer backbones, while state-space (Bhethanabhotla et al., 2024) and xLSTM-based (Auer et al., 2025) models have also been introduced. Zhang & Gilpin (2025) show that zero-shot forecasts of chaotic systems can be competitive with models trained on the target system and that the reconstructed attractor preserves its correlation dimension. They further introduce context parroting ((Zhang & Gilpin, 2026)), a parameter-free retrieval-and-copying forecaster, and show that it can outperform several TSFMs on chaotic forecasting tasks. Retrieval-augmented approaches likewise demonstrate that explicitly supplying similar historical patterns can improve zero-shot forecasting (Ning et al., 2025; Han et al., 2025). These results motivate the question studied here: what different in-context mechanism do TSFMs use to perform a forecast, and what makes similarity-based retrieval beneficial?

Delay embeddings and local approximation. Delay-coordinate reconstruction is a standard tool in nonlinear time-series analysis. Takens' theorem gives conditions under which a delay embedding of a scalar observable is diffeomorphic to the underlying attractor (Takens, 1981), so that local geometry in delay coordinates reflects local geometry in state space. Farmer & Sidorowich (1987) study local approximation and show that fitting a polynomial of order p to neighbors of a query yields a one-step approximation error of order r^{p+1} in the neighborhood radius r. These results provide the

geometric and approximation-theoretic basis for the coverage criterion and the comparison between similarity- and recency-based sample selection developed below.

Mechanistic analyses of in-context learning. In-context learning has been studied in natural language processing where the underlying task family is known, making it possible to ask what computation a model implements from the examples it receives: Von Oswald et al. (2023) provide a concrete algorithmic interpretation in regression, showing that Transformers can implement gradient-descent-like learning within their forward pass; Bhattamishra et al. (2024) study discrete-function learning more broadly and find that Transformers can realize distinct in-context algorithms and adapt their behavior to the sequence of examples provided. For TSFMs, Wiliński et al. (2025) analyze the structure of internal representations, identify learned concepts such as trend and periodicity, and intervene on latent representations to steer forecasts. Besides such behavioral characterization, a complementary line of work asks which internal components causally contribute to the model output. In this setting, attention weights alone need not reflect causal contribution (Jain & Wallace, 2019), whereas activation patching tests internal contribution by replacing an activation and measuring the resulting change in the output (Meng et al., 2022). Our study draws on these mechanistic perspectives for time-series forecasting at the input level: controlled interventions on the context provide a common readout across Transformer, state-space, and xLSTM backbones. We use activation-level analysis as a complementary structural check in Appendix F.

Positioning of this work. We study the in-context behavior of pretrained zero-shot TSFMs with fixed parameters. Rather than asking what information pretraining has encoded in their internal representations or how those representations are organized, we ask how the models use the context of an unseen series at inference time to produce a forecast. We address this question through criteria whose outcomes directly separate competing mechanisms. Each criterion pairs a controlled intervention with a prediction implied by the mechanism under consideration, and each measurement is calibrated against reference forecasters whose behavior is fixed by construction. This allows the readout to be validated before it is interpreted. In addition, we provide formal justification for the proposed diagnostic criteria, discuss their implications for understanding context use in zero-shot forecasting, and utilize the resulting mechanisms to interpret prior observations on zero-shot forecasting, similarity-based retrieval, and long-context utility.

3 PRELIMINARIES

In this section, we introduce the notation and concepts needed to distinguish the in-context mechanisms studied in the following sections. We first formalize a context-only forecaster and its effective weights, and then define the geometric quantities used to determine how those weights are allocated.

Let $y = (y_1, \ldots, y_L)$ be the context and let f map y to a one-step forecast $\hat{y} = f(y)$, with model parameters fixed and y as the only target-series-specific input.

We work in delay coordinates. For an embedding dimension m and lag τ, the delay vector at time t is $z_t = (y_t, y_{t-\tau}, \ldots, y_{t-(m-1)\tau})$, and the *query* is z_L. Under the conditions of Takens' theorem, the map from the attractor to delay space is a diffeomorphism, so proximity in delay space corresponds to proximity in state space.

Definition 1 (Weight profile)**.** Let $B \subseteq \{1, \ldots, L\}$ be a block of indices with indicator $\mathbf{1}_B$, and let σ be the standard deviation of the context. The *effective weight* of B under f at y is

$$W_B(f, y) \;=\; \lim_{a \to 0} \frac{f(y + a\sigma \mathbf{1}_B) - f(y)}{a\sigma}, \tag{1}$$

where the limit exists. If f is linear with fixed coefficients w_i, then $W_B = \sum_{i \in B} w_i$. The collection $\{W_B\}$ over blocks at a given resolution is the *weight profile* of f at y.

Equation 1 is the gradient of f aggregated over a block. The definition is architecture-agnostic and does not assume linearity. The interventions used empirically, however, have finite amplitude. We therefore report the block-response magnitude

$$R_B(a) = \left| f(y + a\sigma \mathbf{1}_B) - f(y) \right|. \tag{2}$$

Fix a metric d on delay space. For a neighborhood size k, write $\mathcal{N}_k(y)$ for the indices of the k delay vectors nearest z_L among those outside the Theiler window, and $\mathcal{S}_k(y) = \{y_{i+1} : i \in \mathcal{N}_k(y)\}$ for the values that followed them.

Definition 2 (Parroting and inference)**.** *f parrots* at y if there exists a neighborhood size k such that $\hat{y} = \sum_{i \in \mathcal{N}_k} w_i \, y_{i+1}$ with convex coefficients $w_i \geq 0$ and $\sum_i w_i = 1$. It *infers* otherwise.

Negative combinations and fitted local maps of high order polynomials are examples of mechanisms that can produce such forecasts. In this sense, parroting restricts the forecast to interpolation among continuations already observed in the context, whereas inference allows various levels of extrapolation beyond them.

Definition 3 (Neighborhood geometry and context weighting)**.** The *coverage radius* is $\rho_k(y) = \max_{i \in \mathcal{N}_k(y)} d(z_i, z_L)$, normalized by the diameter of the attractor. A weight profile is *indexed by phase-space distance* if $W_{\{i\}} = g(d(z_i, z_L))$ for some g, and *indexed by temporal position* if $W_{\{i\}} = g(L - i)$. The *far-end response* at offset Δ and amplitude a is $R_B(a)$ for a block of b samples at that offset. We report its relative value to $a\sigma b/L$, the response of the uniform average.

The coverage radius is small when the context contains a state close to the current one. It is computed from the context alone and can therefore be compared with forecasting error independently of the forecaster.

Our diagnostic criteria in Section 5 build primarily on the following local-approximation result. Write F for the one-step map on delay space.

Lemma 1 (Local approximation, after Farmer & Sidorowich, 1987)**.** *Let F be $(p+1)$-times continuously differentiable on a convex neighborhood U of the query, with $\sup_U \|D^{p+1}F\| \leq M$. Let $k \geq \binom{m+p}{p}$ points of U lie within radius r of the query, in general position in the sense that they do not all lie on a common algebraic surface of degree at most p, and let P be the order-p polynomial fitted to F on those points by least squares. Then*

$$|P(z_L) - F(z_L)| \;\leq\; C(\kappa)\,\frac{M}{(p+1)!}\,r^{\,p+1}, \tag{3}$$

where κ is the condition number of the design matrix of the fit and $C(\kappa)$ depends on the shape of the point cloud but not on r.

The error in Lemma 1 scales as r^{p+1}, linking approximation accuracy to neighborhood radius and providing the basis for the coverage criterion in Section 5.2.

Corollary 1 (Agreement in the dense limit)**.** *Take a sequence of point clouds whose radius r tends to zero and whose shape, after rescaling by r, is held fixed. Then the bound of equation 3 vanishes for every $p \geq 0$. In particular, $p = 0$, which is an equally weighted average of the retrieved successors and hence a convex combination of them, is a consistent estimator of F, and no finite order is distinguished from any other.*

Corollary 1 explains why forecasting accuracy alone cannot separate parroting from inference. The two mechanisms agree as the neighborhood becomes dense, so any empirical distinction between them is necessarily a finite-sample statement and must rely on a property other than prediction error.

4 Experimental setup

Systems and evaluation. The main experiments use scalar observations from the Lorenz system, with Chua and Chen used to test generality across dynamical systems. All series are sampled at 30 observations per Lyapunov time so that temporal context is comparable across systems. We use a common context length of $L = 511$ and evaluate one-step forecasts. The primary Lorenz results are aggregated over five trajectory seeds, with distinct initial conditions and randomly sampled prediction origins.

Models and reference forecasters. We evaluate 17 checkpoints from six families (Chronos-T5, Chronos-Bolt, Chronos-2, TimesFM, TiRex, and Mamba4Cast). Seven reference forecasters with

mechanisms known by construction are tested on the same data (nearest-neighbor successor, phase-space local-linear, trailing-window mean, trailing-window polynomial, uniform average, global linear fit, and flat-far-end reference). Appendix A.3 describes the model interfaces, and Appendix D defines the reference forecasters and gives the dynamical systems, sampling, window construction, and intervention parameters in full.

5 How the models use their context

5.1 Parroting or inference

To characterize the in-context behavior of TSFMs, we first determine whether their use of context is limited to parroting observed continuations or extends to inference beyond them. From the convexity constraint in Definition 2, we derive an architecture-agnostic geometric criterion that distinguishes the two mechanisms without relying on model internals.

Proposition 1 (Convex-hull criterion)**.** *If f parrots at y in the sense of Definition 2, then $\hat{y} \in \operatorname{conv} \mathcal{S}_k(y)$, which for a scalar observable is the interval $[\min \mathcal{S}_k, \max \mathcal{S}_k]$. Consequently, if $\hat{y}$ lies outside that interval for every neighborhood size k, then f does not parrot at y.*

Across the experiments, the reference forecasters separate as expected: the three parroting forecasters never escape the hull, whereas each inference reference does so on a non-zero fraction of windows. Among the TSFMs, every evaluated checkpoint escapes (Figure 1). For 16 of the 17 checkpoints, the mean escape rate ranges from 0.42% to 4.22%, while TimesFM-1.0 reaches 53.32%. Principally, escape rate alone is sufficient for the criterion. As a robustness check, we additionally measure escape magnitude. Among all TSFMs, the median escape magnitude ranges from 0.82% to 45.91% of the attractor diameter; for the discrete-output Chronos-T5 checkpoints, the median escapes correspond to approximately 6–19 output bins. Complete magnitude and robustness results are reported in Appendix E.2.

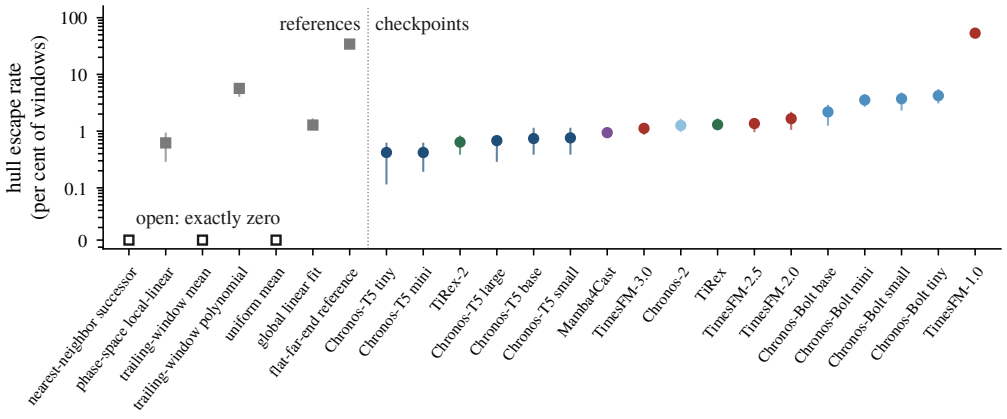

Figure 1: Convex-hull escape rates for the seven reference forecasters and the 17 TSFM checkpoints on Lorenz. Bars span the range across five Lorenz seeds.

These results rule out context parroting as a complete description of the evaluated TSFMs: their zero-shot forecasts can go beyond retrieving previous continuations and interpolating among values already observed in the context.

5.2 How context influence is allocated

Having established what TSFMs do with their context, we next ask how this behavior is realized through context use. With model parameters fixed, the question is equivalent to how influence is allocated across context samples. We contrast two mechanisms: influence organized by phase-space similarity to the current state and influence organized by temporal position. Lemma 1 provides a criterion basis for phase-space-based allocation because the approximation error of a local forecaster depends on the coverage radius of its selected neighborhood.

Proposition 2 (Coverage criterion)**.** *Suppose a forecaster fits an order-p local approximation using the k nearest states in phase space, and that its one-step error follows the scaling of Lemma 1 up to a multiplicative factor whose logarithm is uncorrelated with* $\log \rho_k$*. Then the logarithm of the error has a positive coefficient on* $\log \rho_k$*. Absence of such dependence is therefore inconsistent with this phase-space-based local-approximation mechanism under the stated assumptions.*

The reference forecasters separate as expected: the two phase-space-indexed references have positive coverage exponents of approximately 0.87–1.11, whereas the position-indexed references remain near zero, ranging from -0.13 to 0.03. The 17 TSFM checkpoints similarly concentrate near zero, with exponents from -0.148 to 0.102 (Figure 2). Their forecasting errors therefore show little dependence on the availability of nearby states in phase space.

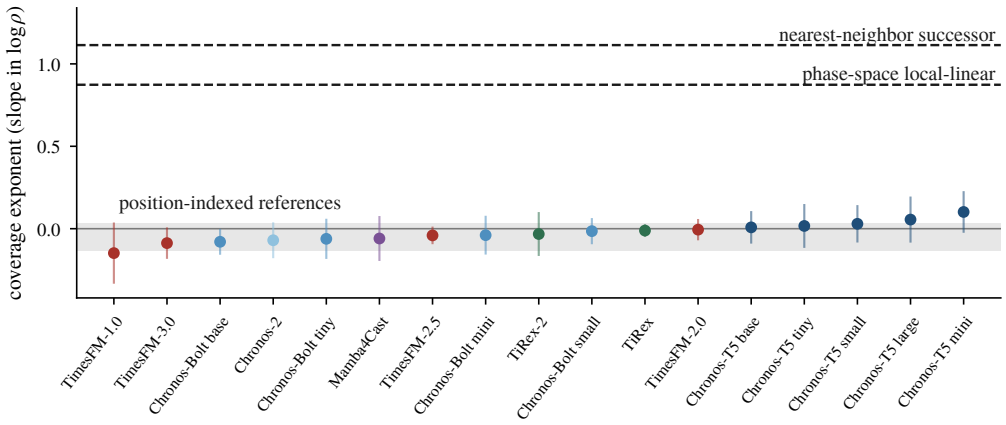

Figure 2: Coverage dependence of one-step forecasting error on Lorenz. Dashed lines denote the phase-space-indexed references and the shaded band represents the position-indexed references.

We further test this distinction by perturbing the block containing the retrieved nearest neighbor and comparing its response with that of an equally sized block at a random position. The phase-space references respond specifically at the retrieved location and have zero response at a random location. In contrast, the retrieved-to-random response ratios of the TSFMs range only from 1.046 to 1.334. Thus, the models do not preferentially respond to the most similar observed state in the manner of the phase-space references. Complete results with temporal-position sensitivity analysis are reported in Appendix E.5.

Next, we characterize how influence varies with temporal distance. A trailing-window mean and a uniform average are both indexed by position but represent opposite extremes of temporal reach. We therefore perturb blocks at increasing temporal offsets and compare the resulting responses with these two references. Blocks overlapping the embedding window of the query are treated separately because they alter the query state rather than probe earlier context.

Proposition 3 (Far-end response calibration)**.** *For the uniform average, displacing a block of b samples by $a\sigma$ changes the forecast by exactly $a\sigma b/L$, at every offset and every amplitude, and the ratio of responses at two amplitudes equals the ratio of the amplitudes. A trailing-window mean over the last k samples has response exactly zero at every offset beyond k.*

Proposition 3 provides two analytic calibrations for the far-end response: the uniform average scales exactly with the displacement, whereas the trailing-window mean is exactly zero beyond its support. The amplitude sweep separately measures how a checkpoint's response changes over the tested amplitudes.

For 15 of the 17 checkpoints, responses decrease strongly with temporal distance, with far-end responses between 0.045 and 0.451 of the uniform reference (Figure 3, left). Mamba4Cast is close to uniform allocation, whereas TimesFM-1.0 shows the opposite pattern, with substantially stronger far-end effect. The corresponding amplitude-scaling ratios are shown in Figure 3 (right) and summarized in Table 1. On the Lorenz attractor, dynamical similarity can recur at widely separated temporal positions, so phase-space relevance is distinct from temporal recency; Appendix E.3 quantifies this separation.

Appendix F provides a complementary architecture-specific check using cross-attention in the Chronos-T5 family. Its temporal profile shows the same qualitative concentration near the prediction point, while increasing model size shifts more attention mass toward distant context. Scaling therefore appears to broaden the temporal range over which context is represented, rather than changing the underlying preference for recent positions. Because attention mass is not itself a measure of causal contribution, this result is consistent with, but does not independently establish, the broader temporal allocation inferred from the input interventions.

Taken together, the diagnostics indicate that context influence is organized primarily by temporal position rather than phase-space similarity and, for most checkpoints, is concentrated substantially more strongly near the prediction point.

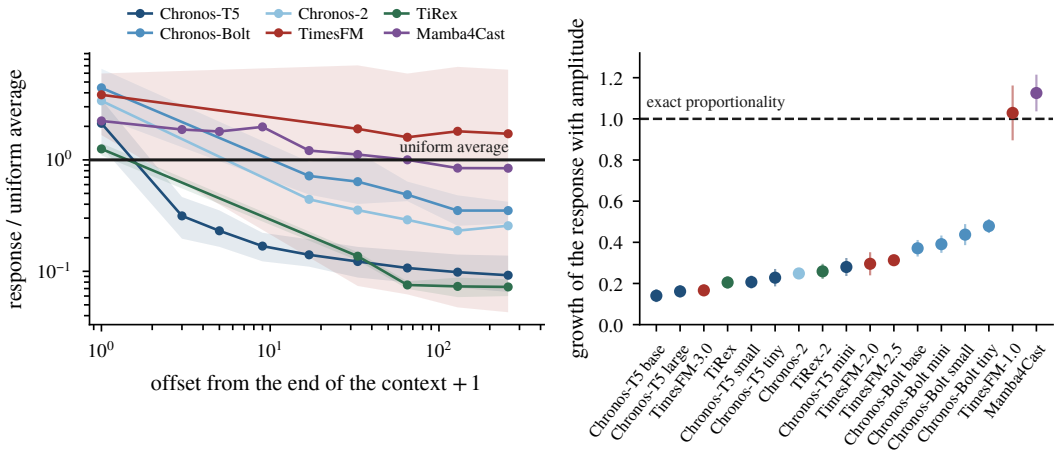

Figure 3: Far-end block-displacement responses on Lorenz. Left: response relative to the uniform reference across temporal offsets. Right: response scaling with perturbation amplitude.

Table 1: Far-end response relative to the uniform reference on Lorenz.

forecaster(s)	**far/uniform**	**amplitude scaling**
Chronos-T5/Bolt/2, TimesFM-2.0/2.5/3.0, TiRex	0.045–0.451	0.140–0.479
Mamba4Cast	0.843	1.126
TimesFM-1.0	6.649	1.029
uniform average	1	1
trailing-window mean	0	–

Understanding this observed allocation pattern also reveals a concrete opportunity for improving how future TSFMs use their context. By construction, selecting the k nearest states in phase space produces the k-sample neighborhood with the smallest possible radius around the query. Its radius therefore cannot exceed that of another k-sample subset.

Theorem 1 (Phase-space selection dominance)**.** *Fix k and p, and let U be a convex neighborhood containing the query, the k nearest delay vectors and the k most recent ones, on which the hypotheses of Lemma 1 hold with constant M. Then the radius of the k nearest delay vectors is no larger than the radius of the k most recent delay vectors, and the bound of equation 3 for the neighborhood chosen by phase-space distance is no larger than the bound for the neighborhood chosen by temporal position, whenever the nearest-neighbor cloud is no worse conditioned than the recent cloud.*

The theorem compares local-approximation bounds rather than realized forecasting errors and therefore identifies an approximation-theoretic opportunity rather than an empirical performance guarantee. Appendix C provides the proof. This result motivates exploring future TSFMs that can exploit phase-space similarity alongside temporal recency, an implication we discuss further in Section 7. Consistent with this opportunity, we demonstrate through a simple empirical comparison that the phase-space local-linear reference achieves a 51.5% lower one-step error than the best trailing polynomial in the tested recency grid. Full results are reported in Appendix D.8.

5.3 WHAT DISTANT CONTEXT SUPPLIES

Existing research has shown that increasing context length can substantially improve zero-shot forecasting on chaotic systems (Zhang & Gilpin, 2025). Given the weak influence of distant context observed above for most models, we ask how it can nevertheless provide an aggregate forecasting benefit and what information that benefit carries.

We first retain only the trailing k samples while keeping the prediction origin and target fixed; implementation and normalization details are provided in Appendix D. Then, we restore the full context length but replace the distant portion with one of four alternatives: another passage from the same system, a shuffled copy of the original far end, Gaussian noise generated with the far end's mean and variance, or a constant equal to the far-end mean. All replacements undergo the same rescaling procedure before being passed through each model's native preprocessing; implementation details are given in Appendix D. Our order intervention complements the block-shuffling experiment of Zhang & Gilpin (2025). They permute contiguous k-sample blocks while keeping the final k context points fixed and compare long-horizon forecast performance with a baseline using only those final k points. We instead shuffle individual samples only in the distant portion, retain a fixed trailing tail, and evaluate one-step point-forecast error together with explicit order-insensitive and order-sensitive reference forecasters.

For an exact permutation, temporal order can be changed while preserving the far-end values exactly, giving the following calibration.

Proposition 4 (Exact-permutation criterion)**.** *If f depends on the far end only through a functional invariant under permutations of its argument, then an exact permutation of the far end leaves the forecast unchanged.*

The shuffled replacement used empirically is subsequently rescaled by the same procedure as the other replacements and therefore is a permutation of the far end only up to a common rescaling about the far-end mean. We use Proposition 4 as an ideal calibration and base the empirical order comparison on fixed-k contrasts between the same-system and shuffled replacements.

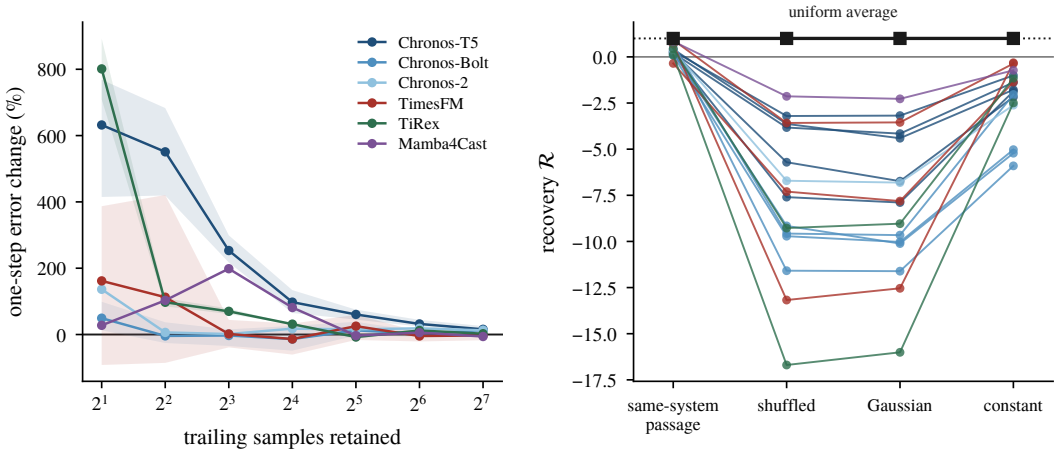

Figure 4: Effect of truncation and far-context replacement on Lorenz. Left: one-step error as the retained trailing context is shortened. Right: recovery under the four far-context replacements.

