DO TIME SERIES FOUNDATION MODELS USE IMPUTED VALUES?

Anonymous authors
Paper under double-blind review

ABSTRACT

Forecasting pipelines impute missing histories for time series foundation models. Do better reconstructions improve forecasts? The answer depends on both computational access and learned use. We probe eighteen checkpoints from eleven families through their *missingness interfaces*, paths combining fills with observation indicators. Ten of the eighteen checkpoints block position-specific fill content on their declared paths. In Chronos-Bolt, masking retains dependence on the context mean and standard deviation but removes this content, explaining why even exact fills recover little forecast accuracy under scattered and block missingness. More generally, an interface bound characterises an imputer's access to the forecast, which is distinct from learned use in paired pretraining experiments. Gap exposure can improve accuracy without responding to fill quality, whereas repair-aware training makes retained content useful. Our retrofit passes fills with observation flags and learns across gap mechanisms and repair qualities. After 5,000 continued-pretraining steps, it reduces zero- and linear-fill degradation relative to complete-context error across three families, with little change on complete histories. On nine GIFT-Eval subsets with recorded gaps, the retrofit lowers median MASE by up to 18% in high-missingness windows. Our code is available at `https://anonymous.4open.science/r/hole-in-input-repro-2E97/README.md`.

1 INTRODUCTION

Time series foundation models (TSFMs) such as Chronos (Ansari et al., 2024), TimesFM (Das et al., 2024), Moirai (Woo et al., 2024) and Tiny Time Mixers (Ekambaram et al., 2024) bring pretrained forecasting to new time series. Yet sensor failures and dropped records leave gaps in their histories. In the released GIFT-Eval files, we count 11.77% of 305M points as explicitly missing (Aksu et al., 2024). Recurrent, attention-based and diffusion imputers reconstruct these histories (Cao et al., 2018; Du et al., 2023; Tashiro et al., 2021), as do recent TSFM adaptations (Le Naour et al., 2026b). A reconstructed history carries both candidate values and their unobserved status. Passing fills as observations loses this distinction, while declaring them missing can remove their content. The forecast impact of a fill thus depends on what the interface preserves and how the model uses it.

Consider supplying the exact values removed from a history, while keeping the model, observations and missing-position indicators fixed. In Chronos-Bolt (Amazon Web Services, 2024), replacing interpolation with these true values leaves forecast error nearly unchanged under scattered and block missingness (Figure 1; App. I). This eliminates reconstruction error without recovering the complete-input forecast. The same sweep looks different on content-retaining paths. Moirai 2.0, which saw one mask pattern in pretraining, recovers part of the shortfall, and Moirai 1.1, trained on diverse patterns, nearly reaches the clean-context error. What reaches the model at a declared missing position?

We call this path the *missingness interface*. Several released interfaces discard or replace declared fills (Table 9). Chronos-Bolt masks after normalisation: exact fills can change the context mean and standard deviation while their temporal content is discarded. Thus fill dependence need not imply temporal access; our bound limits changes carried by the surviving statistics.

This access limit motivates retaining flagged fills at their temporal positions. Whether that access is useful depends on training. Within-family controls show that gap exposure improves accuracy even when fill response stays flat, whereas retained content becomes useful with repair-aware training. Robustness to gaps and use of repairs are thus distinct learned behaviours. We pair content access

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

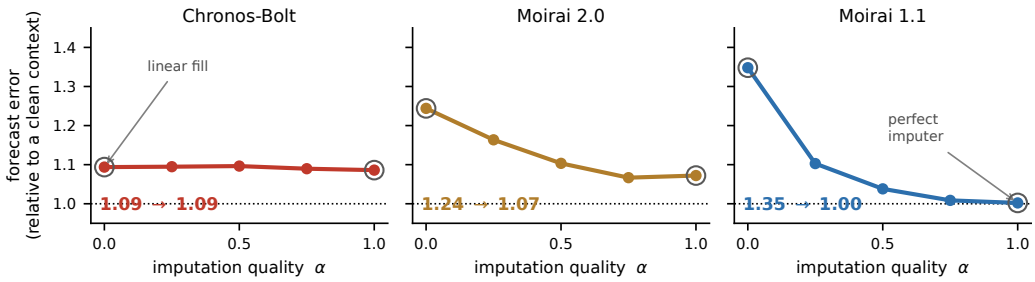

Figure 1: Forecast error of three declared paths as the fill is swept from linear interpolation (α=0) to the *ground truth* (α=1) on identical windows, under block outages at rate 0.7; 150 windows per cell, averaged over ETTh1, ETTm1 and weather. The dotted line is the clean-context error.

with training across gap mechanisms and repair qualities, then average adapted and original weights to limit clean-context changes. Our contributions are:

- **Matched evaluations isolate repair effects.** Matched-window audits show how gap mechanisms, input conventions and target selection change the measured comparison, and we fix all three before testing access and learned use (Section 4.2).
- **Interfaces restrict access to reconstructed content.** We bound fill-reachable forecasts, probe eighteen checkpoints from eleven families, and connect the surviving content to forecast changes through quality sweeps and learned imputers (Sections 3.2–4.4).
- **Training converts retained content into usable repairs.** Within-family controls motivate a content-preserving training recipe that reduces zero- and linear-fill own-clean degradation across three families in 5,000 steps (Sections 4.5–5).

2 Related work

From reconstruction to prediction. Imputation research improves estimates of missing data through generative models such as GAIN and MIWAE (Yoon et al., 2018; Mattei & Frellsen, 2019), and through temporal structure in BRITS, SAITS, CSDI, GP-VAE and GRIN (Cao et al., 2018; Du et al., 2023; Tashiro et al., 2021; Fortuin et al., 2020; Cini et al., 2022). NuwaTS and MoTM extend reconstruction across time-series domains (Cheng et al., 2024; Le Naour et al., 2026a;b). BRITS and SAITS evaluate downstream prediction; GRU-D and NeuMiss incorporate missingness directly into predictors (Che et al., 2018; Le Morvan et al., 2020). Impute-then-predict theory explains why accurate conditional imputation is not necessary for asymptotically optimal prediction with an expressive learner (Le Morvan et al., 2021), with diminishing returns from imputation accuracy for downstream prediction (Morvan & Varoquaux, 2024); probabilistic forecasting can also marginalize missing inputs (Wen et al., 2025). We ask which parts of an external reconstruction survive a released predictor's input processing, then connect this computational access to training on repairs.

Representing gaps and learning from them. Released TSFMs make different choices at declared missing positions: padding tokens, zeroed content with retained indicators, internal interpolation, learned replacement vectors, or value–mask concatenation (Ansari et al., 2024; 2025; Auer et al., 2025; Graf et al., 2026; Feng et al., 2025; Das et al., 2024; Moroshan et al., 2025; Woo et al., 2024; Liu et al., 2025a) (Table 9 lists the paths we probe). These choices determine access to external fills. On the training side, MOMENT uses masked pretraining, while Moirai and TempoPFN expose models to missingness during training (Goswami et al., 2024; Woo et al., 2024; Moroshan et al., 2025). Our fill sweeps and within-family controls separate robustness to incomplete histories from learned use of repairs.

What a missing-data evaluation measures. Missing-data theory characterises mechanisms, testability and observational equivalence (Rubin, 1976; Little, 1988; Bojinov et al., 2020; Molenberghs

et al., 2008). These questions concern what observations identify, while our interface analysis concerns what a specified computation retains. Evaluation must also distinguish point accuracy from uncertainty. Calibration studies measure agreement between interval coverage and nominal frequencies (Adler et al., 2026); conformal prediction provides marginal coverage, including time-series constructions (Lei et al., 2018; Romano et al., 2019; Stankevičiūtė et al., 2021), with missing-data extensions targeting mask-conditional validity (Zaffran et al., 2023; Fan et al., 2025). These distinctions guide our evaluation, and benchmark designs increasingly make them explicit (Qiao et al., 2026): input conventions determine what the model receives, target selection determines which errors count, and coverage assesses uncertainty beyond point accuracy.

3 The missingness interface

3.1 Problem formulation

A context $\mathbf{x}$ contains L past values and the forecast has horizon H. The observation mask m is one at observed positions and zero at holes. The missing-position set and missing fraction are

$$\mathcal{M} = \{t : m_t = 0\}, \qquad r = \frac{|\mathcal{M}|}{L}. \tag{1}$$

Candidate fills complete these positions. A *plain* input passes the context alone, while a *declared* input identifies holes through a mask or native NaN handling. The forecast map f includes preprocessing and the backbone. We vary $\mathbf{x}_{\mathcal{M}}$ at fixed observations, missing positions and covariates. Fill-quality blends interpolate between linear interpolation $\mathbf{x}^{\text{lin}}$ and the true context $\mathbf{x}^{\star}$ at the missing positions:

$$\mathbf{x}_{\mathcal{M}}(\alpha) = (1-\alpha)\mathbf{x}^{\text{lin}}_{\mathcal{M}} + \alpha\mathbf{x}^{\star}_{\mathcal{M}}, \qquad \alpha \in [0,1]. \tag{2}$$

Here α controls fill quality independently of the missing fraction r: $\alpha = 1$ supplies the *oracle* fill and $\alpha = 0$ linear interpolation.

3.2 An interface bound on fill access

To determine what an imputer can change, we track fill dependence through preprocessing. Blocking a fill at its original position need not remove all its influence: normalisation may already have used it. We collect the surviving statistics in s and write any interface that blocks content after normalisation as

$$f(\mathbf{x}) \;=\; \phi\big(\, s(\mathbf{x}),\; m \odot \mathbf{x},\; m \,\big), \tag{3}$$

where $\odot$ is elementwise multiplication and ϕ is the remaining prediction map. With observations and mask fixed, $m \odot \mathbf{x}$ is constant. Every effect of varying fills must therefore pass through s. The number of statistics constrains both the global set of obtainable forecasts and their independent local changes.

Observation 1 (Fill-reachable set)**.** For fixed observations and mask, define the reachable forecasts and fill Jacobian as

$$\mathcal{F} = \{f(\mathbf{x}) : \mathbf{x}_{\mathcal{M}} \in \mathbb{R}^{|\mathcal{M}|}\}, \qquad J_{\mathcal{M}} = \frac{\partial f}{\partial \mathbf{x}_{\mathcal{M}}}. \tag{4}$$

Under equation 3, assume s is locally Lipschitz in fills and $u \mapsto \phi(u, m \odot \mathbf{x}, m)$ is locally Lipschitz on the attained statistics, with $s \in \mathbb{R}^{d_s}$. For an H-dimensional forecast,

$$\dim_{\mathrm{H}}(\mathcal{F}) \le \min(H, |\mathcal{M}|, d_s), \qquad \operatorname{rank}(J_{\mathcal{M}}) \le \min(H, |\mathcal{M}|, d_s), \tag{5}$$

where the rank bound holds wherever the component derivatives exist. If s is independent of fills, $\mathcal{F}$ is a single point. Without the factorisation, the general derivative bound is $\min(H, |\mathcal{M}|)$, which is attained only at full derivative rank.

