WHAT DO ACTIVATION INTERVENTIONS ESTABLISH? CAUSAL CONTROLS FOR AUTOREGRESSIVE TIME-SERIES FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series foundation models (TSFMs) are increasingly studied through interventions on their internal activations; such interventions are also proposed as practical tools, for example to steer a forecast toward a historical crash. When an intervention succeeds, what has it established? Success is usually measured by restoration, the fraction of the gap to a donor's forecast that the edit closes, which shows only that the edited quantity can influence the forecast. It does not show that the effect continues after the edit stops, that the edited site is special, what carries the effect forward, or that the displayed change comes from the edit. We identify these four causal claims and introduce, to our knowledge, the first causal controls for them in autoregressive TSFMs: *releasing the edit and scoring later steps*, *matched edits at other sites*, *replaying the emitted values in an unedited run and simplifying them*, and an *activation-by-scale factorial*; two of them apply to any TSFM. With them, each claim can be tested rather than assumed. On Chronos-T5, each control changes what a successful intervention can be taken to show. On observed traffic and pedestrian series, an edit that removes 82% of the gap while active removes only 7% once it stops. On controlled synthetic series, where a four-step edit does persist (96%), the effect is carried by the emitted forecast values: collapsing them to their mean retains essentially all of the shape improvement, yet with a key–value cache, replay recovers only about 70% of an early-layer edit's effect. In the six public Chronos cases of *time2time*, which reports steering forecasts between market regimes, a case-specific output rescaling produces the displayed direction in all six, while the activation transplant alone moves the median forecast the opposite way in five (three with intervals excluding zero), although it shifts tails and drawdowns toward the donor regime. A successful intervention establishes influence; each stronger causal claim requires the causal control that tests it.

1 INTRODUCTION

Time-series foundation models (TSFMs) such as Chronos are pretrained on diverse series and reused for new forecasting tasks with little or no dataset-specific training (Ansari et al., 2024; Rasul et al., 2024; Das et al., 2024; Woo et al., 2024). A growing body of work asks what these models encode internally. Studies compare learned representations, recover forecast-related concepts, ablate sparse features, and transplant activation statistics between series (Wiliński et al., 2025; Pandey et al., 2025; Mishra, 2026; Sanyal et al., 2025). Many of these studies rest on internal interventions: they replace values the model computes between the input history and the forecast, then observe how the forecast changes. Such interventions are also proposed as practical tools, for example to generate stress-test scenarios by steering a forecast toward a historical crash (time2time; Sanyal et al., 2025). When an intervention succeeds, what has it established?

Activation patching answers part of this question (Meng et al., 2022; Wang et al., 2023; Conmy et al., 2023; Geiger et al., 2025; Makelov et al., 2024; Zhang and Nanda, 2024). Activations from a donor run replace those of a recipient run at a chosen *site* (the output of one model component), and success is usually measured by *restoration*: the fraction of the gap between the recipient's and the donor's outputs that the edit closes. High restoration shows that the edited quantity can causally influence the output. We identify four stronger causal claims (hypotheses about how the effect arises and

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.

travels) that it does not establish: *persistence* (that the effect continues after the edit stops), *location specificity* (that the edited site is special, a concern raised for language models by Hase et al., 2023), *emitted-output sufficiency* (that the values the model outputs during the edit carry the effect), and *numerical attribution* (that the displayed change comes from the edit). Each needs its own causal control: a comparison that tests the specific alternative explanation it faces.

To test these four claims, we study an autoregressive TSFM, Chronos-T5 (Ansari et al., 2024), a widely used open model that scales each series, maps values to a fixed vocabulary of bins, and generates the forecast one value at a time with a T5 encoder–decoder; it is also the model in the public pipeline of *time2time* (Sanyal et al., 2025). Each value emitted during an edited step becomes input to every later step, so a forecast can stay changed after the edit ends, carried by the altered history rather than by any lasting internal state. While an edit is active, its effect is imposed, so only steps after it stops show what the model carries forward; replaying the emitted values in an unedited run separates a lasting internal-state effect from propagation through the history. Decoding and output scaling also transform the forecast before it is shown, so a displayed change need not come from the edit at all. Finally, an edit that works at one site may work equally well at others. We study these questions on controlled synthetic series with known trend and seasonality, observed traffic and pedestrian series (Godahewa et al., 2021), and the NASDAQ-100 windows published with *time2time*.

We therefore propose, for each causal claim, a causal control that tests its main alternative (Table 1). These four causal controls are the core of our approach: *release*, which stops editing and scores only later steps, tests persistence; *matched edits at other sites* test location specificity; *replay and simplification* of the emitted values test emitted-output sufficiency; and an *activation-by-scale factorial*, which separates the edit from output scaling, tests numerical attribution. With them, each claim can be tested rather than assumed.

We make three contributions.

Four causal claims and the first causal controls for them in autoregressive TSFMs (Section 3). We identify four causal claims a successful intervention can be taken to support (*persistence*, *location specificity*, *emitted-output sufficiency*, and *numerical attribution*) and design for each a causal control that tests its main alternative, together with a rule for deciding whether the claim is supported (Table 1). All four controls apply to any autoregressive TSFM, and two of them to any TSFM. The verdicts may differ across models and data; the controls do not.

First evidence that intervention success in an autoregressive TSFM does not imply persistence (Section 4). On observed traffic and pedestrian series, a Chronos-T5 edit that removes 82% of the gap to the donor's forecast while active removes only 7% once it stops. Where effects do persist, as on controlled synthetic series, they can travel entirely through the values already forecast: replaying those values recovers the effect at late decoder sites but only about 70% of it at an early site with a key–value cache. Persistence and its route must be measured, not inferred.

Evidence that the displayed steering direction in a published Chronos pipeline comes from output rescaling, not the activation transplant (Section 5). In the public Chronos pipeline of *time2time* (Sanyal et al., 2025), a study that reports steering forecasts between market regimes, the direction shown in its Chronos figures comes from a case-specific rescaling of the decoded forecast. The activation transplant alone moves the median the opposite way in five of six cases (three with intervals excluding zero), although it shifts tails and drawdowns toward the donor regime.

2 Related Work

Time-series foundation models use several architectures, including decoder-only, encoder–decoder, and masked-encoder models (Rasul et al., 2024; Das et al., 2024; Ansari et al., 2024; Woo et al., 2024). Recent interpretability work asks how these models organize information internally and whether those internal representations can shape forecasts (Wiliński et al., 2025), whether forecast-relevant concepts can be recovered from hidden states (Pandey et al., 2025), whether removing sparse learned features changes model behavior (Mishra, 2026), and whether transplanting activations between series can steer forecasts (Sanyal et al., 2025). Kalnāre et al. (2025) apply activation patching to time-series classification. TimeSAE studies sparse causal explanations for black-box time-series models (Oublal et al., 2026), while Jander et al. (2026) intervene on controlled synthetic data to test whether models preserve known temporal patterns.

More broadly, activation patching, path patching, and causal abstraction use interventions to test where behavior is controlled and which internal computations matter (Vig et al., 2020; Meng et al., 2022; Wang et al., 2023; Goldowsky-Dill et al., 2023; Conmy et al., 2023; Geiger et al., 2025). Our causal controls extend this line of work, which includes hypothesis-driven resampling tests such as causal scrubbing (Chan et al., 2022). Two channels matter especially for autoregressive forecasters. An early edit can persist through the values the model goes on to emit, a route also documented for language models (Maar et al., 2026). Displayed forecasts also pass through numerical decoding, where output scaling can dominate. Three of our four controls target these channels. Prior work shows that conclusions can change with how the baseline input is corrupted or which metric is used, that successful edits can use unusual pathways, and that interventions can create unnatural internal states (Makelov et al., 2024; Zhang and Nanda, 2024; Sun et al., 2025; Grant et al., 2026). Prior TSFM studies check that steering is consistent across random seeds and variants (Wiliński et al., 2025) or ablate learned features (Mishra, 2026); to our knowledge, none tests whether an intervention's effect persists after the edit stops, what carries it, or whether output scaling produces it. We test all three (Sections 4 and 5).

3 Causal Claims and Causal Controls

Following prior work on interpretability illusions in language models (Makelov et al., 2024), we start from the observation that the same intervention success can arise from different mechanisms, and each stronger causal claim asserts one of them. An effect that continues after the edit stops may be carried by a lasting internal state or only by the forecast values already produced; an edit that works at one site may work equally well at others; and a displayed change may come from the edit or from how outputs are rescaled afterwards. For each claim, we identify the competing mechanism and design a causal control: a comparison that varies only the factor separating the two while holding the other conditions fixed. Here, "causal" refers to the model and its forecasting pipeline, not to causal relations in the series being forecast. In causal-inference terms (Pearl, 2009), editing an activation is an intervention, and restoration measures its effect against a counterfactual target, the donor's forecast. Each causal control isolates one factor of that effect (Table 1); replay, for example, intervenes on the emitted values alone, a form of causal mediation analysis (Vig et al., 2020). Figure 1 shows where these mechanisms enter an autoregressive forecast. We use the standard patching terms donor, recipient, clean, corrupt, and restoration (Meng et al., 2022; Zhang and Nanda, 2024) and introduce three procedures of our own: release, replay, and mean-repeat.

3.1 Model Setup and the Activation Write

Our controlled experiments use frozen Chronos-T5-Small (approximately 46M parameters) and Chronos-T5-Base checkpoints (Ansari et al., 2024). Chronos converts numerical time series into discrete tokens and models them with a T5 encoder–decoder architecture (Raffel et al., 2020). Each forecast uses 256 past values to generate the next 32. Our synthetic construction, which we call a *quartet*, contains the four combinations of two drift (trend) values and two seasonal periods, built from the same background components so that drift or seasonality can be changed while everything else is held fixed. The primary comparisons are drift swaps: within each quartet, one series serves as the donor and another as the recipient; the pair shares the seasonal period and background components but differs in drift. A *stream* is one sampled decoding trajectory; streams are paired across conditions by sharing their random draws. A *seed* indexes data generation and sampling randomness (Appendix A.1).

At a given forecast step, a *site* is the output of one model component. Our edit, an *activation write*, replaces the recipient activation at that site with the corresponding activation from a donor run. A continuous write performs this replacement at all 32 forecast steps. We test four terminal interfaces, the last activations before Chronos outputs a distribution over its value bins: the final decoder MLP output; the post-residual state (that output added to the running hidden state); the final-normalization output; and the logits (unnormalized scores over the bins). All four lie after the decoder's last attention layer, so no later position's attention keys or values depend on them. An edit at these sites can therefore affect later steps only through the values it causes the model to emit.

Causal claim	**Control** (alternative tested)	**Supported if**	**Result in this paper**
Persistence: the effect continues after the edit stops	*Release*: stop after m steps, score later steps (effect only while editing)	Released $R \geq 0.50$, lower bound ≥ 0.30 (observed: 0.40, 0.20)	Synthetic: supported (0.96). Observed: not supported (0.07; 0.82 while active)
Location specificity: the edited site is special	*Matched edits* at other candidate sites (works equally well elsewhere)	Edited site ahead by ≥ 0.20; not if all within ± 0.10; else inconclusive	Not supported: all four terminal interfaces within ± 0.10 (only the MLP was in the 30-site screen)
Emitted-output sufficiency: emitted values carry the effect	*Replay* emitted values in an unedited run; *repeat* their mean (lasting internal state)	Replay reproduces release. Simplification: mean-repeat retains $\geq 75\%$	Supported at the terminal site (104% retained); partial at layer 0 with a cache (replay 0.61 vs. release 0.87)
Numerical attribution: the displayed change comes from the edit	*Activation-by-scale factorial* (output rescaling)	The edit alone moves the forecast in the displayed direction	Not supported in *time2time*: rescaling does so in 6/6; the edit alone opposes it in 5/6 (3 with intervals excluding 0)

Table 1: Causal claims and causal controls. Passing a control is necessary but not sufficient for its claim: failing leaves the claim unsupported; passing supports it against the listed alternative for the tested model, data, and sites. Thresholds were prespecified, and no verdict hinges on their exact values: the persistence verdicts hold for any bar from 0.10 to 0.94, retention passes any bar up to 98%, and the ± 0.10 margin could shrink to ± 0.011. R: restoration; m: edited steps.