The effect of truncation is strongly model-dependent (Table 2 and Figure 4, left). Long context is consistently more accurate for Chronos-T5 at $k = 32$, 64, and 128, whereas the other families are more heterogeneous across checkpoints and retained lengths. Thus, longer supplied context can improve one-step accuracy, but its utility is not universal.

For 16 of the 17 checkpoints, the same-system replacement yields lower one-step error than the shuffled replacement; TimesFM-1.0 is the sole exception. Because the two replacements are compared at the same retained length and undergo the same rescaling, this fixed-k contrast shows that forecasting accuracy is sensitive to temporally coherent, system-consistent structure beyond the marginal value distribution. Figure 4 (right) additionally reports recovery as a normalized measure of replacement

performance. Shuffled replacement is also less accurate than truncation in all checkpoints with defined recovery. Computations and complete results are provided in Appendix E.6.

Table 2: Effect of shortening the supplied context on Lorenz. Entries are relative changes in one-step error; positive values indicate better performance than shortened context.

family	$k = 32$	$k = 64$	$k = 128$
Chronos-T5	$+44.8\%$ to $+75.0\%$	$+24.0\%$ to $+44.3\%$	$+5.1\%$ to $+25.4\%$
Chronos-Bolt	-3.2% to $+38.4\%$	-6.7% to $+9.9\%$	$+8.4\%$ to $+16.9\%$
Chronos-2	$+15.5\%$	$+18.7\%$	$+12.5\%$
TimesFM	-16.5% to $+59.4\%$	-21.0% to $+8.5\%$	-17.3% to $+16.2\%$
TiRex	-13.6% to -1.7%	$+9.0\%$ to $+15.2\%$	-0.8% to $+5.7\%$
Mamba4Cast	-3.3%	$+5.7\%$	-6.0%

Together, these results show that weak influence does not preclude aggregate long-context utility. When distant context improves forecasting, the shuffled replacement alters its contribution. Thus, useful distant context contributes temporally organized information beyond distributional statistics.

6 Generality across systems and models

We repeat the analyses across Chua and Chen and across the evaluated model families and backbone classes. In each case, the same three conclusions hold: the models perform inference beyond convex context parroting, their sample influence is governed primarily by temporal position, and distant context contributes temporal structure beyond distributional information. Complete results by system and checkpoint are reported in Appendix E.7.

7 Discussion

Beyond context parroting. The convex-hull criterion rules out convex context parroting as a complete description of the evaluated TSFMs. Their zero-shot behavior therefore goes beyond retrieving previous continuations and interpolating among values already observed in the context. This ability to infer beyond observed continuations provides a mechanistic account of how pretrained TSFMs can produce strong zero-shot forecasts on unseen chaotic systems (Zhang & Gilpin, 2025), without implying that extrapolation alone determines forecasting accuracy.

Limits of current context allocation. The observed preference for temporal position points to a limitation in how current TSFMs use their context. States that are close in phase space can be more informative for local forecasting than states that are merely close in time, yet the evaluated models do not preferentially exploit this structure. This provides a mechanistic interpretation of why explicitly retrieving similar patterns can outperform or improve TSFMs (Zhang & Gilpin, 2026; Ning et al., 2025; Han et al., 2025). Together with Theorem 1, this motivates TSFM context mechanisms that exploit phase-space similarity alongside temporal recency.

Preserving temporal structure. The aggregate contribution of distant context helps reconcile its weak influence with the previously observed benefit of longer context (Zhang & Gilpin, 2025). When distant context is beneficial, preserving its values without their temporal organization is insufficient, showing that its contribution extends beyond distributional information. This suggests that synthetic pretraining data and time-series augmentation should preserve temporal organization rather than match marginal statistics alone. Because our interventions modify inference-time context, this remains a design implication rather than evidence about the causal role of pretraining.

Future directions. A natural next step is to test whether explicitly retrieving, selecting, or upweighting phase-space-similar samples translates the approximation-theoretic advantage identified here into lower forecasting error. Besides, the present study considers univariate forecasting, where context allocation concerns influence across time. In multivariate settings, an additional question is how TSFMs allocate influence across variables, and whether this follows dynamical relevance, statistical dependence, causal structure, or architectural bias.

AI USE STATEMENT

An AI assistant was used solely for language editing. All scientific claims, proofs, methodology, experiments, and interpretations were carried out by the authors.

ETHICS STATEMENT

This work uses only simulated trajectories generated from published dynamical systems and involves no human subjects or personal data.

REPRODUCIBILITY STATEMENT

All dynamical systems, integration settings, assumptions, model checkpoints, and experimental procedures are documented in the appendices. Code and generated trajectories are released in the anonymous repository at `https://anonymous.4open.science/r/TSFM-ICL`.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Bernie Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=gerNCVqqtR.

Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. TiRex: Zero-shot forecasting across long and short horizons with enhanced in-context learning. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 57529–57580. Curran Associates, Inc., 2025. doi: 10.52202/085713-1929. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/5356603f9c47399adfd372f77a677057-Paper-Conference.pdf.

Satwik Bhattamishra, Arkil Patel, Phil Blunsom, and Varun Kanade. Understanding in-context learning in transformers and LLMs by learning to learn discrete functions. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ekeyCgeRfC.

Sathya Kamesh Bhethanabhotla, Omar Swelam, Julien Siems, David Salinas, and Frank Hutter. Mamba4Cast: Efficient zero-shot time series forecasting with state space models. In *NeurIPS Workshop on Time Series in the Age of Large Models*, 2024. URL https://openreview.net/forum?id=YBOQ5HnzI6.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10148–10167. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/das24c.html.

J. Doyne Farmer and John J. Sidorowich. Predicting chaotic time series. *Physical Review Letters*, 59(8):845–848, 1987. doi: 10.1103/PhysRevLett.59.845.

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan O Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21774–21797. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/han25d.html.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Kanghui Ning, Zijie Pan, Yu Liu, Yushan Jiang, James Zhang, Kashif Rasul, Anderson Schneider, Lintao Ma, Yuriy Nevmyvaka, and Dongjin Song. TS-RAG: Retrieval-augmented generation based time series foundation models are stronger zero-shot forecaster. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 163170–163199. Curran Associates, Inc., 2025. doi: 10.52202/085713-5448. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/eed25c037bc08afcbefab6f7a6b700e0-Paper-Conference.pdf.

Floris Takens. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980*, volume 898 of *Lecture Notes in Mathematics*, pp. 366–381, Berlin, Heidelberg, 1981. Springer. doi: 10.1007/BFb0091924.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Michał Wiliński, Mononito Goswami, Willa Potosnak, Nina Żukowska, and Artur Dubrawski. Exploring representations and interventions in time series foundation models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66861–66886. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/wilinski25a.html.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53140–53164. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/woo24a.html.

Yuanzhao Zhang and William Gilpin. Zero-shot forecasting of chaotic systems. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TqYjhJrp9m.

Yuanzhao Zhang and William Gilpin. Context parroting: A simple but tough-to-beat baseline for foundation models in scientific machine learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=EUAXc9Hlvm.

A Background on time-series foundation models

A.1 Two regimes of use

Pretrained time-series models are deployed in two distinct regimes, and the distinction determines what can be inferred from an intervention on the context.

In the first regime, the model is adapted to the target domain before forecasting. Its parameters may be updated through fine-tuning, parameter-efficient adaptation, or further pretraining on data related to the target domain. The resulting forecast can therefore depend jointly on the original pretraining corpus, the adaptation data, and the inference-time context, making these sources difficult to separate through context interventions alone.

In the second regime, the model is applied with its pretrained parameters fixed. No gradient update is performed on the target domain, and the context window provides the information that varies with

the series being forecast. This paper studies this zero-shot regime exclusively. With model parameters held fixed, variation across target series enters through the context, making Definition 1 well posed and allowing controlled changes to the context to be interpreted as changes in the behavior of the forecasting operator.

Every checkpoint in Table 3 is evaluated in this regime. No checkpoint is fine-tuned, and all experiments use the same context length, so comparisons across models are not confounded by different input budgets.

Potential pretraining exposure. The dynamical systems used in our experiments are standard, so potential pretraining exposure requires consideration. Published descriptions of the synthetic components of the relevant pretraining corpora include Gaussian-process kernel compositions, autoregressive processes with trend and seasonal terms, and linear causal graphs, while the real-data components consist of collections of measured time series. These descriptions do not report integrated trajectories of nonlinear differential equations. This is the strongest conclusion supported by the available documentation: corpus descriptions are not equivalent to a complete audit of the training data, and the issue would need to be reconsidered for models explicitly pretrained on simulated dynamical systems. The generality analysis in Section 6 does not depend on the absence of such exposure, because it evaluates whether the same measurements persist across different systems.

A.2 Backbone classes

We group the evaluated checkpoints into three backbone classes: Transformers, state-space models, and xLSTMs. These architectures differ in how information is propagated and combined across temporal positions. All three can produce content-dependent interactions, but they impose different structural biases on how information from earlier positions reaches the forecast.

Transformer. Transformers combine information across positions through self-attention. At each attention layer, representations are projected into queries, keys, and values. A position compares its query with the keys of other positions and forms a weighted combination of their values, so the resulting interaction weights depend on the content of the context. Self-attention itself does not impose a monotonic decay of influence with temporal distance: in principle, a distant position can receive more weight than a recent one when its content is more relevant to the query. Consequently, a weight profile indexed by phase-space distance is representable by this architecture, making the criterion in Proposition 2 an empirical question rather than an architecturally predetermined outcome.

State-space model. A state-space model processes a sequence by propagating a latent state through time. Each new input updates this state, which summarizes information carried from previous positions and is subsequently used to produce the output. In a selective state-space model, the state update depends on the current input, so the model can modulate which information is retained or propagated according to content. At the same time, information from a distant position must pass through a sequence of recurrent state updates before reaching the prediction point, introducing an architectural dependence on temporal position. Such a model can therefore express content-dependent weighting while also carrying a structural preference for more recent information.

xLSTM. xLSTM extends the recurrent structure of the LSTM with enhanced gating and memory mechanisms. Its matrix-memory component stores associations between key and value representations accumulated over time, while forget gates control how strongly information from earlier positions is retained. The current query then retrieves information from this memory according to its content. The contribution of a previous position therefore depends both on its similarity to the current query and on the sequence of forget gates between that position and the prediction point. xLSTM can thus express content-dependent interactions while retaining an explicit recurrent pathway through which influence may decay with temporal distance.

Classes not covered in this study. Two additional architecture classes used in time-series forecasting fall outside the scope of our analysis for different reasons. Mixer and convolutional backbones combine temporal positions using weights that are learned parameters rather than functions of the current input. A profile indexed by phase-space distance is therefore not directly expressible through their temporal mixing, and the outcome of Proposition 2 would be constrained by the architecture

rather than inferred from the measurement. Moreover, no released zero-shot forecasting checkpoint of these types was available to us at a comparable scale. Graph architectures, by contrast, primarily mix information across variables rather than temporal positions. This axis is orthogonal to the present study, and in our univariate setting the graph reduces to a single node.

A.3 THE CHECKPOINTS

Table 3: Evaluated checkpoints. CONTEXT and HORIZON denote the trained limits read directly from each checkpoint configuration at load time rather than from its documentation. When the two disagree, the checkpoint configuration is used. OUTPUT distinguishes discrete token distributions, from which a point forecast must be decoded or sampled, from directly emitted point or quantile forecasts.

family	backbone	checkpoint	context	horizon	output
Chronos (T5)	Transformer	`chronos-t5-tiny`	511	64	discrete tokens
	Transformer	`chronos-t5-mini`	511	64	discrete tokens
	Transformer	`chronos-t5-small`	511	64	discrete tokens
	Transformer	`chronos-t5-base`	511	64	discrete tokens
	Transformer	`chronos-t5-large`	511	64	discrete tokens
Chronos-Bolt	Transformer	`chronos-bolt-tiny`	2048	64	quantiles
	Transformer	`chronos-bolt-mini`	2048	64	quantiles
	Transformer	`chronos-bolt-small`	2048	64	quantiles
	Transformer	`chronos-bolt-base`	2048	64	quantiles
Chronos-2	Transformer	`chronos-2`	8192	1024	quantiles
TimesFM	Transformer	`timesfm-1.0-200m`	512	128	point
	Transformer	`timesfm-2.0-500m`	2048	128	point
	Transformer	`timesfm-2.5-200m`	16384	1024	quantiles
	Transformer	`timesfm-3.0`	2048	1024	quantiles
TiRex	xLSTM	`TiRex`	2048	320	quantiles + mean
	xLSTM	`TiRex-2`	2048	512	quantiles
Mamba4Cast	state-space model	`Mamba4Cast`	512	60	point

Five interface properties differ across model families. Rather than assuming common behavior, we resolve each property when a checkpoint is loaded and verify it through a self-check before running the experiments.

Usable context. The Chronos T5 family appends an end-of-sequence token, which occupies one of the 512 available positions and leaves 511 positions for observations. The other evaluated families append no such token. We therefore use a context length of 511 in all experiments so that each checkpoint receives the same forecasting problem, while reporting the checkpoint-specific maximum context length separately.

Output resolution. The escape magnitude in Section 5.1 is reported in units of the model's output resolution when such a discrete unit exists. For the Chronos-T5 checkpoints, one normalized quantization bin has width $30/4094 \simeq 0.00733$. Continuous-output models have no corresponding discrete unit, so the normalized escape magnitude is undefined and the escape rate alone carries the criterion.

Patch size. Some models process the context in patches. For these models, a displacement shorter than one patch does not correspond to an independent temporal unit: its effect depends on the alignment between the displaced block and the patch boundaries, which would introduce an oscillation at the patch period unrelated to the temporal reach being measured. We therefore set the displacement block to one full patch and restrict offsets to the patch grid. Chronos-Bolt and Chronos-2 use patch size 16, whereas TimesFM and TiRex use patch size 32. The absence of resolution below the patch scale is a property of the model interface and is reported as such.

Point forecast. When a checkpoint emits a predictive distribution, we obtain the point forecast directly from that distribution rather than from sampled forecast paths, preventing Monte Carlo variation from entering a one-step measurement: the exact first-step distribution median for Chronos-T5, the 0.5 quantile for Chronos-Bolt and Chronos-2, and the median quantile for TiRex and TiRex-2. TimesFM and Mamba4Cast provide native point forecasts, which we use directly.

Flags that alter the readout. We disable two optional settings in the TimesFM interface because they would directly modify quantities measured by our criteria. The first clamps the forecast to be non-negative when all input observations are positive. On such a window, the clamp would impose an output constraint that Proposition 1 could incorrectly attribute to the forecasting operator. The second enforces $f(-x) = -f(x)$, which would remove the even component of the response measured by Proposition 3. Neither constraint is a property of the systems studied here.

Input normalization. Each checkpoint retains the native preprocessing of its released forecasting interface. Chronos-T5 uses per-series mean-absolute tokenizer scaling. TimesFM retains its native per-series input normalization (`normalize_inputs=True` explicitly for TimesFM-2.5). Mamba4Cast uses context-wise min–max scaling. Chronos-Bolt, Chronos-2, TiRex, and TiRex-2 are passed through their released pipelines without an additional model-specific normalization in our wrapper. After an intervention, the native preprocessing is applied anew to the resulting input; its statistics are not frozen from the unmodified context.

B Assumptions and validation checks

The analysis relies on six assumptions. The first two concern the setting and finite-sample identifiability, while the remaining four concern whether the geometric quantities and interventions are measured as intended. Each assumption is either satisfied by construction or paired with an empirical check.

(A1) A dynamical law underlies the observations. The analysis assumes that the observed series is generated by an underlying dynamical law that can, in principle, be approximated from observations. This condition holds by construction in our experiments: every series is a scalar observable of a dynamical system integrated from known differential equations. The underlying law therefore exists and is not itself an object of inference.

(A2) Parroting and inference are distinguishable at finite samples. Corollary 1 shows that parroting and inference coincide as the neighborhood radius approaches zero, so the two mechanisms cannot be separated through their limiting prediction error. At finite samples, convexity provides an exact one-sided distinction through Proposition 1. The criterion can therefore establish inference when a forecast escapes every retrieved-successor hull, while failure to escape remains inconclusive.

(A3) Coverage is measurable independently of the forecaster. The coverage radius is computed entirely from the delay embedding of the context and does not depend on the forecasting model. It can therefore be used as a regressor against forecasting error without introducing circularity. *Check.* Nearest-neighbor coverage scales empirically as $\rho \propto L^{-0.554}$ with $R^2 = 0.979$. The reference correlation dimension $D_2 = 2.05$ predicts an exponent $-1/D_2 = -0.488$. Separately, the five sampled Lorenz trajectories have measured correlation dimensions between 1.960 and 1.980 (Table 4).

(A4) The point forecast is free of sampling noise. The one-step criteria compare individual forecasts with deterministic bounds or intervention responses, so Monte Carlo variation in the forecast would enter the measurement directly. Whenever the model interface permits, we therefore read the point forecast directly from the predictive distribution rather than estimate it from sampled forecast paths. *Check.* Across the Chronos-T5 checks, the median spread of repeated sampled-path point forecasts on an unchanged input ranges from 0.0123 to 0.0487, whereas the direct one-step readout has spread exactly 0 for every trajectory. The one-step experiments use the deterministic route.

(A5) The intervention exceeds the input quantization scale. For models with quantized inputs, a displacement smaller than the input quantization step could leave the tokenized context unchanged,

producing a zero response unrelated to the underlying weight profile. *Check.* The Chronos-T5 normalized bin width is 0.0073278. The smallest displacement used in the amplitude sweep is 0.5σ; across the Lorenz contexts, its magnitude after the tokenizer's per-series scaling is at least 0.402, more than 54 bin widths. The displacement is therefore well above the tokenizer resolution.

(A6) The replacements being contrasted are comparable at the boundary. The substitution experiment in Section 5.3 replaces the distant portion of the context. A step at the replacement boundary, or a change in the scale of the input, could produce a model response unrelated to the information being tested. The rescaling applied to every replacement preserves the far-end mean exactly but does not match the level at the boundary, so every replacement introduces a discontinuity next to the retained tail. The fixed-k contrast in Section 5.3 therefore assumes that these effects do not differ systematically between the same-system and shuffled replacements. *Check.* Relative to the median absolute first difference of the unmodified context, the step at the same position in the unmodified context has a median of 1.00 on Lorenz, whereas the median discontinuity, pooled over the seven retained lengths, is 6.2 for the same-system, 5.7 for the shuffled, 5.7 for the Gaussian and 4.8 for the constant replacement. The standard deviation of the replaced portion falls to a median of 0.27 and 0.25 of its original value for the same-system and shuffled replacements. Chen shows the same pattern; on Chua the discontinuities are larger for all four replacements (medians 14.9–16.9) but again comparable between the same-system and shuffled replacements. The same-system and shuffled replacements are therefore comparable in boundary step and in scale.

C Proofs

C.1 Local approximation

Proof of Lemma 1. Write z for a point of U, let $r > 0$ be the radius in the statement, and rescale coordinates about the query by setting

$$u \;=\; \frac{z - z_L}{r}, \tag{4}$$

so that the k selected points satisfy $\|u_i\| \le 1$ and the query is at $u = 0$. Let $\phi(u) \in \mathbb{R}^q$, $q = \binom{m+p}{p}$, collect the monomials in u of degree at most p, ordered with the constant term first, so that $\phi(0) = e_1$.

Let T be the order-p Taylor polynomial of F about z_L, and let $c_T \in \mathbb{R}^q$ be its coefficient vector in the u coordinates. Then $T(z_L + ru) = \phi(u)^\top c_T$, and because the constant term of the Taylor polynomial at its center equals the function value,

$$e_1^\top c_T \;=\; F(z_L). \tag{5}$$

By Taylor's theorem with the Lagrange remainder, each selected point satisfies

$$\big|F(z_i) - \phi(u_i)^\top c_T\big| \;\le\; \frac{\sup_U \|D^{p+1}F\|}{(p+1)!}\,\|z_i - z_L\|^{p+1} \;\le\; \frac{M}{(p+1)!}\,r^{\,p+1}, \tag{6}$$

where the final inequality uses $\|z_i - z_L\| \le r$.

Let $V \in \mathbb{R}^{k\times q}$ have rows $\phi(u_i)^\top$, and let $F \in \mathbb{R}^k$ collect the observed successors $F(z_i)$. The least-squares fit has coefficient vector $c = V^+F$, where V^+ is the Moore–Penrose pseudoinverse. The condition $k \ge q$, together with the assumption that the points do not lie on a degree-p algebraic variety, gives V full column rank. Hence $V^+V = I_q$ and

$$c - c_T \;=\; V^+F - V^+Vc_T \;=\; V^+\big(F - Vc_T\big). \tag{7}$$

The vector $F - Vc_T$ contains the remainders in equation 6, so $\|F - Vc_T\|_2 \le \sqrt{k}\,Mr^{p+1}/(p+1)!$ and therefore

$$\|c - c_T\|_2 \;\le\; \big\|V^+\big\|_2\,\sqrt{k}\,\frac{M}{(p+1)!}\,r^{\,p+1}. \tag{8}$$

The fitted forecast at the query is $P(z_L) = \phi(0)^\top c = e_1^\top c$. Using equation 5,

$$\big|P(z_L) - F(z_L)\big| = \big|e_1^\top(c - c_T)\big| \le \|c - c_T\|_2. \tag{9}$$

Combining this inequality with equation 8 yields equation 3 with

$$C(\kappa) \;=\; \sqrt{k}\,\left\|V^{+}\right\|_2 \;=\; \frac{\sqrt{k}}{\sigma_{\min}(V)}, \tag{10}$$

where $\sigma_{\min}(V)$ denotes the smallest singular value of V.

Finally, $C(\kappa)$ does not depend on r. The matrix V is constructed from the rescaled points u_i, which lie in the unit ball, and therefore depends on the shape of the point cloud after rescaling rather than on its absolute size. Thus $\sigma_{\min}(V)$, and hence $C(\kappa)$, is determined by neighborhood geometry alone. □

Remark 1. Equation 10 makes explicit that the local-approximation bound depends not only on neighborhood radius but also on the shape and conditioning of the selected point cloud. This dependence is why Theorem 1 requires the nearest-neighbor cloud to be no worse conditioned than the recent cloud, $C(\kappa_{\mathrm{sim}}) \le C(\kappa_{\mathrm{rec}})$. The requirement is not automatic. On 600 Lorenz windows with $L = 511$, for $p = 1$ and $k = 8$, the median $C(\kappa)$ is 33 for the nearest-neighbor cloud and 77 for the k most recent admissible delay vectors, and the requirement holds in 81% of windows. At $k = 16$ and $k = 32$, the corresponding proportions are 40% and 38%. The requirement is sufficient but not necessary. Because M and $(p+1)!$ cancel, the ratio of the two bounds, $C(\kappa_{\mathrm{sim}})\, r_{\mathrm{sim}}^{p+1} / (C(\kappa_{\mathrm{rec}})\, r_{\mathrm{rec}}^{p+1})$, is determined by the geometry of the two clouds alone. For $p = 1$ it is below one in 99.8%, 99.7% and 99.8% of windows at $k = 8, 16$ and 32 (medians 0.004, 0.029 and 0.046).