Hausdorff dimension measures the geometric size of the reachable set, while Jacobian rank counts independent local forecast-change directions. The bound follows because a locally Lipschitz map cannot increase dimension, while the derivative factorises through s (App. B.1). Quality sweeps in Section 4.4 test whether accessible changes improve the forecast.

Corollary 2 (The three Chronos-Bolt paths)**.** For fixed $\mathcal{M}$: (a) replacing candidate fills by NaN makes the composed candidate-fill map independent of those fills, with $J_{\mathcal{M}} = 0$; (b) finite fills with an explicit mask give $\operatorname{rank}(J_{\mathcal{M}}) \le \min(H, |\mathcal{M}|, 2)$; (c) finite fills without a mask retain content, with $\operatorname{rank}(J_{\mathcal{M}}) \le \min(H, |\mathcal{M}|)$.

In (b), $s = (\text{loc}, \text{scale})$ contains the completed context's mean and standard deviation. The bound concerns target fills at fixed covariates, matching the intervention in our access probes. Figure 2 illustrates the paths with their measured ranks and responses.

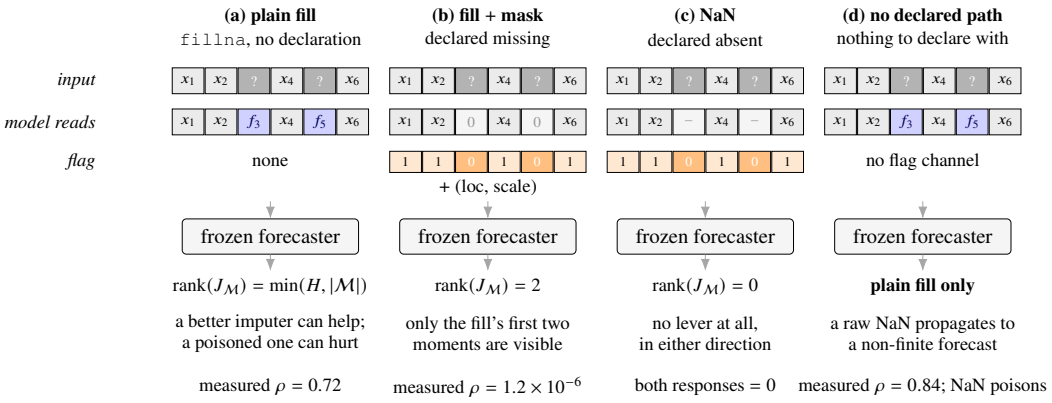

Figure 2: Fill access under selected input paths. Each column reads one shared context with two holes (dark) and shows what the frozen forecaster receives: observed values (gray), fills at their temporal positions (blue), zeros or nothing at the holes, plus the observation flag where one exists. Panels (a)–(c) show Chronos-Bolt: plain fill, fill with an observation mask, and declared absence; (d) shows Timer-base's plain path, its only option. Ranks are measured on 72 series at $r = 0.3$. Both NaN perturbation responses vanish, so ρ is undefined.

3.3 Diagnosing position-specific access

The bound distinguishes access through statistics from access to temporal content. We test this distinction with automatic-differentiation Jacobians and fill permutations. Permuting fills among holes changes their locations but preserves their multiset. Independently redrawn fills provide a response scale:

$$\rho \;=\; \frac{\mathbb{E}\,\|f(\text{permuted fill}) - f(\text{fill})\|}{\mathbb{E}\,\|f(\text{redrawn fill}) - f(\text{fill})\|}. \tag{6}$$

With a positive denominator, the *fill-retention ratio* ρ near one indicates comparable responses, while near zero indicates little position-specific response on that scale. When both responses vanish, ρ is undefined. Permutation also preserves other symmetric statistics, and equation 3 identifies the two-statistic route specifically in Chronos-Bolt. Repeated inference separates sampling noise from the response to changed fills on plain and supported declared paths (App. D.1).

4 Experiments

We first fix the evaluation conditions (Sections 4.1–4.2), then test what released interfaces transmit and whether accessible fills improve forecasts (Sections 4.3–4.4), and finally separate access from learned use (Section 4.5).

4.1 Experimental setup

Data and masks. Controlled experiments impose gaps on nine complete-input benchmarks from standard forecasting collections (7–862 channels) (Godahewa et al., 2021). Recorded-gap experiments use Penmanshiel turbine curtailment (Plumley, 2022), METR-LA sensor outages (Li et al., 2018) and nine GIFT-Eval subsets (Aksu et al., 2024). We count explicit NaNs as missing, while finite zeros and plateaus remain finite values. Unless stated otherwise, learning uses the first 80% of each timeline and evaluation the last 20%, while real-data learning and calibration use half-splits. Synthetic masks use scattered dropout, block outage, value censoring and extreme censoring, with missing fractions $r \in \{0.1, 0.3, 0.5, 0.7\}$ as defined in equation 1. Fill-quality sweeps separately vary

$\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ in equation 2 while keeping the mask fixed. Scattered/block masks are value-independent, while the two censoring mechanisms depend on values (Rubin, 1976).

Model roles and controls. Default context and horizon lengths are $L = 512$ and $H = 64$. We survey eighteen frozen checkpoints in eleven families, with Chronos-Bolt base as the default controlled model, and the Section 3.3 probe covers 31,983 channel-windows (App. C). A released checkpoint with its shipped input path is *vanilla*. Input adapters update embeddings with a frozen backbone. Factorial training tests exposure, while CPT adapts released weights.

Metrics. Our controlled-experiment measure is relative mean squared error (relMSE), each window's missing-input MSE divided by the same configuration's complete-context (*own-clean*) MSE:

$$e \;=\; \text{MSE(missing input)} \,/\, \text{MSE(own-clean input)}, \tag{7}$$

and we report the median ratio over windows. The reference model's ratio is e_{ref}, and the modified interface, adapter or CPT model's ratio is e_{mod}. Each uses its own complete-context error, so 1 means no degradation from clean input. Closure γ measures the fraction of the reference's excess degradation removed:

$$\gamma = \frac{e_{\text{ref}} - e_{\text{mod}}}{e_{\text{ref}} - 1} \times 100\%. \tag{8}$$

We report closure where $e_{\text{ref}} - 1 > 0.02$ (App. G.4.1). The *clean ratio* separately compares complete-context MSEs across configurations. For recorded GIFT-Eval gaps, mean absolute scaled error (MASE) (Hyndman & Koehler, 2006) divides absolute error by within-context one-step persistence error and scores observed targets. Controlled experiments below report relMSE and closure, and recorded-gap experiments report MASE. Appendix E.5 reports relative-MAE checks and Appendix A the baseline reproduction checks.

4.2 Benchmark sensitivity to missingness and input convention

Recorded gaps and input conventions. Before comparing repairs, we fix three measurement choices that otherwise change the answer: the input convention, the missingness mechanism, and the scored windows. GIFT-Eval retains NaN entries in 9 of 28 datasets, and electricity/H contains 19.2% missing points (App. C). On matched windows from these nine datasets, we compare native NaN handling, interpolation, zero fill and forward-fill, scoring observed targets. Above 30% context missingness, the best-to-worst median MASE spread relative to the best reaches 94% for Chronos-Bolt and 106% for Chronos-2 (Figure 24). Input convention therefore changes measured accuracy before the checkpoint or scored windows change. Recorded gaps show the same sensitivity (Figure 3, right).

Missing rate does not determine forecast damage. With linear fills and 70% missing, the three-dataset sweep averages 2.17× clean error under value censoring versus 1.07× under scattered dropout (Figure 3, left), with corresponding ranges of 1.34–22.35× and 1.03–3.07× across nine datasets (App. E.1). On ETTh1, this mechanism difference also affects uncertainty: censoring lowers interval coverage from 0.742 to 0.217 while narrowing width from 4.51 to 3.13, so narrow intervals can accompany severe under-coverage (Figure 3, middle). Mechanism-grouped calibration brings coverage within 0.03 of nominal in 16 of 18 censoring cells without changing point forecasts (App. E.8).

Which windows enter the score? Target selection changes a different part of the comparison. Our earlier Penmanshiel protocol required fully observed targets. Removing that filter and scoring observed target points with a fixed fill excludes 26% of windows, whose median normalised MSE (NMSE) is 17.65 against 1.66 for the retained ones, so the filter lowers the reported median by 15.4% (App. E.6). Input convention determines what the model receives, while the filter determines which cases count, and fixing both isolates the interface's response to changed fills.

4.3 Empirical verification of the interface bound

Three input paths, three access patterns. The bound predicts different fill access even at identical forecasting weights. The three Chronos-Bolt paths test this prediction directly. We probe three datasets, three mechanisms and eight series each (72 series, rate 0.3). The plain-path numerical rank

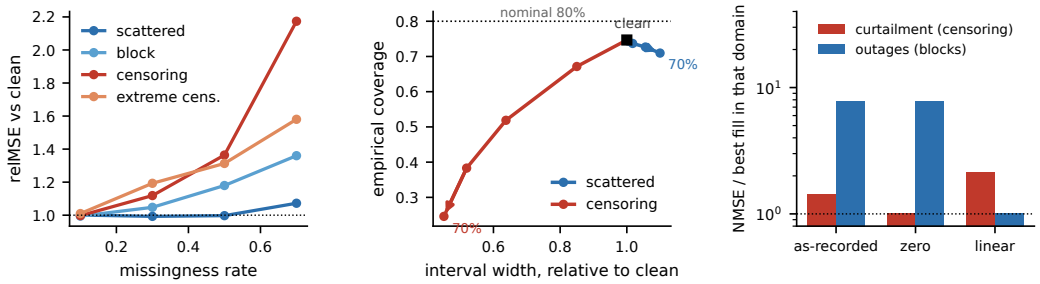

Figure 3: Missingness effects on Chronos-Bolt base under the displayed protocols. Left: relative MSE under four synthetic mechanisms with linear fills, averaged over ETTh1, ETTm1 and weather. Middle: coverage and width of the nominal 80% interval as the missing rate reaches 0.7. Right: NMSE for three fixed fills on Penmanshiel curtailment and METR-LA outages, normalised by each deployment's best tested fill (log scale).

Table 1: **Missingness interfaces in the checkpoint census.** n counts released checkpoints; ranges span checkpoints, not training seeds. ρ is the permutation-to-redraw response ratio over three datasets, three mechanisms and 60 windows at rate 0.3; **–** means both responses vanish (magnitudes: Table 10). relMSE decl. is linear-fill own-clean relative degradation at 70% missing, not absolute accuracy. †Sampled forecasts need self-noise comparison (Appendix D.1). Timer-base lacks declaration support and returns NaN forecasts on raw NaN input.