3.2 Counterfactual Target, Scale, and Restoration

The patched forecast answers a counterfactual question (Pearl, 2009): what the recipient would have forecast if the edited activation were set to its value on the donor input. The donor's own forecast is the *counterfactual target*, the forecast a fully successful edit would reproduce. Let f_{clean}, f_{corrupt}, and f_{patched} denote the donor (target) forecast, the unedited recipient (baseline) forecast, and the edited recipient forecast. Here, "corrupt" is simply the unedited recipient, not an artificially corrupted input. Chronos divides each input series by its mean absolute value, its MeanScale, before tokenizing. For each comparison q we fix one scale, $s_q = \text{mean}\,|x_q|$, where x_q is the 256-value input history of the donor in the controlled synthetic studies and of the recipient in the observed-pair study (with one exception; Appendix B.9), and use it for tokenization, decoding, and error normalization, so numerical scaling does not enter the activation contrast. We measure the distance between forecasts u and v as scale-normalized mean absolute error over the $H = 32$ forecast steps,

$$d(u, v) = \frac{1}{H} \sum_{h=1}^{H} \frac{|u_h - v_h|}{s_q},$$

and define *restoration* as the fraction of the baseline-to-target distance removed by the edit, analogous to the clean–corrupt normalization of logit-difference patching effects (Zhang and Nanda, 2024):

$$R = 1 - \frac{\sum_i d(f_{\text{patched},i}, f_{\text{clean},i})}{\sum_i d(f_{\text{corrupt},i}, f_{\text{clean},i})}. \tag{1}$$

$R = 1$ means complete repair toward the model counterfactual, $R = 0$ no repair, and $R < 0$ movement away from it. Here i indexes eligible quartets (defined below) in the controlled studies or donor–recipient pairs in the observed-data study. Sampled streams are combined within each unit before summing and are not treated as independent experiments. Restoration measures agreement with a model counterfactual, not predictive accuracy.

A quartet is *eligible* in the ratio-based controlled experiments when $d(f_{\text{corrupt}}, f_{\text{clean}}) \geq 0.25$, so restoration is never computed from a baseline already near the target; in the site screen of Section 4.1, every threshold up to 1.0 retains all quartets and leaves the estimate unchanged (Appendix B.1). The mean-repeat experiment (Section 4.2) instead uses error differences and applies no eligibility threshold. Observed-pair eligibility is defined in Section 4.3; Appendix B.9 lists all populations.

3.3 Release, Replay, and Uncertainty

Release stops copying donor activations after the first m forecast steps and evaluates only the remaining forecast steps, positions $m + 1, \ldots, H$, which form the *released suffix*. The values emitted during the first m steps form the *prefix*; restricting Eq. 1 to the released suffix gives $R_{\text{release}}(m)$. *Replay* feeds that prefix into a separate recipient run with no activation editing and then lets the model continue normally. *Exact replay* uses the activation-generated prefix unchanged, while simplified

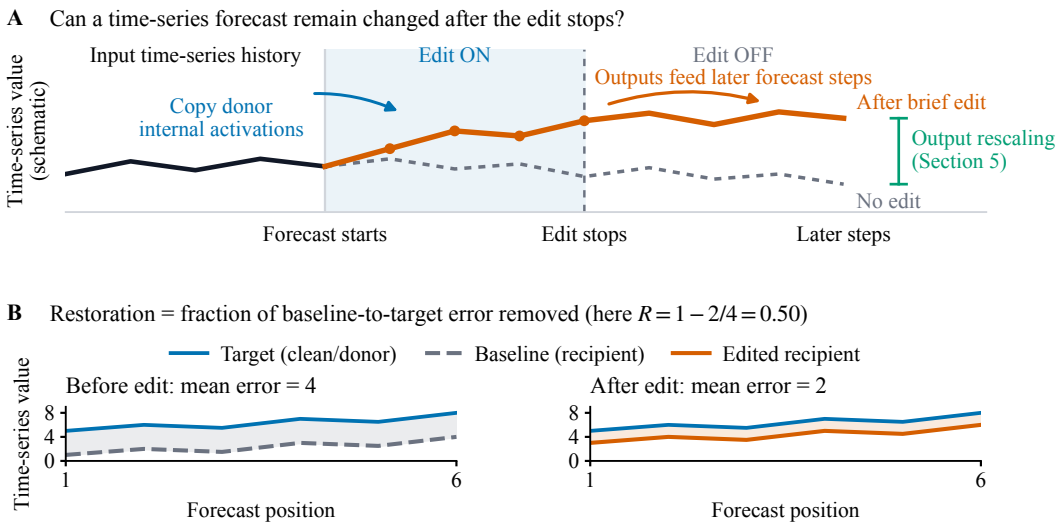


Figure 1: Where an intervention's effect can travel, and how restoration is measured (schematic). **A:** Donor activations are written during the first m forecast steps; after the edit stops, the emitted values stay in the history that later steps condition on, and output rescaling acts after decoding. **B:** Restoration is the fraction of the baseline-to-target error the edit removes, measured against a model-generated target rather than the actual future.

replay modifies it. Unless stated otherwise (Section 4.4), our Chronos execution is cache-free: decoder state is recomputed from the recipient context and the full emitted prefix at every step. Matching the context, prefix, fixed scale, and random draws therefore determines the continuation; see Eq. 2 (Appendix B.3). Exact replay tests whether the emitted prefix suffices to carry the effect under fixed recipient conditions; simplifying the prefix asks which of its properties are needed (Figure 1A).

For uncertainty estimates, we resample quartets in the controlled studies and donor–recipient pairs in the observed-data study, keeping all associated streams and conditions grouped within them. Reported central 90% intervals are the 5th–95th bootstrap percentiles; each endpoint also serves as the corresponding one-sided 95% bound for prespecified tests. Full sampling and eligibility details are in Appendix B.9.

Applicability beyond Chronos. None of the controls depends on Chronos. Release and replay need a forecaster whose outputs re-enter as inputs, one value at a time as in Chronos (Ansari et al., 2024) and Lag-Llama (Rasul et al., 2024) or in blocks as in decoder-only models such as TimesFM (Das et al., 2024) and Toto (Cohen et al., 2024); the cached-decoding variant (Section 4.4) needs a key–value cache; matched edits need only other candidate sites; and the activation-by-scale factorial (Table 1) applies to any pipeline that rescales its outputs, including models that forecast the whole horizon at once. Each control presupposes an intervention that works while active; on Lag-Llama and Toto, our writes either did not work while active or did not persist (Appendix D.1).

4 Dissecting the Chronos Intervention

We first use the synthetic construction, which lets us set the donor, recipient, target, write duration, and decoding randomness separately. Section 4.3 then asks whether the persistence found there holds on observed traffic and pedestrian pairs.

4.1 Establishing an Effective Write and Matched Interface Controls

We screen 30 component outputs as candidate write sites: the attention and MLP outputs of every encoder and decoder layer (residual-stream sites: Section 4.4). Replacing the final decoder MLP output at every forecast step moves the recipient forecast almost exactly onto the donor target ($R = 0.9999$ in each of three seeds; lower bounds above 0.9997). It is the only screened site to pass

the prespecified $R \geq 0.90$ criterion (cf. the illustrative 90% recovery target of Heimersheim and Nanda, 2024; all 24 quartets eligible); the next-best site reaches only 0.074, so any threshold between the two selects the same site (Appendix B.1). On newly generated quartets, it also beats size-matched control edits along other activation directions by at least 0.635 in restoration (Appendix B.1).

Location specificity. Matched writes at the three other terminal interfaces, which the screen did not include, perform equivalently: the 90% intervals of their paired differences from the MLP write lie within the prespecified ± 0.10 margin, supporting equivalence under the confidence-interval criterion (Lakens, 2017; Appendix B.2).

4.2 Persistence Survives Prefix Simplification

Persistence and emitted-output sufficiency. Writing only forecast positions 1–4 removes 96.2% of the recipient-to-target error over positions 5–32, although no donor activations are written at those positions ($R_{\text{release}}(4) = 0.9624$; central 90% interval $[0.9492, 0.9736]$), passing the prespecified bar ($R_{\text{release}} \geq 0.50$, lower bound ≥ 0.30). Persistence also appears in Chronos-T5-Base (0.700 after release; Figure 3B, Appendix B.8). On positions 17–32, which no write of $m \leq 16$ steps edits (Figure 3A), restoration at $m = 4$ is 0.948.

Exact replay, forcing the four emitted values into an otherwise unedited run, reproduces positions 5–32 exactly, and forcing the recipient's original prefix reproduces its original suffix. This equality follows from the determinism described in Section 3.3, so it is an implementation check rather than a result. The informative question is how much of the prefix's structure is needed. *Mean-repeat*, a simplification we introduce, replaces the four emitted values with four copies of their mean, quantized at the same fixed scale, and lets the model continue without further editing. The recipient history remains available, so a constant prefix need not produce a constant continuation.

We test mean-repeat on 30 newly generated quartets, each under four conditions: upward or downward drift swap, crossed with either the original recipient history or an endpoint-matched history, shifted by a constant so its final value equals the donor's. Endpoint matching removes most of the level gap. This gives 120 conditions with 40 paired streams each; uncertainty comes from resampling quartets. Because mean-repeat changes the prefix's last value, from which the continuation proceeds, it can shift the continuation's level, so we measure shape error after subtracting each suffix's mean.

Let E_{baseline}, E_{mean}, and E_{write} denote the corresponding pooled centered-shape errors. We report $\text{Retention} = (E_{\text{baseline}} - E_{\text{mean}})/(E_{\text{baseline}} - E_{\text{write}})$, the ratio of pooled improvements rather than an average of condition-level ratios.

The original write improves centered-shape error by 0.340 (one-sided 95% lower bound 0.241), and mean-repeat retains 104.3% of that improvement (central 90% interval 98.7–107.8%), passing the prespecified 75% criterion (Figure 2C). The interval includes 100%, so mean-repeat is not shown to outperform the original prefix, and the pooled result need not hold in every condition. A smaller study of three simplified prefixes was inconclusive (Appendix B.6).

Prefixes are also target-specific: released restoration is 0.63 toward the target that generated the prefix but -0.81 toward the opposite-drift one (Appendix B.4).

4.3 Strong Restoration Does Not Imply Persistence on Observed Pairs

We next form 240 candidate donor–recipient pairs from distinct observed traffic and pedestrian histories (Godahewa et al., 2021), grouped by whether the two histories differ by a level shift or a demand ramp, and apply the same continuous-versus-release comparison. As in the controlled study, the donor's forecast is the counterfactual target. A pair is eligible when its unedited donor and recipient forecasts differ by at least 0.25, the same threshold as in Section 3.2. Among the 173 eligible pairs, continuous writing across all 32 positions restores 0.824 toward the target (one-sided 95% lower bound 0.768), whereas writing only positions 1–4 yields 0.073 over the released suffix (central 90% pair-bootstrap interval $[0.054, 0.096]$), failing even the lower bar prespecified for observed pairs ($R \geq 0.40$, lower bound ≥ 0.20). Scored on the same window (positions 5–32, post hoc), continuous writing restores 0.831 $[0.734, 0.926]$ against 0.073 after release, a paired difference of 0.758 $[0.667, 0.853]$ (Appendix B.5). A follow-up specified before decoding varies the write duration on the same pairs (Figure 3A): restoration on positions 17–32 rises from 0.01 at $m = 1$ to 0.17 at

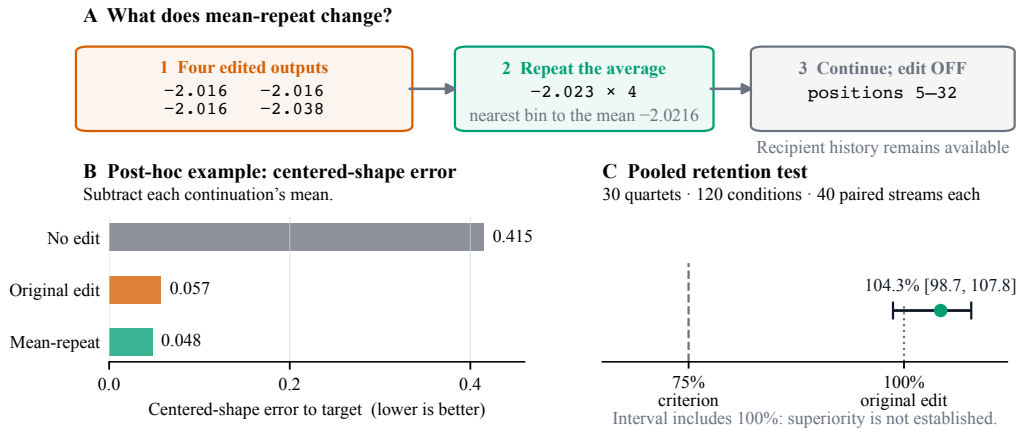


Figure 2: Mean-repeat attribution. **A:** Replace the four edited outputs by their tokenized mean and continue unedited. **B:** Representative centered-shape error. **C:** Mean-repeat retains 104.3% $[98.7\%, 107.8\%]$ of the original edit's pooled improvement, passing the prespecified 75% criterion.