Proof of Corollary 1. Fix $p \ge 0$ and hold the rescaled shape of the point cloud fixed as its radius shrinks, so that $C(\kappa)$ remains constant along the limit. Neither $C(\kappa)$ nor M depends on r, and the right-hand side of equation 3 is therefore a constant multiple of r^{p+1}, which vanishes as $r \to 0$. Hence $P(z_L) \to F(z_L)$ for every finite p.

For $p = 0$, the monomial vector is the constant 1, the design matrix is a column of ones, and the least-squares fit is the mean of the observed successors. This mean is a convex combination with equal weights. The $p = 0$ estimator is therefore both consistent and convex, and no finite polynomial order is distinguished in the dense limit. □

C.2 Convex-hull criterion

Lemma 2. *For a finite set $S \subset \mathbb{R}^n$, the set of convex combinations of the elements of S is exactly* $\operatorname{conv} S$. *For $n = 1$,* $\operatorname{conv} S = [\min S, \max S]$.

Proof. The set of convex combinations of elements of S is convex and contains S, and therefore contains $\operatorname{conv} S$, the smallest convex set containing S. Conversely, by Carathéodory's theorem, every point in $\operatorname{conv} S$ can be written as a convex combination of at most $n + 1$ elements of S. The two sets are therefore equal. For $n = 1$, every convex combination lies between $\min S$ and $\max S$, and every point in this interval is a convex combination of the two endpoints. □

Proof of Proposition 1. If f parrots at y, Definition 2 gives at least one neighborhood size k for which

$$\hat{y} = \sum_{i \in \mathcal{N}_k} w_i y_{i+1}, \qquad w_i \ge 0, \qquad \sum_i w_i = 1.$$

Lemma 2 therefore gives $\hat{y} \in \operatorname{conv} \mathcal{S}_k(y)$. Hence, if the forecast lies outside $\operatorname{conv} \mathcal{S}_k(y)$ for every tested neighborhood size, no such k exists and the forecaster does not parrot at that window.

The nearest-neighbor sets are nested: $\mathcal{N}_{k_1} \subseteq \mathcal{N}_{k_2}$ for $k_1 < k_2$. Their successor hulls are therefore nested as well. The experimental radius grid reaches the full retrieved-successor set, so escape from the largest hull is sufficient to establish escape from every smaller hull. □

Remark 2. The criterion is one-sided. A forecaster that performs inference may still produce a forecast inside the convex hull on a particular window, so failure to escape is not evidence of parroting. This motivates reporting escape rates and headroom across many windows.

C.3 COVERAGE CRITERION

Proof of Proposition 2. Under the proposition's hypothesis, write the one-step error as

$$E = A\,\rho_k^{p+1},$$

where the multiplicative factor A collects the remaining terms and $\log A$ is uncorrelated with $\log \rho_k$. Taking logarithms gives

$$\log E = \log A + (p+1)\log \rho_k.$$

The population coefficient on $\log \rho_k$ is therefore $p + 1 > 0$. Consequently, absence of a positive dependence is inconsistent with this phase-space-based local-approximation mechanism under the stated assumptions. □

Remark 3. The empirical coverage axis uses the nearest-neighbor radius ρ_1 stored with each window, rather than the radius of the complete fitted neighborhood. We therefore use the exponent as a calibrated mechanism diagnostic: the phase-space-indexed references must show positive dependence on this common coverage proxy, whereas forecasters whose errors do not track local coverage should not. The empirical coefficient is not interpreted as an estimate of the polynomial order p.

C.4 FAR-END RESPONSE CALIBRATION

Proof of Proposition 3. For the uniform average $f(y) = L^{-1}\sum_{i=1}^{L} y_i$, displacing a block B with $|B| = b$ gives

$$f(y + a\sigma \mathbf{1}_B) - f(y) = a\sigma b/L,$$

independently of the block offset. The response is exactly proportional to the amplitude.

For the trailing-window mean $f(y) = k^{-1}\sum_{i>L-k} y_i$, a block outside the final k samples contributes no term to the sum, so the response is exactly zero. □

C.5 PHASE-SPACE SELECTION DOMINANCE

Lemma 3. *For any k-element subset S of admissible historical delay vectors, define $r(S) = \max_{i\in S} d(z_i, z_L)$. Then*

$$r(\mathcal{N}_k) = \min_{|S|=k} r(S).$$

In particular, the radius of the k nearest delay vectors is no larger than the radius of the k most recent admissible delay vectors.

Proof. Order all available distances increasingly as $d_{(1)} \le d_{(2)} \le \cdots$. Since $\mathcal{N}_k$ contains the indices of the k smallest distances, $r(\mathcal{N}_k) = d_{(k)}$. Every other k-element subset contains at least one point at distance at least $d_{(k)}$ and therefore has radius no smaller than $d_{(k)}$. □

Proof of Theorem 1. Let $r_{\mathrm{sim}} = r(\mathcal{N}_k)$ and let r_{rec} be the radius of the k most recent admissible delay vectors. Lemma 3 gives $r_{\mathrm{sim}} \le r_{\mathrm{rec}}$. Applying Lemma 1 to the two point clouds gives

$$E_{\mathrm{sim}} \le C(\kappa_{\mathrm{sim}})\frac{M}{(p+1)!}r_{\mathrm{sim}}^{p+1}, \qquad E_{\mathrm{rec}} \le C(\kappa_{\mathrm{rec}})\frac{M}{(p+1)!}r_{\mathrm{rec}}^{p+1}.$$

When $C(\kappa_{\mathrm{sim}}) \le C(\kappa_{\mathrm{rec}})$, the first upper bound is no larger than the second. □

Remark 4 (Scope of the dominance result). Theorem 1 compares the local-approximation bounds supplied by Lemma 1, not realized forecasting errors. Its conclusion isolates the geometric advantage of selecting a smaller-radius neighborhood under matched fit order and conditioning; it does not guarantee that a particular similarity-based forecaster will outperform a particular recency-based forecaster.

C.6 EXACT-PERMUTATION CRITERION

Lemma 4. *Suppose $f(y) = h\big(T(y_{1:L-k}),\, y_{L-k+1:L}\big)$ for some function h and some T invariant under permutations of its argument. Then, for every permutation π of $\{1, \dots, L-k\}$,*

$$f\big(y_{\pi(1)}, \dots, y_{\pi(L-k)}, y_{L-k+1}, \dots, y_L\big) = f(y).$$

Proof. Permutation invariance leaves the first argument of h unchanged, while the retained tail is identical. Both arguments of h are therefore the same before and after permutation. □

Proof of Proposition 4. An exact permutation leaves the permutation-invariant summary of the far end and the retained tail unchanged. The forecast is therefore unchanged by Lemma 4. □

D EXPERIMENTAL PROCEDURE

D.1 DYNAMICAL SYSTEMS

We evaluate three three-dimensional chaotic systems, each integrated directly from its defining equations. In every case, the first state coordinate is used as the scalar observable.

Lorenz.

$$\dot{x} = \sigma(y - x), \qquad \dot{y} = x(\rho - z) - y, \qquad \dot{z} = xy - \beta z,$$

with $\sigma = 10$, $\rho = 28$, $\beta = 8/3$, default initial condition $(1, 1, 1)$, and integration step 2×10^{-3}. Reference values are $\lambda_{\max} = 0.9056$ and $D_2 = 2.05$.

Chen.

$$\dot{x} = a(y - x), \qquad \dot{y} = (c - a)x - xz + cy, \qquad \dot{z} = xy - bz,$$

with $a = 35$, $b = 3$, $c = 28$, default initial condition $(-0.1, 0.5, -0.6)$, and integration step 10^{-3}. Reference values are $\lambda_{\max} = 2.014$ and $D_2 = 2.15$.

Chua. We use the cubic form

$$\dot{x} = \alpha(y - h(x)), \qquad \dot{y} = x - y + z, \qquad \dot{z} = -\beta y, \qquad h(x) = c_1 x^3 + c_0 x,$$

with $\alpha = 12.8$, $\beta = 19.1$, $c_0 = -1/7$, $c_1 = 2/7$, default initial condition $(0.1, 0, -0.1)$, and integration step 2×10^{-3}. Reference values are $\lambda_{\max} = 0.4466$ and $D_2 = 2.01$. We use the cubic rather than the piecewise-linear form because the latter has a discontinuous Jacobian at its breakpoints.

Table 4: Trajectories used in the analysis. The correlation dimensions are measured from the sampled trajectories; the Lyapunov exponents are the values used to set the sampling interval.

system	**trajectories**	$\lambda_{\max}$	Δt	**measured** D_{corr}	**embedding** $(m, \tau, \text{Theiler})$
Lorenz	5	0.9056	0.036808	1.960–1.980	$(3, 4, 26)$
Chen	1	2.0140	0.016551	2.168	$(3, 6, 33)$
Chua	1	0.4466	0.074638	2.028	$(3, 7, 29)$

Delay embedding. We fix the embedding dimension at $m = 3$ for all three systems, a common low-dimensional reconstruction choice for scalar observations of three-dimensional chaotic flows. The delay τ and Theiler window are calibrated separately for each system.

D.2 SAMPLING CALIBRATION

Sampling is normalized by Lyapunov time. With n samples per Lyapunov time,

$$\Delta t = \frac{1}{n\lambda_{\max}}.$$

All experiments use $n = 30$, so a context of $L = 511$ observations spans the same number of Lyapunov times across systems. Each trajectory contains $30{,}000$ sampled observations after discarding a transient of 20 Lyapunov times.

D.3 WINDOW CONSTRUCTION

Prediction origins are sampled uniformly without replacement from each trajectory. The windows are therefore non-rolling and use distinct prediction origins, but they are not required to be disjoint in time. Prediction origins are held fixed when context length changes, so shorter-context conditions reveal less of the same preceding trajectory while preserving the forecasting target.

For each Lorenz trajectory, the convex-hull experiment uses $1{,}000$ one-step windows. The coverage experiment uses 250 windows. The block-displacement and targeted nearest-neighbor perturbation experiments use 120 windows, truncation uses 300, and far-end substitution uses 200. The same five Lorenz trajectories are used throughout. The cross-system check uses $1{,}000$ one-step windows and 120 perturbation windows for each of Chen and Chua.

D.4 REFERENCE FORECASTERS

The seven reference forecasters are not treated as competing forecasting methods. Their mechanisms are fixed by construction, so they provide known readouts against which each diagnostic can be calibrated. All references receive the same complete context as the TSFMs.

The *nearest-neighbor successor* retrieves the admissible delay vector closest to the query in phase space and returns its observed successor,

$$i^\star = \arg\min_i d(z_i, z_L), \qquad \hat{y} = y_{i^\star+1}. \tag{11}$$

It is phase-space-indexed and convex: with one retrieved state, the forecast is exactly one previously observed continuation. The *phase-space local-linear* reference retrieves the k nearest admissible states and fits

$$y_{i+1} = a + b^\top (z_i - z_L) \tag{12}$$

by least squares, evaluating the fitted map at z_L, so that $\hat{y} = a$. We use $k = 8$ for the three-dimensional embeddings in this study. This reference is also phase-space-indexed but can extrapolate beyond the retrieved successors.

The *trailing-window mean* assigns equal weight to the final $k = 8$ samples and zero weight elsewhere. The *trailing-window polynomial* fits a quadratic polynomial to the same final eight observations and extrapolates one step. These references depend on temporal position rather than phase-space similarity and respectively provide convex and non-convex short-range controls. The *uniform average* assigns weight $1/L$ to every context sample, providing the analytic uniform reference for the block-displacement probe. The *global linear fit* estimates one affine delay-space map from all admissible context pairs and evaluates that map at the query; it uses the full context without nearest-state retrieval.

Finally, the *flat-far-end reference* is

$$\hat{y} = 2\bar{y}_{\text{near}} - \bar{y}_{\text{far}}, \tag{13}$$

where $\bar{y}_{\text{near}}$ is the mean of the final eight samples and $\bar{y}_{\text{far}}$ is the mean of the portion preceding the final 64 samples. Its negative far-end coefficient makes it non-convex, while its dependence on the distant region only through a mean makes it invariant to permutations of that region. It therefore separates convexity from temporal-order dependence.

Table 5 summarizes the resulting Lorenz calibration. The three convex references—nearest-neighbor successor, trailing-window mean, and uniform average—must remain inside the retrieved-successor hull. The first two rows provide the phase-space-indexed calibration for coverage and targeted perturbation, while the trailing-window mean and uniform average give the two analytic extremes of temporal reach.

Table 5: Reference forecasters and their calibrated mechanism readouts on Lorenz. Coverage values are fitted exponents on the empirical coverage axis. The far-end responses of the uniform average and trailing-window mean are analytic.

forecaster	sample selection	escape (%)	coverage exp.	targeted response	far end
nearest-neighbor successor	phase-space similarity	0.00	$+1.114$	0.0845 vs. 0	retrieved state
phase-space local-linear	phase-space similarity	0.62	$+0.873$	0.0208 vs. 0	retrieved neighborhood
trailing-window mean	temporal recency	0.00	-0.046	–	0 outside support
trailing-window polynomial	temporal recency	5.66	-0.132	–	short support
uniform average	pooled	0.00	$+0.034$	ratio 1	1 exactly
global linear fit	pooled	1.28	-0.041	–	measured
flat-far-end reference	temporal regions	34.24	$+0.001$	ratio 1	broad and flat

D.5 Reference calibration

The convex-hull references establish the one-sided calibration required by Proposition 1: all three convex references have exactly zero escapes, whereas each of the four inference references escapes on a non-zero fraction of Lorenz windows. The coverage readout separates the two phase-space-indexed references, with positive exponents, from the five references that do not retrieve by phase-space similarity, whose exponents remain near zero. For the block-displacement probe, the uniform average has its exact uniform response at every offset, while the trailing-window mean has zero response outside its eight-sample support. These known behaviors anchor the interpretation of the corresponding TSFM measurements below.

D.6 Noise floor

Before the one-step experiments, repeated evaluations of an unchanged input are used to determine whether the model interface itself introduces stochastic variation. Across the Chronos-T5 checks, the sampled-path median noise floor is 0.0123–0.0487, while the direct one-step readout has spread 0. We therefore use the deterministic direct-output route for all one-step criteria whenever the interface permits it.

D.7 Coverage computation

Each window stores the nearest-neighbor coverage statistic

$$\rho_1 = \min_i d(z_i, z_L),$$

normalized by the attractor diameter and computed with the same delay embedding and Theiler exclusion used by the retrieval procedure. This is the horizontal axis used by the coverage analysis. It is a model-independent proxy for local sampling density rather than the radius of the complete neighborhood used by every local fit. Its interpretation is therefore calibrated against the reference forecasters in Table 5.

For Lorenz, the median reference value across the coverage windows is $\rho_{\mathrm{ref}} = 0.01495$.

D.8 Recency order-grid comparison

To illustrate the forecasting opportunity discussed after Theorem 1, we compare polynomial extrapolation from recent observations with the phase-space local-linear reference. This experiment expands the fixed trailing-window polynomial reference by varying both the number of recent observations and the polynomial order. The comparison consistently uses the five Lorenz seeds with context length $L = 511$ and a one-step prediction horizon.

Trailing-polynomial grid. For each context $y = (y_1, \ldots, y_L)$, a trailing polynomial uses only the final k observations. Writing their relative time coordinates as $j = -k+1, \ldots, 0$, we fit

$$\widehat{\boldsymbol{c}}_{k,p} \in \operatorname*{arg\,min}_{\boldsymbol{c} \in \mathbb{R}^{p+1}} \sum_{j=-k+1}^{0} \left(y_{L+j} - \sum_{r=0}^{p} c_r j^r \right)^2, \tag{14}$$

Table 6: Recency order-grid comparison on Lorenz. Entries are mean relative one-step errors across five seeds, normalized within each seed by the nearest-neighbor successor reference error. Lower values are better. A dash denotes an undefined fit with $k < p + 1$. Boldface marks the best tested trailing-polynomial configuration. The phase-space local-linear reference has a mean relative error of 0.0509.

k	$p = 1$	$p = 2$	$p = 3$	$p = 4$	$p = 5$
2	0.9093	–	–	–	–
3	1.4900	0.3296	–	–	–
4	2.1755	0.5535	0.1918	–	–
6	3.7419	1.2740	0.5196	0.2204	**0.1050**
8	5.4557	2.2771	1.1079	0.5596	0.3107
12	8.5394	4.9816	3.1002	1.9708	1.3895
16	10.2142	7.5646	6.0631	4.2785	3.4956
24	12.7621	11.5286	10.3078	9.1370	7.8853
32	13.2133	11.9091	11.9011	11.0022	10.6531
64	13.0560	13.7185	11.0185	12.5951	13.2376

and extrapolate to the next time point,

$$\widehat{y}_{L+1}^{(k,p)} = \sum_{r=0}^{p} \widehat{c}_{k,p,r}. \tag{15}$$

We evaluate $k \in \{2, 3, 4, 6, 8, 12, 16, 24, 32, 64\}$ and $p \in \{1, 2, 3, 4, 5\}$, retaining the combinations with $k \geq p + 1$. Thus, the grid contains 44 valid configurations. The prediction origin and target remain fixed as k and p vary.

The phase-space local-linear reference instead selects the eight nearest admissible delay vectors and fits an affine map from those vectors to their observed successors, as described in Appendix D. Its prediction is the fitted map evaluated at the current query. The trailing polynomials therefore fit the scalar trajectory as a function of time, whereas the phase-space reference fits a local successor map in delay space.

Error normalization and aggregation. Within each seed, we normalize the one-step error summary of each forecaster by the corresponding nearest-neighbor successor reference error. The local-linear error is evaluated at the seed-specific reference coverage radius using the coverage analysis described in Appendix E.3. We then average the relative errors across the five seeds. Accordingly, the reported values are means of within-seed error ratios, rather than ratios of errors pooled across seeds. A value below one indicates a lower error than the nearest-neighbor successor reference.

Results and interpretation. The configuration $(k, p) = (6, 5)$ attains the lowest relative error among the tested trailing polynomials in every seed, with a cross-seed mean of 0.1050 (Table 6). Increasing the trailing-window length does not systematically improve prediction: for example, at $p = 5$, the mean relative error rises from 0.1050 at $k = 6$ to 0.3107 at $k = 8$ and 1.3895 at $k = 12$. The result therefore depends on both the temporal extent of the fit and its polynomial order.

The phase-space local-linear reference achieves a mean relative error of 0.0509, approximately 51.5% lower than that of the best tested trailing polynomial. This percentage compares the two cross-seed mean relative errors; it is not an average of seedwise percentage reductions. Both values describe reference forecasters, rather than a TSFM checkpoint or an improvement obtained by modifying a pretrained model.

These results illustrate the empirical opportunity motivating Theorem 1, but do not directly test its bound comparison. The comparison demonstrates that a simple phase-space reference can outperform every trailing polynomial in the tested grid on these Lorenz data; it does not establish a general performance guarantee or isolate sample selection as the sole cause of the difference.

D.9 BLOCK-DISPLACEMENT PROBE

For a block of consecutive samples at offset Δ from the end of the context, we add $a\sigma$ to each sample in the block, where σ is the standard deviation of the full context. The default unpatched block size is 16, offsets are $\{0, 2, 4, 8, 16, 32, 64, 128, 256\}$, and amplitudes are $\{0.5, 2, 5\}\sigma$. Each trajectory contributes 120 windows. Models with patch size 16 use block size 16 and offsets $\{0, 16, 32, 64, 128, 256\}$; models with patch size 32 use block size 32 and offsets $\{0, 32, 64, 128, 256\}$.

The amplitude sweep characterizes how the response scales with the imposed displacement across the tested amplitudes. The reported far-end summary uses the largest amplitude, 5σ.

D.10 NEAREST-NEIGHBOR PERTURBATION

The targeted variant perturbs the block containing the retrieved nearest neighbor and compares its response with an equally sized block at a random admissible location. A third control places the block at the end of the context and measures the response to the retained tail. The phase-space references provide the location-specific calibration: the nearest-neighbor successor and phase-space local-linear references respond at the retrieved location but exactly zero at a random location, whereas the uniform average responds equally at retrieved and random locations. We therefore summarize the model readout by the retrieved-to-random response ratio rather than assume a particular functional form for dependence on phase-space distance.

D.11 TRUNCATION AND FAR-END SUBSTITUTION

Truncation. The model is provided the trailing $k \in \{2, 4, 8, 16, 32, 64, 128, 256\}$ samples. Before forecasting, a truncated segment x is multiplied by

$$c = \frac{\overline{|y|}}{\overline{|x|}},$$

where y is the complete context, and the resulting forecast is divided by c before the error is computed. This is an intervention-side rescaling. The input is then passed through the model family's native preprocessing. Each trajectory contributes 300 prediction origins.

Far-end substitution. The experiment retains a tail of $k \in \{2, 4, 8, 16, 32, 64, 128\}$ samples and replaces the earlier context. The four modes are another segment from the same system, a shuffled copy of the original far end, a Gaussian surrogate generated with the far end's mean and variance, and a constant equal to the far-end mean. For the shuffled mode, individual samples of the far end are permuted uniformly at random before the common rescaling step. Each trajectory contributes 200 prediction origins, and recovery is reported only where truncation increases the error by more than 10% of the full-context error.

Algorithm 1 specifies the shuffled replacement. The intermediate context y^{perm} is an exact permutation of the far end and provides the calibration in Proposition 4. The empirical intervention uses y^{shuffle}, obtained by rescaling the permuted far end about its mean. Here, $\overline{v}$ denotes the arithmetic mean of a vector v, and $\|$ denotes concatenation.

The same rescaling is applied to every replacement before forecasting. The implementation first shifts the replacement to the level of the tail's first b samples and then restores the far-end mean; the two shifts cancel exactly, so Algorithm 1 states the net operation. The rescaling preserves the mean of the original far end exactly and, for the shuffled replacement, the empirical distribution of the far-end values up to the factor g about that mean. It does not match the level at the boundary, and it multiplies the variance of the replacement by g^2, so the mean-absolute scale of the complete context changes. The retained tail itself is unchanged, and each model then applies its native preprocessing to the resulting input.

Algorithm 1 Far-end shuffling with slope-matched rescaling

Require: Context $y = (y_1, \ldots, y_L)$; retained tail length $2 \le k \le L - 2$
Ensure: Exact-permutation context y^{perm} and shuffled context y^{shuffle}

$n \leftarrow L - k$
$F \leftarrow (y_1, \ldots, y_n)$
$T \leftarrow (y_{n+1}, \ldots, y_L)$
Uniform samplewise permutation
Sample π uniformly from all permutations of $\{1, \ldots, n\}$
$R \leftarrow (F_{\pi(1)}, \ldots, F_{\pi(n)})$
$y^{\mathrm{perm}} \leftarrow R \| T$
Slope-matched rescaling about the far-end mean
$b \leftarrow \min\{8, n, k\}$
$s_R \leftarrow \frac{1}{b-1} \sum_{j=n-b+1}^{n-1} (R_{j+1} - R_j)$
$s_T \leftarrow \frac{1}{b-1} \sum_{j=1}^{b-1} (T_{j+1} - T_j)$
if $|s_R| < 10^{-12}$ **then**
    $g \leftarrow 1$
else
    $g \leftarrow \min\{4, \max\{0.25, s_T / s_R\}\}$
end if
$R \leftarrow \big(g(R_j - \overline{R}) + \overline{F}\big)_{j=1}^{n}$
$y^{\mathrm{shuffle}} \leftarrow R \| T$
return $(y^{\mathrm{perm}}, y^{\mathrm{shuffle}})$

E COMPLETE EXPERIMENTAL RESULTS

Unless stated otherwise, entries below aggregate the five Lorenz trajectory seeds. Each numeric cell is reported as mean ± standard deviation [standard error] across seeds. A dash denotes a quantity that is undefined for that forecaster or is outside its model-specific evaluation grid. Cross-system experiments use one trajectory per system and are therefore reported without cross-seed error bars. The tables retain checkpoint-level values, while the accompanying text highlights the mechanism-level patterns used in the main paper.