Family	Declared path	n	Behaviour	ρ plain	ρ declared	relMSE decl.
Chronos-Bolt	mask/NaN	4	discards	0.87–0.97	**0.00**	1.57
Chronos-2	NaN	1	discards	0.90	–	1.52
Chronos-T5	NaN	2	discards	0.91–0.95	–	–
TiRex	NaN	1	discards	0.96	–	1.59
FlowState	NaN	1	discards	0.85	–	1.30
TempoPFN	NaN token	1	discards	0.84	–	–
TimesFM	NaN	1	overwrites	0.86	–	1.55
Moirai 2.0	obs. mask	1	retains	0.97	0.95	1.60
Moirai 1.1/MoE†	obs. mask	5	retains	0.97–0.99	0.94–0.97	–
Timer-base	none	1	no declared path	0.84	N/A	–

reaches $\min(H, |\mathcal{M}|)$ with median $\rho = 0.72$. Explicit masking gives rank two on all 72 series and $\rho = 1.2 \times 10^{-6}$, and replacing candidate fills by NaN makes the map constant (Figure 2). Declaration therefore keeps statistic-level changes but removes the position-specific response available on the plain path, exactly as the bound predicts.

Declared paths across families. Table 1 applies this access classification across released families: their interfaces retain, overwrite or block content through different paths and surviving statistics. The census also records unsupported declarations (N/A) and non-finite outputs. Appendix B.1.2 extends these probes across model sizes.

4.4 Fill sensitivity and covariate use

A quality sweep isolates accessible response. The oracle endpoint removes imputation error while preserving the interface. With Chronos-Bolt, windows and masks fixed, we sweep declared fills from interpolation ($\alpha = 0$) to true values ($\alpha = 1$). The change is 3.8–5.2% of that path's excess over clean under scattered/block missingness, versus 6.7–40% under censoring: the surviving mean and standard-deviation route can transmit useful corrections, especially under censoring, while perfect fills leave much of the degradation unchanged.

Retaining temporal content changes this response. We restore content and adapt the input embedding for 600 steps with the backbone frozen, comparing full-context and observed-only

normalisation. Taking the better variant in each cell gives median oracle closure 53%, positive in 22/23 training-dataset cells. Training on ETTh1, ETTm1 and weather, then freezing the adapters, yields a positive best-variant closure in all 44 held-out cells (median 73%), and in 59/67 cells at $\alpha = 0.25$ (Figure 4; App. H.2). At $\alpha = 1$ the plain path is the clean input while the restored path stays slightly above its own clean error.

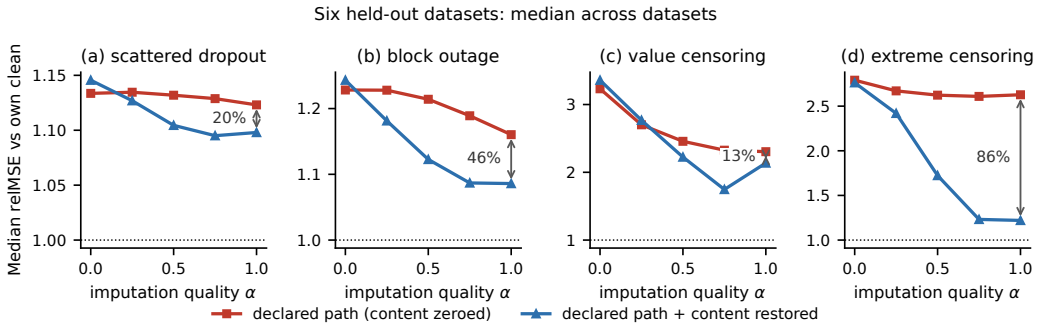

Figure 4: Fill-quality sweeps at missing fraction $r = 0.7$ on six held-out datasets: ETTh2, ETTm2, electricity, traffic, exchange and illness. Points are medians across datasets of within-dataset median own-clean relMSE; panels (a)–(d) identify the mechanisms. The restored curve takes the lower of the two normalisation variants at each α (a pointwise envelope). Annotations give endpoint closure at $\alpha = 1$. Appendix H.2 reports the full grid and clean effects.

Do learned imputers realise that response? SAITS (Du et al., 2023) and BRITS (Cao et al., 2018) supply estimated fills rather than controlled blends. Table 28 compares the two interfaces on 300 windows, three datasets, four mechanisms and rates 0.3/0.7, each imputer with a mechanism-matched *all* variant and an *mb* variant trained on scattered/block masks. SAITS gives positive median closure in both, while BRITS gives negative medians despite sometimes lower imputation RMSE, so lower reconstruction error does not order the benefit of restoration. The oracle row reaches 40.2% closure versus 10.6% for the best tested learned variant, leaving a gap to the oracle in this evaluation. That gap tracks the structure of repair error rather than its magnitude (App. H.4).

Restoring content lowers own-clean relMSE in 13 of 16 ETT cells and leaves the 8 weather cells unchanged (App. H.3). Section 4.5 shows the benefit depends on learned use, not access alone.

Bad fills and covariates use the same path. An accessible path admits a fill's errors along with its content. A search within the observed value range attains median error multipliers of 3.06× (plain fill), 1.18× (declared rank-two access) and 1.00× (NaN) on METR-LA, and raises error on Chronos-2 and Moirai 2.0 as well (App. F.1). Covariates reach the forecast through the same route: under Chronos-2 block outages at rate 0.7, four observed neighbours give filled-target median relMSE 1.357 against 1.023 with declaration, so accuracy gains can reflect the surviving observations rather than the repairs (App. F.2).

4.5 Separating content access from learned use

Because the restored adapters change both access and training, a paired 2×2 factorial within Chronos-Bolt tiny (8.7M) separates their effects by crossing vanilla and restored interfaces with filtered and mechanism-diverse pretraining at matched corpus and budget. Filtered arms share weights because their interfaces coincide on clean inputs (App. G.1). Table 2 pairs closure with fill response.

Training can improve robustness without content access. Vanilla augmentation reaches 88.7% ± 5.3 oracle closure while fill-quality response stays flat: absolute median slopes on scattered/block masks are 0.011 ± 0.003 before augmentation and 0.006 ± 0.005 afterward across three seeds. This gain reflects gap robustness without learned fill response.

Table 2: Missingness robustness and useful fill access are distinct training outcomes. Entries are means ± sample standard deviations over three paired seeds of Chronos-Bolt tiny. Response: median absolute quality-sweep slope $|s|$ (vanilla arms) or permutation ratio ρ on scattered/block masks (restored arms). Closure uses oracle fills on three datasets, retaining cells with same-seed baseline excess > 0.02; floor counts cells no worse than baseline at $\alpha = 0$. Clean: median complete-context MSE ratio over nine datasets; attack: median range-constrained damage. Appendix G.2 reports all nine datasets and three seeds.

Interface / training	Fill response	Closure (%)	Floor (/24)	Clean	Attack
Vanilla / filtered	$\|s\| = 0.011 \pm 0.003$	–	–	1.000 ± 0.000	1.06 ± 0.02
Vanilla / mech.-diverse	$\|s\| = 0.006 \pm 0.005$	88.7 ± 5.3	19.0 ± 1.7	1.106 ± 0.073	1.21 ± 0.15
Restored / filtered	0.97 ± 0.02 / 0.83 ± 0.02	-47.6 ± 24.9	7.7 ± 1.2	1.000 ± 0.000	1.32 ± 0.03
Restored / mech.-diverse	0.97 ± 0.02 / 0.72 ± 0.01	72.7 ± 16.1	18.3 ± 1.2	1.033 ± 0.023	1.18 ± 0.03
Restored / fill-div. (block)	0.98 ± 0.01 / 0.78 ± 0.01	93.8 ± 2.5	20.7 ± 1.2	0.949 ± 0.035	1.28 ± 0.07

Training makes retained content useful. Filtered training exposes the model to clean inputs. Restoring content afterward gives −47.6% ± 24.9 closure. With mechanism-diverse training, that same restored interface reaches 72.7% ± 16.1, permutation response $\rho = 0.852 \pm 0.034$ and clean ratio 1.033 ± 0.023. Unlike robustness alone, this combination of improved accuracy and position-specific response supports learned use of repairs.

Released models show the same distinction: Moirai 2.0's (Liu et al., 2025a) declared-path relMSE falls from 1.37 to 1.18 on ETTh1 at rate 0.7 under scattered dropout when the fill becomes exact, while Chronos-Bolt stays flat. ETTm1 and weather repeat the contrast, and under censoring both models transmit large corrections (Table 30; App. I).

Exposure composition. With interface, corpus, sampler and 15,000 steps fixed, a nine-dataset, three-seed comparison gives 2.7% ± 63.0 closure for declaration-only training, 94.4% ± 5.5 for fill-diverse training and 97.6% ± 2.6 for the full four-mechanism mixture (App. G.2). Repair diversity, not gap exposure alone, drives the gain and motivates the varied fills used in the retrofit.

5 A CONTENT-PRESERVING RETROFIT

The controls in Section 4.5 show that access alone does not make fills useful. The retrofit therefore pairs a content-preserving path with repair-aware training.

Retaining fills with flags. The bound motivates retaining temporal fill content alongside flags that distinguish repairs from measurements. For Chronos-Bolt and Chronos-2, we remove masked-content zeroing, passing fills and flags to the embedding with unchanged normalisation. Moirai 2.0 already retains both, preserving the values and their reconstructed status.

Learning across fill qualities. Content access permits a response but does not make it useful. We therefore vary repair content as well as gap locations during training, blending linear interpolation with the true context as in equation 2. Complete training windows supply the $\alpha = 1$ *oracle reference*. Half the windows receive synthetic gaps, sampled uniformly from the four mechanisms of Section 4.1 with $r \sim \text{Uniform}[0.05, 0.7]$. Half the affected windows use zero fills with flags to teach prediction from holes, while half use blends with uniform α to expose repairs of varying quality. Zero fill differs from interpolation at $\alpha = 0$. Training targets remain observed (App. G.1). Inference supplies fills and flags.

Adapting released checkpoints. We adapt forecasting weights with 5,000 continued-pretraining (CPT) steps on the public evaluation-excluded Chronos subset, then average trained and original parameters with equal weights to limit clean-context changes (WiSE-FT; Wortsman et al., 2022):

$$\theta_{\text{retrofit}} = \tfrac{1}{2}\theta_{\text{CPT}} + \tfrac{1}{2}\theta_{\text{pretrained}}. \tag{9}$$

Appendices G.4–G.6 report the per-family configurations and seeds.