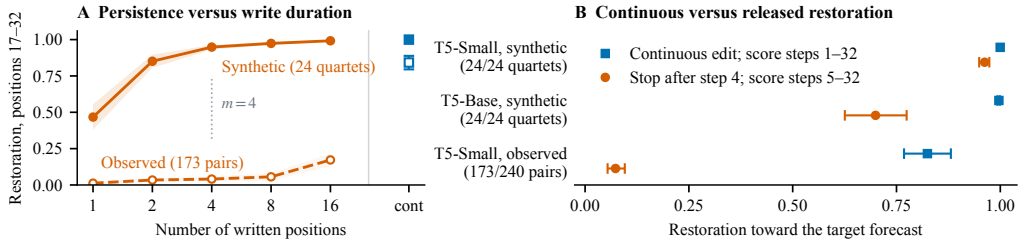


Figure 3: Persistence has to be tested after release. **A:** Released restoration on positions 17–32 versus write duration, for synthetic quartets (24) and observed pairs (173). Blue squares at "cont": continuous writing, scored on the same positions (filled: synthetic; open: observed). **B:** Continuous versus released restoration by setting; bars are central 90% intervals where available. Squares and circles use different scoring windows and are not paired.

$m = 16$ $[0.13, 0.21]$, still far below 0.84 for continuous writing on the same window. Even writing half the horizon does not produce the persistence that four steps produce on synthetic series, so persistence has to be measured after release rather than assumed from intervention success.

4.4 Upstream Sites with a Key–Value Cache

At terminal sites, a released effect can travel only through the emitted values (Section 3.1). At upstream decoder sites with cached decoding, keys and values computed during the edited steps remain available afterwards, giving a second route. On fresh synthetic quartets, with the design fixed before decoding (Appendix B.10), we cross the prefix tokens (original or emitted) with the cache for the edited steps (unedited or edited); at the terminal site the two caches coincide bitwise, as they must. After decoder layer 4 (of layers 0–5), replay recovers nearly all of the released effect (0.944 of 0.962; release minus replay $+0.018$ $[0.003, 0.033]$, small but nonzero). After layer 0 it does not: release restores 0.871 but replay only 0.605. The edited cache adds nothing on its own (-0.012 $[-0.045, 0.021]$) but 0.278 $[0.173, 0.388]$ together with the emitted tokens, so replay alone would under-attribute persistence there by 0.27 $[0.18, 0.35]$ (derived: release minus replay).

5 Separating the Transplant from Output Rescaling

Numerical attribution. The study we reanalyze, *time2time* (hereafter the source; Sanyal et al., 2025), asks, for financial series, whether TSFMs "internalize semantic concepts such as market regimes, or merely fit curves." Its activation transplantation imposes a donor series' per-feature activation mean

Case	Joint direction	s_d/s_c	α	s_r/s_c	Scale-only ÷ joint	Transplant-only ÷ joint
A 17C←08X	↓	0.265	$+0.1$	0.927	1.981	-1.059
B 07C←00X	↓	1.529	-0.2	0.894	1.423	-0.473
C 19C←20X	↓	1.760	-0.1	0.924	1.505	-0.546
D 08X←17C	↑	3.772	$+0.2$	1.554	1.039	-0.025
E 00X←07C	↑	0.654	-1.8	1.623	1.119	-0.073
F 20X←19C	↑	0.568	-0.4	1.173	0.921	$+0.067$

Table 2: Output-scale operation and factorial contrasts for the six public *time2time* Chronos cases (recipient←donor; C = calm, X = crash; numbers are the source's case IDs, not necessarily years). α is the public code's hard-coded weight in $s_r = \alpha s_d + (1-\alpha)s_c$; the terminal median is the median forecast at the final step. *Joint direction:* the direction in which both operations together (as in the public code) move the terminal median. s_d/s_c and s_r/s_c are the donor's and the public code's scale relative to the recipient's. The last two columns give the changes from each operation alone as fractions of the joint change; these are accounting ratios, not causal shares.

and standard deviation, computed across the sequence, on a recipient's hidden states (the source's "style" and "target" series are our donor and recipient). It reports bidirectional steering: crash statistics in calm contexts induce downturn forecasts, and calm statistics in crash contexts suppress them, which it interprets as evidence for steerable semantic representations. Its experiments center on Toto (Cohen et al., 2024), with Chronos-T5 variants replicating key results; we reanalyze the six public Chronos cases, each pairing a calm and a crash window of the NASDAQ-100 index.

We do not test the broader representational claim. We ask a narrower question that the public pipeline makes answerable: can the displayed direction of the Chronos intervention, the direction shown in the source's Chronos figures, be attributed to activation transplantation? The public pipeline performs two operations. First, it transplants the activation mean and standard deviation at encoder layer 1 (zero-based). Second, it decodes the forecast at scale s_r instead of the recipient's own scale s_c, where $s_r = \alpha s_d + (1-\alpha)s_c$ and s_c, s_d are the recipient's and donor's MeanScale values (Section 3.2). The source's method section describes only the first operation. The code hard-codes six case-specific α values and states no rule for choosing them (Appendix C.1). A before-and-after comparison applies both operations together, so their effects are confounded: it cannot say which one sets the direction.

We cross transplantation on/off with rescaling on/off in all six public cases, using the source's 128-value context and 64-step horizon. This gives four forecasts: the *baseline* (neither operation), *transplant-only* (mean and standard deviation transplanted, recipient's own scale kept), *scale-only* (rescaled by s_r, no transplant), and *joint* (both, as in the public code). We reproduce the joint pipeline bitwise in every case and summarize each forecast by its median over 200 paired samples at the final forecast position, which we call the terminal median. Measured from the baseline, the joint change splits exactly into three parts: joint change = scale-only change + transplant-only change + interaction, where the interaction is whatever the two operations do together beyond their separate effects (Appendix C.1).

Transplantation alone moves the terminal median opposite to the joint intervention in five of six cases (Table 2, Figure 4). In case E, a crash recipient with a calm donor, the joint pipeline raises the terminal median from 2089.1 to 3252.3, apparently steering the crash forecast toward calm. Transplantation alone lowers it to 2003.8, while rescaling alone raises it to 3390.7. Across cases, the scale-only change is 0.92–1.98 times the joint change. A post-hoc paired bootstrap over the 200 decoding samples per case gives transplant-only intervals that exclude zero in the three large opposing cases (A–C) but not in the two small ones (D, E); the attribution below does not rest on those two.

Rescaling alone moves the forecast in the displayed direction in every case, and in each of the 200 samples. Because rescaling multiplies every decoded value by s_r/s_c, the scale-only change is exactly $(s_r/s_c - 1)$ times the baseline value, so its sign depends only on whether s_r is above or below s_c. In all six cases, α places s_r on the side of s_c that produces the displayed direction (Table 2). Moving toward the donor's scale ($0 < \alpha < 1$) would do so only in cases A and D; in the other four the code's α is negative, moving the scale away from the donor. Because these case-specific α values set the displayed Chronos direction, the similar size of the Toto and Chronos effects reported by the source is not independent evidence of cross-architecture generality.

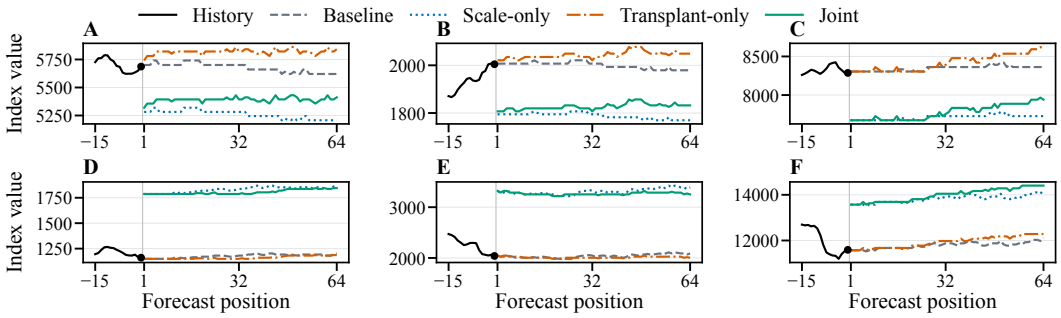


Figure 4: Activation-by-scale factorial for the six public Chronos cases: median baseline, scale-only, transplant-only, and joint forecasts after the 128-step history. Panel letters match the cases in Table 2, which gives the contrasts. Top row: crash donor, calm recipient; bottom row: the reverse.

Rescaling also determines where the displayed effect occurs. At the step from the last observed value to the first forecast value, the scale-only jump is at least as large as the joint jump in every case; in Figure 4 this is the gap between the end of the black history and the start of the joint (green) forecast. For the three calm recipients with crash donors, the joint forecasts end below the historical endpoint, yet their net movement within the horizon is upward (last minus first forecast value: 93.0, 24.9, 266.5). The displayed downturn therefore occurs at the boundary between history and forecast rather than over the forecast horizon.

The transplant is not inert, however; its effect appears in other statistics. In two further post-hoc comparisons (metrics fixed before computation), a crash-donor transplant alone deepens the median within-horizon drawdown (the largest peak-to-trough fall within the 64-step forecast) by 3.0–6.1 percentage points and lowers the terminal 5th percentile; a calm-donor transplant raises the 5th percentile and makes drawdowns 5.0–10.2 points shallower (Appendix C.1). Because rescaling leaves percentage drawdowns unchanged, the drawdown shifts are the joint pipeline's entire effect on that metric. A follow-up transplanting the mean and standard deviation separately locates these effects in the means. Transplanting only the donor's per-feature means reproduces the donor-direction drawdown shift in all six cases and the 5th-percentile shift in five; transplanting only its standard deviations does so in none. The two interact in some cases, but these effects travel mainly through the means rather than through imported activation variance. Separating the two operations therefore changes the interpretation, not just the magnitude: output rescaling sets the displayed median-level direction, while the transplant shifts the lower tail and within-horizon drawdowns toward the donor. These controls cannot decide whether the transplanted means encode regime-specific information or other donor properties.

6 Discussion and Conclusion

We set out to ask what a successful intervention in an autoregressive TSFM establishes. Persistence had to be measured: edits that moved the forecast while active left little trace on observed series once they stopped. Its route had to be tested: replaying the emitted values accounted for persistence at late decoder sites but not at an early one. Numerical attribution required separating the edit from output scaling, which set the displayed direction in the public *time2time* Chronos pipeline. None of this rules out meaningful internal representations; it shows that intervention success alone cannot establish them. We recommend reporting these causal controls alongside restoration, so that each claim is tested rather than assumed. The observations are specific to these pipelines, but the controls apply to any autoregressive TSFM: a successful intervention establishes influence, and each stronger causal claim requires the causal control that tests it.

Limitations. These controls do not identify the unedited model's computation. Every effect we report comes from Chronos-T5 (Small, and Base for persistence): on Lag-Llama and Toto, our writes either did not work while active or did not persist (Section 3). Observed-pair results are scoped to the tested populations (Appendix D), and the *time2time* reanalysis does not cover its Toto experiments.

Reproducibility Statement. Confirmatory thresholds and analysis plans were fixed experiment by experiment, rather than for the full program in advance. These specifications were internally versioned rather than externally preregistered. Failed attempts, known outcome exposure, and later corrections are disclosed in Appendix D; Appendix A specifies inference and eligibility. From the evidence-package root, `bash scripts/regenerate_headline_artifacts.sh` rebuilds the figures, tables, the five post-hoc analyses (the observed-pair dependence sensitivity, the paired same-window release comparison, the six-case transplant-sign bootstrap, the six-case quantile/drawdown contrasts, and the six-case moment split), and the analyses of two follow-up experiments specified before decoding (the observed-pair write-duration curve and the cached upstream-site comparison) without model inference.

AI Use Statement. Generative AI tools assisted with code development, debugging, LaTeX preparation, manuscript editing, and aspects of experimental planning and analysis. The authors reviewed the resulting code, analyses, and text, verified reported results against the experimental records, and retain responsibility for the manuscript's accuracy, interpretation, and conclusions.

REFERENCES

A. F. Ansari, L. Stella, C. Turkmen, X. Zhang, P. Mercado, H. Shen, O. Shchur, S. S. Rangapuram, S. Pineda Arango, S. Kapoor, J. Zschiegner, D. C. Maddix, H. Wang, M. W. Mahoney, K. Torkkola, A. G. Wilson, M. Bohlke-Schneider, and Y. Wang. Chronos: Learning the Language of Time Series. arXiv:2403.07815, 2024.