E.1 CHECKPOINT-LEVEL SUMMARY

Table 7: Checkpoint-level summary of the principal Lorenz readouts. Escape rates are percentages. Far/uniform is the far-end response ratio relative to the uniform reference. Keep-8 cost is the relative increase in one-step error from the substitution/truncation grid, expressed as a percentage. Cells report mean ± SD [SE].

checkpoint	escape (%)	coverage exp.	retr./random	far/uniform	amp. scaling	keep-8 cost (%)
Chronos-T5 tiny	0.4200 ± 0.2387 [0.1068]	0.0166 ± 0.1266 [0.0566]	1.175 ± 0.195 [0.087]	0.1165 ± 0.0088 [0.0039]	0.2281 ± 0.0370 [0.0165]	254.0 ± 46.6 [20.8]
Chronos-T5 mini	0.4200 ± 0.1643 [0.0735]	0.1015 ± 0.1198 [0.0536]	1.218 ± 0.191 [0.086]	0.1399 ± 0.0159 [0.0071]	0.2800 ± 0.0384 [0.0172]	238.3 ± 42.8 [19.1]
Chronos-T5 small	0.7600 ± 0.2608 [0.1166]	0.0296 ± 0.1070 [0.0478]	1.239 ± 0.253 [0.113]	0.0810 ± 0.0047 [0.0021]	0.2069 ± 0.0258 [0.0116]	295.8 ± 41.9 [18.8]
Chronos-T5 base	0.7400 ± 0.2702 [0.1208]	0.00810 ± 0.09215 [0.04121]	1.206 ± 0.176 [0.079]	0.0682 ± 0.0087 [0.0039]	0.1405 ± 0.0258 [0.0116]	204.2 ± 23.2 [10.4]
Chronos-T5 large	0.6800 ± 0.2168 [0.0970]	0.0553 ± 0.1333 [0.0596]	1.334 ± 0.240 [0.107]	0.0705 ± 0.0054 [0.0024]	0.1616 ± 0.0238 [0.0106]	251.9 ± 56.0 [25.0]
Chronos-Bolt tiny	4.220 ± 0.958 [0.428]	-0.0615 ± 0.1154 [0.0516]	1.119 ± 0.261 [0.117]	0.4510 ± 0.0166 [0.0074]	0.4790 ± 0.0286 [0.0128]	-35.22 ± 5.74 [2.57]
Chronos-Bolt mini	3.540 ± 0.573 [0.256]	-0.0393 ± 0.1112 [0.0497]	1.046 ± 0.101 [0.045]	0.3423 ± 0.0113 [0.0051]	0.3909 ± 0.0368 [0.0164]	-7.068 ± 6.896 [3.084]
Chronos-Bolt small	3.720 ± 0.853 [0.381]	-0.0149 ± 0.0728 [0.0326]	1.142 ± 0.066 [0.030]	0.3482 ± 0.0355 [0.0159]	0.4374 ± 0.0458 [0.0205]	13.44 ± 4.17 [1.86]
Chronos-Bolt base	2.180 ± 0.614 [0.275]	-0.0794 ± 0.0722 [0.0323]	1.122 ± 0.179 [0.080]	0.2620 ± 0.0200 [0.0090]	0.3711 ± 0.0349 [0.0156]	5.972 ± 20.886 [9.340]
Chronos-2	1.260 ± 0.270 [0.121]	-0.0701 ± 0.1023 [0.0458]	1.277 ± 0.148 [0.066]	0.2440 ± 0.0127 [0.0057]	0.2487 ± 0.0209 [0.0093]	2.498 ± 15.209 [6.802]
TimesFM-1.0	53.32 ± 1.86 [0.83]	-0.1481 ± 0.1793 [0.0802]	1.078 ± 0.117 [0.052]	6.649 ± 1.159 [0.518]	1.029 ± 0.128 [0.057]	-31.36 ± 17.96 [8.03]
TimesFM-2.0	1.660 ± 0.404 [0.181]	-0.00580 ± 0.05826 [0.02606]	1.058 ± 0.085 [0.038]	0.2373 ± 0.0270 [0.0121]	0.2956 ± 0.0508 [0.0227]	-40.63 ± 5.83 [2.61]
TimesFM-2.5	1.360 ± 0.288 [0.129]	-0.0407 ± 0.0469 [0.0210]	1.065 ± 0.099 [0.044]	0.1142 ± 0.0098 [0.0044]	0.3130 ± 0.0125 [0.0056]	27.86 ± 19.13 [8.56]
TimesFM-3.0	1.120 ± 0.205 [0.092]	-0.0874 ± 0.0890 [0.0398]	1.253 ± 0.372 [0.167]	0.0452 ± 0.0015 [0.0007]	0.1665 ± 0.0225 [0.0101]	37.87 ± 16.22 [7.25]
TiRex	1.300 ± 0.212 [0.095]	-0.0113 ± 0.0319 [0.0143]	1.147 ± 0.238 [0.106]	0.0593 ± 0.0069 [0.0031]	0.2050 ± 0.0240 [0.0107]	55.14 ± 15.57 [6.96]
TiRex-2	0.6400 ± 0.1817 [0.0812]	-0.0323 ± 0.1267 [0.0566]	1.200 ± 0.202 [0.090]	0.0863 ± 0.0084 [0.0038]	0.2588 ± 0.0314 [0.0141]	80.90 ± 10.46 [4.68]
Mamba4Cast	0.9400 ± 0.1517 [0.0678]	-0.0594 ± 0.1299 [0.0581]	1.224 ± 0.301 [0.135]	0.8429 ± 0.0664 [0.0297]	1.126 ± 0.084 [0.037]	200.8 ± 26.2 [11.7]

Table 7 provides a compact view of the four diagnostics before they are expanded below. Every checkpoint has a positive convex-hull escape rate, whereas the coverage exponents remain close to zero and the targeted retrieved-to-random ratios remain close to one. Far-end response is below the uniform reference for 15 of the 17 checkpoints; Mamba4Cast is close to uniform, while TimesFM-1.0 assigns substantially stronger influence to distant positions. The keep-8 column also shows that the aggregate value of long context is model-dependent rather than determined by the local influence profile alone.

E.2 CONVEX-HULL CRITERION

Table 8: Convex-hull calibration for the seven reference forecasters. Median escape and headroom are in percent of the attractor diameter.

reference	escape (%)	median escape (%)	median headroom (%)
nearest-neighbor successor	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	33.46 ± 0.66 [0.29]
phase-space local-linear	0.6200 ± 0.2387 [0.1068]	1.26 ± 0.51 [0.23]	33.53 ± 0.57 [0.26]
trailing-window mean	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	30.81 ± 0.83 [0.37]
trailing-window polynomial	5.660 ± 0.971 [0.434]	4.41 ± 0.75 [0.34]	35.74 ± 0.74 [0.33]
uniform average	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	45.22 ± 0.46 [0.21]
global linear fit	1.280 ± 0.192 [0.086]	2.24 ± 0.47 [0.21]	35.29 ± 0.71 [0.32]
flat-far-end reference	34.24 ± 1.59 [0.71]	10.76 ± 0.49 [0.22]	24.71 ± 0.68 [0.30]

Table 9: Convex-hull results for all TSFM checkpoints. Escape rate is in percent. Escape margin is in quantization-bin units and is defined only for Chronos-T5.

checkpoint	escape (%)	median escape (%)	median headroom (%)	escape margin (bins)
Chronos-T5 tiny	0.4200 ± 0.2387 [0.1068]	1.70 ± 1.90 [0.85]	34.49 ± 0.54 [0.24]	12.38 ± 13.83 [6.19]
Chronos-T5 mini	0.4200 ± 0.1643 [0.0735]	2.62 ± 1.73 [0.78]	33.92 ± 0.77 [0.34]	19.04 ± 12.60 [5.64]
Chronos-T5 small	0.7600 ± 0.2608 [0.1166]	0.824 ± 0.192 [0.086]	33.67 ± 0.68 [0.30]	5.989 ± 1.388 [0.621]
Chronos-T5 base	0.7400 ± 0.2702 [0.1208]	1.24 ± 0.37 [0.17]	33.70 ± 0.59 [0.26]	9.041 ± 2.707 [1.210]
Chronos-T5 large	0.6800 ± 0.2168 [0.0970]	1.64 ± 0.59 [0.26]	33.66 ± 0.62 [0.28]	11.92 ± 4.28 [1.91]
Chronos-Bolt tiny	4.220 ± 0.958 [0.428]	3.01 ± 0.60 [0.27]	32.99 ± 0.72 [0.32]	–
Chronos-Bolt mini	3.540 ± 0.573 [0.256]	2.71 ± 0.62 [0.28]	33.93 ± 0.58 [0.26]	–
Chronos-Bolt small	3.720 ± 0.853 [0.381]	2.88 ± 0.35 [0.16]	32.87 ± 0.59 [0.26]	–
Chronos-Bolt base	2.180 ± 0.614 [0.275]	2.61 ± 0.59 [0.26]	33.77 ± 0.73 [0.32]	–
Chronos-2	1.260 ± 0.270 [0.121]	1.59 ± 0.60 [0.27]	33.70 ± 0.91 [0.41]	–
TimesFM-1.0	53.32 ± 1.86 [0.83]	45.91 ± 3.48 [1.56]	27.17 ± 0.65 [0.29]	–
TimesFM-2.0	1.660 ± 0.404 [0.181]	2.58 ± 0.89 [0.40]	33.05 ± 0.69 [0.31]	–
TimesFM-2.5	1.360 ± 0.288 [0.129]	2.51 ± 0.98 [0.44]	33.60 ± 0.70 [0.31]	–
TimesFM-3.0	1.120 ± 0.205 [0.092]	2.52 ± 0.77 [0.34]	33.79 ± 0.62 [0.28]	–
TiRex	1.300 ± 0.212 [0.095]	2.36 ± 0.75 [0.34]	33.81 ± 0.46 [0.20]	–
TiRex-2	0.6400 ± 0.1817 [0.0812]	1.45 ± 0.65 [0.29]	33.70 ± 0.66 [0.29]	–
Mamba4Cast	0.9400 ± 0.1517 [0.0678]	1.61 ± 0.98 [0.44]	33.82 ± 0.68 [0.30]	–

The convex references remain exactly inside the hull, while each inference reference has a positive escape rate. All 17 TSFM checkpoints also escape on at least some windows. For 16 checkpoints the mean rate is 0.42%–4.22%; TimesFM-1.0 reaches 53.32%. Escape magnitudes are not confined to numerical resolution: across checkpoints the median magnitude ranges from 0.824% to 45.91% of the attractor diameter, and the Chronos-T5 medians correspond to approximately 6–19 output bins. Table 10 provides the corresponding independent-window resampling check.

Theiler-window sensitivity. We recompute the retrieved-successor hull after admitting the temporally adjacent states excluded by the Theiler window. Across the $5{,}000$ Lorenz windows, the hull expands in 9.6% of cases and the median change in hull width is zero. This criterion does not by itself exclude reliance on the immediate past, but the justification on TSFMs remains valid.

E.3 COVERAGE DEPENDENCE

The reference radius varies slightly across the five trajectories because it is computed from each trajectory-specific geometry; across the five Lorenz trajectories, its mean is 0.01495.

The primary one-step exponent cleanly separates the calibration mechanisms. The nearest-neighbor successor and phase-space local-linear references have positive exponents of 1.114 and 0.877, whereas the remaining references range from -0.131 to 0.037. The 17 TSFM exponents lie be-

Table 10: Additional rare-event robustness: escape-rate range across five independently resampled window sets on one Lorenz trajectory, with the median escape magnitude across those window sets.

forecaster	min escape (%)	max escape (%)	median magnitude (%)
nearest-neighbor successor	0.00	0.00	0.00
phase-space local-linear	0.40	0.90	1.64
Chronos-T5 tiny	0.30	0.70	1.61
Chronos-T5 mini	0.20	0.70	3.76
Chronos-T5 small	0.40	1.10	1.35
Chronos-T5 base	0.60	0.90	1.59
Chronos-T5 large	0.60	0.80	2.52
Chronos-Bolt tiny	2.80	4.30	2.99
Chronos-Bolt mini	2.40	3.60	3.30
Chronos-Bolt small	2.80	3.80	2.71
Chronos-Bolt base	1.50	2.20	2.26
Chronos-2	0.70	1.10	2.29
TimesFM-1.0	51.40	55.80	46.20
TimesFM-2.0	1.30	1.90	3.18
TimesFM-2.5	1.10	1.60	2.55
TimesFM-3.0	0.90	1.30	3.09
TiRex	1.00	1.60	2.05
TiRex-2	0.50	0.70	1.22
Mamba4Cast	0.20	1.10	1.67

Table 11: Coverage calibration for the reference forecasters.

reference	exp.	R^2	$E(\rho_{\mathrm{ref}})$	E/E_{parrot}
nearest-neighbor successor	$1.114 \pm 0.108\,[0.048]$	$0.4493 \pm 0.0694\,[0.0310]$	$0.3595 \pm 0.0212\,[0.0095]$	$1.000 \pm 0.000\,[0.000]$
phase-space local-linear	$0.8772 \pm 0.1971\,[0.0881]$	$0.1717 \pm 0.0547\,[0.0245]$	$0.0169 \pm 0.0008\,[0.0004]$	$0.0471 \pm 0.0050\,[0.0022]$
trailing-window mean	$-0.0411 \pm 0.0934\,[0.0418]$	$0.00471 \pm 0.00580\,[0.00259]$	$3.407 \pm 0.169\,[0.075]$	$9.507 \pm 0.809\,[0.362]$
trailing-window polynomial	$-0.1313 \pm 0.0404\,[0.0181]$	$0.00624 \pm 0.00362\,[0.00162]$	$0.7385 \pm 0.0469\,[0.0210]$	$2.062 \pm 0.212\,[0.095]$
uniform average	$0.0367 \pm 0.1064\,[0.0476]$	$0.00453 \pm 0.00162\,[0.00073]$	$4.350 \pm 0.298\,[0.133]$	$12.14 \pm 1.17\,[0.52]$
global linear fit	$-0.0421 \pm 0.0763\,[0.0341]$	$0.00345 \pm 0.00263\,[0.00118]$	$0.6053 \pm 0.0288\,[0.0129]$	$1.690 \pm 0.154\,[0.069]$
flat-far-end reference	$-0.00137 \pm 0.06124\,[0.02739]$	$0.00111 \pm 0.00182\,[0.00081]$	$4.473 \pm 0.530\,[0.237]$	$12.42 \pm 0.92\,[0.41]$

Table 12: Coverage results for all TSFM checkpoints.

checkpoint	exp.	R^2	$E(\rho_{\mathrm{ref}})$	E/E_{parrot}
Chronos-T5 tiny	$0.0166 \pm 0.1266\,[0.0566]$	$0.00541 \pm 0.00605\,[0.00271]$	$0.2423 \pm 0.0178\,[0.0080]$	$0.6761 \pm 0.0676\,[0.0302]$
Chronos-T5 mini	$0.1015 \pm 0.1198\,[0.0536]$	$0.00764 \pm 0.00643\,[0.00288]$	$0.2016 \pm 0.0202\,[0.0090]$	$0.5629 \pm 0.0699\,[0.0312]$
Chronos-T5 small	$0.0296 \pm 0.1070\,[0.0478]$	$0.00360 \pm 0.00364\,[0.00163]$	$0.1308 \pm 0.0093\,[0.0042]$	$0.3648 \pm 0.0318\,[0.0142]$
Chronos-T5 base	$0.00810 \pm 0.09215\,[0.04121]$	$0.00260 \pm 0.00257\,[0.00115]$	$0.1210 \pm 0.0066\,[0.0029]$	$0.3372 \pm 0.0212\,[0.0095]$
Chronos-T5 large	$0.0553 \pm 0.1333\,[0.0596]$	$0.00652 \pm 0.01185\,[0.00530]$	$0.1235 \pm 0.0112\,[0.0050]$	$0.3451 \pm 0.0419\,[0.0187]$
Chronos-Bolt tiny	$-0.0615 \pm 0.1154\,[0.0516]$	$0.00587 \pm 0.00886\,[0.00396]$	$0.7840 \pm 0.0207\,[0.0093]$	$2.189 \pm 0.176\,[0.079]$
Chronos-Bolt mini	$-0.0393 \pm 0.1112\,[0.0497]$	$0.00449 \pm 0.00347\,[0.00155]$	$0.6246 \pm 0.0655\,[0.0293]$	$1.746 \pm 0.240\,[0.107]$
Chronos-Bolt small	$-0.0149 \pm 0.0728\,[0.0326]$	$0.00185 \pm 0.00167\,[0.00075]$	$0.5933 \pm 0.0463\,[0.0207]$	$1.657 \pm 0.183\,[0.082]$
Chronos-Bolt base	$-0.0794 \pm 0.0722\,[0.0323]$	$0.00549 \pm 0.00734\,[0.00328]$	$0.4858 \pm 0.0481\,[0.0215]$	$1.360 \pm 0.204\,[0.091]$
Chronos-2	$-0.0701 \pm 0.1023\,[0.0458]$	$0.00510 \pm 0.00502\,[0.00224]$	$0.3238 \pm 0.0177\,[0.0079]$	$0.9040 \pm 0.0820\,[0.0367]$
TimesFM-1.0	$-0.1481 \pm 0.1793\,[0.0802]$	$0.00652 \pm 0.01066\,[0.00477]$	$18.50 \pm 1.56\,[0.70]$	$51.79 \pm 7.09\,[3.17]$
TimesFM-2.0	$-0.00580 \pm 0.05826\,[0.02606]$	$0.00108 \pm 0.00087\,[0.00039]$	$0.3983 \pm 0.0276\,[0.0123]$	$1.114 \pm 0.134\,[0.060]$
TimesFM-2.5	$-0.0407 \pm 0.0469\,[0.0210]$	$0.00157 \pm 0.00195\,[0.00087]$	$0.2134 \pm 0.0106\,[0.0047]$	$0.5952 \pm 0.0456\,[0.0204]$
TimesFM-3.0	$-0.0874 \pm 0.0890\,[0.0398]$	$0.00668 \pm 0.00753\,[0.00337]$	$0.1142 \pm 0.0021\,[0.0009]$	$0.3186 \pm 0.0208\,[0.0093]$
TiRex	$-0.0113 \pm 0.0319\,[0.0143]$	$0.00035 \pm 0.00043\,[0.00019]$	$0.1384 \pm 0.0080\,[0.0036]$	$0.3855 \pm 0.0235\,[0.0105]$
TiRex-2	$-0.0323 \pm 0.1267\,[0.0566]$	$0.00487 \pm 0.00506\,[0.00226]$	$0.1704 \pm 0.0175\,[0.0078]$	$0.4757 \pm 0.0596\,[0.0267]$
Mamba4Cast	$-0.0594 \pm 0.1299\,[0.0581]$	$0.00669 \pm 0.00554\,[0.00248]$	$0.7591 \pm 0.0858\,[0.0384]$	$2.180 \pm 0.213\,[0.095]$

tween -0.148 and 0.102, placing their coverage dependence near the non-retrieval references rather than the phase-space-indexed references.

Phase-space similarity versus temporal recency. For the comparison between phase-space and temporal allocation to be informative, phase-space proximity must be empirically distinguishable from temporal recency on the tested dynamics. We verify this directly without applying a Theiler exclusion. For 300 windows from each system, we retrieve the nearest historical state in phase space and record how far back it lies in the context. Because no temporal exclusion is imposed for this analysis, Figure 5 reflects the geometry of the trajectories rather than a filtering rule.

Table 13: Temporal separation of nearest phase-space states, computed without a Theiler exclusion. T is the adjacent-state scale used as the Theiler window in the forecasting experiments; $d_{\mathrm{adj}}/d_{\mathrm{nn}}$ compares the phase-space distance of the immediately preceding state with that of the actual nearest state.

system	T	**median offset**	$\Pr(\Delta \le T)$	$\Pr(\Delta > 128)$	$d_{\mathrm{adj}}/d_{\mathrm{nn}}$
Lorenz	26	229.0	4.7%	64.7%	4.11
Chua	29	159.5	28.3%	64.7%	1.64
Chen	33	205.5	15.7%	62.7%	2.17

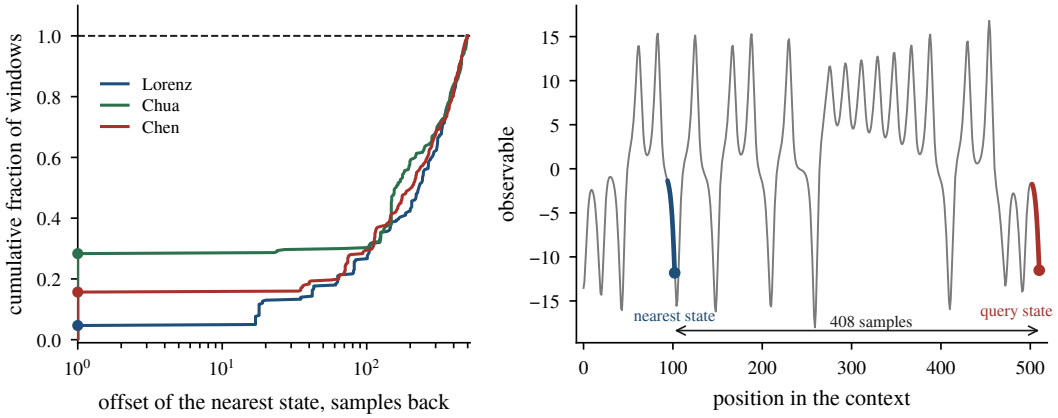

Figure 5: Phase-space proximity is distinct from temporal recency. Left: the empirical distribution of the temporal offset of the nearest phase-space state for Lorenz, Chua, and Chen, computed without a Theiler exclusion. A purely recency-based selection would choose offset one, while a recurrence occurring uniformly over the available context would follow the diagonal. Right: a representative Lorenz window in which the query state and its nearest phase-space recurrence are separated by 408 samples.

The distinction is substantial in all three systems. The median nearest-state offset is 159.5–229 samples, and 62.7%–64.7% of the retrieved states lie more than 128 samples before the query. Only 4.7% of Lorenz, 15.7% of Chen, and 28.3% of Chua windows place the nearest state within the adjacent-state scale T. The reported adjacent-to-nearest phase-space distance ratio is 1.64–4.11 across the three systems. Thus, the dynamical analogue most relevant to a phase-space-indexed forecaster is commonly far from the prediction point in temporal position.