Table 3: **Relative degradation after CPT.** Single-run own-clean relMSE at $r = 0.7$, median over windows, four mechanisms and nine datasets. Clean change is the median clean-MSE change; control is equal-budget CPT without missingness. Oracle is diagnostic; bold marks each family's lowest ratio. Per-dataset results appear in Table 31.

Model	Arm	Clean change	Zero	Linear	Oracle
Chronos-Bolt	Vanilla	–	3.51	1.57	1.51
	Control	+4.3%	9.80	1.64	1.83
	CPT	+4.2%	**2.26**	**1.51**	**1.22**
Chronos-2	Vanilla	–	2.10	1.52	1.54
	Control	+5.2%	20.16	1.50	**1.24**
	CPT	−0.7%	**1.66**	**1.44**	1.29
Moirai 2.0	Vanilla	–	11.56	1.60	1.18
	Control	+2.5%	11.94	1.61	1.14
	CPT	+1.1%	**2.96**	**1.49**	**1.11**

Table 4: **CPT on nine GIFT-Eval subsets by context missingness.** Median MASE uses an in-window naive-1 reference, observed targets, random windows and linear fills. Strata pool series and datasets; additional results appear in Tables 8 and 15.

model	arm	< 1%	1–5%	5–15%	15–30%	> 30%
Chronos-Bolt	vanilla	0.661	0.854	1.367	1.738	1.730
	+CPT (ours)	0.700	0.853	1.275	1.430	1.445
Chronos-2	vanilla	0.629	0.860	1.445	2.020	1.822
	+CPT (ours)	0.627	0.860	1.373	1.694	1.495
Moirai 2.0	vanilla	0.633	0.892	1.829	2.406	1.938
	+CPT (ours)	0.641	0.873	1.745	2.068	1.795
windows n		4,308	2,327	625	208	257

Results and transfer. The retrofit lowers zero- and linear-fill own-clean degradation across three families (Table 3). On the absolute scale, Chronos-Bolt wins 74% of windows under zero fill (62% linear, 61% oracle; App. G.4.1). Against equal-budget CPT without missingness, its zero-fill relMSE is 2.26 versus 9.80; Chronos-2 and Moirai 2.0 show the same control gap (App. G.8).

Retained content becomes responsive. Fill permutation changes retrofit attention and forecasts but leaves vanilla Chronos-Bolt invariant (Figures 26–27; App. I), and Chronos-2 reaches $\rho = 0.880 \pm 0.023$. Moirai 2.0 reaches 54.3% ± 1.3 closure at clean ratio 1.016, with weight interpolation cutting its negative-closure cells from 31% to 9% (App. G.6).

Transfer to recorded gaps. With linear fills on nine GIFT-Eval subsets (Aksu et al., 2024), median MASE falls 16–18% for Chronos-Bolt and Chronos-2 and 7–14% for Moirai 2.0 in the 15–30% and > 30% groups (Table 4), and per-window win rates rise from 0.43 and 0.39 near clean to 0.56–0.71 above 15% missingness. The transfer benefit therefore concentrates in incomplete histories, while pooled over all windows, Chronos-Bolt's median MASE is 5% higher (mean 9.6%).

Interface design determines what CPT can adapt. Native-path CPT leaves TimesFM (Das et al., 2024) and TempoPFN (Moroshan et al., 2025) invariant to declared fills; Timer-base (Liu et al., 2024) lacks a declaration input (App. G.9). Chronos-2's observed-only NaN path wins 26/36 cells against the zero-plus-flag retrofit (App. G.5); performance also degrades beyond the training rate (App. D.8). CPT can learn from fills only along a path that exposes them.

6 Conclusion

Even an exact fill may yield little forecast benefit when its temporal content is discarded or the model has not learned to use it. Our bound and paired controls separate these barriers. Retaining flagged fills and training across repair qualities addresses both: a 5,000-step retrofit reduces fill-induced degradation across three TSFM families and improves high-missingness recorded forecasts. Imputation should therefore be judged by forecast benefit, not reconstruction accuracy alone.

Reproducibility Statement

The census and baseline sweeps use public frozen checkpoints. Adaptation, from-scratch pretraining and CPT corpora, configurations and seeds are specified in the appendices, with stored-anchor and released-inference-path checks in Appendices A, D.4 and G.4. The analysis package contains the missingness mechanisms, evaluation scripts and result summaries behind each reported comparison.

AI Use Statement

Generative AI was used for language polishing and writing assistance. The authors designed and conducted all experiments, verified all claims, and take full responsibility for the final content.

References

Coen Adler, Yuxin Chang, Felix Draxler, Samar Abdi, and Padhraic Smyth. Beyond accuracy: Are time series foundation models well-calibrated? In *ICLR*, 2026. URL https://openreview.net/forum?id=nGBN7UjHcy.

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Gift-eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024. URL https://arxiv.org/abs/2410.10393.

Alexander Alexandrov, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Danielle C. Maddix, Syama Rangapuram, David Salinas, Jasper Schulz, Lorenzo Stella, Ali Caner Türkmen, and Yuyang Wang. Gluonts: Probabilistic and neural time series modeling in python. *Journal of Machine Learning Research*, 21(116):1–6, 2020. URL http://jmlr.org/papers/v21/19-820.html.

Amazon Web Services. Chronos-Bolt: Fast and Accurate Zero-Shot Forecasting. Hugging Face model card, https://huggingface.co/amazon/chronos-bolt-base, 2024. Released 2024-11-25.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024. URL https://arxiv.org/abs/2403.07815.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025. URL https://arxiv.org/abs/2510.15821.

Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. Tirex: Zero-shot forecasting across long and short horizons with enhanced in-context learning. In *NeurIPS*, volume 38, pp. 64179–64230, 2025. doi: 10.52202/085713-1929. URL https://doi.org/10.52202/085713-1929.

Iavor I Bojinov, Natesh S Pillai, and Donald B Rubin. Diagnosing missing always at random in multivariate data. *Biometrika*, 107(1):246–253, 2020. doi: 10.1093/biomet/asz061. URL https://doi.org/10.1093/biomet/asz061.

Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. In *NeurIPS*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/734e6bfcd358e25ac1db0a4241b95651-Paper.pdf.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8 (1):6085, 2018. doi: 10.1038/s41598-018-24271-9. URL https://doi.org/10.1038/s41598-018-24271-9.

Jinguo Cheng, Chunwei Yang, Wanlin Cai, Yuxuan Liang, Qingsong Wen, and Yuankai Wu. Nuwats: a foundation model mending every incomplete time series. *arXiv preprint arXiv:2405.15317*, 2024. URL https://arxiv.org/abs/2405.15317.

Andrea Cini, Ivan Marisca, and Cesare Alippi. Filling the g_ap_s: Multivariate time series imputation by graph neural networks. In *ICLR*, 2022. URL https://openreview.net/forum?id=kOu3-S3wJ7.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *ICML*, volume 235, pp. 10148–10167, 2024. URL https://proceedings.mlr.press/v235/das24c.html.

Wenjie Du, David Côté, and Yan Liu. Saits: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619, 2023. doi: 10.1016/j.eswa.2023.119619. URL https://doi.org/10.1016/j.eswa.2023.119619.

Vijay Ekambaram, Arindam Jati, Pankaj Dayama, Sumanta Mukherjee, Nam H. Nguyen, Wesley M. Gifford, Chandra Reddy, and Jayant Kalagnanam. Tiny time mixers (ttms): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. In *NeurIPS*, volume 37, pp. 74147–74181, 2024. doi: 10.52202/079017-2359. URL https://doi.org/10.52202/079017-2359.

Jiarong Fan, Juhyun Park, Thi Phuong Thuy Vo, and Nicolas Brunel. Weighted conformal prediction provides adaptive and valid mask-conditional coverage for general missing data mechanisms. *arXiv preprint arXiv:2512.14221*, 2025. URL https://arxiv.org/abs/2512.14221.

Kun Feng, Shaocheng Lan, Yuchen Fang, Wenchao He, Sihan Lu, Shuqi Gu, Lintao Ma, Xingyu Lu, and Kan Ren. Kairos: Toward adaptive and parameter-efficient time series foundation models. *arXiv preprint arXiv:2509.25826*, 2025. URL https://arxiv.org/abs/2509.25826.

Vincent Fortuin, Dmitry Baranchuk, Gunnar Raetsch, and Stephan Mandt. Gp-vae: Deep probabilistic time series imputation. In *AISTATS*, volume 108, pp. 1651–1661, 2020. URL https://proceedings.mlr.press/v108/fortuin20a.html.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. In *NeurIPS Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2105.06643.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *ICML*, volume 235, pp. 16115–16152, 2024. URL https://proceedings.mlr.press/v235/goswami24a.html.

Lars Graf, Thomas Ortner, Stanisław Woźniak, and Angeliki Pantazi. Flowstate: Sampling-rate-equivariant time-series forecasting. In *ICML*, 2026. URL https://arxiv.org/abs/2508.05287.

Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006. doi: 10.1016/j.ijforecast.2006.03.001. URL https://doi.org/10.1016/j.ijforecast.2006.03.001.

L. Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1 (1):38–53, 1973. doi: 10.1214/aos/1193342380. URL https://doi.org/10.1214/aos/1193342380.

Marine Le Morvan, Julie Josse, Thomas Moreau, Erwan Scornet, and Gael Varoquaux. Neumiss networks: differentiable programming for supervised learning with missing values. In *NeurIPS*, volume 33, pp. 5980–5990, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/42ae1544956fbe6e09242e6cd752444c-Paper.pdf.

Marine Le Morvan, Julie Josse, Erwan Scornet, and Gael Varoquaux. What's a good imputation to predict with missing values? In *NeurIPS*, volume 34, pp. 11530–11540, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/5fe8fdc79ce292c39c5f209d734b7206-Paper.pdf.

Etienne Le Naour, Tahar Nabil, and Ghislain Agoua. Motm: Towards a foundation model for time series imputation based on continuous modeling. In *Advanced Analytics and Learning on Temporal Data*, pp. 70–86, 2026a. doi: 10.1007/978-3-032-15535-1_5. URL https://doi.org/10.1007/978-3-032-15535-1_5.

Etienne Le Naour, Tahar Nabil, Adrien Petralia, and Ghislain Agoua. Are time-indexed foundation models the future of time series imputation? *Transactions on Machine Learning Research*, 2026b. URL https://openreview.net/forum?id=cTk56KpsP5.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116. URL https://doi.org/10.1080/01621459.2017.1307116.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *ICLR*, 2018. URL https://arxiv.org/abs/1707.01926.

Roderick J. A. Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202, 1988. doi: 10.1080/01621459.1988.10478722. URL https://doi.org/10.1080/01621459.1988.10478722.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. *arXiv preprint arXiv:2511.11698*, 2025a. URL https://arxiv.org/abs/2511.11698.