L. Chan, A. Garriga-Alonso, N. Goldowsky-Dill, R. Greenblatt, J. Nitishinskaya, A. Radhakrishnan, B. Shlegeris, and N. Thomas. Causal Scrubbing: A Method for Rigorously Testing Interpretability Hypotheses. AI Alignment Forum, 2022.

B. Cohen, E. Khwaja, K. Wang, C. Masson, E. Ramé, Y. Doubli, and O. Abou-Amal. Toto: Time Series Optimized Transformer for Observability. arXiv:2407.07874, 2024.

A. Conmy, A. N. Mavor-Parker, A. Lynch, S. Heimersheim, and A. Garriga-Alonso. Towards Automated Circuit Discovery for Mechanistic Interpretability. NeurIPS 2023; arXiv:2304.14997.

A. Das, W. Kong, R. Sen, and Y. Zhou. A Decoder-Only Foundation Model for Time-Series Forecasting. arXiv:2310.10688, 2024.

Datadog. Toto: Timeseries-Optimized Transformer for Observability, v0.2.0, 2026.

A. Geiger, D. Ibeling, A. Zur, M. Chaudhary, S. Chauhan, J. Huang, A. Arora, Z. Wu, N. Goodman, C. Potts, and T. Icard. Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability. Journal of Machine Learning Research, 26(83):1–64, 2025.

R. Godahewa, C. Bergmeir, G. I. Webb, R. J. Hyndman, and P. Montero-Manso. Monash Time Series Forecasting Archive. arXiv:2105.06643, 2021.

N. Goldowsky-Dill, C. MacLeod, L. Sato, and A. Arora. Localizing Model Behavior with Path Patching. arXiv:2304.05969, 2023.

S. Grant, S. J. Han, A. Tartaglini, and C. Potts. Addressing Divergent Representations from Causal Interventions on Neural Networks. ICLR 2026.

P. Hase, M. Bansal, B. Kim, and A. Ghandeharioun. Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models. NeurIPS 2023; arXiv:2301.04213.

S. Heimersheim and N. Nanda. How to use and interpret activation patching. arXiv:2404.15255, 2024.

M. Jander, W. van Heeswijk, and M. Mes. Causal Analysis for Time Series Foundation Models. arXiv:2608.24303, 2026.

M. Kalnāre, S. Kitharidis, T. Bäck, and N. van Stein. Mechanistic Interpretability for Transformer-based Time Series Classification. arXiv:2511.21514, 2025.

D. Lakens. Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. Social Psychological and Personality Science, 8(4):355–362, 2017.

J. Maar, D. Paperno, C. McDougall, and N. Nanda. What's the Plan? Metrics for Implicit Planning in LLMs and Their Application to Rhyme Generation and Question Answering. ICLR 2026.

A. Makelov, G. Lange, A. Geiger, and N. Nanda. Is This the Subspace You Are Looking for? An Interpretability Illusion for Subspace Activation Patching. ICLR 2024.

K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and Editing Factual Associations in GPT. NeurIPS 2022; arXiv:2202.05262.

A. Mishra. Dissecting Chronos: Sparse Autoencoders Reveal Causal Feature Hierarchies in Time Series Foundation Models. arXiv:2603.10071, 2026.

K. Oublal, Q. Bouniot, Q. Gan, S. Clémençon, and Z. Akata. TimeSAE: Causal Sparse Decoding for Faithful Explanations of Black-Box Time Series Models. International Conference on Machine Learning, 2026; arXiv:2601.09776.

A. Pandey, A. Neog, and G. Jajoo. On the Internal Semantics of Time-Series Foundation Models. BERT2S Workshop at NeurIPS, 2025; arXiv:2511.15324.

J. Pearl. Causality: Models, Reasoning, and Inference. Cambridge University Press, second edition, 2009.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. JMLR, 21(140):1–67, 2020.

K. Rasul, A. Ashok, A. R. Williams, H. Ghonia, R. Bhagwatkar, A. Khorasani, M. J. Darvishi Bayazi, G. Adamopoulos, R. Riachi, N. Hassen, M. Biloš, S. Garg, A. Schneider, N. Chapados, A. Drouin, V. Zantedeschi, Y. Nevmyvaka, and I. Rish. Lag-Llama: Towards Foundation Models for Probabilistic Time Series Forecasting. arXiv:2310.08278, 2024.

M. Rothmann, N. Li, G. Chen, G. Y. H. Chi, R. Temple, and H.-H. Tsou. Design and Analysis of Non-Inferiority Mortality Trials in Oncology. Statistics in Medicine, 22(2):239–264, 2003.

D. Sanyal, A. Nagpal, D. Kumar, M. Mandal, and S. Deshpande. time2time: Causal Intervention in Hidden States to Simulate Rare Events in Time Series Foundation Models. NeurIPS 2025 BERT2S Workshop; arXiv:2509.05801, 2025.

J. Sun, J. Huang, S. Baskaran, K. D'Oosterlinck, C. Potts, M. Sklar, and A. Geiger. HyperDAS: Towards Automating Mechanistic Interpretability with Hypernetworks. ICLR 2025.

J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, Y. Singer, and S. Shieber. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. NeurIPS 2020.

K. Wang, A. Variengien, A. Conmy, B. Shlegeris, and J. Steinhardt. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small. ICLR 2023; arXiv:2211.00593.

M. Wiliński, M. Goswami, W. Potosnak, N. Żukowska, and A. Dubrawski. Exploring Representations and Interventions in Time Series Foundation Models. Proceedings of the 42nd International Conference on Machine Learning, PMLR 267:66861–66886, 2025.

G. Woo, C. Liu, A. Kumar, C. Xiong, S. Savarese, and D. Sahoo. Unified Training of Universal Time Series Forecasting Transformers. arXiv:2402.02592, 2024.

F. Zhang and N. Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. ICLR 2024; arXiv:2309.16042.

A Experimental Units and Inference

Appendix B gives the Chronos construction and controls; Appendix C gives the public-transplant reproduction and scale decomposition. Supplementary scope statements and material qualifications are in Appendix D; the artifact index is in the evidence package.

A.1 Experimental Units and Shared Vocabulary

A *quartet* is one synthetic construction containing the four combinations of two drift values and two seasonal periods. Its series share the sampled background components. A drift-swap comparison changes drift while holding the seasonal construction fixed; the paired clean computation supplies the donor and counterfactual target. Quartet members and their repeated forecasts are not independent replications.

A *stream* is one sampled decoding trajectory. Paired common-random-number (CRN) streams use matched random-number transitions across intervention arms. A *panel* is the collection of constructions and conditions evaluated by one experiment; a fresh panel uses new constructions, whereas reuse is identified explicitly. A *seed* indexes data construction and, where applicable, sampling randomness, not an independently trained checkpoint. Forecast positions, doses, and streams remain nested within their construction. An *arm* is one intervention variant within an experiment (e.g., write, replay, mean-repeat); an edit's *dose* is its size relative to the native activation.

A.2 Inference, Eligibility, and Scientific Status

Unless stated otherwise, bootstraps use 2,000 resamples. The notation 5/95 denotes the fifth and 95th percentiles: together they form a central 90% interval, or individually one-sided 95% lower/upper bounds. The notation 2.5/97.5 denotes a two-sided 95% percentile interval. Controlled Chronos experiments resample constructions within seed, retaining their conditions, arms, and streams together. Paired arms share resampling draws; ratio estimators are recomputed from aggregated distances rather than averaged from per-stream ratios. Observed-pair experiments resample donor–recipient pairs; Table 7 records the experiment-specific populations and qualifications. LB denotes a lower bound.

The ratio-based controlled restoration experiments require damage $d(f_{\text{corrupt}}, f_{\text{clean}}) \geq 0.25$. The fresh mean-repeat retention study instead includes every implementation-valid condition and uses denominator-free error differences. Damage eligibility is separate from the MeanScale normalization and from the observed-pair eligibility of Appendix B.5. Eligibility counts describe the retained population.

Some follow-up tests are interpreted only when a prespecified starting effect is large enough. We call that starting requirement an *anchor*. If it fails, the follow-up is reported descriptively and does not support the planned confirmatory claim. A passed component of an otherwise incomplete experiment supports only that component; failure of one criterion is not evidence that every related hypothesis is false. The evidence package's artifact index records the exact decision and provenance status.

B Chronos Intervention: Construction and Controls

These details support Section 4. The main release test uses the frozen final decoder MLP of Chronos-T5-Small, 256 context values, a 32-position horizon, three construction seeds, eight quartets per seed, and 20 paired sampled streams per quartet. Writing the first four positions is the primary release condition; evaluation uses only the remaining 28 positions. The prespecified release criterion requires a restoration point of at least 0.50 and a one-sided 95% lower bound of at least 0.30. Restoration of at least 0.50 means the released forecast has closed at least half of the baseline-to-target gap, so in pooled distance it is no farther from the donor's forecast than from the recipient's unedited forecast; the one-sided lower-bound requirement additionally supports restoration of at least 0.30. The matched terminal-interface comparison reuses this panel and is not an independent replication.

Representative construction in Figure 5. Both panels use the eligible construction whose denominator-free genuine released-suffix improvement is closest to the median over all 24 eli-

gible constructions, breaking ties by lowest record index. This descriptive selection rule was fixed after outcomes were observed. Each curve is the pointwise median over 20 paired streams; panel B first centers every sampled path by its own positions 5–32 mean.

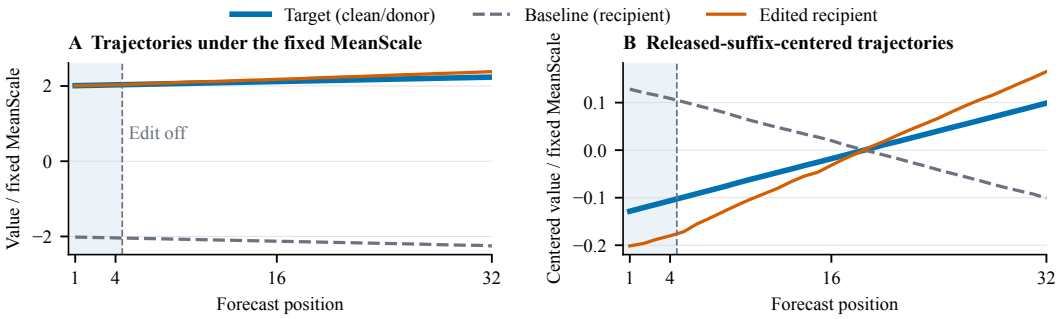


Figure 5: Post-hoc illustrative views of the same four-position write and release. **A:** Trajectories divided by the fixed MeanScale. **B:** Each sampled trajectory is centered by its own released-suffix mean before pointwise medians are taken across 20 streams. Shading marks the four written positions; the remaining forecast steps are unedited. This selected construction is explanatory, not additional evidence.

B.1 Component Screen and Denominator Sensitivity

The screen asks which coarse component output can restore the controlled forecast under a common donor write. The prespecified screen criterion is $R \geq 0.90$. Among the 30 screened outputs, the final MLP reaches 0.9999 in every seed (one-sided lower bounds above 0.9997); the next-best site reaches 0.0736, and the best other decoder MLP reaches 0.0331. Tests on newly generated quartets at the frozen site exceed 32 matched random-subspace, bottom-rank, ridge-probe, and attention controls by 0.635–1.035. All 24/24 quartets in each of these sampled experiments are eligible. The sensitivity analysis asks whether the near-ceiling score could be an artifact of its denominator reduction.

Each of three construction seeds contributes eight quartets, with 20 paired clean and patched streams per quartet. Their patched–clean distances are reduced by the within-quartet median, while corrupt–clean damage uses the matching stream-0 clean and corrupt forecasts. Within each seed, the eight quartet distances are summed before recomputing Eq. 1. The 2,000-resample quartet bootstrap recomputes the ratio and reports the fifth percentile as a one-sided 95% lower bound. This coupling and reduction are specific to the original screen.

The archived run retains all patched–clean distances but only that corrupt–clean distance, so it cannot support an exact symmetric-stream re-estimate. It does provide a conservative bound. Across the 24 retained quartets, the largest median patched–clean distance is 0.001146. Thus any quartet that remains admissible under an alternative denominator reduction has restoration of at least $1 - 0.001146/0.25 = 0.9954$. Replacing all eight denominators in each seed by the minimum admissible value gives aggregate lower bounds of 0.9984, 0.9988, and 0.9988. Under the archived stream-0 reduction, the smallest corrupt–clean distance is 1.058, so every threshold up to 1.0 retains all 24 quartets and leaves the reported estimate unchanged. These checks establish that denominator magnitude cannot explain the near-ceiling result; because the other corrupt streams were not retained, they do not establish how many quartets a symmetric-stream admissibility rule would retain.