E.4 BLOCK-DISPLACEMENT PROFILE

Table 14: Block-displacement calibration for references with block size 16. Responses use the 5σ amplitude.

reference	$\Delta = 0$	$\Delta = 2$	$\Delta = 4$	$\Delta = 8$	$\Delta = 16$	$\Delta = 32$	$\Delta = 64$	$\Delta = 128$	$\Delta = 256$	far	far/unif.	amp. scaling
trailing-window mean	4.375 ± 0.000 [0.000]	–	–	–	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	–
uniform average	0.1566 ± 0.0000 [0.0000]	–	–	0.1566 ± 0.0000 [0.0000]	0.1566 ± 0.0000 [0.0000]	0.1566 ± 0.0000 [0.0000]	0.1566 ± 0.0000 [0.0000]	0.1566 ± 0.0000 [0.0000]	0.1566 ± 0.0000 [0.0000]	0.1566 ± 0.0000 [0.0000]	1.000 ± 0.000 [0.000]	1.000 ± 0.000 [0.000]
global linear fit	0.8494 ± 0.0081 [0.0036]	–	–	0.0546 ± 0.0025 [0.0011]	0.0546 ± 0.0017 [0.0007]	0.0573 ± 0.0023 [0.0010]	0.0562 ± 0.0017 [0.0008]	0.0550 ± 0.0022 [0.0010]	0.0556 ± 0.0014 [0.0006]	0.3552 ± 0.0089 [0.0040]	3.738 ± 0.405 [0.181]	–
trailing-window polynomial	0.6250 ± 0.0000 [0.0000]	–	–	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	–

Table 15: Block-displacement calibration for references with block size 32. Responses use the 5σ amplitude.

reference	$\Delta = 0$	$\Delta = 2$	$\Delta = 4$	$\Delta = 8$	$\Delta = 16$	$\Delta = 32$	$\Delta = 64$	$\Delta = 128$	$\Delta = 256$	far	far/unif.	amp. scaling
trailing-window mean	4.375 ± 0.000 [0.000]	–	–	–	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	–
uniform average	0.3131 ± 0.0000 [0.0000]	–	–	–	0.3131 ± 0.0000 [0.0000]	0.3131 ± 0.0000 [0.0000]	0.3131 ± 0.0000 [0.0000]	0.3131 ± 0.0000 [0.0000]	0.3131 ± 0.0000 [0.0000]	1.000 ± 0.000 [0.000]	1.000 ± 0.000 [0.000]	1.000 ± 0.000 [0.000]
global linear fit	0.7695 ± 0.0109 [0.0049]	–	–	–	0.0469 ± 0.0034 [0.0015]	0.0444 ± 0.0026 [0.0011]	0.0448 ± 0.0043 [0.0019]	0.0453 ± 0.0051 [0.0023]	0.0450 ± 0.0045 [0.0020]	0.1438 ± 0.0143 [0.0064]	2.350 ± 0.263 [0.118]	–
trailing-window polynomial	0.6250 ± 0.0000 [0.0000]	–	–	–	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	–

Table 16: Block-displacement responses for all TSFM checkpoints. Responses use the 5σ amplitude; block size is shown because it follows the model interface.

checkpoint	b	amp. scaling	$\Delta = 0$	$\Delta = 2$	$\Delta = 4$	$\Delta = 8$	$\Delta = 16$	$\Delta = 32$	$\Delta = 64$	$\Delta = 128$	$\Delta = 256$
Chronos-T5 tiny	16	0.2281 ± 0.0370 [0.0165]	0.2125 ± 0.0173 [0.0077]	0.0727 ± 0.0056 [0.0025]	0.0552 ± 0.0047 [0.0021]	0.0347 ± 0.0066 [0.0030]	0.0259 ± 0.0033 [0.0015]	0.0259 ± 0.0051 [0.0023]	0.0211 ± 0.0015 [0.0007]	0.0192 ± 0.0015 [0.0007]	0.0173 ± 0.0025 [0.0011]
Chronos-T5 mini	16	0.2800 ± 0.0384 [0.0172]	0.2045 ± 0.0143 [0.0064]	0.0583 ± 0.0093 [0.0041]	0.0429 ± 0.0033 [0.0015]	0.0344 ± 0.0045 [0.0020]	0.0304 ± 0.0051 [0.0023]	0.0248 ± 0.0014 [0.0006]	0.0240 ± 0.0024 [0.0011]	0.0221 ± 0.0037 [0.0017]	0.0217 ± 0.0014 [0.0006]
Chronos-T5 small	16	0.2069 ± 0.0258 [0.0116]	0.3297 ± 0.0584 [0.0261]	0.0412 ± 0.0062 [0.0028]	0.0308 ± 0.0061 [0.0027]	0.0233 ± 0.0029 [0.0013]	0.0192 ± 0.0014 [0.0006]	0.0163 ± 0.0022 [0.0010]	0.0134 ± 0.0019 [0.0009]	0.0127 ± 0.0007 [0.0003]	0.0127 ± 0.0010 [0.0004]
Chronos-T5 base	16	0.1405 ± 0.0258 [0.0116]	0.5448 ± 0.0999 [0.0447]	0.0307 ± 0.0054 [0.0024]	0.0260 ± 0.0027 [0.0012]	0.0192 ± 0.0011 [0.0005]	0.0172 ± 0.0016 [0.0007]	0.0150 ± 0.0022 [0.0010]	0.0128 ± 0.0010 [0.0004]	0.0111 ± 0.0008 [0.0004]	0.0102 ± 0.0022 [0.0010]
Chronos-T5 large	16	0.1616 ± 0.0238 [0.0106]	0.3756 ± 0.0330 [0.1147]	0.0431 ± 0.0067 [0.0030]	0.0259 ± 0.0028 [0.0013]	0.0201 ± 0.0027 [0.0012]	0.0174 ± 0.0008 [0.0003]	0.0138 ± 0.0007 [0.0003]	0.0124 ± 0.0006 [0.0003]	0.0118 ± 0.0015 [0.0007]	0.0103 ± 0.0007 [0.0003]
Chronos-Bolt tiny	16	0.4790 ± 0.0286 [0.0128]	1.029 ± 0.163 [0.073]	–	–	–	–	0.1502 ± 0.0134 [0.0060]	0.1831 ± 0.0204 [0.0091]	0.1000 ± 0.0069 [0.0031]	0.0753 ± 0.0068 [0.0030]	0.0659 ± 0.0080 [0.0036]
Chronos-Bolt mini	16	0.3909 ± 0.0368 [0.0164]	0.7006 ± 0.0483 [0.0216]	–	–	–	–	0.1178 ± 0.0114 [0.0051]	0.0733 ± 0.0116 [0.0052]	0.0677 ± 0.0071 [0.0032]	0.0535 ± 0.0045 [0.0020]	0.0537 ± 0.0028 [0.0013]
Chronos-Bolt small	16	0.4374 ± 0.0458 [0.0205]	0.3784 ± 0.0623 [0.0279]	–	–	–	–	0.1053 ± 0.0105 [0.0047]	0.0799 ± 0.0122 [0.0055]	0.0667 ± 0.0051 [0.0023]	0.0509 ± 0.0081 [0.0036]	0.0581 ± 0.0035 [0.0016]
Chronos-Bolt base	16	0.3711 ± 0.0349 [0.0156]	0.6687 ± 0.0652 [0.0291]	–	–	–	–	0.0763 ± 0.0104 [0.0046]	0.0629 ± 0.0116 [0.0052]	0.0705 ± 0.0072 [0.0032]	0.0400 ± 0.0016 [0.0007]	0.0420 ± 0.0066 [0.0029]
Chronos-2	16	0.2487 ± 0.0209 [0.0093]	0.5310 ± 0.0661 [0.0296]	–	–	–	–	0.0693 ± 0.0080 [0.0036]	0.0555 ± 0.0055 [0.0025]	0.0454 ± 0.0058 [0.0026]	0.0363 ± 0.0040 [0.0018]	0.0401 ± 0.0014 [0.0006]
TimesFM-1.0	32	1.029 ± 0.128 [0.057]	1.868 ± 0.332 [0.149]	–	–	–	–	–	2.209 ± 0.274 [0.122]	1.866 ± 0.385 [0.172]	2.143 ± 0.353 [0.158]	2.021 ± 0.414 [0.185]
TimesFM-2.0	32	0.2956 ± 0.0508 [0.0227]	1.619 ± 0.110 [0.049]	–	–	–	–	–	0.0831 ± 0.0126 [0.0057]	0.0813 ± 0.0139 [0.0062]	0.0720 ± 0.0079 [0.0035]	0.0766 ± 0.0106 [0.0047]
TimesFM-2.5	32	0.3130 ± 0.0125 [0.0056]	0.7990 ± 0.0552 [0.0247]	–	–	–	–	–	0.0631 ± 0.0077 [0.0034]	0.0322 ± 0.0025 [0.0011]	0.0324 ± 0.0037 [0.0016]	0.0391 ± 0.0032 [0.0014]
TiRex	32	0.1665 ± 0.0225 [0.0101]	0.5231 ± 0.0330 [0.0148]	–	–	–	–	–	0.0231 ± 0.0024 [0.0011]	0.0194 ± 0.0025 [0.0011]	0.0148 ± 0.0010 [0.0004]	0.0134 ± 0.0003 [0.0001]
TiRex-2	32	0.2050 ± 0.0240 [0.0107]	0.4422 ± 0.0220 [0.0098]	–	–	–	–	–	0.0390 ± 0.0039 [0.0017]	0.0218 ± 0.0029 [0.0013]	0.0184 ± 0.0034 [0.0015]	0.0188 ± 0.0020 [0.0009]
MambaCast	16	0.2588 ± 0.0314 [0.0141]	0.3438 ± 0.0210 [0.0094]	–	–	–	–	–	0.0465 ± 0.0026 [0.0012]	0.0255 ± 0.0030 [0.0013]	0.0275 ± 0.0030 [0.0013]	0.0266 ± 0.0028 [0.0012]
	16	1.126 ± 0.084 [0.037]	0.3503 ± 0.0397 [0.0178]	0.2933 ± 0.0302 [0.0135]	0.2817 ± 0.0406 [0.0181]	0.3097 ± 0.0344 [0.0154]	0.1898 ± 0.0181 [0.0081]	0.1750 ± 0.0266 [0.0119]	0.1567 ± 0.0138 [0.0062]	0.1321 ± 0.0163 [0.0073]	0.1318 ± 0.0155 [0.0069]

Tables 14–16 report the complete offset profiles at 5σ together with the amplitude-scaling statistic.

The reference profiles establish the two temporal extremes used in Proposition 3: the trailing-window mean is exactly zero beyond its support, while the uniform average is constant across offsets and has far/uniform ratio 1. Among the TSFMs, 15 checkpoints have far/uniform ratios between 0.045 and 0.451 and sublinear amplitude scaling between 0.140 and 0.479, indicating substantially weaker distant influence than the uniform reference. Mamba4Cast is much broader (0.843), while TimesFM-1.0 reverses the dominant profile with far/uniform ratio 6.649. Responses at offsets that overlap the query embedding are not interpreted as distant-context influence.

Zero-mean perturbation robustness. We also evaluate a zero-mean block perturbation, which removes the net block shift while preserving a local disturbance. Table 17 reports the results for all 17 checkpoints together with the four probe references.

Table 17: Zero-mean perturbation robustness. Cells report mean ± SD [SE] across five trajectories.

forecaster	far response	far/unif.	amp. scaling
trailing-window mean	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	–
uniform average	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	–
global linear fit	0.0771 ± 0.0087 [0.0039]	0.4920 ± 0.0563 [0.0252]	3.148 ± 0.188 [0.084]
trailing-window polynomial	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	–
Chronos-T5 tiny	0.0198 ± 0.0014 [0.0006]	0.1280 ± 0.0084 [0.0037]	0.3200 ± 0.0158 [0.0071]
Chronos-T5 mini	0.0241 ± 0.0025 [0.0011]	0.1540 ± 0.0167 [0.0075]	0.3680 ± 0.0164 [0.0073]
Chronos-T5 small	0.0136 ± 0.0010 [0.0004]	0.0880 ± 0.0045 [0.0020]	0.2420 ± 0.0295 [0.0132]
Chronos-T5 base	0.0123 ± 0.0014 [0.0006]	0.0780 ± 0.0084 [0.0037]	0.1840 ± 0.0288 [0.0129]
Chronos-T5 large	0.0110 ± 0.0005 [0.0002]	0.0700 ± 0.0000 [0.0000]	0.1740 ± 0.0055 [0.0024]
Chronos-Bolt tiny	0.0671 ± 0.0051 [0.0023]	0.4280 ± 0.0349 [0.0156]	0.4500 ± 0.0464 [0.0207]
Chronos-Bolt mini	0.0499 ± 0.0046 [0.0021]	0.3180 ± 0.0295 [0.0132]	0.3380 ± 0.0363 [0.0162]
Chronos-Bolt small	0.0563 ± 0.0061 [0.0027]	0.3600 ± 0.0412 [0.0184]	0.4340 ± 0.1064 [0.0476]
Chronos-Bolt base	0.0470 ± 0.0015 [0.0007]	0.3000 ± 0.0141 [0.0063]	0.3220 ± 0.0192 [0.0086]
Chronos-2	0.0384 ± 0.0015 [0.0007]	0.2460 ± 0.0089 [0.0040]	0.2280 ± 0.0130 [0.0058]
TimesFM-1.0	2.120 ± 0.332 [0.148]	6.768 ± 1.058 [0.473]	0.6220 ± 0.0756 [0.0338]
TimesFM-2.0	0.0651 ± 0.0082 [0.0037]	0.2080 ± 0.0249 [0.0111]	0.1580 ± 0.0192 [0.0086]
TimesFM-2.5	0.0325 ± 0.0039 [0.0018]	0.1020 ± 0.0130 [0.0058]	0.1740 ± 0.0297 [0.0133]
TimesFM-3.0	0.0151 ± 0.0008 [0.0004]	0.0500 ± 0.0000 [0.0000]	0.1500 ± 0.0187 [0.0084]
TiRex	0.0253 ± 0.0023 [0.0011]	0.0809 ± 0.0075 [0.0034]	0.274 ± 0.040 [0.018]
TiRex-2	0.0263 ± 0.0016 [0.0007]	0.0860 ± 0.0055 [0.0024]	0.2340 ± 0.0152 [0.0068]
Mamba4Cast	0.1769 ± 0.0141 [0.0063]	1.130 ± 0.088 [0.039]	1.452 ± 0.090 [0.040]

The zero-mean perturbation preserves the same qualitative separation: the short-window references remain zero in the far field, while most TSFM checkpoints remain well below the uniform scale. TimesFM-1.0 again shows a large distant response and Mamba4Cast remains broad. This robustness check is secondary to the constant-displacement probe used for the main calibration.

Fine-offset robustness of temporal decay. The fine-offset sweep uses the weaker fine-grid perturbation (two-sample blocks at the same 5σ amplitude) and reports offsets 0–11 explicitly, together with a far-end summary. This sweep is intended to resolve behavior below the default block width.

The two-sample sweep preserves the dominant temporal pattern: responses remain more concentrated near the prediction point for 16 of the 17 checkpoints with available measurements, with TimesFM-1.0 again the exception. The absolute far-end ratio, however, increases substantially relative to the 16- or 32-sample probe and exceeds the uniform reference for 12 checkpoints. Thus, the far-end response depends on block size, while its decay with temporal distance is substantially more stable.

Table 18: Fine-offset responses for all TSFM checkpoints. Each cell reports mean ± SD [SE] over the five trajectories.

checkpoint	0	1	2	3	4	5	6	7	8	9	10	11	far
Chronos-T5 tiny	0.1547 ± 0.0232 [0.0104]	0.1412 ± 0.0189 [0.0085]	0.0693 ± 0.0079 [0.0035]	0.0463 ± 0.0045 [0.0020]	0.0347 ± 0.0046 [0.0020]	0.0282 ± 0.0030 [0.0014]	0.0271 ± 0.0035 [0.0016]	0.0247 ± 0.0026 [0.0012]	0.0220 ± 0.0039 [0.0017]	0.0204 ± 0.0022 [0.0010]	0.0179 ± 0.0017 [0.0008]	0.0173 ± 0.0020 [0.0009]	0.0142 ± 0.0020 [0.0009]
Chronos-T5 mini	0.1165 ± 0.0172 [0.0077]	0.0753 ± 0.0064 [0.0028]	0.0454 ± 0.0046 [0.0021]	0.0345 ± 0.0027 [0.0012]	0.0256 ± 0.0023 [0.0010]	0.0234 ± 0.0025 [0.0011]	0.0204 ± 0.0045 [0.0020]	0.0174 ± 0.0017 [0.0008]	0.0170 ± 0.0034 [0.0015]	0.0148 ± 0.0013 [0.0006]	0.0144 ± 0.0014 [0.0006]	0.0135 ± 0.0033 [0.0015]	0.0112 ± 0.0019 [0.0008]
Chronos-T5 small	0.1165 ± 0.0172 [0.0077]	0.0743 ± 0.0064 [0.0028]	0.0456 ± 0.0047 [0.0021]	0.0336 ± 0.0027 [0.0012]	0.0256 ± 0.0023 [0.0010]	0.0224 ± 0.0025 [0.0011]	0.0204 ± 0.0045 [0.0020]	0.0174 ± 0.0017 [0.0008]	0.0170 ± 0.0034 [0.0015]	0.0148 ± 0.0013 [0.0006]	0.0144 ± 0.0014 [0.0006]	0.0135 ± 0.0033 [0.0015]	0.0112 ± 0.0019 [0.0008]
Chronos-T5 base	0.1199 ± 0.0186 [0.0085]	0.0345 ± 0.0034 [0.0015]	0.0163 ± 0.0014 [0.0006]	0.0152 ± 0.0016 [0.0007]	0.0144 ± 0.0015 [0.0007]	0.0148 ± 0.0013 [0.0006]	0.0148 ± 0.0009 [0.0004]	0.0143 ± 0.0016 [0.0007]	0.0137 ± 0.0005 [0.0002]	0.0131 ± 0.0017 [0.0008]	0.0122 ± 0.0012 [0.0005]	0.0124 ± 0.0006 [0.0002]	0.0100 ± 0.0007 [0.0003]
Chronos-T5 large	0.0751 ± 0.0110 [0.0049]	0.0274 ± 0.0035 [0.0016]	0.0173 ± 0.0033 [0.0015]	0.0148 ± 0.0029 [0.0013]	0.0131 ± 0.0019 [0.0009]	0.0129 ± 0.0013 [0.0006]	0.0132 ± 0.0020 [0.0009]	0.0117 ± 0.0014 [0.0006]	0.0101 ± 0.0019 [0.0008]	0.00986 ± 0.0244 [0.0109]	0.0104 ± 0.0024 [0.0011]	0.0110 ± 0.0018 [0.0008]	0.00824 ± 0.00044 [0.0020]
Chronos-Bolt tiny	0.3031 ± 0.0221 [0.0099]	0.1862 ± 0.0106 [0.0047]	0.1783 ± 0.0130 [0.0058]	0.1492 ± 0.0198 [0.0089]	0.1419 ± 0.0186 [0.0083]	0.1361 ± 0.0149 [0.0067]	0.1212 ± 0.0229 [0.0102]	0.1291 ± 0.0172 [0.0077]	0.1528 ± 0.0049 [0.0022]	0.1321 ± 0.0092 [0.0041]	0.1419 ± 0.0085 [0.0038]	0.1077 ± 0.0054 [0.0024]	0.0950 ± 0.0087 [0.0039]
Chronos-Bolt mini	0.4200 ± 0.0321 [0.0233]	0.2083 ± 0.0204 [0.0091]	0.2220 ± 0.0281 [0.0126]	0.2240 ± 0.0473 [0.0211]	0.1649 ± 0.0196 [0.0088]	0.1305 ± 0.0180 [0.0080]	0.1119 ± 0.0119 [0.0053]	0.1174 ± 0.0154 [0.0069]	0.1169 ± 0.0100 [0.0045]	0.1024 ± 0.0170 [0.0076]	0.1262 ± 0.0175 [0.0078]	0.1116 ± 0.0095 [0.0042]	0.0819 ± 0.0044 [0.0020]
Chronos-Bolt small	0.2477 ± 0.0222 [0.0099]	0.1642 ± 0.0109 [0.0049]	0.1375 ± 0.0059 [0.0026]	0.1290 ± 0.0079 [0.0035]	0.1136 ± 0.0098 [0.0044]	0.1033 ± 0.0092 [0.0041]	0.1014 ± 0.0140 [0.0063]	0.1144 ± 0.0169 [0.0075]	0.1194 ± 0.0228 [0.0102]	0.1085 ± 0.0141 [0.0063]	0.0952 ± 0.0146 [0.0065]	0.0959 ± 0.0134 [0.0060]	0.0270 ± 0.0028 [0.0012]
Chronos-Bolt base	0.4944 ± 0.0254 [0.0114]	0.1366 ± 0.0122 [0.0054]	0.1346 ± 0.0068 [0.0031]	0.1477 ± 0.0129 [0.0058]	0.1107 ± 0.0134 [0.0060]	0.1023 ± 0.0103 [0.0046]	0.1223 ± 0.0082 [0.0037]	0.0962 ± 0.0067 [0.0030]	0.0890 ± 0.0080 [0.0036]	0.0940 ± 0.0073 [0.0033]	0.0798 ± 0.0062 [0.0028]	0.0789 ± 0.0071 [0.0032]	0.0660 ± 0.0031 [0.0014]
Chronos-2	0.3933 ± 0.0462 [0.0207]	0.1481 ± 0.0220 [0.0098]	0.1205 ± 0.0194 [0.0087]	0.1224 ± 0.0127 [0.0057]	0.1343 ± 0.0096 [0.0043]	0.1115 ± 0.0126 [0.0056]	0.1018 ± 0.0068 [0.0030]	0.1118 ± 0.0125 [0.0056]	0.1028 ± 0.0088 [0.0039]	0.1050 ± 0.0118 [0.0053]	0.1072 ± 0.0054 [0.0024]	0.1057 ± 0.0056 [0.0025]	0.0324 ± 0.0034 [0.0015]
TimesFM-1.0	1.400 ± 0.153 [0.068]	1.227 ± 0.065 [0.029]	1.218 ± 0.103 [0.046]	1.360 ± 0.150 [0.067]	1.212 ± 0.137 [0.061]	1.175 ± 0.096 [0.043]	1.360 ± 0.113 [0.051]	1.252 ± 0.120 [0.054]	1.095 ± 0.175 [0.078]	1.474 ± 0.169 [0.075]	1.237 ± 0.133 [0.059]	1.279 ± 0.222 [0.099]	1.331 ± 0.194 [0.087]
TimesFM-2.0	0.7093 ± 0.0368 [0.0165]	0.2600 ± 0.0259 [0.0116]	0.1423 ± 0.0236 [0.0114]	0.1584 ± 0.0260 [0.0116]	0.0943 ± 0.0146 [0.0065]	0.0855 ± 0.0039 [0.0018]	0.0888 ± 0.0076 [0.0034]	0.0825 ± 0.0089 [0.0040]	0.0805 ± 0.0093 [0.0042]	0.0920 ± 0.0190 [0.0085]	0.0878 ± 0.0142 [0.0063]	0.0884 ± 0.0112 [0.0050]	0.0842 ± 0.0072 [0.0032]
TimesFM-2.5	1.277 ± 0.020 [0.009]	0.1242 ± 0.0079 [0.0035]	0.1154 ± 0.0113 [0.0051]	0.0692 ± 0.0051 [0.0023]	0.1035 ± 0.0165 [0.0073]	0.0758 ± 0.0105 [0.0047]	0.0652 ± 0.0116 [0.0052]	0.0589 ± 0.0078 [0.0035]	0.1201 ± 0.0091 [0.0041]	0.0538 ± 0.0085 [0.0038]	0.1011 ± 0.0097 [0.0043]	0.0498 ± 0.0068 [0.0030]	0.0783 ± 0.0045 [0.0020]
TimesFM-3.0	0.5429 ± 0.0290 [0.0116]	0.1362 ± 0.0098 [0.0044]	0.1154 ± 0.0061 [0.0027]	0.0696 ± 0.0041 [0.0018]	0.0755 ± 0.0078 [0.0035]	0.0752 ± 0.0051 [0.0023]	0.0653 ± 0.0070 [0.0031]	0.0534 ± 0.0085 [0.0038]	0.0547 ± 0.0041 [0.0018]	0.0493 ± 0.0051 [0.0023]	0.0577 ± 0.0040 [0.0018]	0.0498 ± 0.0062 [0.0028]	0.0729 ± 0.0041 [0.0018]
TiRex	0.3230 ± 0.0427 [0.0191]	0.1746 ± 0.0125 [0.0056]	0.0848 ± 0.0088 [0.0036]	0.0694 ± 0.0061 [0.0027]	0.0762 ± 0.0064 [0.0029]	0.0496 ± 0.0047 [0.0021]	0.0438 ± 0.0043 [0.0019]	0.0418 ± 0.0043 [0.0019]	0.0331 ± 0.0034 [0.0015]	0.0299 ± 0.0021 [0.0009]	0.0332 ± 0.0028 [0.0013]	0.0426 ± 0.0056 [0.0025]	0.0203 ± 0.0022 [0.0010]
TiRex-2	0.4024 ± 0.0294 [0.0132]	0.1029 ± 0.0115 [0.0052]	0.0541 ± 0.0060 [0.0027]	0.0441 ± 0.0031 [0.0014]	0.0420 ± 0.0050 [0.0022]	0.0475 ± 0.0017 [0.0008]	0.0548 ± 0.0073 [0.0033]	0.0449 ± 0.0065 [0.0029]	0.0420 ± 0.0042 [0.0019]	0.0412 ± 0.0043 [0.0019]	0.0400 ± 0.0049 [0.0022]	0.0396 ± 0.0034 [0.0015]	0.0307 ± 0.0029 [0.0013]
MambaCast	0.7163 ± 0.0691 [0.0309]	0.5332 ± 0.0466 [0.0209]	0.2322 ± 0.0170 [0.0076]	0.2366 ± 0.0279 [0.0125]	0.1793 ± 0.0221 [0.0099]	0.4258 ± 0.0361 [0.0161]	0.2353 ± 0.0194 [0.0087]	0.2427 ± 0.0394 [0.0176]	0.2431 ± 0.0126 [0.0056]	0.2855 ± 0.0207 [0.0092]	0.2251 ± 0.0287 [0.0128]	0.3406 ± 0.0311 [0.0139]	0.1697 ± 0.0138 [0.0062]