Xu Liu, Juncheng Liu, Gerald Woo, Taha Aksu, Yuxuan Liang, Roger Zimmermann, Chenghao Liu, Junnan Li, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Moirai-MoE: Empowering time series foundation models with sparse mixture of experts. In *ICML*, volume 267, pp. 38940–38962, 2025b. URL https://proceedings.mlr.press/v267/liu25an.html.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In *ICML*, volume 235, pp. 32369–32399, 2024. URL https://proceedings.mlr.press/v235/liu24cb.html.

Pierre-Alexandre Mattei and Jes Frellsen. MIWAE: Deep generative modelling and imputation of incomplete data sets. In *ICML*, volume 97, pp. 4413–4423, 2019. URL https://proceedings.mlr.press/v97/mattei19a.html.

Geert Molenberghs, Caroline Beunckens, Cristina Sotto, and Michael G. Kenward. Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(2):371–388, 2008. doi: 10.1111/j.1467-9868.2007.00640.x. URL https://doi.org/10.1111/j.1467-9868.2007.00640.x.

Vladyslav Moroshan, Julien Siems, Arber Zela, Timur Carstensen, and Frank Hutter. Tempopfn: Synthetic pre-training of linear rnns for zero-shot time series forecasting. *arXiv preprint arXiv:2510.25502*, 2025. URL https://arxiv.org/abs/2510.25502.

Marine Le Morvan and Gaël Varoquaux. Imputation for Prediction: Beware of Diminishing Returns, 2024. URL https://arxiv.org/abs/2407.19804.

Charlie Plumley. Penmanshiel wind farm data. Zenodo, 2022. URL https://doi.org/10.5281/zenodo.5946808.

Zhongzheng Qiao, Sheng Pan, Anni Wang, Viktoriya Zhukova, Yong Liu, Xudong Jiang, Qingsong Wen, Mingsheng Long, Ming Jin, and Chenghao Liu. It's time: Towards the next generation of time series forecasting benchmarks. In *ICML*, 2026. URL https://arxiv.org/abs/2602.12147.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. In *NeurIPS*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/5103c3584b063c431bd1268e9b5e76fb-Paper.pdf.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. doi: 10.1093/biomet/63.3.581. URL https://doi.org/10.1093/biomet/63.3.581.

Kamilė Stankevičiūtė, Ahmed M. Alaa, and Mihaela van der Schaar. Conformal Time-Series Forecasting. In *NeurIPS*, volume 34, 2021.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. In *NeurIPS*, volume 34, pp. 24804–24816, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/cfe8504bda37b575c70ee1a8276f3486-Paper.pdf.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Chen Wang, Mingsheng Long, and Jianmin Wang. Deep time series models: A comprehensive survey and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2026. doi: 10.1109/tpami.2026.3690845. URL https://doi.org/10.1109/tpami.2026.3690845.

Honglin Wen, Pierre Pinson, Jie Gu, and Zhijian Jin. Marginalize, rather than impute: Probabilistic wind power forecasting with incomplete data. *arXiv preprint arXiv:2403.03631*, 2025. URL https://arxiv.org/abs/2403.03631.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *ICML*, volume 235, pp. 53140–53164, 2024. URL https://proceedings.mlr.press/v235/woo24a.html.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *CVPR*, pp. 7949–7961, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wortsman_Robust_Fine-Tuning_of_Zero-Shot_Models_CVPR_2022_paper.html.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GAIN: Missing data imputation using generative adversarial nets. In *ICML*, volume 80, pp. 5689–5698, 2018. URL https://proceedings.mlr.press/v80/yoon18a.html.

Margaux Zaffran, Aymeric Dieuleveut, Julie Josse, and Yaniv Romano. Conformal prediction with missing values. In *ICML*, volume 202, pp. 40578–40604, 2023. URL https://proceedings.mlr.press/v202/zaffran23a.html.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):11106–11115, 2021. doi: 10.1609/aaai.v35i12.17325. URL https://doi.org/10.1609/aaai.v35i12.17325.

A Numerical reproduction of the reference comparisons

The interface interventions change how a fixed forecaster receives its input. Their reference computation must therefore agree with the original inference path before the input is changed. Table 5 checks this agreement and reproduces stored synthetic and real-data settings, using a 5% tolerance. Synthetic values agree to five decimal places; real-data deviations are $|\Delta| \leq 0.0001\%$. The switchable encoder also matches the vanilla clean-input forward pass to 5.7×10^{-6}. These paired computations provide the reference outputs for the interface interventions; each experiment specifies its scoring windows, metric and aggregation.

Table 5: Reference errors reproduced under the same configurations. The final row compares the interface-switchable encoder with the vanilla clean-input forward pass.

Source	Cell	Metric	Stored	Re-run	$\lvert\Delta\rvert$
Synthetic	ETTh1 clean	MSE	11.199030	11.199030	0.0000%
Synthetic	ETTh1 censoring 0.5, linear	MSE	16.050925	16.050925	0.0000%
Synthetic	ETTh1 scattered 0.3, zero	MSE	12.359439	12.359439	0.0000%
Penmanshiel	censored, as-recorded	NMSE	40.28983	40.28983	0.0000%
Penmanshiel	censored, zero fill	NMSE	28.42657	28.42657	0.0000%
Penmanshiel	censored, linear fill	NMSE	60.16071	60.16070	0.0000%
Penmanshiel	control, clean	NMSE	9.91381	9.91381	0.0000%
METR-LA	missing, as-recorded	NMSE	21.19334	21.19334	0.0000%
METR-LA	missing, linear fill	NMSE	2.74913	2.74913	0.0000%
METR-LA	missing, NaN	NMSE	1.99908	1.99908	0.0001%
METR-LA	control, clean	NMSE	3.19901	3.19901	0.0000%

Re-implementation check. Our interface-switchable encoder under plain on a clean context matches the vanilla forward pass to $\max|\Delta| = 5.7 \times 10^{-6}$.

B Interface access, statistical information and fitted risk

Section 3.2 asks which forecast changes remain available when fills vary. Turning that access question into a prediction-error statement requires specifying both the available information and the predictor. We first prove the interface bound, then separate two statistical questions: whether the observed input distinguishes candidate data-generating laws, and how much predictive information is lost under one known law. The final comparison returns to fitted predictors and the relative errors actually measured.

B.1 Interface access and statistical information

B.1.1 Surviving statistics bound the reachable forecasts

The factorisation in equation 3 reduces every fill intervention to a change in the statistics that survive the interface. Fix the observed values and the observation mask, with $m_t = 1$ for observed positions and $\mathcal{M} = \{t : m_t = 0\}$. Under equation 3, the fill enters the forecast only through $s \in \mathbb{R}^{d_s}$. The chain rule gives, wherever the component derivatives exist,

$$\operatorname{rank}(J_{\mathcal{M}}) \;\le\; \min\left\{H, \operatorname{rank}\left(\frac{\partial s}{\partial \mathbf{x}_{\mathcal{M}}}\right)\right\} \;\le\; \min(H, |\mathcal{M}|, d_s). \tag{10}$$

When both the statistics map and the map from attained statistics to forecasts are locally Lipschitz, the Hausdorff dimension of the global reachable set is also at most $\min(H, |\mathcal{M}|, d_s)$. Apply dimension monotonicity first to the statistics image and then to its forecast image; a countable cover by Lipschitz neighbourhoods gives the global bound. The Hausdorff bound describes the whole image, whereas the derivative rank measures variation at a particular fill. Neither upper bound asserts attainment.

If observed-only statistics and the blocked content path are both independent of the fill, the forecast is constant in the fill. Replacing every candidate fill by NaN before inference likewise defines a constant candidate-fill map, rather than a derivative with respect to a NaN value. Full-context statistics can leave a route of dimension at most d_s. Passing finite fills into the content path removes this restriction, leaving the general derivative bound $\min(H, |\mathcal{M}|)$; the full ranks in Section 3.3 are empirical attainment on the tested Chronos-Bolt windows.

This bound identifies a computational restriction, not an error floor. A constant map makes all fills forecast-equivalent with other inputs fixed; a low-dimensional nonconstant map may still reach an accurate forecast. Separately supplied covariates can affect the forecast through other routes, whose derivatives are outside the fill bound. Prediction error consequently requires an additional specification of information and reference risk.

B.1.2 Persistence across windows, sizes and inference paths

The empirical question is whether the identified restriction persists when the evaluation window or inference procedure changes. The larger gradient-free study covers nine datasets and 31,983 series,

with $\rho \in [0.79, 0.99]$ on the plain path and at most 9.6×10^{-6} on the explicit-mask path (App. E.1). Zero NaN responses indicate invariance, with undefined zero-over-zero ratios. Identical-input calls give no sampling variability in the deterministic families (App. D.1). Longer-horizon probes then test a different issue: whether autoregressive updates preserve the single-block factorisation (App. D.2). Extending the dimension bound requires every update and mask transformation to factor through the same statistics; the rollout is therefore evaluated as a composed prediction map.

The audit classifies declared inputs by the route left open: internal replacement, finite content with an observation flag, observed-only statistics with blocked content, or discarded, unsupported, or failing input. TimesFM overwrites missing positions through internal interpolation after a NaN declaration. Moirai 2.0 retains finite fills and observation flags in its content path. In the audited TiRex, FlowState and Chronos-2 paths, observed-only normalisation and blocked missing content remove the statistical route as well. TempoPFN replaces missing positions with a learned token; both perturbation responses are exactly zero (Table 10), yielding structural nonresponse and undefined ρ. The audited decoder-only interfaces have no declared path. Raw NaN input produces a non-finite forecast in the tested Timer checkpoint, an input failure distinct from finite fill-invariant forecasts on supported declaration paths. This classification keeps fill-transmitting, fill-discarding, unsupported, and failing interfaces as separate cases.

The scale sweep asks whether increasing capacity changes the access path or the fitted relative degradation. Across four Chronos-Bolt sizes spanning 24× in parameters, the explicit-mask probe gives $\rho \le 2.2 \times 10^{-6}$; the tested Moirai sizes retain permutation sensitivity. The access pattern persists across the sampled sizes. On this ladder, relative degradation is flat or mildly worsening, while the spread across fill conventions grows from 18.4% to 32.3% (App. D.3). Increasing size thus leaves the measured access class unchanged while changing the fitted response to input conventions. Accuracy here is measured relative to each checkpoint's clean performance.

B.1.3 When the observed input cannot distinguish two laws

Even an unrestricted observed-input predictor faces uncertainty if the available observation is compatible with different target laws. Let $(\mathbf{x}, y)$ be a context–target pair with $\mathbf{x} \in \mathbb{R}^L$ and a square-integrable scalar target y. The observation is $Z = (m, \{x_t : m_t = 1\})$. A method in the following statistical experiment is a single measurable map $q : Z \mapsto \hat{y}$, chosen without knowing which of two candidate laws generated the data. Its only law-dependent input is the test observation Z. In particular, this experiment supplies no labelled training sample that could distinguish the laws.