B.2 Matched Terminal Interfaces

The plan fixed two criteria for downstream edits matched for donor information, edit magnitude, exposed positions, temporal rank, and decoding randomness: an MLP restoration advantage of at least 0.20 (not met) and equivalence within ± 0.10 (met).

Let Y_d^s and Y_r^s collect donor and recipient rows at terminal interface s. The matched edit is $\widetilde{E}_s = \alpha_s \, \mathrm{SVD}_{k_m}(Y_d^s - Y_r^s)$ from the $m \times 512$ recipient and donor current rows, with $k_m = \min(4, m)$ and α_s chosen so $\|\widetilde{E}_s\|_F / \|Y_r^s\|_F = \rho_{\mathrm{MLP}}$, the corresponding truncated-edit ratio at the MLP. Here

SVD_{k_m} retains the leading k_m singular directions, $\|\cdot\|_F$ is the Frobenius norm, and ρ_{MLP} is the MLP edit-to-activation norm ratio being matched. Each exposed row is set to $Y_r^s(t) + \widetilde{E}_s(t)$, matching the exposure window, donor, temporal rank, relative dose, and decoding streams across interfaces. For logits, the 512-dimensional final-normalization edit is mapped through the frozen $d_{\mathrm{model}}^{-1/2}$ scaling, where $d_{\mathrm{model}} = 512$ is the hidden width, and the language-model head; the 4096-dimensional interface gains no degrees of freedom. The comparison tests operational equivalence within its prespecified band, not literal representational identity. The four interfaces lie on one chain after the last attention layer, so the comparison asks whether the selected interface is privileged among equally informed alternatives, not whether upstream sites are special.

At $m = 4$, suffix restoration is 0.9624 at the MLP, 0.9628 after the residual connection, and 0.9657 at final normalization and logits. Paired MLP-minus-interface differences are -0.0005 ($[-0.0063, 0.0047]$) after the residual connection and -0.0034 ($[-0.0107, 0.0024]$) at final normalization and logits. All central 90% intervals lie inside the prespecified $[-0.10, 0.10]$ equivalence band. These comparisons reuse the release panel; they are not independent replications.

B.3 Prefix Replay and Scrambled Controls

Continuing state. The continuing state consists of the recipient context C, emitted prefix P, encoder state, decoder cache, randomness U, and scale. Let $Y_{m+1:H}$ be the released forecast suffix and let G_θ be the fixed model-and-decoder computation with parameters θ. In our cache-free execution, encoder state is reconstructed from C and the numerical scale is fixed within a comparison, so the suffix depends only on the recipient context, forced prefix, and later random draws:

$$Y_{m+1:H} = G_\theta(C, P_{1:m}, U_{m+1:H}). \tag{2}$$

Matching these arguments predicts replay equality. Replay checks this execution contract; the write-and-release test measures how much the earlier edit changes the continuation.

The release and prefix-control experiments use fresh drift-swap quartets (Figure 6). The clean (donor) and corrupt (recipient) series share the same seasonal construction but differ in the specified drift component; the donor is the clean computation whose forecast defines the restoration target. Forcing a prefix samples each exposed step and then overwrites the emitted token, preserving the random-number-state transition; after position m, forcing stops and its schedule is discarded. The emitted tokens remain in the decoder's autoregressive history, and decoding continues without a hook or donor schedule.

Each randomized control keeps the absolute sizes of the genuine prefix's token displacements but permutes those sizes and assigns predetermined positive or negative directions, adjusting any move that would leave the valid token range. A control qualifies when its mean displacement is within 5% of the genuine prefix and, at $m = 4$, it shares at most one token with that prefix. Thus it matches edit magnitude, but not how probable the prefix is under the decoder. Eight candidates are attempted per control; the prespecified panel gate requires at least 16/20 qualifying streams in at least 7/8 quartets per seed. At $m = 4$, 477/480 streams and 24/24 quartets qualify.

The positive scrambled-control contrast is the *target-prefix advantage*, not restoration by the scrambled controls. Its value is 0.9437 with a 90% interval $[0.8153, 1.0659]$, whereas scrambled-prefix restoration is -0.0016. This comparison supports an advantage over the tested dose-matched controls. It does not settle donor identity or order specificity, and the controls are not likelihood-matched.

Evaluation qualification. Outcomes for one of the 24 evaluation quartets were visible during development. The specification and implementation were unchanged afterward, but complete outcome blinding is not claimed.

B.4 Directional Assignment with an Identical Prefix Multiset

This directional-assignment test is a new sequential experiment, not a revision of the earlier prefix-test conclusions. Each construction contains one corrupt recipient and four targets with drift $\{-1.5, -0.75, 0.75, 1.5\}$. Four genuine prefixes are produced by four-position activation writes. Correct assignment compares each released continuation with its generating target; the fixed opposite assignment maps $-1.5 \leftrightarrow 1.5$ and $-0.75 \leftrightarrow 0.75$. Because the two assignments permute the same

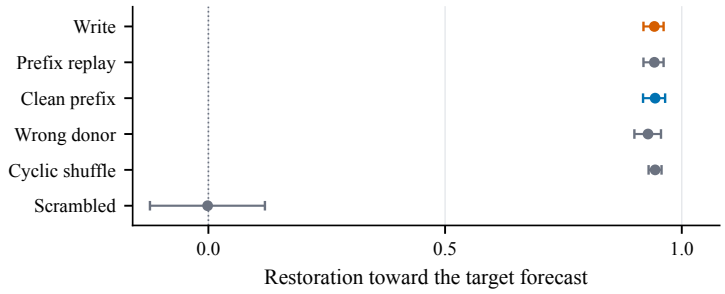


Figure 6: Earlier four-position prefix controls. **Write:** the original activation write. **Prefix replay:** its four emitted values forced into an unedited recipient run. **Clean prefix:** the clean target's own first four forecast values. **Wrong donor:** a clean-forecast prefix from another quartet in the same seed and stream, selected to match displacement magnitude. **Cyclic shuffle:** the genuine prefix cyclically reordered. **Scrambled:** a randomized prefix matched to the genuine prefix's displacement magnitude, but not its decoder probability. Restoration 1 means reaching the target forecast; 0 means no improvement over the recipient baseline. Wrong-donor and cyclic-shuffle prefixes also restore strongly, so this experiment does not establish donor identity or order specificity. Intervals are 90% quartet-cluster intervals.

four prefixes within every stream, the multiset of prefix negative log-likelihoods (NLLs), doses, token histograms, and all other prefix-only statistics is identical.

The outer unit is the shared-components construction. Within each unit and target, raw suffix distances are summed over 20 streams and positions 5–32 before restoration is recomputed; the four target restorations are then equally weighted. The paired contrast resamples 12 units within each of three construction seeds with shared multiplicities across assignments (5,000 draws). All 12 units qualify in every seed. Correct restoration is 0.6261 (one-sided 95% LB 0.6094), opposite-assignment restoration is -0.8088, and the advantage is 1.4348 (LB 1.4122; per-seed points $1.436/1.440/1.428$).

This assignment test concerns the original activation-generated prefixes, not mean-repeat prefixes, and does not establish unique donor identity or order specificity. The preceding fresh multi-target experiment ended at its prespecified starting-effect gate ($R_{\text{correct}} = 0.6923 < 0.80$); that failure remains unchanged and is reported in the supplementary study record in the evidence package. This follow-up uses separately prespecified equal-target weighting and criteria. Saved target-distance blocks support reanalysis; raw target trajectories were not retained.

B.5 Observed-Pair Write and Release

This experiment uses naturally occurring traffic and pedestrian series rather than synthetic drift swaps (Godahewa et al., 2021). Four strata cross traffic-hourly or pedestrian-count data with level-shift or demand-ramp regimes. The frozen panel contains 240 donor–recipient pairs from three pairing seeds; all 480 window roles are globally distinct and exclude every window used by the earlier observed-pair studies. Eligibility is fixed before intervention endpoints as mean donor–recipient forecast distance at least 0.25 over all 20 paired streams and 32 positions. In total, 173/240 pairs qualify, with at least 11 in each of the twelve stratum–seed cells. The pair is the resampling unit; streams and positions remain nested.

The all-position donor write passes its full-horizon criterion (prespecified: point ≥ 0.60, one-sided 95% LB ≥ 0.50): restoration is 0.8245 (one-sided 95% LB 0.7682), with per-seed points $0.8620/0.7755/0.8522$. Four-position release is 0.0733 (LB 0.0542, upper bound 0.0957). Under the prespecified pair bootstrap, that upper bound is below both the 0.20 lower-bound requirement and the prespecified 0.40 point-restoration target. This bar is lower than the synthetic one (0.50, lower bound 0.30), which makes a conclusion of insufficient persistence harder to reach; release fails even this lower bar. The result does not imply the absence of any post-release effect.

Distinct windows can still share a source series: the panel contains 480 distinct window roles but only 167 source series. A post-hoc source-incidence graph joins each pair's donor and recipient series; its 240 scheduled edges form 33 connected components, 28 containing eligible pairs. A 10,000-

draw component-weight bootstrap gives a released-suffix 5/95 interval of $[0.0601, 0.1033]$, and the maximum leave-one-component point change is 0.0255. This sensitivity moves every pair sharing either source together, including bridges through ineligible scheduled pairs; it corroborates weak persistence but does not replace the prespecified pair analysis. Exact replay, clamping, generator-state equality, and zero post-release donor access hold bitwise in the cache-free graph.

Write-duration follow-up. Specified and hashed before decoding (specification hash and timestamp recorded in the evidence package). The $m = 4$ and $m = 8$ release arms reproduce the previously saved arms bitwise; the $m = 8$ suffix value had been observed before this specification and is not newly confirmatory. The $m = 32$ arm equals the continuous write by implementation, and replay of the $m = 16$ prefix is exact. Restoration on positions 17–32 is 0.012, 0.034, 0.041, 0.056, and 0.172 for $m = 1, 2, 4, 8, 16$, against 0.842 for continuous writing on the same window (Table 3). The prespecified contrast $m = 16$ minus $m = 4$ is 0.132 $[0.089, 0.175]$; the $m = 16$ upper bound 0.214 lies below this follow-up's prespecified 0.50 bar (and below the 0.40 observed-pair release target); and per-stratum values at $m = 16$ range from 0.146 to 0.198.

Table 3: Write-duration follow-up on the 173 eligible observed pairs (specification fixed and hashed before decoding): released-suffix and fixed-window restoration versus write duration, point [5th, 95th] from a seed-stratified pair bootstrap (2,000 resamples, recorded seeds). Intervals for $m = 4$ differ slightly from Appendix B.5 and Table 9 because each analysis uses its own bootstrap.

Written positions	$R_{\text{suffix}}(m{+}1{\text{–}}32)$	$R_{17\text{–}32}$
1	0.029 [0.011, 0.048]	0.012 [−0.000, 0.025]
2	0.065 [0.044, 0.086]	0.034 [0.020, 0.049]
4	0.073 [0.052, 0.093]	0.041 [0.026, 0.057]
8	0.084 [0.057, 0.114]	0.056 [0.029, 0.082]
16	0.172 [0.132, 0.213]	0.172 [0.131, 0.214]
32 (continuous)	n/a	0.842 [0.795, 0.888]

Paired same-window comparison (post-hoc). The decoded artifact retains, for every eligible pair, the 20 paired 32-position forecasts of the continuous write, the four-position release, the unedited baseline, and the donor reference under the pair's fixed scale. Scoring the continuous write on the released window of the same 173 eligible pairs gives restoration 0.8313 (per-seed points $0.8672/0.7936/0.8453$; central 90% pair-bootstrap interval $[0.7336, 0.9261]$), against 0.0733 for the four-position release on the identical window, units, and streams; the paired difference is 0.7579 $[0.6671, 0.8532]$. This derived comparison reuses the prespecified aggregation and the prespecified pair-bootstrap design (2,000 seed-stratified resamples carrying all streams with each resampled pair, shared draws across arms, recorded seed); it does not amend the prespecified full-horizon and released-suffix estimands. As a pipeline check, the script first reproduces both prespecified endpoints from the saved decoding artifact. In the original analysis, the continuous-write anchor was computed on positions 5–32 rather than the prespecified full horizon; a documented correction moved it to the full horizon ($0.8313 \to 0.8245$; the released-suffix estimand and the reported conclusion are unchanged). The suffix score recomputed here equals that earlier value. The extraction script and its hash are recorded in the evidence package.