Table 19: Fine-offset calibration for the trailing-mean and uniform references.

reference	0	1	2	3	4	5	6	7	8	9	10	11	far
trailing-window mean	1.250 ± 0.000 [0.000]	1.250 ± 0.000 [0.000]	1.250 ± 0.000 [0.000]	1.250 ± 0.000 [0.000]	1.250 ± 0.000 [0.000]	1.250 ± 0.000 [0.000]	0.6250 ± 0.0000 [0.0000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]
uniform average	0.0196 ± 0.0000 [0.0000]	0.0196 ± 0.0000 [0.0000]	0.0196 ± 0.0000 [0.0000]	0.0196 ± 0.0000 [0.0000]	0.0196 ± 0.0000 [0.0000]	0.0196 ± 0.0000 [0.0000]	0.0196 ± 0.0000 [0.0000]	0.0196 ± 0.0000 [0.0000]	0.0196 ± 0.0000 [0.0000]	0.0196 ± 0.0000 [0.0000]	0.0196 ± 0.0000 [0.0000]	0.0196 ± 0.0000 [0.0000]	0.0196 ± 0.0000 [0.0000]

E.5 TARGETED NEAREST-NEIGHBOR PERTURBATION

Table 20: Targeted-perturbation calibration for block size 16.

reference	retrieved	random	tail	retr./random
nearest-neighbor successor	0.0809 ± 0.0146 [0.0065]	0.00000 ± 0.00000 [0.00000]	0.7220 ± 0.4026 [0.1801]	∞
phase-space local-linear	0.0145 ± 0.0128 [0.0057]	0.00000 ± 0.00000 [0.00000]	1.842 ± 0.065 [0.029]	∞
trailing-window mean	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	2.000 ± 0.000 [0.000]	∞
trailing-window polynomial	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	2.000 ± 0.000 [0.000]	∞
uniform average	0.0626 ± 0.0000 [0.0000]	0.0626 ± 0.0000 [0.0000]	0.0626 ± 0.0000 [0.0000]	1.000 ± 0.000 [0.000]
global linear fit	0.0158 ± 0.0014 [0.0006]	0.0153 ± 0.0012 [0.0005]	1.927 ± 0.003 [0.001]	1.039 ± 0.118 [0.053]
flat-far-end reference	0.0716 ± 0.0000 [0.0000]	0.0716 ± 0.0000 [0.0000]	4.000 ± 0.000 [0.000]	1.000 ± 0.000 [0.000]

Table 21: Targeted-perturbation calibration for block size 32.

reference	retrieved	random	tail	retr./random
nearest-neighbor successor	0.0912 ± 0.0164 [0.0073]	0.00000 ± 0.00000 [0.00000]	1.029 ± 0.285 [0.128]	∞
phase-space local-linear	0.0324 ± 0.0281 [0.0126]	0.00000 ± 0.00000 [0.00000]	1.927 ± 0.033 [0.015]	∞
trailing-window mean	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	2.000 ± 0.000 [0.000]	∞
trailing-window polynomial	0.00000 ± 0.00000 [0.00000]	0.00000 ± 0.00000 [0.00000]	2.000 ± 0.000 [0.000]	∞
uniform average	0.1252 ± 0.0000 [0.0000]	0.1252 ± 0.0000 [0.0000]	0.1252 ± 0.0000 [0.0000]	1.000 ± 0.000 [0.000]
global linear fit	0.0173 ± 0.0023 [0.0010]	0.0129 ± 0.0015 [0.0007]	1.938 ± 0.005 [0.002]	1.345 ± 0.190 [0.085]
flat-far-end reference	0.1432 ± 0.0000 [0.0000]	0.1432 ± 0.0000 [0.0000]	4.000 ± 0.000 [0.000]	1.000 ± 0.000 [0.000]

Table 22: Targeted nearest-neighbor perturbation for all TSFM checkpoints.

checkpoint	b	retrieved	random	tail	retr./random
Chronos-T5 tiny	16	0.0130 ± 0.0020 [0.0009]	0.0111 ± 0.0007 [0.0003]	1.953 ± 0.015 [0.007]	1.175 ± 0.195 [0.087]
Chronos-T5 mini	16	0.0177 ± 0.0015 [0.0007]	0.0148 ± 0.0021 [0.0009]	1.950 ± 0.011 [0.005]	1.218 ± 0.191 [0.086]
Chronos-T5 small	16	0.0134 ± 0.0014 [0.0006]	0.0110 ± 0.0018 [0.0008]	1.987 ± 0.005 [0.002]	1.239 ± 0.253 [0.113]
Chronos-T5 base	16	0.0119 ± 0.0019 [0.0009]	0.00995 ± 0.00121 [0.00054]	1.985 ± 0.002 [0.001]	1.206 ± 0.176 [0.079]
Chronos-T5 large	16	0.0119 ± 0.0016 [0.0007]	0.00904 ± 0.00105 [0.00047]	1.978 ± 0.004 [0.002]	1.334 ± 0.240 [0.107]
Chronos-Bolt tiny	16	0.0424 ± 0.0043 [0.0019]	0.0389 ± 0.0057 [0.0025]	1.957 ± 0.013 [0.006]	1.119 ± 0.261 [0.117]
Chronos-Bolt mini	16	0.0373 ± 0.0030 [0.0013]	0.0359 ± 0.0042 [0.0019]	2.027 ± 0.022 [0.010]	1.046 ± 0.101 [0.045]
Chronos-Bolt small	16	0.0342 ± 0.0018 [0.0008]	0.0300 ± 0.0017 [0.0007]	2.039 ± 0.017 [0.008]	1.142 ± 0.066 [0.030]
Chronos-Bolt base	16	0.0299 ± 0.0033 [0.0015]	0.0269 ± 0.0025 [0.0011]	1.969 ± 0.017 [0.008]	1.122 ± 0.179 [0.080]
Chronos-2	16	0.0351 ± 0.0015 [0.0007]	0.0277 ± 0.0032 [0.0014]	2.032 ± 0.018 [0.008]	1.277 ± 0.148 [0.066]
TimesFM-1.0	32	0.6793 ± 0.0746 [0.0334]	0.6332 ± 0.0703 [0.0314]	3.257 ± 0.108 [0.048]	1.078 ± 0.117 [0.052]
TimesFM-2.0	32	0.0599 ± 0.0033 [0.0015]	0.0569 ± 0.0047 [0.0021]	2.046 ± 0.013 [0.006]	1.058 ± 0.085 [0.038]
TimesFM-2.5	32	0.0326 ± 0.0031 [0.0014]	0.0307 ± 0.0022 [0.0010]	1.994 ± 0.008 [0.003]	1.065 ± 0.099 [0.044]
TimesFM-3.0	32	0.0143 ± 0.0020 [0.0009]	0.0119 ± 0.0021 [0.0009]	1.981 ± 0.004 [0.002]	1.253 ± 0.372 [0.167]
TiRex	32	0.0133 ± 0.0018 [0.0008]	0.0119 ± 0.0025 [0.0011]	2.010 ± 0.003 [0.001]	1.147 ± 0.238 [0.106]
TiRex-2	32	0.0225 ± 0.0026 [0.0011]	0.0189 ± 0.0014 [0.0006]	2.014 ± 0.016 [0.007]	1.200 ± 0.202 [0.090]
Mamba4Cast	16	0.0683 ± 0.0120 [0.0054]	0.0567 ± 0.0060 [0.0027]	1.786 ± 0.018 [0.008]	1.224 ± 0.301 [0.135]

The targeted intervention sharply separates the phase-space references from location-insensitive controls. For both block sizes, the nearest-neighbor successor and phase-space local-linear references respond at the retrieved location and exactly zero at a random location, whereas the uniform average and flat-far-end reference give retrieved-to-random ratio 1. In contrast, the 17 TSFM ratios range only from 1.046 to 1.334.

The targeted perturbation uses amplitude 2σ. The retrieved-block offsets have quartiles 91, 210, and 340 for 16-sample blocks, compared with 131, 230, and 350 for the randomly drawn blocks (75, 194, and 324 against 127, 224, and 339 for 32-sample blocks). As a sensitivity analysis, we evaluate an offset-only response using each checkpoint's measured displacement profile; this gives retrieved-to-random ratios of 1.000–1.024. Because this calibration uses the separately measured displacement profile, we treat it as a sensitivity check rather than an exact position-matched control. The models therefore show at most a mild preference for the retrieved nearest state, consistent with the near-zero coverage exponents in Appendix E.3.

E.6 TRUNCATION AND FAR-CONTEXT REPLACEMENT

As an auxiliary normalized summary, we also report recovery

$$\mathcal{R} = \frac{E_{\mathrm{trunc}} - E_{\mathrm{repl}}}{E_{\mathrm{trunc}} - E_{\mathrm{full}}}, \tag{16}$$

only where truncation increases the error by more than 10% of the full-context error. Because both the denominator and the set of usable retained lengths vary across checkpoints, recovery is not used for the primary cross-model mechanism comparison.

The truncation results show that long-context utility is strongly model-dependent. At $k = 8$, error increases by approximately 204%–296% for Chronos-T5, 55%–81% for TiRex, and 201% for Mamba4Cast, while Chronos-Bolt and TimesFM contain checkpoints for which truncation improves the forecast. TimesFM-1.0 has negative truncation cost at every retained length in the replacement grid, so distant context provides no positive forecasting benefit there and its recovery statistic is undefined.

The shuffled replacement gives the most direct test of temporal organization. Under this replacement, the uniform average and flat-far-end reference have reported recovery equal to 1, consistent with their dependence on the far-end mean, which the intervention preserves. All 16 Lorenz checkpoints with defined recovery have negative aggregated recovery under the shuffled replacement, ranging from approximately -16.7 to -2.1, while TimesFM-1.0 has no retained length satisfying the recovery eligibility criterion. Same-system replacement gives more heterogeneous results: it supplies a coherent segment from the same system without preserving the original dynamical connection to the retained tail. Recovery provides an auxiliary summary of relative forecasting performance; its magnitude is not interpreted as an amount of information recovered.

Table 23: Truncation sweep for reference forecasters. Entries are raw one-step errors.

reference	E_{full}	$k=2$	$k=4$	$k=8$	$k=16$	$k=32$	$k=64$	$k=128$	$k=256$	flat from
nearest-neighbor successor	$0.3964 \pm 0.0225\,[0.0101]$	–	–	–	–	–	$4.479 \pm 0.686\,[0.307]$	$1.056 \pm 0.115\,[0.051]$	$0.5914 \pm 0.0546\,[0.0244]$	–
trailing-window mean	$4.631 \pm 0.188\,[0.084]$	$1.643 \pm 0.051\,[0.023]$	$2.669 \pm 0.025\,[0.011]$	$4.631 \pm 0.188\,[0.084]$	$4.631 \pm 0.188\,[0.084]$	$4.631 \pm 0.188\,[0.084]$	$4.631 \pm 0.188\,[0.084]$	$4.631 \pm 0.188\,[0.084]$	$4.631 \pm 0.188\,[0.084]$	8
trailing-window polynomial	$0.8962 \pm 0.0962\,[0.0430]$	$0.3174 \pm 0.0219\,[0.0098]$	$0.2058 \pm 0.0258\,[0.0116]$	$0.8962 \pm 0.0962\,[0.0430]$	$0.8962 \pm 0.0962\,[0.0430]$	$0.8962 \pm 0.0962\,[0.0430]$	$0.8962 \pm 0.0962\,[0.0430]$	$0.8962 \pm 0.0962\,[0.0430]$	$0.8962 \pm 0.0962\,[0.0430]$	8

Table 24: Truncation sweep for all TSFM checkpoints. Entries are relative error changes $100(E_k/E_{\text{full}} - 1)$ in percent.

checkpoint	$k=2$	$k=4$	$k=8$	$k=16$	$k=32$	$k=64$	$k=128$	$k=256$	flat from
Chronos-T5 tiny	$578.6 \pm 18.6\,[8.3]$	$554.7 \pm 45.9\,[20.5]$	$254.2 \pm 33.5\,[15.0]$	$133.9 \pm 12.7\,[5.7]$	$62.66 \pm 13.78\,[6.16]$	$26.90 \pm 9.97\,[4.46]$	$5.060 \pm 10.375\,[4.640]$	$-3.982 \pm 10.284\,[4.599]$	128
Chronos-T5 mini	$414.8 \pm 56.5\,[25.3]$	$419.3 \pm 50.4\,[22.6]$	$233.6 \pm 42.9\,[19.2]$	$94.43 \pm 11.69\,[5.23]$	$75.04 \pm 11.96\,[5.35]$	$28.70 \pm 5.36\,[2.40]$	$16.48 \pm 8.10\,[3.62]$	$7.576 \pm 5.002\,[2.237]$	128
Chronos-T5 small	$686.3 \pm 59.9\,[26.8]$	$683.1 \pm 70.4\,[31.5]$	$299.2 \pm 48.5\,[21.7]$	$104.7 \pm 18.7\,[8.3]$	$70.94 \pm 22.16\,[9.91]$	$35.82 \pm 13.65\,[6.11]$	$14.00 \pm 15.38\,[6.88]$	$4.482 \pm 8.715\,[3.898]$	128
Chronos-T5 base	$711.8 \pm 71.5\,[32.0]$	$513.3 \pm 37.7\,[16.9]$	$217.4 \pm 20.8\,[9.3]$	$64.85 \pm 13.24\,[5.92]$	$44.76 \pm 6.28\,[2.81]$	$24.01 \pm 7.02\,[3.14]$	$17.49 \pm 16.17\,[7.23]$	$4.240 \pm 5.886\,[2.632]$	128
Chronos-T5 large	$768.4 \pm 84.4\,[37.7]$	$584.4 \pm 76.2\,[34.1]$	$262.3 \pm 21.8\,[9.7]$	$90.47 \pm 21.81\,[9.75]$	$49.14 \pm 6.39\,[2.86]$	$44.26 \pm 15.08\,[6.74]$	$25.42 \pm 8.97\,[4.01]$	$11.07 \pm 6.90\,[3.09]$	128
Chronos-Bolt tiny	$12.56 \pm 14.01\,[6.27]$	$-19.19 \pm 8.57\,[3.83]$	$-34.65 \pm 6.10\,[2.73]$	$-48.71 \pm 3.84\,[1.72]$	$-3.236 \pm 10.337\,[4.623]$	$-3.014 \pm 10.235\,[4.577]$	$8.413 \pm 6.270\,[2.804]$	$-8.126 \pm 7.121\,[3.184]$	64
Chronos-Bolt mini	$37.99 \pm 14.10\,[6.30]$	$-25.70 \pm 4.30\,[1.92]$	$-2.177 \pm 7.717\,[3.451]$	$-29.28 \pm 10.55\,[4.72]$	$13.24 \pm 8.89\,[3.97]$	$9.906 \pm 11.503\,[5.144]$	$10.89 \pm 10.78\,[4.82]$	$2.747 \pm 6.726\,[3.008]$	96
Chronos-Bolt small	$47.75 \pm 9.14\,[4.09]$	$-8.994 \pm 2.770\,[1.239]$	$15.63 \pm 5.72\,[2.56]$	$-2.992 \pm 7.545\,[3.374]$	$2.275 \pm 14.746\,[6.595]$	$-6.681 \pm 10.119\,[4.525]$	$16.74 \pm 9.59\,[4.29]$	$4.448 \pm 11.381\,[5.090]$	128
Chronos-Bolt base	$98.88 \pm 19.01\,[8.50]$	$37.00 \pm 16.68\,[7.46]$	$9.460 \pm 14.985\,[6.702]$	$26.21 \pm 14.35\,[6.42]$	$38.42 \pm 19.24\,[8.60]$	$9.488 \pm 15.842\,[7.085]$	$16.88 \pm 7.13\,[3.19]$	$14.37 \pm 17.77\,[7.95]$	96
Chronos-2	$136.4 \pm 21.8\,[9.7]$	$6.325 \pm 9.559\,[4.275]$	$1.415 \pm 8.983\,[4.017]$	$17.62 \pm 6.07\,[2.71]$	$15.52 \pm 6.20\,[2.77]$	$18.73 \pm 13.27\,[5.93]$	$12.52 \pm 9.15\,[4.09]$	$1.052 \pm 9.980\,[4.463]$	128
TimesFM-1.0	$-91.91 \pm 2.01\,[0.90]$	$-85.60 \pm 3.17\,[1.42]$	$-24.53 \pm 20.00\,[8.94]$	$-60.60 \pm 5.11\,[2.29]$	$-16.49 \pm 7.79\,[3.48]$	$-20.99 \pm 11.59\,[5.18]$	$-17.29 \pm 15.44\,[6.91]$	$-6.240 \pm 11.874\,[5.310]$	128
TimesFM-2.0	$196.2 \pm 26.6\,[11.9]$	$-4.358 \pm 4.328\,[1.936]$	$-38.86 \pm 6.72\,[3.01]$	$-33.28 \pm 3.29\,[1.47]$	$30.95 \pm 19.21\,[8.59]$	$-12.46 \pm 8.51\,[3.80]$	$-2.126 \pm 12.549\,[5.612]$	$-1.663 \pm 14.321\,[6.404]$	128
TimesFM-2.5	$387.0 \pm 62.9\,[28.1]$	$119.2 \pm 32.5\,[14.5]$	$25.94 \pm 15.45\,[6.91]$	$33.07 \pm 15.51\,[6.93]$	$27.24 \pm 14.97\,[6.69]$	$6.878 \pm 4.544\,[2.032]$	$-7.631 \pm 7.037\,[3.147]$	$4.291 \pm 9.026\,[4.036]$	–
TimesFM-3.0	$155.8 \pm 22.6\,[10.1]$	$420.9 \pm 45.9\,[20.5]$	$44.17 \pm 16.53\,[7.39]$	$6.521 \pm 10.823\,[4.840]$	$59.38 \pm 13.22\,[5.91]$	$8.500 \pm 9.924\,[4.438]$	$16.18 \pm 13.89\,[6.21]$	$14.00 \pm 7.07\,[3.16]$	64
TiRex	$709.3 \pm 60.7\,[27.2]$	$84.97 \pm 8.40\,[3.75]$	$61.56 \pm 15.34\,[6.86]$	$31.23 \pm 9.43\,[4.22]$	$-13.55 \pm 4.48\,[2.00]$	$15.24 \pm 12.77\,[5.71]$	$5.721 \pm 7.710\,[3.448]$	$8.426 \pm 8.313\,[3.718]$	64
TiRex-2	$893.0 \pm 46.2\,[20.7]$	$109.5 \pm 6.4\,[2.9]$	$78.27 \pm 10.73\,[4.80]$	$31.36 \pm 6.29\,[2.81]$	$-1.664 \pm 6.527\,[2.919]$	$9.023 \pm 3.966\,[1.774]$	$-0.7777 \pm 7.2097\,[3.2243]$	$0.0760 \pm 8.0250\,[3.5889]$	–
Mamba4Cast	$27.70 \pm 7.15\,[3.20]$	$103.0 \pm 13.3\,[5.9]$	$198.5 \pm 26.2\,[11.7]$	$81.42 \pm 25.70\,[11.49]$	$-3.255 \pm 8.767\,[3.921]$	$5.743 \pm 11.661\,[5.215]$	$-6.006 \pm 7.112\,[3.181]$	$-3.063 \pm 8.629\,[3.859]$	64

Table 25: Far-context replacement/truncation cost sweep for reference forecasters. Entries are the relative cost, expressed as percentages.

reference	$k=2$	$k=4$	$k=8$	$k=16$	$k=32$	$k=64$	$k=128$
nearest-neighbor successor	–	–	–	–	–	$901.2 \pm 270.0\,[120.8]$	$152.4 \pm 39.8\,[17.8]$
phase-space local-linear	–	–	–	–	–	$4056.8 \pm 531.8\,[237.8]$	$665.0 \pm 111.4\,[49.8]$
trailing-window mean	$-64.34 \pm 1.76\,[0.79]$	$-41.81 \pm 3.00\,[1.34]$	$0.00000 \pm 0.00000\,[0.00000]$	$0.00000 \pm 0.00000\,[0.00000]$	$0.00000 \pm 0.00000\,[0.00000]$	$0.00000 \pm 0.00000\,[0.00000]$	$0.00000 \pm 0.00000\,[0.00000]$
trailing-window polynomial	$-64.65 \pm 1.69\,[0.76]$	$-76.93 \pm 2.26\,[1.01]$	$0.00000 \pm 0.00000\,[0.00000]$	$0.00000 \pm 0.00000\,[0.00000]$	$0.00000 \pm 0.00000\,[0.00000]$	$0.00000 \pm 0.00000\,[0.00000]$	$0.00000 \pm 0.00000\,[0.00000]$
uniform average	$-70.19 \pm 1.82\,[0.82]$	$-51.34 \pm 3.25\,[1.45]$	$-16.37 \pm 3.47\,[1.55]$	$6.003 \pm 10.261\,[4.589]$	$2.204 \pm 9.738\,[4.355]$	$-2.796 \pm 7.512\,[3.359]$	$-4.859 \pm 5.368\,[2.400]$
global linear fit	–	–	–	$-52.05 \pm 6.14\,[2.75]$	$-3.025 \pm 12.670\,[5.666]$	$9.353 \pm 5.728\,[2.562]$	$3.820 \pm 8.451\,[3.779]$
flat-far-end reference	$-71.89 \pm 1.67\,[0.75]$	$-54.15 \pm 2.55\,[1.14]$	$-20.95 \pm 7.09\,[3.17]$	$-20.95 \pm 7.09\,[3.17]$	$-20.95 \pm 7.09\,[3.17]$	$-20.95 \pm 7.09\,[3.17]$	$14.53 \pm 6.55\,[2.93]$