For a law P, let $m_P(Z) = \mathbb{E}_P[y \mid Z]$. The excess risk relative to the observed-input Bayes predictor is

$$\mathcal{E}_P(q) \;=\; \mathbb{E}_P[(q(Z) - y)^2] - \inf_{q'} \mathbb{E}_P[(q'(Z) - y)^2] \;=\; \mathbb{E}_P[(q(Z) - m_P(Z))^2]. \tag{11}$$

Proposition 3 (Two-law non-identifiability)**.** Let P_0, P_1 induce the same distribution μ of Z. For every shared measurable predictor q in the statistical experiment above,

$$\max_{j \in \{0,1\}} \mathcal{E}_{P_j}(q) \;\ge\; \tfrac{1}{4}\mathbb{E}_\mu\big[(m_{P_0}(Z) - m_{P_1}(Z))^2\big]. \tag{12}$$

Proof. Both excess risks are integrals against μ. At fixed Z, set $a = q(Z) - m_{P_0}(Z)$ and $b = q(Z) - m_{P_1}(Z)$. Since $(a-b)^2 \le 2(a^2 + b^2)$,

$$\max_j \mathcal{E}_{P_j}(q) \ge \tfrac{1}{2}\big(\mathcal{E}_{P_0}(q) + \mathcal{E}_{P_1}(q)\big) = \tfrac{1}{2}\int (a^2 + b^2)\, d\mu \ge \tfrac{1}{4}\int (m_{P_0} - m_{P_1})^2\, d\mu. \qquad \square$$

The same q faces both laws, so taking its infimum preserves the bound:

$$\inf_q \max_{j \in \{0,1\}} \mathcal{E}_{P_j}(q) \;\ge\; \tfrac{1}{4}\mathbb{E}_\mu\big[(m_{P_0}(Z) - m_{P_1}(Z))^2\big]. \tag{13}$$

The obstruction in this two-point argument (Le Cam, 1973) is the shared predictor's inability to distinguish the laws from Z. A learning procedure also observes its training sample, so the corresponding lower bound would require the laws to remain indistinguishable on that full input. The experiment above supplies only the test observation.

B.1.4 Censoring preserves ambiguity about the hidden tail

We instantiate the two-law condition by changing only the unobserved tail while preserving the distribution of every observed value and mask. Let $y = \alpha x_L + \varepsilon$, with finite second moments for x_L and ε. The noise has a common zero-mean law and is independent of the entire context. Let all other coordinates be observed, independent of x_L, and identically distributed under both candidate laws. Censor x_L when $x_L > \tau$, so $m_L = \mathbf{1}\{x_L \le \tau\}$. Choose two laws for x_L that agree on $(-\infty, \tau]$ and place the same mass p above τ, but have different conditional tail laws G_0, G_1. Both laws give the same observed-value and mask distribution, but different means on the censored event:

$$m_{P_j}(Z) = \alpha \mathbb{E}_{G_j}[x_L], \qquad \Delta\mu = \mathbb{E}_{G_0}[x_L] - \mathbb{E}_{G_1}[x_L].$$

The conditional means agree on the observed event, giving

$$\max_j \mathcal{E}_{P_j}(q) \;\ge\; \tfrac{1}{4} p \alpha^2 (\Delta\mu)^2 \qquad \text{for every shared predictor } q. \tag{14}$$

Here $\Delta\mu$ measures the separation of tail means compatible with the same observed law. If the tail is confined to $(\tau, B]$, pairs can have mean separation approaching $B - \tau$, yielding bounds approaching $p\alpha^2(B-\tau)^2/4$. The construction specifies a family of indistinguishable laws; it does not estimate that separation in the benchmark datasets.

Independent dropout in this construction. Instead delete x_L independently with probability $p < 1$, retaining the same known target relation and common noise law. The observed x_L values identify its full marginal law because each value has the same positive observation probability. Two candidate laws in this restricted family with the same distribution of Z must therefore have the same m_P, so the two-point bound is zero. This removes the tail non-identifiability of this construction. A missing test coordinate can still carry predictive information even when its distribution is known.

B.1.5 Information loss under a fixed law

For a fixed law, compare nested information sets. Let $W = (\mathbf{x}, m)$ retain the complete context and the mask, while Z retains only observed values and the mask. Define $R^{\mathrm{B}}_W = \mathbb{E}[(y - \mathbb{E}[y \mid W])^2]$ and $R^{\mathrm{B}}_Z = \mathbb{E}[(y - \mathbb{E}[y \mid Z])^2]$. Conditional-expectation orthogonality gives

$$R^{\mathrm{B}}_Z - R^{\mathrm{B}}_W = \mathbb{E}\big[(\mathbb{E}[y \mid W] - \mathbb{E}[y \mid Z])^2\big] \ge 0.$$

Indeed, expand $y - \mathbb{E}[y \mid Z]$ as $(y - \mathbb{E}[y \mid W]) + (\mathbb{E}[y \mid W] - \mathbb{E}[y \mid Z])$; the cross term vanishes on conditioning on W. This identity quantifies information lost by hiding context values under one law. When m is deterministic from $\mathbf{x}$, or is conditionally independent of y given $\mathbf{x}$, $\mathbb{E}[y \mid W] = \mathbb{E}[y \mid \mathbf{x}]$ and R^{B}_W equals the complete-context Bayes risk.

Independent dropout can give a strictly positive difference: take x_L uniformly in $\{-1, 1\}$, $y = x_L$, and hide x_L independently with probability p. The complete-input Bayes risk is zero and the observed-input Bayes risk is p. Information loss persists even when the law is known.

These references clarify the role of the fitted direct predictor in Appendix B.2. Its population risk is at least R^{B}_Z, and its held-out error estimates that risk with sampling uncertainty. It measures what a specified training procedure achieves. The two-law bound and fixed-law Bayes-risk difference instead describe what the available information permits; neither is estimated by the predictor's reported clean-relative degradation.

B.2 Comparing fitted predictors and fills

The empirical comparison separates changing a fill from changing the predictor itself. We compare a fixed fill, a learned fill passed through the frozen forecaster, and a direct predictor trained on the observed context and mask. For block outages, the learned-fill score and contrast columns use the explicit-mask path; the other mechanisms use the plain path. For each mechanism and rate, let r_{fixed}, r_{learned}, and r_{direct} denote aggregate relative scores, each using that arm's own clean-context reference. Define the empirical contrasts on this aggregate scale:

$$\begin{aligned} d_{\mathrm{direct}} &= r_{\mathrm{direct}} - 1, \\ d_{\mathrm{learned}} &= r_{\mathrm{learned}} - r_{\mathrm{direct}}, \\ d_{\mathrm{fixed}} &= r_{\mathrm{fixed}} - r_{\mathrm{learned}}. \end{aligned} \tag{15}$$

Table 6: **Fitted-predictor error contrasts.** The comparisons around equation 15 are empirical, not certified optima. The learned column uses the mask-conditioned fill for block outages and the plain fill network otherwise, matching the stored contrasts: (i) direct minus one, (ii) learned minus direct, and (iii) fixed minus learned. Negative values indicate lower relative scores for the first arm in the stated difference. **Bold** marks the largest contrast. Chronos-Bolt base, $L = 512$, $H = 64$.

		measured relMSE			empirical contrasts		
Mechanism	rate	best fixed	learned fill	direct	(i) direct	(ii) learned	(iii) fixed
scattered dropout	0.3	1.013	1.026	1.013	0.013	**0.013**	−0.013
	0.7	1.121	1.187	1.057	0.057	**0.130**	−0.066
block outage	0.3	1.020	1.029	1.036	**0.036**	−0.007	−0.009
	0.7	1.086	1.121	1.094	**0.094**	0.027	−0.035
value censoring	0.3	1.193	1.094	1.102	**0.102**	−0.008	0.099
	0.7	2.875	1.672	1.308	0.308	0.364	**1.203**
extreme censoring	0.3	1.210	1.095	1.101	0.101	−0.006	**0.115**
	0.7	1.827	1.646	1.325	**0.325**	0.320	0.182

These contrasts telescope to $r_{\text{fixed}} - 1$, organising relative degradation into differences between fitted alternatives rather than decomposing Bayes risk. The block-outage learned-fill scores are 1.029 and 1.121 at rates 0.3 and 0.7, respectively; the contrast columns are computed from these scores and the corresponding fixed/direct scores. These contrasts compare relative degradation, not absolute errors.

Under scattered dropout and block outages at rate 0.7, the best fixed fill's reported excess is 0.09–0.12, and the learned fill has a higher reported score. Thus, under this own-clean scale, the tested fixed-fill menu is stronger than the learned fill for these two mechanisms. Routing among those fixed fills also yields no gain in this evaluation. For these masks, the comparison provides no empirical reason to replace the selected fixed fill with the tested learned alternative.

Value censoring at rate 0.7 gives a larger positive fixed contrast. The table reports total excess 1.875 and direct, learned, and fixed contrasts of 0.308, 0.364, and **1.203**. The learned fill also beats the best fixed fill on 81% of windows. Unlike the scattered and block comparisons, this result identifies a useful role for the learned fill under censoring: it lowers relative degradation and wins on most windows. The tested alternatives provide attained scores, not an infimum over all possible fills.

The direct predictor's reported relative degradation is 0.31 under value censoring and 0.33 under extreme censoring, compared with 0.06–0.09 under scattered and block missingness. Censoring remains harder for this fitted predictor even without the frozen forecasting interface.

Transfer to real curtailment. On Penmanshiel curtailment, the learned fill wins on 62% of held-out windows, with a median per-window error ratio corresponding to a 6% reduction relative to the selected fixed fill (paired sign test and per-window ratios in Appendix H.5). This comparison uses the selected real-data fixed fill as its baseline, so its reduction answers a real-data baseline comparison and is distinct from synthetic excess over clean.

C Datasets and models

The data comparisons serve complementary purposes. Synthetic masks allow the same history to be compared before and after a known intervention; recorded gaps test the input conventions encountered in deployment. The checkpoint census then identifies which parts of those inputs each released model retains.

Datasets and splits. Table 7 lists nine standard benchmarks for the synthetic mechanisms (ETTh1/2, ETTm1/2, weather, electricity, traffic, exchange, illness; 7 to 862 channels), two real-missingness deployments (Penmanshiel wind-farm curtailment and METR-LA traffic-sensor outages), and nine GIFT-Eval subsets with recorded gaps. Per-subset missingness appears in Appendix C. The ETT benchmarks originate with Informer (Zhou et al., 2021); these standard forecasting datasets are also documented in the TSLib benchmark collection (Wang et al., 2026). Windows come from the last

20% of each timeline and learned components are fitted on the first 80% under the default synthetic protocol. Real-data learning and calibration use half-splits (Appendices H.5 and E.8.3).