B.6 Endpoint-Matched Prefix Attribution

Construction and interventions. Endpoint matching adds one constant to the corrupt drift-swap context and sets its final historical value exactly to the clean endpoint. The eight intervention variants are: clean; corrupt; the genuine four-position write-and-release; replay of its exact four emitted values; four repeats of the genuine prefix's final value; four repeats of its mean; a four-point linear prefix ending at the matched endpoint; and a control that forces the corrupt recipient's original four-value prefix. Reduced prefixes (the simplified prefixes of Section 4.2) are quantized under the same positive recipient scale; each forced variant samples before overwriting and installs no activation edit.

An initial 3-seed by 8-quartet evaluation, with 20 paired streams, retained only 2/8, 3/8, and 4/8 damage-eligible quartets by seed, below the frozen 80% floor. It cannot answer its prespecified attribution question. The paired and fresh-panel studies instead use denominator-free errors. They

report the jump from the last observed value to forecast position 1, the jump from written position 4 to released position 5, and adjacent changes entirely within the forecast suffix (positions 6–32) separately.

Paired follow-up. The paired follow-up reuses all 24 constructions scheduled in the initial experiment, comparing their unshifted and endpoint-matched variants at a fixed clean-series scale. Its new analysis was fixed before this evaluation; reuse means it is not an independent-panel replication. It decodes one shared clean variant and seven intervention variants per construction version, for 15 variants per stream. Denominator-free errors over positions 5–32 use mean absolute error (MAE) for value, centered MAE after subtracting each suffix's own mean, and adjacent-increment MAE over positions 6–32. For variant a, $G_a = d(\mathrm{corrupt}, \mathrm{clean}) - d(a, \mathrm{clean})$; for reduction r, $\Delta_r = d(r, \mathrm{clean}) - d(\mathrm{genuine}, \mathrm{clean})$. Ten thousand seed-stratified quartet-cluster resamples retain both construction versions, all variants, and 20 streams within each unit.

Endpoint matching reduces corrupt value error by 3.562 scale-normalized units (Table 4), with its entire 90% interval away from zero. Genuine matched improvement is positive in every seed ($0.685/0.306/0.233$) and is 0.408 overall ($[0.150, 0.759]$). The three reduced-prefix value disadvantages are 0.016, 0.011, and -0.021 for final-repeat, mean-repeat, and endpoint-linear prefixes. Every increment interval lies inside ± 0.025, as does mean-repeat centered shape, but every reduction crosses a value or centered-shape margin. No interval excludes zero. The joint equivalence criterion therefore fails, while neither a richer-prefix advantage nor a difference between genuine and reduced prefixes is established.

Table 4: Baseline and activation-write errors before and after endpoint matching, over all 24 constructions and 20 nested streams. An arm is one intervention variant. d is scale-normalized absolute error from the paired clean forecast, and $G = d(\mathrm{corrupt}) - d(\mathrm{arm})$; positive G is improvement. Value compares forecast levels; centered compares shape after removing each suffix mean; increment compares adjacent changes within positions 6–32. These are descriptive components; reduced-prefix equivalence is tested in Table 5.

Construction	Variant	d_{value}	G_{value}	G_{centered}	$G_{\mathrm{increment}}$
Unshifted	Corrupt	4.155	0.000	0.000	0.000
Unshifted	Genuine write	0.164	3.991	0.300	0.187
Endpoint-matched	Corrupt	0.593	0.000	0.000	0.000
Endpoint-matched	Genuine write	0.185	0.408	0.121	0.050

Table 5: Direct endpoint-matched reduced-prefix disadvantages $\Delta = d(\mathrm{reduced}) - d(\mathrm{genuine})$, with 90% seed-stratified quartet-cluster intervals. Positive values favor the genuine prefix. Simultaneous equivalence requires every interval in a row to lie within $[-0.025, 0.025]$.

Reduction	Δ_{value}	$\Delta_{\mathrm{centered}}$	$\Delta_{\mathrm{increment}}$
Final repeat	0.016 $[-0.023, 0.052]$	0.014 $[-0.011, 0.036]$	0.003 $[-0.002, 0.008]$
Mean repeat	0.011 $[-0.031, 0.049]$	0.002 $[-0.023, 0.024]$	0.001 $[-0.005, 0.008]$
Endpoint-linear	-0.021 $[-0.050, 0.002]$	-0.009 $[-0.030, 0.007]$	-0.002 $[-0.004, 0.001]$

B.7 Mean-Repeat Retention Across Endpoint Matching and Drift Direction

Token displacement. In a post-hoc check of all 4,800 streams, mean-repeat changes a median of three of the four tokens, yet moves each prefix by a median of only 0.44% of its distance from the recipient's own prefix: a structural simplification rather than a large perturbation.

Illustrative sample in Figure 2. The example is selected after outcomes were observed and is not additional statistical evidence. Among the 120 valid condition evaluations, choose the one whose stream-mean (G, Δ) is closest in squared Euclidean distance to the pooled $(\mathbb{E}[G], \mathbb{E}[\Delta])$; then choose the paired stream closest to that condition's (G, Δ) means. Ties use seed, quartet index, condition

name, and stream index. This selects seed 0, quartet 4, endpoint-matched/upward perturbation, stream 9 (zero-based indices). All plotted values come from the saved decoded forecasts under their fixed clean-series scale. Each centered curve subtracts that sample's own positions 5–32 mean. Selection uses both observed effects and does not imply that every condition resembles the example. The supplementary figure record, `figures/fig_mean_repeat_example.json`, stores the selection rule, source hash, and plotted values.

The fresh panel contains 30 independent quartets (ten per construction seed), each contributing four conditions and 40 paired streams. Endpoint matching (off/on) is crossed with corrupt drift-perturbation direction (down/up). Each quartet has one clean context with drift d_c; corrupt drifts are $d_c - \Delta d$ and $d_c + \Delta d$, with the magnitude fixed by the construction family. If g is the clean-minus-corrupt historical endpoint gap before matching, matching shifts the corrupt context by g; the unmatched condition leaves it unchanged. Because g can depend on drift direction, this is not an independent manipulation of signed gap magnitude and drift. The clean-series scale $s_q = \text{mean}\,|x_{\text{clean}}|$ is fixed across conditions, arms, tokenization, and distances. Identical clean inputs, scale, and decoding randomness permit one shared clean forecast per quartet and stream. Checkpoint, decoding, four-position write, and the eight arm definitions follow Appendix B.6; the fresh evaluation configuration is indexed in the evidence package described in Appendix D.

All 120/120 conditions pass implementation validity and enter the equally weighted primary analysis; no baseline-damage or response-based filtering is used. Centered error is computed separately for each sampled path over positions 5–32, after subtracting that path's own suffix mean, and only then averaged across streams. For each condition, let $G = d_{\text{centered}}(\text{corrupt}) - d_{\text{centered}}(\text{genuine})$ and $\Delta = d_{\text{centered}}(\text{mean-repeat}) - d_{\text{centered}}(\text{genuine})$. Mean-repeat improvement is thus $G - \Delta$, and retaining at least 75% requires

$$\mathbb{E}[(G - \Delta) - 0.75G] = \mathbb{E}[0.25G - \Delta] \geq 0.$$

The prespecified 75% retention test (T1) requires a positive one-sided 95% lower bound on $\mathbb{E}[0.25G - \Delta]$; the population positive control (T2) requires the lower bound on $\mathbb{E}[G]$ to exceed 0.05 in clean-series-scale units. Equivalently, simplifying the prefix may cost at most a quarter of the original pooled improvement; this follows the fraction-retention formulation of non-inferiority testing (Rothmann et al., 2003). Internal-increment retention over positions 6–32 is supporting and non-gating (T3). Ten thousand seed-stratified bootstrap draws resample quartets, carrying all four conditions, arms, and streams together.

Table 6: Pooled retention and its components: 30 independent quartets, 120/120 valid conditions, and 40 paired streams per condition. Retention is pooled mean-repeat improvement divided by pooled original-edit improvement. Bounds are the 5th and 95th percentiles of 10,000 seed-stratified quartet-cluster draws: individually one-sided 95% bounds, together a central 90% interval. T1 is the 75%-retention test, T2 checks that the original edit has positive centered-shape improvement, and T3 is a supporting adjacent-increment retention check. LB means one-sided 95% lower bound.

Component	Point	Lower bound	Upper bound	Prespecified criterion
$\mathbb{E}[G]$ (T2)	0.3396	0.2411	0.4401	LB > 0.05
$\mathbb{E}[\Delta]$	-0.0144	-0.0325	0.0032	Reported component
$\mathbb{E}[0.25G - \Delta]$ (T1)	0.0993	0.0589	0.1406	LB > 0
Pooled retention $\widehat{R}$	1.043	0.987	1.078	Reported estimate
$\mathbb{E}[G_{\text{inc}}]$	0.1965	0.1321	0.2632	Reported component
$\mathbb{E}[\Delta_{\text{inc}}]$	-0.0092	-0.0160	-0.0028	Reported component
$\mathbb{E}[0.25G_{\text{inc}} - \Delta_{\text{inc}}]$ (T3)	0.0583	0.0366	0.0811	LB > 0; supporting

The decision rule checks coverage (all 120 conditions valid), then the genuine-effect test T2, then the retention test T1; all pass, as does the supporting increment check T3 (Table 6). The pooled retention ratio $\widehat{R} = 1 - \sum \Delta / \sum G$ is 1.043 (central 90% interval $[0.987, 1.078]$); all 10,000 bootstrap denominators are positive. Because the interval includes 1, mean-repeat is not shown to be superior. Per-cell genuine improvements range from 0.105 to 0.761 and are reported only as heterogeneity. The pooled pass does not establish retention in every cell, equivalent dynamics, absence of richer prefix information, or value-metric equivalence, and it leaves the paired study's joint-equivalence result (Appendix B.6) unchanged.

B.8 Base Replication

On the same synthetic construction, Chronos-T5-Base gives 0.997 restoration under continuous writing and 0.700 after a four-position release (one-sided 95% LB 0.626) (Section 4.2), passing the prespecified release criterion; exact replay reproduces the suffix. Its forced-prefix restoration, 0.698, is below the 0.80 starting effect required for prefix-specificity tests, so no prefix-specificity claim is made for Base.

B.9 Sampling and Eligibility

Table 7 gives populations and exceptions to the inference conventions in Appendix A. Directional assignment uses 5,000 construction resamples; the paired endpoint and fresh retention studies use 10,000.

Scale. The controlled synthetic studies (Appendices B.1–B.3, B.6–B.8, and B.10) fix the clean (donor) series' MeanScale; directional assignment (Appendix B.4) pools the MeanScale over its five contexts; the observed-pair studies (Appendix B.5 and Table 9) fix the recipient's.

B.10 Cached Upstream Sites with a Key–Value Cache

Cached upstream sites (follow-up; specified and hashed before decoding). Under cached decoding, the design crosses prefix tokens (recipient baseline or emitted under the edit) with the cache for the edited steps (unedited or edited): baseline, replay, cache-only (edited cache with baseline tokens forced), and release. Restoration shares the baseline denominator, so the token route is the replay restoration, the cache route the cache-only restoration, and the interaction the release restoration minus their sum. A pilot on fresh development quartets screened the residual-stream outputs of decoder layers (blocks in the code) 0–4 against a continuous-write anchor of 0.90 and, by a rule fixed in advance, selected the shallowest passing site (layer 0) and layer 4; the final decoder MLP served as control. At the control site, cache-only equals baseline and release equals replay bitwise on every stream (8,640 synthetic and 62,280 observed checks); replay is deterministic; cached and cache-free baselines give identical tokens on all 352 checked streams (maximum logit difference 3.1×10^{-5}). The prespecified reading rule (interval within ± 0.10 of zero) classifies the cache-only route as negligible at both upstream sites on the synthetic panel; the interaction at layer 0 is 0.278 $[0.173, 0.388]$ (Table 8; observed pairs in Table 9); the release-minus-replay contrasts, derived from the same prespecified estimands on the same shared draws, are 0.27 $[0.18, 0.35]$ at layer 0 and 0.018 $[0.003, 0.033]$ at layer 4. Observed-pair results are descriptive because continuous-write anchors fall below 0.90 at layer 0 and the control site; on observed pairs, persistence stays weak at every site (0.07–0.12 released restoration at $m = 4$). An in-run bootstrap resampler drew only first-seed clusters; point estimates were unaffected, and all synthetic intervals were recomputed from the saved records after the fix; the observed-pair run already used the fixed resampler.