Table 26: Far-context replacement/truncation cost sweep for all TSFM checkpoints. Entries are the relative cost, expressed as percentages.

checkpoint	$k = 2$	$k = 4$	$k = 8$	$k = 16$	$k = 32$	$k = 64$	$k = 128$
Chronos-T5 tiny	$569.1 \pm 87.7\,[39.2]$	$541.3 \pm 79.3\,[35.5]$	$254.0 \pm 46.6\,[20.8]$	$138.4 \pm 22.9\,[10.3]$	$59.32 \pm 14.20\,[6.35]$	$26.92 \pm 13.41\,[6.00]$	$6.468 \pm 10.227\,[4.574]$
Chronos-T5 mini	$417.1 \pm 68.1\,[30.4]$	$415.0 \pm 45.0\,[20.1]$	$238.3 \pm 42.8\,[19.1]$	$94.03 \pm 16.68\,[7.46]$	$74.41 \pm 10.00\,[4.47]$	$28.68 \pm 13.35\,[5.97]$	$15.09 \pm 15.06\,[6.73]$
Chronos-T5 small	$676.9 \pm 96.0\,[43.0]$	$683.7 \pm 101.5\,[45.4]$	$295.8 \pm 41.9\,[18.8]$	$95.99 \pm 10.16\,[4.54]$	$64.74 \pm 16.62\,[7.43]$	$24.64 \pm 12.02\,[5.37]$	$8.727 \pm 13.362\,[5.976]$
Chronos-T5 base	$714.8 \pm 101.0\,[45.2]$	$509.7 \pm 56.3\,[25.2]$	$204.2 \pm 23.2\,[10.4]$	$60.21 \pm 17.14\,[7.66]$	$50.38 \pm 4.18\,[1.87]$	$23.18 \pm 9.81\,[4.39]$	$15.80 \pm 22.20\,[9.93]$
Chronos-T5 large	$783.7 \pm 149.8\,[67.0]$	$594.6 \pm 112.4\,[50.3]$	$251.9 \pm 56.0\,[25.0]$	$95.63 \pm 34.18\,[15.29]$	$51.30 \pm 11.59\,[5.19]$	$40.45 \pm 18.10\,[8.09]$	$25.78 \pm 13.94\,[6.23]$
Chronos-Bolt tiny	$12.19 \pm 14.17\,[6.34]$	$-16.07 \pm 11.99\,[5.36]$	$-35.22 \pm 5.74\,[2.57]$	$-48.50 \pm 6.46\,[2.89]$	$-2.041 \pm 13.025\,[5.825]$	$-3.200 \pm 12.091\,[5.407]$	$5.401 \pm 10.902\,[4.875]$
Chronos-Bolt mini	$38.06 \pm 14.60\,[6.53]$	$-27.46 \pm 7.19\,[3.21]$	$-7.068 \pm 6.896\,[3.084]$	$-30.77 \pm 5.01\,[2.24]$	$12.65 \pm 10.09\,[4.51]$	$9.908 \pm 15.756\,[7.046]$	$13.15 \pm 9.70\,[4.34]$
Chronos-Bolt small	$47.89 \pm 2.97\,[1.33]$	$-9.089 \pm 4.798\,[2.146]$	$13.44 \pm 4.17\,[1.86]$	$-5.986 \pm 5.025\,[2.247]$	$1.591 \pm 11.280\,[5.045]$	$-8.977 \pm 8.237\,[3.684]$	$13.45 \pm 7.50\,[3.35]$
Chronos-Bolt base	$91.37 \pm 31.94\,[14.29]$	$32.97 \pm 24.81\,[11.10]$	$5.972 \pm 20.886\,[9.340]$	$22.31 \pm 22.65\,[10.13]$	$36.08 \pm 28.32\,[12.67]$	$7.601 \pm 16.251\,[7.268]$	$13.83 \pm 11.70\,[5.23]$
Chronos-2	$133.5 \pm 25.2\,[11.3]$	$6.475 \pm 16.178\,[7.235]$	$2.498 \pm 15.209\,[6.802]$	$16.75 \pm 4.63\,[2.07]$	$14.80 \pm 6.56\,[2.93]$	$22.47 \pm 20.62\,[9.22]$	$11.07 \pm 10.98\,[4.91]$
TimesFM-1.0	$-92.06 \pm 1.45\,[0.65]$	$-85.73 \pm 3.53\,[1.58]$	$-31.36 \pm 17.96\,[8.03]$	$-63.04 \pm 5.24\,[2.34]$	$-17.64 \pm 4.08\,[1.82]$	$-22.69 \pm 11.63\,[5.20]$	$-22.71 \pm 15.82\,[7.07]$
TimesFM-2.0	$188.4 \pm 38.4\,[17.2]$	$-6.157 \pm 4.554\,[2.037]$	$-40.63 \pm 5.83\,[2.61]$	$-32.91 \pm 3.66\,[1.64]$	$32.98 \pm 22.93\,[10.25]$	$-12.68 \pm 7.50\,[3.35]$	$-5.071 \pm 18.047\,[8.071]$
TimesFM-2.5	$397.6 \pm 108.3\,[48.4]$	$127.5 \pm 48.3\,[21.6]$	$27.86 \pm 19.13\,[8.56]$	$39.89 \pm 25.60\,[11.45]$	$33.22 \pm 22.67\,[10.14]$	$10.14 \pm 9.04\,[4.04]$	$-5.276 \pm 13.534\,[6.053]$
TimesFM-3.0	$160.2 \pm 28.1\,[12.5]$	$427.8 \pm 47.0\,[21.0]$	$37.87 \pm 16.22\,[7.25]$	$7.461 \pm 10.099\,[4.516]$	$69.54 \pm 14.67\,[6.56]$	$8.009 \pm 17.720\,[7.925]$	$19.28 \pm 13.28\,[5.94]$
TiRex	$701.4 \pm 78.8\,[35.2]$	$81.18 \pm 13.86\,[6.20]$	$55.14 \pm 15.57\,[6.96]$	$31.45 \pm 8.67\,[3.88]$	$-14.11 \pm 2.60\,[1.16]$	$15.32 \pm 13.93\,[6.23]$	$4.111 \pm 2.082\,[0.931]$
TiRex-2	$926.1 \pm 33.7\,[15.1]$	$117.1 \pm 7.6\,[3.4]$	$80.90 \pm 10.46\,[4.68]$	$33.87 \pm 4.43\,[1.98]$	$1.891 \pm 8.549\,[3.823]$	$10.17 \pm 6.87\,[3.07]$	$-0.2799 \pm 11.8895\,[5.3172]$
Mamba4Cast	$25.42 \pm 10.70\,[4.79]$	$95.59 \pm 17.87\,[7.99]$	$200.8 \pm 26.2\,[11.7]$	$84.60 \pm 15.50\,[6.93]$	$-4.460 \pm 5.350\,[2.393]$	$4.457 \pm 13.713\,[6.133]$	$-11.82 \pm 11.81\,[5.28]$

Table 27: Recovery under far-context replacement for reference forecasters. Each recovery is followed by the mean number of usable retained-length settings in parentheses.

reference	**same-system**	**shuffled**	**Gaussian**	**constant**
nearest-neighbor successor	$0.3838 \pm 0.0732\,[0.0327]$ (2.0)	$0.2020 \pm 0.0480\,[0.0215]$ (2.0)	$0.1861 \pm 0.0802\,[0.0358]$ (2.0)	$0.2731 \pm 0.0410\,[0.0183]$ (2.0)
phase-space local-linear	$0.0918 \pm 0.0882\,[0.0394]$ (2.0)	$-2.877 \pm 0.784\,[0.351]$ (2.0)	$-3.232 \pm 0.683\,[0.305]$ (2.0)	$-0.3217 \pm 0.1702\,[0.0761]$ (2.0)
trailing-window mean	– (0.0)	– (0.0)	– (0.0)	– (0.0)
trailing-window polynomial	– (0.0)	– (0.0)	– (0.0)	– (0.0)
uniform average	$1.000 \pm 0.000\,[0.000]$ (0.4)	$1.000 \pm 0.000\,[0.000]$ (0.4)	$1.000 \pm 0.000\,[0.000]$ (0.4)	$1.000 \pm 0.000\,[0.000]$ (0.4)
global linear fit	$1.031 \pm 0.330\,[0.190]$ (1.2)	$-15.29 \pm 1.67\,[0.97]$ (1.2)	$-14.23 \pm 3.30\,[1.91]$ (1.2)	$0.4423 \pm 0.2028\,[0.1171]$ (1.2)
flat-far-end reference	$1.000 \pm 0.000\,[0.000]$ (0.8)	$1.000 \pm 0.000\,[0.000]$ (0.8)	$1.000 \pm 0.000\,[0.000]$ (0.8)	$1.000 \pm 0.000\,[0.000]$ (0.8)

Table 28: Recovery under far-context replacement for all TSFM checkpoints. Each recovery is followed by the mean number of usable retained-length settings in parentheses.

checkpoint	same-system	shuffled	Gaussian	constant
Chronos-T5 tiny	0.3180 ± 0.0725 [0.0324] (6.2)	-3.203 ± 0.633 [0.283] (6.2)	-3.182 ± 0.676 [0.302] (6.2)	-1.006 ± 0.334 [0.149] (6.2)
Chronos-T5 mini	0.2122 ± 0.0941 [0.0421] (6.4)	-3.828 ± 0.451 [0.202] (6.4)	-4.158 ± 0.531 [0.237] (6.4)	-1.385 ± 0.246 [0.110] (6.4)
Chronos-T5 small	0.2485 ± 0.1198 [0.0536] (6.2)	-5.708 ± 1.228 [0.549] (6.2)	-6.735 ± 1.133 [0.507] (6.2)	-2.176 ± 0.912 [0.408] (6.2)
Chronos-T5 base	0.1619 ± 0.1324 [0.0592] (6.6)	-7.599 ± 2.112 [0.945] (6.6)	-7.891 ± 1.867 [0.835] (6.6)	-1.871 ± 0.517 [0.231] (6.6)
Chronos-T5 large	0.4142 ± 0.0819 [0.0366] (6.8)	-3.631 ± 0.455 [0.204] (6.8)	-4.407 ± 0.593 [0.265] (6.8)	-1.782 ± 0.401 [0.179] (6.8)
Chronos-Bolt tiny	0.2558 ± 0.7919 [0.3960] (1.2)	-9.158 ± 6.758 [3.379] (1.2)	-10.11 ± 6.41 [3.20] (1.2)	-5.218 ± 5.452 [2.726] (1.2)
Chronos-Bolt mini	0.0841 ± 0.3397 [0.1519] (2.8)	-9.715 ± 5.356 [2.395] (2.8)	-10.03 ± 5.23 [2.34] (2.8)	-5.027 ± 4.061 [1.816] (2.8)
Chronos-Bolt small	0.6681 ± 0.5387 [0.2409] (2.4)	-11.59 ± 3.83 [1.71] (2.4)	-11.61 ± 3.87 [1.73] (2.4)	-5.905 ± 2.482 [1.110] (2.4)
Chronos-Bolt base	0.6605 ± 0.3569 [0.1596] (5.0)	-9.573 ± 5.073 [2.269] (5.0)	-9.658 ± 4.808 [2.150] (5.0)	-2.039 ± 0.988 [0.442] (5.0)
Chronos-2	0.3443 ± 0.6190 [0.2768] (4.6)	-6.711 ± 2.489 [1.113] (4.6)	-6.812 ± 2.442 [1.092] (4.6)	-2.614 ± 1.173 [0.525] (4.6)
TimesFM-1.0	– (0.0)	– (0.0)	– (0.0)	– (0.0)
TimesFM-2.0	0.9174 ± 0.3513 [0.1571] (2.2)	-3.580 ± 1.613 [0.721] (2.2)	-3.549 ± 1.398 [0.625] (2.2)	-0.3621 ± 0.3088 [0.1381] (2.2)
TimesFM-2.5	0.6894 ± 0.2112 [0.0945] (5.4)	-13.18 ± 5.57 [2.49] (5.4)	-12.54 ± 5.36 [2.40] (5.4)	-0.3189 ± 0.6390 [0.2857] (5.4)
TimesFM-3.0	-0.3524 ± 0.6401 [0.2862] (5.8)	-7.299 ± 2.664 [1.191] (5.8)	-7.814 ± 2.755 [1.232] (5.8)	-1.366 ± 0.173 [0.078] (5.8)
TiRex	0.1020 ± 0.2839 [0.1270] (4.6)	-16.70 ± 2.70 [1.21] (4.6)	-16.01 ± 3.40 [1.52] (4.6)	-2.494 ± 0.388 [0.174] (4.6)
TiRex-2	0.4448 ± 0.2088 [0.0934] (5.2)	-9.275 ± 1.406 [0.629] (5.2)	-9.038 ± 0.944 [0.422] (5.2)	-1.156 ± 0.183 [0.082] (5.2)
Mamba4Cast	0.8423 ± 0.0472 [0.0211] (4.2)	-2.136 ± 0.487 [0.218] (4.2)	-2.274 ± 0.533 [0.238] (4.2)	-0.7283 ± 0.5486 [0.2453] (4.2)

E.7 CROSS-SYSTEM RESULTS

Table 29: Chen results for all 17 checkpoints.

checkpoint	escape (%)	coverage exp.	far/unif.	amp. scaling	same-system	shuffled	Gaussian	constant
Chronos-T5 tiny	1.400	0.220	0.126	0.305	0.408	−2.698	−3.198	−1.777
Chronos-T5 mini	1.000	0.267	0.143	0.342	0.309	−4.621	−4.973	−3.143
Chronos-T5 small	0.900	0.274	0.085	0.225	−0.210	−11.050	−11.499	−10.976
Chronos-T5 base	1.000	0.337	0.056	0.129	−0.231	−10.710	−11.026	−13.947
Chronos-T5 large	0.800	0.216	0.065	0.161	0.396	−7.025	−7.165	−5.690
Chronos-Bolt tiny	1.700	0.121	0.801	0.671	–	–	–	–
Chronos-Bolt mini	2.200	0.040	0.401	0.375	−0.456	−5.444	−6.269	−4.405
Chronos-Bolt small	2.700	0.040	0.389	0.409	0.100	−21.737	−22.892	−13.557
Chronos-Bolt base	2.100	0.106	0.385	0.410	0.247	−8.080	−9.601	−2.930
Chronos-2	1.100	0.121	0.261	0.234	−0.468	−5.513	−5.010	−1.042
TimesFM-1.0	37.100	0.016	5.237	0.822	2.177	2.433	2.890	3.092
TimesFM-2.0	0.900	0.048	0.329	0.451	0.585	−2.227	−1.909	−0.844
TimesFM-2.5	1.000	0.054	0.101	0.272	0.047	−9.661	−9.075	−2.399
TimesFM-3.0	0.900	0.205	0.038	0.210	0.005	−6.275	−7.365	−1.228
TiRex	1.000	0.191	0.061	0.229	0.165	−6.965	−7.312	−1.928
TiRex-2	1.100	0.114	0.077	0.256	0.228	−5.805	−6.747	−0.687
Mamba4Cast	0.300	0.217	0.997	1.342	0.176	−4.423	−4.633	−1.397

Table 30: Chua results for all 17 checkpoints.

checkpoint	escape (%)	coverage exp.	far/unif.	amp. scaling	same-system	shuffled	Gaussian	constant
Chronos-T5 tiny	1.400	0.178	0.073	0.184	−0.271	−3.551	−6.410	−4.897
Chronos-T5 mini	0.600	0.260	0.091	0.160	−0.517	−3.677	−9.745	−4.279
Chronos-T5 small	1.000	0.208	0.064	0.178	−0.366	−4.878	−7.134	−5.742
Chronos-T5 base	1.000	0.075	0.048	0.128	−0.304	−8.617	−16.115	−8.884
Chronos-T5 large	1.000	0.125	0.044	0.117	−0.357	−8.998	−17.357	−14.468
Chronos-Bolt tiny	2.000	0.047	0.248	0.369	−5.394	−50.617	−54.530	−42.344
Chronos-Bolt mini	2.100	−0.030	0.256	0.335	−0.767	−18.085	−19.172	−11.975
Chronos-Bolt small	1.200	−0.032	0.243	0.402	−2.337	−37.041	−41.074	−24.355
Chronos-Bolt base	2.700	−0.022	0.210	0.491	−0.527	−27.341	−29.666	−7.738
Chronos-2	0.300	0.180	0.164	0.201	−1.313	−6.014	−10.828	−4.902
TimesFM-1.0	69.100	−0.100	5.137	0.488	–	–	–	–
TimesFM-2.0	2.000	0.043	0.149	0.214	−0.246	−10.217	−9.996	−3.842
TimesFM-2.5	0.800	0.026	0.070	0.186	0.454	−6.007	−11.824	−0.992
TimesFM-3.0	0.900	0.084	0.022	0.174	−1.055	−4.813	−4.858	−0.952
TiRex	0.200	0.107	0.031	0.153	−0.883	−7.926	−9.135	−3.140
TiRex-2	0.900	0.027	0.055	0.229	−2.115	−14.091	−17.126	−1.645
Mamba4Cast	2.000	0.016	0.639	0.854	0.072	−5.054	−4.866	−1.907

Across Chen and Chua, convex-hull escape remains positive for all 34 system–checkpoint pairs. Coverage exponents are somewhat more system-dependent than on Lorenz, ranging from 0.016 to 0.337 on Chen and from -0.100 to 0.260 on Chua, but they remain far below the phase-space reference exponent. Far/uniform is below 1 for 32 of the 34 pairs; TimesFM-1.0 is the repeated exception, with values above 5 on both systems. Recovery under the shuffled replacement is negative and worse than same-system passage in 31 of 32 defined pairs. The cross-system results therefore preserve the main qualitative pattern (Figure 6).

F STRUCTURAL ANALYSIS OF CHRONOS-T5

The input interventions above apply across architectures. Chronos-T5 additionally exposes encoder–decoder attention and intermediate activations, allowing an architecture-specific check of the temporal allocation observed behaviorally. Unless stated otherwise, cells are mean $\pm$ SD [SE] across the same five Lorenz trajectories. Cross-attention mass and linear probes are observational and ask whether distant-context information is represented or decodable from internal states. Edge cuts, encoder-position ablations, and activation patching instead intervene on internal routes and test whether changing those routes changes the forecast. Neither class of analysis uniquely identifies a mechanism or circuit, so these measurements are used as complementary structural evidence rather than as substitutes for the architecture-agnostic input diagnostics.

The attention profile changes systematically with model scale. The fraction of cross-attention assigned within 64 samples of the prediction point decreases from 41.5% for Chronos-T5 tiny to 23.2% for Chronos-T5 large, with the same monotonic decrease within 4 and 16 samples. Larger checkpoints therefore spread a greater share of cross-attention to distant context, while all five re-

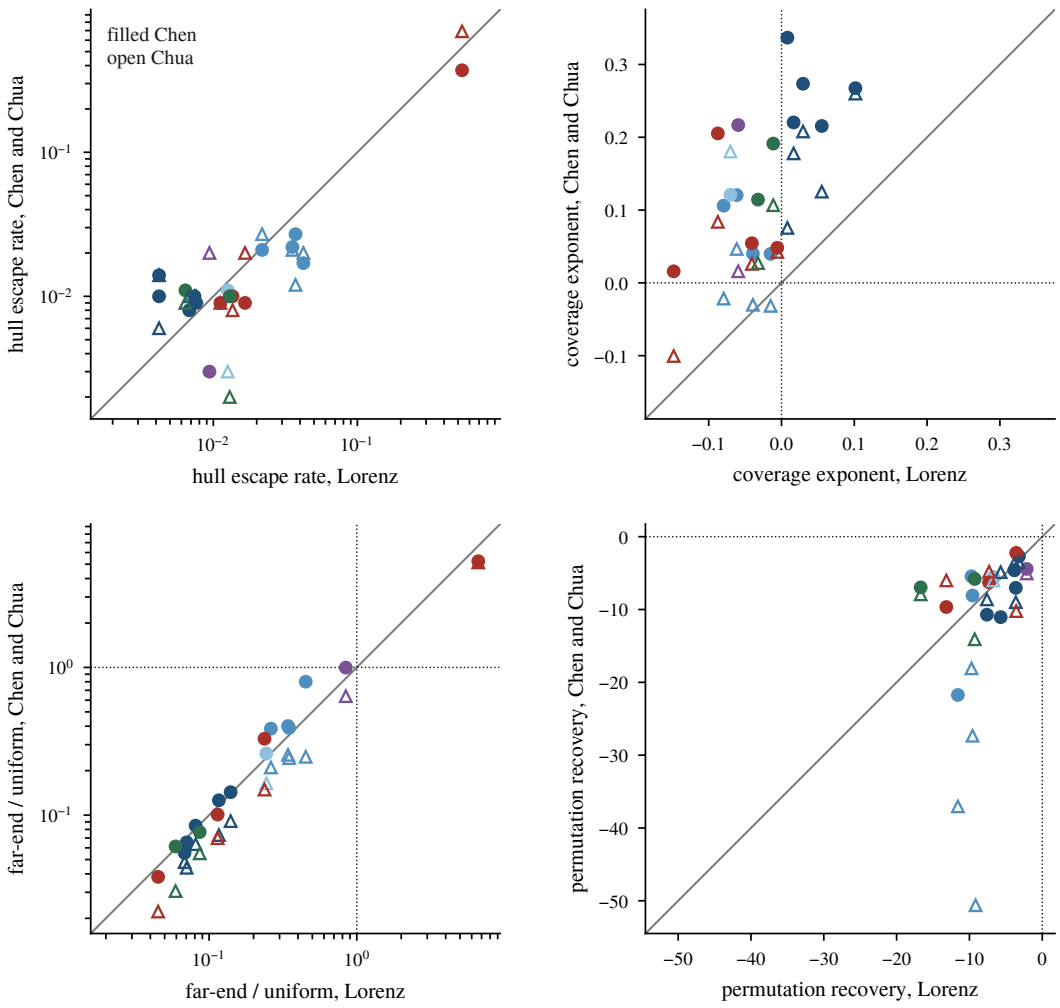

Figure 6: Each criterion's Lorenz reading against the same reading on Chen (filled) and Chua (open), one point per checkpoint, with the line of equality. A point on the line kept its reading under a change of system. Dotted lines mark the value that separates the two mechanisms on each axis. The far-end response ratio, lower left, is reproduced most closely; the coverage exponent, upper right, is slightly larger on both systems but remains far below the reference value of 1.11.