Table 7: **Dataset configurations for the main probe and evaluation grids.** A 'series' is one channel of one window. Channel counts written *a* of *b* denote a fixed seeded subsample. These lengths describe the listed grids, not every experiment: calibration uses a separate split and configuration (Appendix E.8). **Bold** marks Penmanshiel's distinct context and horizon. Table 8 gives missingness for each GIFT-Eval subset.

			source		lengths		evaluation	
Dataset	Domain	Freq.	steps	chan.	*L*	*H*	win.	series
Synthetic mechanisms applied to a clean source								
ETTh1	electr. transformer	1 h	17,420	7	512	64	300	2,100
ETTh2		1 h	17,420	7	512	64	300	2,100
ETTm1		15 min	69,680	7	512	64	300	2,100
ETTm2		15 min	69,680	7	512	64	300	2,100
weather	meteorology	10 min	52,696	21	512	64	300	6,166
electricity	household load	1 h	26,304	24 of 321	512	64	300	6,900
traffic	road occupancy	1 h	17,544	24 of 862	512	64	300	7,200
exchange rate	finance	1 d	7,588	8	512	64	300	2,400
illness	public health	1 w	966	7	512	64	131	917
Real missingness, no mask applied by us								
Penmanshiel	wind-farm curtailm.	10 min	–	–	**144**	**24**	602	–
METR-LA	traffic-sensor out.	5 min	–	–	512	64	618	–
GIFT-Eval (9 subsets)	mixed	5 min–1 d	–	–	512	64	7,725	–

Recorded gaps and finite placeholders. Counting missing observations requires distinguishing explicit markers from values that could also be genuine measurements. Table 8 makes this distinction in the distributed files. Its upper panel describes the nine GIFT-Eval subsets used in Tables 4 and 16: nine of 28 datasets contain NaNs, accounting for 11.77% of 305M points overall.

The audit counts only explicit NaNs as missing observations. The lower panel therefore describes NaN-free classical benchmarks that contain zero-runs and constant plateaus. The UCI electricity original contains 20.15% zeros; 212 of 370 clients first record a nonzero value after 2011-01-01, including 161 after a full year. Such patterns can reflect delayed activation, genuine zero load, or recording problems, and their values alone do not identify missing observations. Two library paths turn non-finite entries into finite values rather than flags: `np.nan_to_num` in TSLib's segmentation loader (Wang et al., 2026) and the configurable `dummy_value` of GluonTS's splitter (Alexandrov et al., 2020). The zero-run counts describe the released values; assigning their origin would require information beyond those values.

Checkpoint coverage and interface persistence. Table 1 groups the permutation probe by family; Table 10 lists all eighteen checkpoints and their sizes. Within each tested family, the declared-path classification is unchanged across sizes. The Timer checkpoint belongs to the original Timer model (Liu et al., 2024); Moirai-MoE uses the sparse-expert architecture (Liu et al., 2025b).

D Measuring access and locating its computational restrictions

The census in Section 4.3 classifies the input route available to a fill. Interpreting that classification requires separating genuine input response from sampling variability and from changes introduced by autoregressive rollout. The controls below establish these distinctions, compare access across model sizes and families, and locate an additional decoder restriction. Finally, independently varying repair and covariate quality tests whether content access supports substitution between sources.

Table 8: **NaNs and value-pattern statistics in distributed benchmark files.** The upper panel measures NaNs in GIFT-Eval. The lower panel reports zero-runs and plateaus in NaN-free classical benchmarks. These patterns can arise from preprocessing or operational behaviour and are not counts of identified missing values. Worst-channel and corpus-wide fractions are labelled.

	% points NaN	% series with NaN	zero-runs (worst channel)	plateaus (worst channel)
GIFT-Eval, missing-containing subsets				
electricity/H	19.2	57.3	–	–
kdd_cup_2018_with_missing/H	17.1	100	–	–
restaurant	13.8	98.8	–	–
bitbrains_rnd/5T	13.1	100	–	–
bitbrains_fast_storage/5T	4.5	94.3	–	–
car_parts_with_missing	4.5	6.2	–	–
temperature_rain_with_missing	2.6	96.5	–	–
hierarchical_sales/D	1.5	100	–	–
jena_weather/H	0.01	100	–	–
whole corpus (28 datasets, 305M points)	11.77	–	–	–
Classical benchmarks, published NaN-*free*				
ETTh1	–	–	1.2%	3.0%
ETTh2	–	–	28.5%	37.3%
ETTm1	–	–	1.2%	3.2%
ETTm2	–	–	29.1%	62.0%
weather	–	–	95.5%	93.8%
electricity (TSLib, 2016–19)	–	–	75.7%	76.3%
electricity (UCI original)	–	–	20.15% of all points are zeros	
traffic	–	–	13.6%	8.7%
exchange	–	–	0.0%	11.3%
illness	–	–	0.0%	0.0%

Table 9: **What fills a gap, and what reaches the forecaster?** TSFM rows describe the tested declared paths (Appendix C), except Timer's plain path. Keep means retaining position-specific external fills while flagging gaps. Bound means a fill-reachable forecast bound; Train means forecaster training across external fill quality, not missing-rate augmentation. ✓/× indicate presence/absence; – marks a property inapplicable to imputers. Sources, in row order: (Ansari et al., 2024); (Amazon Web Services, 2024); (Ansari et al., 2025; Auer et al., 2025; Graf et al., 2026); (Moroshan et al., 2025); (Das et al., 2024); (Woo et al., 2024; Liu et al., 2025a;b); (Liu et al., 2024); (Cao et al., 2018; Du et al., 2023).

Model / path	Value at a gap	Processing	Keep	Bound	Train
Chronos-T5	Padding token	Token replacement and attention mask	×	×	×
Chronos-Bolt (finite fill + mask)	External fill, then zero	Fills affect scaling, then are zeroed before embedding	×	×	×
Chronos-2 / TiRex / FlowState	Zero	Missing content zeroed; indicators retained	×	×	×
TempoPFN	Learned token	Token replacement; observed-only scaling	×	×	×
TimesFM 2.5	Internal linear interpolation	NaNs replaced upstream; module mask is padding-only	×	×	×
Moirai 1.1 / 2.0 / MoE	External fill	Values and observation mask retained	✓	×	×
Timer-base (plain)	External fill	Values passed without a missingness flag	×	×	×
BRITS / SAITS	Learned estimates	Recurrent / attention-based reconstruction	–	×	×
Retrofit (ours)	Zero / linear at test	Fills and flags retained; blend/zero training	✓	✓	✓

Table 10: **Input-path responses for eighteen checkpoints.** Protocol: three datasets, three mechanisms, 60 windows, missing rate 0.3, $L = 512$, $H = 64$. Plain and declared columns report ρ and the absolute permutation/redraw response magnitudes, each aggregated over the same cells. For Chronos-Bolt, declared statistics describe the finite-fill mask path. A nonzero redraw response with negligible permutation response is consistent with that path's statistics-only dependence. Both responses vanish for content-free deterministic endpoints; their ratio is undefined and is shown as –. Timer-base has no declaration endpoint; its non-finite raw-NaN result is reported separately from ρ. †For sampled checkpoints, compare ρ with the measured self-response baseline (Appendix D.1).

			plain			declared		
Checkpoint	Params	Declared path	ρ	perm.	redraw	ρ	perm.	redraw
discards the fill content once a hole is declared								
Chronos-Bolt tiny	8.7M	mask/NaN	0.87	0.167	0.236	**0.00**	0.000	0.038
Chronos-Bolt mini	21M	mask/NaN	0.96	0.204	0.229	**0.00**	0.000	0.042
Chronos-Bolt small	48M	mask/NaN	0.97	0.199	0.224	**0.00**	0.000	0.052
Chronos-Bolt base	205M	mask/NaN	0.91	0.212	0.235	**0.00**	0.000	0.087
Chronos-2	120M	NaN	0.90	0.188	0.242	–	0.000	0.000
Chronos-T5 small	46M	NaN	0.95	0.358	0.413	–	0.000	0.000
Chronos-T5 base	200M	NaN	0.91	0.368	0.391	–	0.000	0.000
TiRex 1.1	35M	NaN	0.96	0.214	0.232	–	0.000	0.000
FlowState r1	9.1M	NaN	0.85	0.168	0.217	–	0.000	0.000
TempoPFN	38M	NaN token	0.84	0.208	0.249	–	0.000	0.000
overwrites the fill with its own interpolation								
TimesFM 2.5	200M	NaN	0.86	0.207	0.235	–	0.000	0.000
retains the fill content after a declaration								
Moirai 2.0 small	11M	obs. mask	0.97	0.251	0.266	0.95	0.263	0.326
Moirai 1.1 small†	14M	obs. mask	0.98	0.450	0.463	0.97	0.448	0.463
Moirai 1.1 base†	91M	obs. mask	0.97	0.380	0.429	0.97	0.384	0.427
Moirai 1.1 large†	311M	obs. mask	0.97	0.388	0.440	0.94	0.386	0.423
Moirai-MoE small†	117M	obs. mask	0.97	0.405	0.464	0.97	0.401	0.466
Moirai-MoE base†	935M	obs. mask	0.99	0.428	0.445	0.97	0.420	0.438
ships no way to declare a hole at all								
Timer-base	84M	none	0.84	0.259	0.326	N/A	–	–

D.1 Separating fill response from forecast sampling

The probe of equation 6 compares sensitivity to fill permutation with sensitivity to fill redraw. For deterministic forecasts and a nonzero denominator, a small ratio indicates little response to the tested permutations; a permutation-invariant statistic can still depend on the fill. When both perturbations leave the forecast unchanged, the ratio is undefined. For sampled forecasts, repeated-call variability enters both terms, motivating an identical-input noise control.

An identical-input noise control. Two calls on the *same* context measure sampling variability:

$$\rho_{\text{self}} = \frac{\mathbb{E}\,\|f(\mathbf{x};\xi_1) - f(\mathbf{x};\xi_2)\|}{\mathbb{E}\,\|f(\text{redrawn fill}) - f(\mathbf{x})\|}, \tag{16}$$

Here ξ_1 and ξ_2 denote independent sampling draws. The control uses the same datasets, mechanisms, windows, masks and seeds as the permutation probe. Table 11 compares its empirical level with ρ. Comparing the paired means asks whether the permutation response exceeds the variability already present without changing the input. This is an empirical noise control, not a significance test.