Arm specification from the implementation. Every arm starts with an empty decoder cache and the same recipient encoder output; no arm reuses a cache snapshot from another trajectory. Baseline feeds its sampled recipient prefix with no write. Release samples its prefix under the write while building that trajectory's cache. Replay forces release's emitted prefix in an unedited run and rebuilds the cache without a write. Cache-only forces baseline's emitted prefix while applying the write and rebuilds the cache under that combination. At an upstream residual site i, the write replaces only the current-position activation at steps $0, \ldots, m-1$: it never overwrites an existing cache entry, but the downstream self-attention key and value rows in decoder blocks $i+1, \ldots, 5$ are computed from the altered state at those positions. A final-MLP write affects no cache entry. Cross-attention keys and values come from the fixed recipient encoder output and are unchanged. The hook is removed before the first suffix forward, which is unedited and consumes the last prefix token together with the arm-specific cache. Every arm initializes generators from the same per-stream seeds and makes one sampling call per position; forced-prefix arms sample before overwriting the token, preserving shared draw advancement.

Table 7: Exceptions and eligibility for the main Chronos experiments. Passing validity checks is distinct from satisfying a scientific anchor.

Experiment	Outer population	Qualification
Screen / new quartets / release / interfaces	24 quartets (8/seed), 20 streams	All eligible; interfaces reuse the release panel and are paired controls, not replications.
Prefix controls	24 quartets, 20 streams	477/480 streams match at $m = 4$; scrambled arms match dose but not likelihood.
Directional assignment	36 units, four targets, 20 streams	All qualify; correct/opposite conditions share the exact prefix multiset.
Base release	24 quartets, 20 streams	Release is confirmatory; prefix specificity is not established because its starting-effect criterion fails.
Observed-pair release	173/240 eligible pairs, 20 streams	Pair bootstrap was fixed before evaluation; the post-hoc 33-component source sensitivity keeps both pair roles clustered.
Mean-repeat retention	30 quartets, four conditions each, 40 streams	All 120 scheduled conditions valid; no damage or recovery filtering; quartet-cluster inference, not 120 independent replications.
Write-duration follow-up	173 eligible pairs, 20 streams, $m \in \{1, 2, 4, 8, 16\}$ and continuous	Specified before decoding; the $m = 8$ value was observed earlier.
Cached upstream sites	24 fresh synthetic quartets; 173 observed pairs	Sites selected on separate development quartets; observed-pair results descriptive.

Table 8: Token route versus cache route at $m = 4$ on fresh synthetic quartets (24 eligible; design fixed and hashed before decoding). Each column is the restoration named in the text, as point [5th, 95th] from a seed-stratified quartet-cluster bootstrap (2,000 resamples, shared draws across arms, recorded seeds). The anchor is the continuous write scored on positions 5–32.

Site	Anchor	Release	Replay	Cache-only	Interaction
Layer 0 (residual)	0.981 [0.955, 0.997]	0.871 [0.790, 0.935]	0.605 [0.478, 0.730]	−0.012 [−0.045, 0.021]	0.278 [0.173, 0.388]
Layer 4 (residual)	0.999 [0.999, 1.000]	0.962 [0.949, 0.974]	0.944 [0.922, 0.969]	0.011 [−0.001, 0.025]	0.007 [−0.011, 0.023]
Final MLP (control)	1.000 [0.999, 1.000]	0.939 [0.914, 0.964]	0.939 [0.914, 0.964]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]

Table 9: Token route versus cache route at $m = 4$ on the 173 eligible observed pairs (descriptive; pair-cluster bootstrap, 2,000 resamples, shared draws, recorded seeds; the pairs and their scale are those of Section 4.3).

Site	Anchor	Release	Replay	Cache-only	Interaction
Layer 0 (residual)	0.649 [0.622, 0.679]	0.122 [0.096, 0.151]	0.054 [0.038, 0.071]	0.063 [0.046, 0.083]	0.006 [−0.009, 0.020]
Layer 4 (residual)	0.923 [0.908, 0.939]	0.110 [0.084, 0.136]	0.096 [0.072, 0.120]	0.008 [0.004, 0.011]	0.007 [0.000, 0.013]
Final MLP (control)	0.831 [0.780, 0.884]	0.073 [0.053, 0.095]	0.073 [0.053, 0.095]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]

C Public Moment Transplant: Reproduction and Factorial Attribution

These details support Section 5. Reproducing the public pipeline and attributing its displayed output answer different questions.

We reproduce the six public NASDAQ examples using Chronos-T5-Small.

C.1 Public Transplant and Activation-by-Scale Factorial

The public code uses *content/target* for the recipient and *style* for the donor. To avoid conflict with Sections 3–4.3, we call them recipient and donor below; s_r denotes the public code's output scale. The pinned source is `generate_chronos_predictions.py` at commit `1c98680c3daee88c621663946215ca62e9d8e482`.

Figure 4 uses letters A–F for the six public pairs. In recipient←donor notation, the source-ID crosswalk is A: 17C←08X; B: 07C←00X; C: 19C←20X; D: 08X←17C; E: 00X←07C; F: 20X←19C. In these IDs, C means calm and X means crash. Numeric prefixes are source case names rather than necessarily observation years. The source paper and code give different windows for all six cases. For the five cases other than 20X, their start dates agree but their end dates differ. The largest discrepancy is 20X: the source paper's Table 1 lists 30 January–6 June 2020, whereas the public code configures 27 December 2021–4 October 2022 and uses its first 128 observations as model input.

At encoder layer 1 (zero-based), excluding the end-of-sequence (EOS) row, the intervention maps recipient activation x to $(x - \mu_c)(\sigma_c + \epsilon)^{-1}(\sigma_d + \epsilon) + \mu_d$ using sequence-axis means μ_c, μ_d and unbiased standard deviations σ_c, σ_d for recipient and donor, with $\epsilon = 10^{-5}$. The public code uses PyTorch's default unbiased standard deviation, whereas Eq. 2 of Sanyal et al. (2025) writes the $1/T$ form; its Figure 2 states layer 8 for both models, whereas the executable Chronos path uses encoder layer 1 (zero-based). Our reproduction follows that executable path. The public code's forecast then uses output scale $\alpha s_d + (1 - \alpha)s_c$ with a case-specific α. The six values are hard-coded in the same file as `alphas=[0.1, -0.2, -0.1, 0.2, -1.8, -0.4]`, with no comment or stated selection rule.

The factorial crosses transplant off/on with recipient/public-code scale on the exact six public cases, using common normalized samples across scale conditions. For each cell it reports terminal position relative to history, history-to-first-forecast transition, and within-forecast movement. The joint effect, output-scale-only effect, transplant-only effect, and interaction obey the additive identity exactly (Table 10). The public-code-path and baseline reproductions are exact; verification details are indexed in the evidence package described in Appendix D. Because these are six fixed public examples, the decomposition is descriptive and carries no population interval.

Let $f_{as,H}$ be the terminal median, with transplant off/on indexed by $a = 0/1$ and recipient/public-code scale by $s = 0/1$. The joint (J), scale-only (S), transplant-only (T), and interaction (I) contrasts satisfy

$$J = f_{11,H} - f_{00,H}, \qquad S = f_{01,H} - f_{00,H}, \quad T = f_{10,H} - f_{00,H},$$
$$I = f_{11,H} - f_{10,H} - f_{01,H} + f_{00,H}, \qquad J = S + T + I.$$

In words: joint change = scale-only change + transplant-only change + interaction. The reference is the unedited terminal median. Subtracting a common historical endpoint gives the same contrasts. Their signed ratios are accounting quantities, not shares of semantic information or unique causal responsibility.

For 00X←07C ($\alpha = -1.8$), the public code's scale 4719.2 exceeds both the recipient scale 2907.6 and the donor scale 1901.2; the operation is not always an interpolation.

Table 10: Exact six-case activation-by-output-scale decomposition. Case IDs are recipient←donor; C/X denote calm/crash, and all paths are terminal medians over 200 paired samples. The public code calls the recipient content/target and the donor style; here s_c, s_d, and s_r denote recipient, donor, and public-code output scales. J is the joint terminal change from the unedited baseline, S the output-scale-only change, T the transplant-only change, and $I = J - S - T$ the interaction remainder. Values are rounded only for display.

Case	α	s_c	s_d	s_r	s_r/s_c
17C←08X	+0.1	5476.0	1451.8	5073.6	0.927
07C←00X	-0.2	1901.2	2907.6	1699.9	0.894
19C←20X	-0.1	7866.7	13845.3	7268.8	0.924
08X←17C	+0.2	1451.8	5476.0	2256.7	1.554
00X←07C	-1.8	2907.6	1901.2	4719.2	1.623
20X←19C	-0.4	13845.3	7866.7	16236.7	1.173

Case	Baseline $f_{00,H}$	Output scale only $f_{01,H}$	Transplant only $f_{10,H}$	Joint $f_{11,H}$
17C←08X	5620.6	5207.5	5841.4	5412.1
07C←00X	1979.2	1769.7	2048.9	1832.0
19C←20X	8362.7	7727.1	8593.4	7940.3
08X←17C	1202.8	1869.5	1186.8	1844.7
00X←07C	2089.1	3390.7	2003.8	3252.3
20X←19C	12129.8	14225.0	12282.1	14403.5

Case	Joint J	Scale-only S	Transplant-only T	Interaction I
17C←08X	-208.5	-413.0	220.8	-16.2
07C←00X	-147.2	-209.6	69.7	-7.4
19C←20X	-422.4	-635.6	230.7	-17.5
08X←17C	642.0	666.8	-16.0	-8.9
00X←07C	1163.2	1301.6	-85.3	-53.1
20X←19C	2273.7	2095.1	152.3	26.3

Scale identity and scoring windows. For positive output scales, let q_a be the normalized terminal median under transplant state a. Then $S = (s_1 - s_0)q_0$, $T = s_0(q_1 - q_0)$, and $I = (s_1 - s_0)(q_1 - q_0)$. Multiplicative decoding therefore permits interaction and scale-only/joint ratios above one; the ratios are not shares of causal responsibility. Transplant-only terminal contrasts oppose the joint direction in all three calm-from-crash cases and two of three crash-from-calm cases. At the history-to-forecast transition, the signed scale-only/joint ratio is at least one in every case. All three joint calm-from-crash forecasts have positive net within-horizon movement ($f_H - f_1 = 92.99, 24.93, 266.45$) despite ending below the historical endpoint. Their displayed downturn is at the boundary transition; positive net movement neither implies monotonic increase nor excludes an intermediate drawdown.

Sample-level sign stability (post-hoc). The decoded artifact retains the 200 paired normalized sample paths per cell, so the stability of the terminal contrasts under decoding randomness can be assessed from the saved artifact without new model inference. A paired percentile bootstrap over the 200 shared sample indices (2,000 resamples, recorded seed) gives transplant-only central 90% intervals excluding zero in the three large opposing contrasts (cases A–C) and including zero in the two small opposing contrasts (D, E) and in the agreeing case F (Table 11). The terminal medians therefore oppose the joint direction in five cases, but the opposition is resolved at the sample level in three. The scale-only sign is determined by construction: $S_j = (s_r - s_c)\, q_{0,j}$ with every baseline terminal sample positive, so the scale-only change matches the displayed direction in all 200 paired samples in each of the six cases, and likewise at the history-to-forecast transition. The attribution statement therefore does not rest on the two unresolved transplant-only signs; it follows from the exact accounting identity and from the scale-only change, which supplies 92–198% of the joint terminal change and at least 100% of the transition change. The analysis is derived and descriptive; the extraction script and its hash are recorded in the evidence package.

Table 11: Post-hoc sample-level stability of the terminal contrasts. T is the transplant-only terminal change (median over 200 paired samples); the interval is a paired percentile bootstrap over the shared sample indices (2,000 resamples). S counts the samples whose scale-only change moves in the displayed joint direction; the count is 200/200 in every case because $S_j = (s_r - s_c)q_{0,j}$ with all baseline terminal samples positive. Derived from the saved factorial artifact; six fixed public cases, no cross-case inference.