Table 31: Cross-attention mass near the prediction point.

checkpoint	≤ 4	≤ 16	≤ 64	**EOS mass**
Chronos-T5 tiny	22.4%	27.0%	41.5%	2.7%
Chronos-T5 mini	13.6%	17.2%	32.5%	10.0%
Chronos-T5 small	9.9%	13.4%	27.7%	7.0%
Chronos-T5 base	9.2%	12.6%	25.1%	3.6%
Chronos-T5 large	8.4%	11.5%	23.2%	4.9%

tain their strongest aggregate concentration near the prediction point. This is consistent with scaling broadening the temporal range represented by the model rather than reversing the recent-position preference.

For activation patching, a distant-context corruption first creates a clean– corrupted prediction gap. Patching a clean activation into the corrupted run at one site measures the fraction of that gap restored after the remaining network is recomputed. In Table 32, *gap* is the median clean–corrupted prediction difference normalized by the signal standard deviation; *noise floor* is the spread under re-

peated evaluation of an unchanged input; *scale ratio* is the corrupted-to-clean mean-absolute context scale; and *seam* is the substitution boundary step relative to a typical first difference. *Peak recovery* is the largest recovery on the encoder measurement path, while *direct contribution* is the recovery at the final encoder measurement site, after which no further encoder mixing can redistribute the patched state.

Table 32: Activation-patching scalar diagnostics for Chronos-T5.

checkpoint	gap	noise floor	scale ratio	seam	peak recovery	direct contrib.
Chronos-T5 tiny	0.0129 ± 0.0037 [0.0016]	0.00000 ± 0.00000 [0.00000]	1.000 ± 0.000 [0.000]	0.9446 ± 0.1874 [0.0838]	0.1758 ± 0.1042 [0.0466]	0.0845 ± 0.1824 [0.0816]
Chronos-T5 mini	0.0129 ± 0.0041 [0.0019]	0.00000 ± 0.00000 [0.00000]	1.000 ± 0.000 [0.000]	0.9446 ± 0.1874 [0.0838]	0.0883 ± 0.0649 [0.0290]	0.00495 ± 0.08151 [0.03645]
Chronos-T5 small	0.0137 ± 0.0020 [0.0009]	0.00000 ± 0.00000 [0.00000]	1.000 ± 0.000 [0.000]	0.9446 ± 0.1874 [0.0838]	0.2358 ± 0.1124 [0.0503]	-0.0891 ± 0.0614 [0.0275]
Chronos-T5 base	0.00694 ± 0.00207 [0.00092]	0.00000 ± 0.00000 [0.00000]	1.000 ± 0.000 [0.000]	0.9446 ± 0.1874 [0.0838]	0.4598 ± 0.1617 [0.0723]	0.1309 ± 0.1403 [0.0628]
Chronos-T5 large	0.00627 ± 0.00101 [0.00045]	0.00000 ± 0.00000 [0.00000]	1.000 ± 0.000 [0.000]	0.9446 ± 0.1874 [0.0838]	0.4345 ± 0.1934 [0.0865]	0.1353 ± 0.0914 [0.0409]

The corruption is well controlled: the scale ratio is exactly 1 for every checkpoint, the seam remains near ordinary trajectory variation, and the measured noise floor is zero under the deterministic readout. Peak encoder recovery is modest for tiny through small and larger for base and large, but it does not define a single common layer at which distant information enters or leaves the encoder.

Entry and exit sites. Using a recovery threshold of 0.2 as a descriptive summary, no stable common entry or exit layer is identified across trajectories. Entry is often not identified for the tiny and mini checkpoints; when it is identified in the larger checkpoints, it occurs at different encoder depths across trajectories. An exit before the final encoder boundary is likewise not consistently identified. We therefore interpret the full layer-wise recovery curves below, rather than assign a single localized circuit to distant-context processing.

Table 33: Layer-wise activation-patching recovery for Chronos-T5 tiny.

site	fraction restored
`embed`	0.00000 ± 0.00000 [0.00000]
`enc0`	0.00075 ± 0.00672 [0.00300]
`enc1`	0.1138 ± 0.0405 [0.0181]
`enc2`	0.0853 ± 0.0583 [0.0261]
`enc3`	0.0845 ± 0.1824 [0.0816]
`cross0`	0.4622 ± 0.2525 [0.1129]
`dec0`	0.4622 ± 0.2525 [0.1129]
`cross1`	1.032 ± 0.262 [0.117]
`dec1`	1.032 ± 0.262 [0.117]
`cross2`	1.185 ± 0.304 [0.136]
`dec2`	1.185 ± 0.304 [0.136]
`cross3`	1.000 ± 0.000 [0.000]
`dec3`	1.000 ± 0.000 [0.000]

Table 34: Layer-wise activation-patching recovery for Chronos-T5 mini.

site	fraction restored
`embed`	0.00000 ± 0.00000 [0.00000]
`enc0`	0.00343 ± 0.00985 [0.00440]
`enc1`	0.0768 ± 0.0722 [0.0323]
`enc2`	0.0708 ± 0.0437 [0.0196]
`enc3`	0.00495 ± 0.08151 [0.03645]
`cross0`	0.3623 ± 0.1221 [0.0546]
`dec0`	0.3623 ± 0.1221 [0.0546]
`cross1`	0.8975 ± 0.3695 [0.1653]
`dec1`	0.8975 ± 0.3695 [0.1653]
`cross2`	1.029 ± 0.115 [0.051]
`dec2`	1.029 ± 0.115 [0.051]
`cross3`	1.000 ± 0.000 [0.000]
`dec3`	1.000 ± 0.000 [0.000]

Table 35: Layer-wise activation-patching recovery for Chronos-T5 small.

site	fraction restored
embed	$0.00000 \pm 0.00000\,[0.00000]$
enc0	$-0.00016 \pm 0.00585\,[0.00262]$
enc1	$0.0714 \pm 0.0317\,[0.0142]$
enc2	$0.1626 \pm 0.0669\,[0.0299]$
enc3	$0.2333 \pm 0.1158\,[0.0518]$
enc4	$0.0191 \pm 0.0595\,[0.0266]$
enc5	$-0.0891 \pm 0.0614\,[0.0275]$
cross0	$0.2638 \pm 0.0745\,[0.0333]$
dec0	$0.2638 \pm 0.0745\,[0.0333]$
cross1	$0.3812 \pm 0.1939\,[0.0867]$
dec1	$0.3812 \pm 0.1939\,[0.0867]$
cross2	$0.6459 \pm 0.1450\,[0.0648]$
dec2	$0.6459 \pm 0.1450\,[0.0648]$
cross3	$1.039 \pm 0.110\,[0.049]$
dec3	$1.039 \pm 0.110\,[0.049]$
cross4	$1.033 \pm 0.061\,[0.027]$
dec4	$1.033 \pm 0.061\,[0.027]$
cross5	$1.000 \pm 0.000\,[0.000]$
dec5	$1.000 \pm 0.000\,[0.000]$

Table 36: Layer-wise activation-patching recovery for Chronos-T5 base.

site	fraction restored
embed	$0.00000 \pm 0.00000\ [0.00000]$
enc0	$-0.00329 \pm 0.00651\ [0.00291]$
enc1	$0.0277 \pm 0.0320\ [0.0143]$
enc2	$0.0707 \pm 0.0690\ [0.0309]$
enc3	$0.1831 \pm 0.1074\ [0.0480]$
enc4	$0.3369 \pm 0.1656\ [0.0741]$
enc5	$0.4063 \pm 0.1602\ [0.0717]$
enc6	$0.4534 \pm 0.1611\ [0.0721]$
enc7	$0.3920 \pm 0.1292\ [0.0578]$
enc8	$0.2872 \pm 0.1083\ [0.0484]$
enc9	$0.2445 \pm 0.1713\ [0.0766]$
enc10	$0.2117 \pm 0.1987\ [0.0889]$
enc11	$0.1309 \pm 0.1403\ [0.0628]$
cross0	$0.0984 \pm 0.1037\ [0.0464]$
dec0	$0.0984 \pm 0.1037\ [0.0464]$
cross1	$0.0619 \pm 0.1023\ [0.0457]$
dec1	$0.0619 \pm 0.1023\ [0.0457]$
cross2	$0.1682 \pm 0.1292\ [0.0578]$
dec2	$0.1682 \pm 0.1292\ [0.0578]$
cross3	$0.3546 \pm 0.1547\ [0.0692]$
dec3	$0.3546 \pm 0.1547\ [0.0692]$
cross4	$0.6041 \pm 0.1042\ [0.0466]$
dec4	$0.6041 \pm 0.1042\ [0.0466]$
cross5	$0.7576 \pm 0.1442\ [0.0645]$
dec5	$0.7576 \pm 0.1442\ [0.0645]$
cross6	$1.006 \pm 0.167\ [0.075]$
dec6	$1.006 \pm 0.167\ [0.075]$
cross7	$1.037 \pm 0.134\ [0.060]$
dec7	$1.037 \pm 0.134\ [0.060]$
cross8	$1.091 \pm 0.103\ [0.046]$
dec8	$1.091 \pm 0.103\ [0.046]$
cross9	$0.9916 \pm 0.0697\ [0.0311]$
dec9	$0.9916 \pm 0.0697\ [0.0311]$
cross10	$0.9823 \pm 0.0330\ [0.0148]$
dec10	$0.9823 \pm 0.0330\ [0.0148]$
cross11	$1.000 \pm 0.000\ [0.000]$
dec11	$1.000 \pm 0.000\ [0.000]$

Table 37: Layer-wise activation-patching recovery for Chronos-T5 large.

site	fraction restored	site	fraction restored
embed	0.00000 ± 0.00000 [0.00000]	dec11	0.8291 ± 0.0723 [0.0323]
enc0	-0.00332 ± 0.00656 [0.00293]	enc12	0.3600 ± 0.2269 [0.1015]
enc1	-0.0294 ± 0.0184 [0.0082]	enc13	0.3656 ± 0.2277 [0.1018]
enc2	-0.0423 ± 0.0355 [0.0159]	enc14	0.3212 ± 0.1674 [0.0748]
enc3	-0.0849 ± 0.0626 [0.0280]	enc15	0.2684 ± 0.1700 [0.0760]
enc4	-0.1127 ± 0.1366 [0.0611]	enc16	0.2805 ± 0.1546 [0.0691]
enc5	-0.0631 ± 0.1176 [0.0526]	enc17	0.2756 ± 0.1234 [0.0552]
enc6	-0.0360 ± 0.1225 [0.0548]	enc18	0.2384 ± 0.1058 [0.0473]
enc7	-0.0350 ± 0.0972 [0.0435]	enc19	0.1986 ± 0.1232 [0.0551]
enc8	0.0371 ± 0.1208 [0.0540]	enc20	0.1661 ± 0.0836 [0.0374]
enc9	0.1092 ± 0.1631 [0.0730]	enc21	0.2461 ± 0.0932 [0.0417]
enc10	0.1210 ± 0.1774 [0.0793]	enc22	0.1790 ± 0.0714 [0.0320]
enc11	0.2431 ± 0.1492 [0.0667]	enc23	0.1353 ± 0.0914 [0.0409]
cross0	0.0394 ± 0.0146 [0.0065]	cross12	0.9590 ± 0.1110 [0.0496]
dec0	0.0394 ± 0.0146 [0.0065]	dec12	0.9590 ± 0.1110 [0.0496]
cross1	0.1414 ± 0.0829 [0.0371]	cross13	1.130 ± 0.086 [0.039]
dec1	0.1414 ± 0.0829 [0.0371]	dec13	1.130 ± 0.086 [0.039]
cross2	0.1287 ± 0.0747 [0.0334]	cross14	1.140 ± 0.092 [0.041]
dec2	0.1287 ± 0.0747 [0.0334]	dec14	1.140 ± 0.092 [0.041]
cross3	0.1498 ± 0.0734 [0.0328]	cross15	1.201 ± 0.056 [0.025]
dec3	0.1498 ± 0.0734 [0.0328]	dec15	1.201 ± 0.056 [0.025]
cross4	0.2213 ± 0.1011 [0.0452]	cross16	1.202 ± 0.063 [0.028]
dec4	0.2213 ± 0.1011 [0.0452]	dec16	1.202 ± 0.063 [0.028]
cross5	0.2462 ± 0.0678 [0.0303]	cross17	1.137 ± 0.036 [0.016]
dec5	0.2462 ± 0.0678 [0.0303]	dec17	1.137 ± 0.036 [0.016]
cross6	0.3357 ± 0.0924 [0.0413]	cross18	1.117 ± 0.064 [0.029]
dec6	0.3357 ± 0.0924 [0.0413]	dec18	1.117 ± 0.064 [0.029]
cross7	0.4393 ± 0.1239 [0.0554]	cross19	1.042 ± 0.026 [0.012]
dec7	0.4393 ± 0.1239 [0.0554]	dec19	1.042 ± 0.026 [0.012]
cross8	0.4558 ± 0.1077 [0.0482]	cross20	1.015 ± 0.008 [0.003]
dec8	0.4558 ± 0.1077 [0.0482]	dec20	1.015 ± 0.008 [0.003]
cross9	0.5125 ± 0.1133 [0.0507]	cross21	1.002 ± 0.006 [0.003]
dec9	0.5125 ± 0.1133 [0.0507]	dec21	1.002 ± 0.006 [0.003]
cross10	0.6355 ± 0.1240 [0.0555]	cross22	0.9982 ± 0.0036 [0.0016]
dec10	0.6355 ± 0.1240 [0.0555]	dec22	0.9982 ± 0.0036 [0.0016]
cross11	0.8291 ± 0.0723 [0.0323]	cross23	1.000 ± 0.000 [0.000]
		dec23	1.000 ± 0.000 [0.000]

Across the five layer-wise tables, decoder-side recovery rises toward one as the patched activation approaches the output, providing a plumbing check that the intervention reaches the prediction. Encoder-side recovery is more distributed and model-dependent. In particular, the base and large checkpoints show substantial recovery across several intermediate encoder layers rather than a single sharp site, consistent with distant-context information being represented through a distributed encoder pathway.

The next two interventions test whether distant encoder positions causally reach the decoder. The cross-attention edge cut leaves encoder computation unchanged and masks decoder access beyond a specified offset. Encoder-position ablation instead replaces distant encoder states, so it combines whether information was encoded with whether the decoder reads it.

Table 38: Cross-attention edge-cut output shift.

checkpoint	0	2	4	8	16	32	64	128	256	511
Chronos-T5 tiny	0.0138 ± 0.0049 [0.0022]	0.0110 ± 0.0022 [0.0010]	0.0104 ± 0.0015 [0.0007]	0.0107 ± 0.0012 [0.0005]	0.00899 ± 0.00131 [0.00059]	0.00724 ± 0.00217 [0.00097]	0.00467 ± 0.00086 [0.00038]	0.00260 ± 0.00037 [0.00016]	0.00133 ± 0.00031 [0.00014]	–
Chronos-T5 mini	0.0123 ± 0.0019 [0.0008]	0.0124 ± 0.0024 [0.0011]	0.0125 ± 0.0025 [0.0011]	0.0114 ± 0.0016 [0.0007]	0.00921 ± 0.00131 [0.00059]	0.00902 ± 0.00152 [0.00068]	0.00591 ± 0.00122 [0.00055]	0.00433 ± 0.00141 [0.00063]	0.00225 ± 0.00061 [0.00027]	–
Chronos-T5 small	0.0153 ± 0.0037 [0.0017]	0.0102 ± 0.0014 [0.0006]	0.00827 ± 0.00156 [0.00070]	0.00730 ± 0.00152 [0.00068]	0.00533 ± 0.00094 [0.00042]	0.00452 ± 0.00055 [0.00025]	0.00361 ± 0.00077 [0.00034]	0.00298 ± 0.00101 [0.00045]	0.00141 ± 0.00051 [0.00023]	–
Chronos-T5 base	0.0102 ± 0.0021 [0.0009]	0.00720 ± 0.00064 [0.00029]	0.00690 ± 0.00055 [0.00024]	0.00648 ± 0.00092 [0.00041]	0.00465 ± 0.00084 [0.00038]	0.00481 ± 0.00023 [0.00010]	0.00371 ± 0.00055 [0.00025]	0.00348 ± 0.00109 [0.00049]	0.00177 ± 0.00083 [0.00037]	–
Chronos-T5 large	0.0121 ± 0.0031 [0.0014]	0.00893 ± 0.00143 [0.00064]	0.00840 ± 0.00068 [0.00030]	0.00668 ± 0.00078 [0.00035]	0.00524 ± 0.00128 [0.00057]	0.00502 ± 0.00125 [0.00056]	0.00398 ± 0.00046 [0.00020]	0.00288 ± 0.00061 [0.00027]	0.00154 ± 0.00029 [0.00013]	–

Table 39: Encoder-position ablation output shift.

checkpoint	0	2	4	8	16	32	64	128	256	511
Chronos-T5 tiny	0.0141 ± 0.0041 [0.0018]	0.0267 ± 0.0047 [0.0021]	0.0354 ± 0.0059 [0.0026]	0.0367 ± 0.0108 [0.0048]	0.0265 ± 0.0053 [0.0024]	0.0160 ± 0.0049 [0.0022]	0.00899 ± 0.00158 [0.00070]	0.00411 ± 0.00080 [0.00036]	0.00155 ± 0.00012 [0.00005]	–
Chronos-T5 mini	0.0165 ± 0.0033 [0.0015]	0.0216 ± 0.0036 [0.0016]	0.0230 ± 0.0049 [0.0022]	0.0230 ± 0.0056 [0.0025]	0.0186 ± 0.0052 [0.0023]	0.0116 ± 0.0036 [0.0016]	0.00755 ± 0.00168 [0.00075]	0.00468 ± 0.00150 [0.00067]	0.00246 ± 0.00059 [0.00026]	–
Chronos-T5 small	0.0196 ± 0.0024 [0.0011]	0.0188 ± 0.0017 [0.0008]	0.0193 ± 0.0015 [0.0007]	0.0168 ± 0.0027 [0.0012]	0.0117 ± 0.0020 [0.0009]	0.00807 ± 0.00094 [0.00042]	0.00519 ± 0.00018 [0.00008]	0.00325 ± 0.00049 [0.00022]	0.00152 ± 0.00023 [0.00010]	–
Chronos-T5 base	0.0150 ± 0.0012 [0.0005]	0.0114 ± 0.0009 [0.0004]	0.0135 ± 0.0013 [0.0006]	0.0139 ± 0.0019 [0.0009]	0.0116 ± 0.0032 [0.0014]	0.00791 ± 0.00178 [0.00079]	0.00440 ± 0.00054 [0.00024]	0.00368 ± 0.00059 [0.00026]	0.00189 ± 0.00053 [0.00024]	–
Chronos-T5 large	0.0143 ± 0.0028 [0.0013]	0.0146 ± 0.0019 [0.0008]	0.0157 ± 0.0023 [0.0010]	0.0148 ± 0.0019 [0.0008]	0.0103 ± 0.0013 [0.0006]	0.00791 ± 0.00073 [0.00033]	0.00501 ± 0.00087 [0.00039]	0.00361 ± 0.00065 [0.00029]	0.00180 ± 0.00035 [0.00016]	–

Figure 7 collects the structural readouts across model sizes. Their common pattern is compatible with the architecture-agnostic finding in Section 5.2: recent positions dominate the temporal allocation, while increasing model size distributes more internal attention to distant context. Taken together, these results show that distant-context information is decodable in the encoder and that intervening on routes carrying it can alter the prediction. They do not identify a unique layer, circuit, or functional weighting mechanism; the mechanism-level conclusions therefore continue to rest on the architecture-agnostic input interventions.

For both interventions, output shifts generally decrease as the cut or ablation moves farther from the prediction point. The edge-cut result is the cleaner causal readout of decoder access because the encoder states themselves are held fixed. Its decay with temporal offset is consistent with the recent-position concentration observed in the input perturbations, while the non-zero distant shifts show that the decoder is not strictly limited to a short hard window.

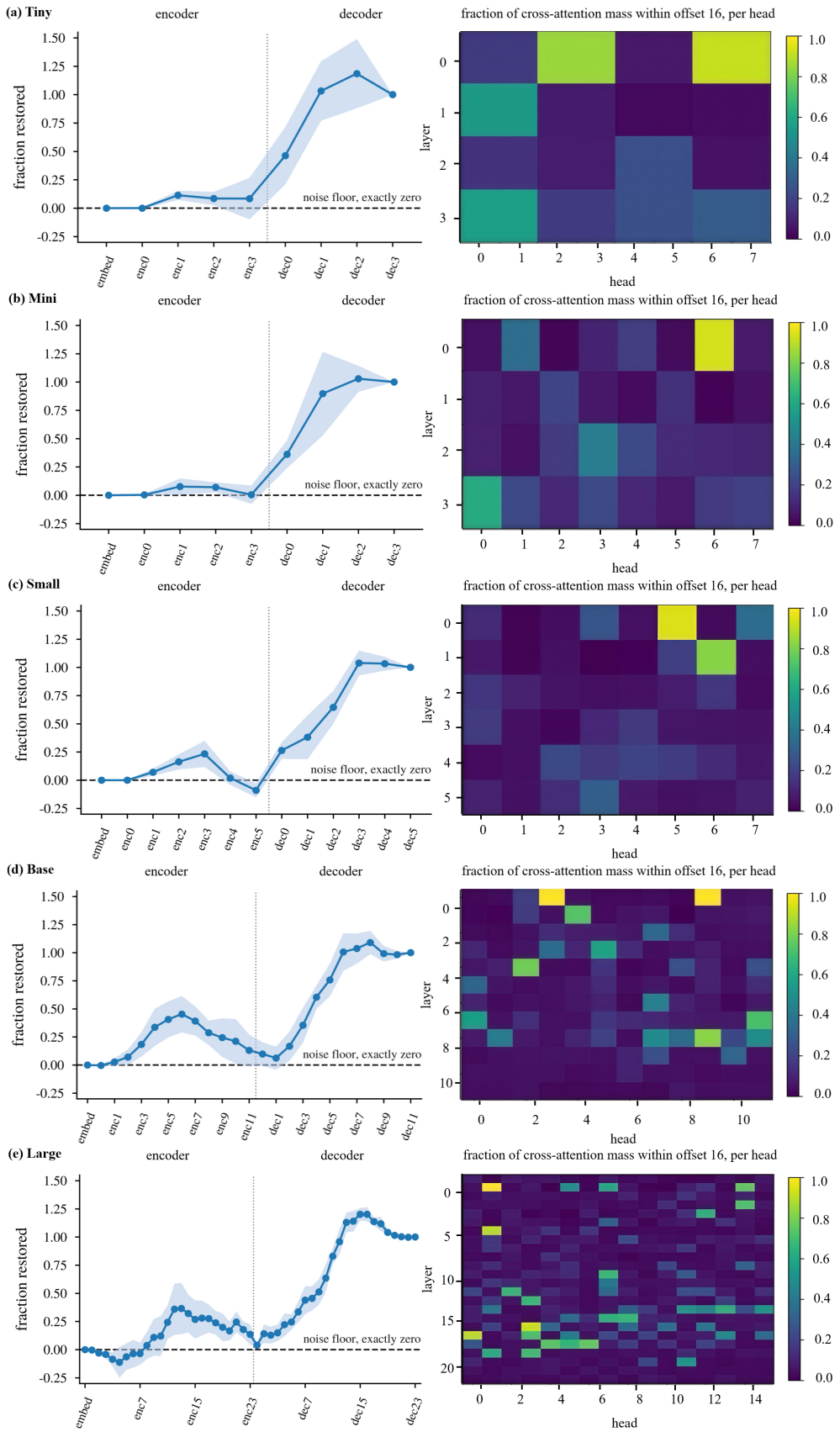

Figure 7: Chronos-T5 structural readouts for (a) tiny, (b) mini, (c) small, (d) base, and (e) large.