Case	T median	90% paired bootstrap	Excludes 0	S in displayed direction
A 17C←08X	$+220.8$	$[40.1, 281.0]$	yes	200/200
B 07C←00X	$+69.7$	$[20.9, 111.5]$	yes	200/200
C 19C←20X	$+230.7$	$[86.5, 403.7]$	yes	200/200
D 08X←17C	-16.0	$[-53.2, 31.9]$	no	200/200
E 00X←07C	-85.3	$[-181.2, 53.3]$	no	200/200
F 20X←19C	$+152.3$	$[-101.5, 558.3]$	no	200/200

Terminal quantiles and within-horizon drawdowns (post-hoc). Because the saved artifact retains the full sample paths, two further metric families probe dependence on the reported statistic; their list was fixed before these contrasts were computed. First, lower quantiles at the terminal position, the lower edges of the 50% and 90% bands displayed by the source. Quantiles are scale-equivariant under a positive multiplier, so the scale-only sign rule holds at every quantile by construction; the scale-only change supplies 0.21–0.95 of the joint quantile change across cases and quantiles. The transplant-only quantile contrasts show that with a crash donor the transplant alone deepens the lower tail (T at $q = 0.05$: -479.8, -280.2, -692.1, paired 90% intervals excluding zero) and with a calm donor raises it, even though at the median the transplant-only contrast opposes the displayed direction in five of the six cases. The $q = 0.05$ percentile rests on about ten of the 200 samples, so its intervals are correspondingly wide. Second, the percentage maximum drawdown of each 64-step forecast path is invariant to positive rescaling, so the scale-only and interaction terms are exactly zero and the joint effect equals the transplant-only effect; the transplant deepens median drawdowns by 3.0–6.1 percentage points with a crash donor and shallows them by 5.0–10.2 with a calm donor (Table 12). The summary is two-sided: the transplant shifts the lower tail and within-horizon drawdowns toward the donor regime (deeper with crash donors, shallower with calm donors), while the displayed median-level direction comes from the output-scale operation. The moment split below locates these shifts mainly in the transplanted means. This also measures the intermediate drawdown possibility left open above.

Table 12: Post-hoc metric dependence for the six public cases; the metric list was fixed before these contrasts were computed. T is the transplant-only terminal change at quantile q in display units (paired percentile bootstrap over the 200 shared sample indices, 2,000 resamples; the q=0.05 percentile rests on about ten samples and its intervals are wide). Drawdown is the within-horizon percentage maximum drawdown of each 64-step forecast path, in percent units, with per-cell medians before and after transplantation and the median contrast; it is invariant to positive rescaling, so the scale-only and interaction terms are exactly zero. Derived from the saved factorial artifact; six fixed public cases, no cross-case inference.

Case	T at q=0.05	T at q=0.25	Drawdown (off→on; Δ, 90% interval)
A 17C←08X	-479.8 $[-598.2, -321.2]$	-120.4 $[-240.9, 0.0]$	$5.3 \to 9.1$; $+3.8$ $[2.7, 5.0]$
B 07C←00X	-280.2 $[-361.7, -223.7]$	-111.5 $[-139.4, -52.3]$	$6.1 \to 12.2$; $+6.1$ $[4.5, 7.9]$
C 19C←20X	-692.1 $[-1095.8, -360.2]$	-230.7 $[-346.0, 115.3]$	$4.1 \to 7.1$; $+3.0$ $[2.2, 3.6]$
D 08X←17C	$+170.3$ $[52.2, 281.5]$	$+21.3$ $[-18.6, 53.2]$	$12.1 \to 7.1$; -5.1 $[-6.8, -4.2]$
E 00X←07C	$+786.6$ $[509.5, 991.3]$	$+250.5$ $[111.6, 405.0]$	$23.4 \to 13.2$; -10.2 $[-13.0, -7.2]$
F 20X←19C	$+1735.7$ $[918.6, 2842.1]$	$+532.9$ $[177.6, 913.5]$	$12.8 \to 7.8$; -5.0 $[-6.7, -3.7]$

Moment split (post-hoc; specified and hashed before computation). To locate the transplant-only tail and drawdown effects, we split the transplant into its two moments. Mean-only replaces the recipient's per-feature sequence means with the donor's, $x' = x - \mu_c + \mu_d$. Std-only rescales the recipient's centered activations to the donor's standard deviations while keeping the recipient's means, $x' = (x - \mu_c)(\sigma_c + \epsilon)^{-1}(\sigma_d + \epsilon) + \mu_c$. Both variants use the public path's layer, edited rows,

end-of-sequence handling, statistics, ϵ, checkpoint, and sampling. Before any decode, we recorded the SHA-256 hash of a specification fixing the conditions, metrics, contrasts, bootstrap design, and reading rule. Two positive controls reproduce the saved artifact bitwise in all six cases: replacing neither moment gives the baseline cell, and replacing both gives the transplant-only cell. The two arm functions pre-exist in the frozen reproduction code and were executed there on earlier synthetic panels with different data, sampling, and statistics (only summary dispersion statistics were recorded); the six-case conditions above were not computed before the specification was hashed. An earlier std-only arm in the same code omits ϵ from both standard deviations; the follow-up uses the ϵ-matched form above and leaves that file unchanged. A variant carries an effect when its paired 90% interval excludes zero in the donor direction. Because decoded values are quantized, interval endpoints often fall exactly at zero; such intervals are not counted as excluding zero. Mean-only carries the drawdown shift in 6/6 cases (Table 13) and the 5th-percentile shift in 5/6. Std-only carries neither in any case, and its one drawdown interval that excludes zero (case D) points away from the donor direction. The full transplant is not always the sum of its parts: the non-additivity $\Delta_{\text{full}} - \Delta_{\text{mean}} - \Delta_{\text{std}}$ excludes zero in 6 of 24 case–metric cells, most strongly in case C. Full-transplant point estimates match Table 12 exactly; their intervals differ slightly because this follow-up uses its own prespecified bootstrap draws. The split locates the effect in the transplanted means but does not identify what those means encode.

Table 13: Moment split for the six public cases (post-hoc; conditions, metrics, and reading rule fixed and hashed before computation). Changes without rescaling (at the recipient scale) for the full transplant, mean-only, and std-only: terminal 5th percentile (display units) and within-horizon percentage maximum drawdown (percentage points), each shown as point [5th, 95th] from a paired percentile bootstrap over the 200 shared sample indices (2,000 resamples). Full-transplant points equal Table 12, while intervals use this follow-up's draws. The $q = 0.25$ and median contrasts and the non-additivity terms are in the evidence package.

Δ *at* q=0.05 *(display units)*

Case	Full	Mean-only	Std-only
A	-479.8 [$-598.2, -321.2$]	-560.1 [$-724.7, -361.3$]	-38.1 [$-122.4, +120.4$]
B	-280.2 [$-361.7, -223.7$]	-169.4 [$-348.5, -85.0$]	-29.3 [$-83.6, +41.8$]
C	-692.1 [$-1095.8, -354.7$]	-285.5 [$-640.2, +40.5$]	$+60.6$ [$-230.7, +297.0$]
D	$+170.3$ [$+54.2, +260.8$]	$+160.2$ [$+43.1, +275.1$]	$+20.8$ [$-74.5, +117.6$]
E	$+786.6$ [$+507.3, +1021.1$]	$+740.8$ [$+315.5, +959.3$]	$+127.9$ [$-39.4, +383.7$]
F	$+1735.7$ [$+913.5, +2938.6$]	$+1431.2$ [$+715.6, +2735.6$]	$+700.4$ [$-304.5, +1908.5$]

Δ *within-horizon drawdown (percentage points)*

Case	Full	Mean-only	Std-only
A	$+3.8$ [$+2.7, +5.0$]	$+3.6$ [$+2.5, +4.8$]	-0.4 [$-0.9, +0.0$]
B	$+6.1$ [$+4.4, +7.8$]	$+4.0$ [$+2.8, +5.7$]	$+0.2$ [$-0.7, +1.1$]
C	$+3.0$ [$+2.2, +3.6$]	$+1.4$ [$+0.7, +1.8$]	$+0.0$ [$-0.3, +0.5$]
D	-5.1 [$-6.8, -4.0$]	-4.9 [$-6.5, -3.7$]	$+2.4$ [$+0.5, +3.5$]
E	-10.2 [$-13.2, -7.2$]	-9.5 [$-12.4, -6.6$]	$+0.1$ [$-2.5, +2.3$]
F	-5.0 [$-6.7, -3.7$]	-4.7 [$-6.4, -3.5$]	-1.0 [$-2.4, +0.0$]

D Study Scope, Provenance, and Qualifications

D.1 Scope of Supporting Experiments

Structural tests. Supporting experiments asked whether the all-step write at the final decoder MLP could be reproduced by a simpler or shared part of the donor's activations. Writing only the donor's mean activation at that site (a different experiment from the mean-only transplant in Section 5) restored the target in 0/12 cells, against 10/12 for the full donor write. Frequency-based components and pooled low-dimensional subspaces also missed their prespecified transfer criteria; in one subspace test the required starting effect held in only 2 of 3 seeds. These tests do not identify a shared or naturally occurring hidden subspace, do not rule one out, and do not bear on the emitted-prefix results of Section 4.

Architecture extensions. We also tried release on Lag-Llama (Rasul et al., 2024) and Toto (Cohen et al., 2024; Datadog, 2026). In both main experiments, a write at every step did not reach the required starting effect, so persistence could not be tested. In a separate Toto study, a write worked while active but failed its persistence criterion after a single edited step; longer writes and value-versus-cache comparisons there are descriptive and show no particular persistence route. These results neither establish nor rule out persistence on these architectures and do not rank them. Full methods and numbers are in the evidence package.

Exploratory directional evaluation. A separate exploratory experiment on event direction used a ratio whose denominator changed sign across conditions, so its intended reading is unreliable. It is reported descriptively in the evidence package and is not used as evidence in Section 5.

D.2 Sequential Specification

After the initial screen selected the final decoder MLP, outcomes motivated successive questions. Each experiment's panels, quantities to be estimated, thresholds, and decision rules were fixed before its evaluation; the full program was not specified in advance. Development runs are excluded from confirmatory evidence. Records are internally versioned, without external timestamps.

D.3 Material Qualifications

Table 14: Material qualifications and their consequences. Artifact codes are preserved for evidence-package lookup; Appendix locations identify the corresponding PDF discussion.

Study	Qualification	Consequence for reported evidence
Prefix controls (artifact EP-B3; App. B.3)	One evaluation outcome was visible during development.	Blinding was incomplete; criteria and implementation remained unchanged.
Directional assignment (artifact EP-DPR; App. B.4)	Raw target trajectories not retained; likelihood-stratified diagnostic excluded.	Saved target-distance blocks support reanalysis; the diagnostic supplies no primary evidence.
Observed-pair release (artifact EP-NWR; App. B.5)	An initial anchor computation scored the full write on positions 5–32; a documented correction moved it to the prespecified full horizon. Source-dependence sensitivity and the paired same-window comparison are post-hoc.	The correction changed the anchor value, not the released-suffix estimand or the reported conclusion; prespecified pair inference remains distinct from both post-hoc analyses.
Initial endpoint matching (artifact EP-B11; App. B.6)	Too few damage-eligible constructions.	The prespecified attribution question remains unanswered.
Paired endpoint study (artifact EP-B13; App. B.6)	Reuses EP-B11 constructions.	Not an independent-panel replication; new-analysis timing is disclosed.
Public-code reproduction (artifact EP-B7; App. C.1)	Historical runner bodies and decode timestamps are unavailable; the sample-level sign bootstrap, the quantile/drawdown contrasts, and the moment split are post-hoc.	This limits historical execution reconstruction but does not affect the bitwise reproduction of the public code; all are descriptive within the six fixed cases. The moment-split arm functions had been run earlier on different synthetic panels; the six-case split was specified and hashed before computation.
Toto	Some historical verification records are missing; an earlier Toto attempt to compare value and cache routes was not fully implemented.	Only the final Toto write-and-release study in Appendix D.1 is supported, not the earlier attempt.
Mean-repeat retention (artifact EP-B14; App. B.7)	Specification fixed before evaluation; additional checks followed outcome access.	Original verification record incomplete; later checks documented separately.

continued on next page

Table 14: Material qualifications and their consequences (continued)

Study	Qualification	Consequence for reported evidence
Cached upstream sites (artifact EP-CACHE_ROUTE; App. B.10)	An in-run bootstrap resampler drew only first-seed clusters; the runner's path setup changed between the pilot and evaluation launches (path-only edit).	Point estimates were unaffected; all synthetic intervals were recomputed from the saved records after the fix, and the observed-pair run used the fixed resampler; the path edit touched no scientific semantics.
Cached upstream sites (earlier artifact EP-B8)	Reuses the EP-B8 cached-execution code read-only; EP-B8 is an earlier terminal-site cached experiment on a different panel realization, and 2 of 24 evaluation construction signature tuples coincide with B8's.	Code lineage only; EP-B8's earlier terminal-site result (bitwise token-route mediation) is independent evidence consistent with the bitwise control here; the panels are largely distinct (2 of 24 construction signatures coincide).
Write-duration follow-up (artifact EP-NWR-DOSE; App. B.5)	Two launches failed before any artifact was written (runner bugs); the $m = 8$ suffix value had been observed before specification; the continuous arm is the $m = 32$ release by implementation.	All bitwise controls passed on all 173 pairs in the clean launch; prespecified reading rules used only post-specification computations; previously observed values are labeled as such.