Sensitivity above the sampled control. The deterministic families have zero repeated-call differences. For the five sampled Moirai checkpoints, declared ρ is 0.95–0.99, above the corresponding self-control levels of 0.47–0.61. The permutation response exceeds repeated-call variability in each sampled checkpoint. This makes the input-sensitivity comparison interpretable

Table 11: Permutation sensitivity and repeated-call variability. Recorded zeros remain as stored; zero redraw response makes the ratio undefined, requiring interpretation of absolute responses.

Checkpoint	sampling	path	ρ	ρ_{self}
Chronos-Bolt base	deterministic	plain	0.907	0.0000
		mask	0.000	0.0000
		NaN	0.000	0.0000
Chronos-2	deterministic	plain	0.898	0.0000
		NaN	0.000	0.0000
Moirai 1.1 small	20 samples	plain	0.969	0.564
		declared	0.979	0.573
Moirai 1.1 base	20 samples	plain	0.962	0.461
		declared	0.976	0.498
Moirai 1.1 large	20 samples	plain	0.959	0.454
		declared	0.949	0.466
Moirai-MoE small	20 samples	plain	0.972	0.575
		declared	0.959	0.597
Moirai-MoE base	20 samples	plain	0.974	0.689
		declared	0.991	0.605

across deterministic and sampled forecasts. Differences between Moirai generations still include their distinct architectures and training recipes (Section 4.5).

Distinguishing statistical response from discard. A zero permutation response has two different explanations: the model may depend on permutation-invariant statistics, or it may discard the fill entirely. Redrawing fills separates these cases, and inspecting the input operation identifies the surviving route. When redraw also produces no change, ρ is undefined. An unsupported declaration is marked as N/A.

D.2 Autoregressive rollout changes the composed response

Section 3.3 measures the probe at $H = 64$, one native Chronos-Bolt block. We rerun it unchanged at $H \in \{24, 128, 192\}$: the first is a slice of the native block; the last two are autoregressive median rollouts through the same head. Because the vanilla pipeline does not accept an explicit observation mask, the mask path rolls forward manually, extending the mask with ones over appended forecasts. Each ratio uses the same rollout procedure for its numerator and denominator. We evaluate three datasets and three mechanisms, with 60 windows each at nominal rate 0.3.

Within the native block, the declared path's ρ is at most 4.6×10^{-6} at both $H = 24$ and $H = 64$ on every dataset and mechanism. Under rollout the same probe grows with the number of blocks (up to 0.32 at $H = 128$ and 0.46 at $H = 192$ on scattered dropout), while the plain path stays at $\rho \approx 0.9$–1.1. The NaN path has zero absolute response at every horizon; its recorded zero ratio represents an undefined zero-over-zero case. The native block preserves near-invariance to fill arrangement, but the autoregressive procedure develops a measurable response. These results identify the unit to which the access claim applies: one prediction block and a repeated rollout are different maps. Extending the block-level dimension bound requires every autoregressive and mask update to factor through the same statistics. The measured rollout response therefore needs a separate access analysis.

D.3 Increasing scale preserves the declared access class

We compare four released Chronos-Bolt sizes within the same architectural family to test whether capacity changes the declared access path or the fitted relative degradation. The explicit-mask response remains at $\rho \leq 2.2 \times 10^{-6}$ across the 24× parameter span (App. B.1.2). Figure 5 shows that larger checkpoints do not consistently reduce degradation relative to their own clean forecasts. In this ladder, increased capacity retains the same input restriction.

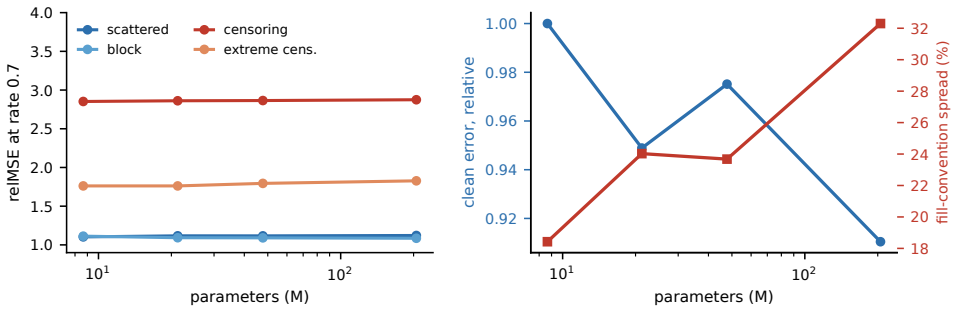

Figure 5: Relative degradation is flat or worsening across a 24× parameter span.

D.4 What the comparison between Moirai generations holds fixed

D.4.1 Moirai 2.0: preserving the module's value-and-flag input

The comparison needs external fills to reach the released module without being replaced upstream. The sweep in Table 30 therefore uses a wrapper around Moirai 2.0. We compare its clean-input computation with the library's inference path, including the GluonTS transform, instance splitter and batching before the same forward call. On identical clean windows, the forecasts agree exactly, max $|\Delta| = 0.0$. Moirai 2.0 emits quantiles deterministically, so no sampling noise enters this comparison (unlike Moirai 1.1, whose forecasts are sampled). The clean agreement validates the forward wrapper on complete windows. The wrapper's *declared* arm is a separate input-contract check: it passes the fill and observation mask together, which is the released module's own convention (`input_tokens = cat([scaled_target, observed_mask])`), and bypasses the inference pipeline's default causal-mean imputation. Thus the clean comparison checks the forward call, while the declared experiment deliberately varies the content delivered at missing positions instead of reproducing the pipeline's default fill.

D.4.2 Moirai 1.1: sampling and patch-size controls

Moirai 1.1 introduces two additional choices: sampled output and inference-time patch size. We assess both before interpreting its quality sweep against Moirai 2.0.

Sampling-relative agreement. Moirai 1.1's sampled forecasts require a repeated-library reference rather than bit-exact output equality. We compare our direct call to the library's `create_predictor`/GluonTS path *and* to a second run of the library path against itself, on identical windows: ours-vs-library gives correlation 0.975 and maximum relative deviation 2.14, and library-vs-library across two seeds gives 0.973 and 1.77. In these sampled paired windows, the wrapper agrees with the library about as well as the library agrees with itself on these windows. The comparison establishes finite-sample consistency relative to the library's own prediction variability.

Patch size and relative-degradation closure. Our wrapper runs 1.1 at the fixed `patch_size` = 32 used in the comparison. The library also offers 16, 64 and an `auto` setting that selects per series; `auto` raises a shape error under our pinned version and is not part of the fixed-size comparison. The sweep slope in α varies across fixed patch sizes, partly because the starting degradation differs. At `patch_size` = 16, three of Moirai 1.1's four cells are already at or below the clean-context error at $\alpha = 0$, leaving no positive excess to close. For cells with positive excess, we report the share closed relative to that configuration's own clean error:

cell (rate 0.7)	patch 16	patch 32	patch 64
ETTh1 scattered	–	96.2%	105.1%
ETTh1 block	92.9%	98.4%	99.4%
ETTm1 scattered	–	97.5%	99.1%
ETTm1 block	–	97.1%	95.8%

("–" marks a cell whose $\alpha = 0$ error is already at or below clean, so the excess is not positive and the ratio is undefined.) Closure lies in [92.9%, 105.1%] in all nine positive-excess cells, across a 4× range of patch size. Initial degradation varies with patch size: under 70% missingness, ETTh1 block outages give 1.09 at patch 16 and 1.81 at patch 32. All nine eligible cells improve with better fills.

Shared access, different recovery. Under the same protocol, Moirai 1.1 recovers to near its own clean error at $\alpha = 1$ (dataset-averaged 1.348 → 1.002 under block outages at rate 0.7 and 1.092 → 0.999 under scattered dropout; on ETTh1 alone 1.804 → 1.002 and 1.167 → 1.000), whereas Moirai 2.0 recovers to 1.03–1.18. Both generations therefore expose a fill-responsive endpoint, but access alone does not determine how fully the endpoint recovers. The contrast is consistent with their different pretraining recipes: 1.1 pairs a masked encoder with variable-length context and prediction windows (Woo et al., 2024), whereas 2.0 uses a single scattered patch-mask mechanism with a built-in fill (Liu et al., 2025a). Because architecture and implementation also change, the generation contrast establishes that access alone does not specify recovery. The within-family factorial in Section 4.5 then isolates interface and training changes within one model.

D.5 Patch size changes degradation without a token-count explanation

Varying Moirai 1.1's inference-time patch size keeps its weights fixed. Smaller patches generally give lower relative degradation but higher clean-context error in the measured summaries. The two readouts describe different effects of this choice. Absolute corrupted error would require paired errors on a shared scale, because multiplying aggregate clean MSE by aggregate relMSE does not recover it.

A contiguous gap of length G intersects approximately $G/P + 1$ of the L/P tokens, suggesting that patch size could act through the corrupted-token fraction. The tested linear regressions provide little support for this explanation. Regressing relMSE on the corrupted-token fraction gives $R^2 = 0.08$, on partially-corrupted tokens 0.05, on fully missing tokens 0.00, and on the absolute count of intact tokens 0.06; the raw missingness rate gives 0.15. Patch size therefore changes the observed degradation, but these token-count summaries explain little of that change in the tested linear models.

D.6 TempoPFN separates missingness exposure from repair access

TempoPFN uses online NaN-pattern and censoring augmentation, with rates and run lengths sampled from GIFT-Eval's NaN statistics. Its interface replaces values at NaN positions with a learned token (`channel_embeddings[nan_mask] = nan_embedding`) and computes scaler statistics over observed points only. This path cannot accept a user fill: permutation and redraw both change the forecast by exactly 0.0, making their ratio undefined. The α sweep is flat in all 48 cells.

The declared path is at or below its own clean error in several scattered-dropout and block-outage cells, reaching 0.927 on ETTm1 at rate 0.3, and has lower relMSE than plain fill in 11 of 24 cells at $\alpha = 0$. These scores show prediction from incomplete inputs despite complete fill discard. The released checkpoint combines its architecture and augmentation recipe; the within-family controls test the training contribution separately.

Under value censoring at rate 0.7, interval coverage falls to 0.258–0.381. Unlike Chronos-Bolt, TempoPFN does not narrow its intervals as coverage falls. Coverage loss can accompany different width responses; calibration and fill access are distinct.

For interpreting the TempoPFN probe, two implementation facts matter. TempoPFN runs in bf16 because its kernel rejects fp32, and repeated identical calls give max $|\Delta| = 0.0$. The weather plain-path probe is therefore not a sampling-noise effect: a scaler floor makes a few near-constant series inflate the redraw mean, while the median remains in the family range.

D.7 From input corruption to decoder access

The input bound explains what enters the encoder, but an encoder can also reconstruct missing values from observed positions. The following interventions ask where corruption changes the computation and whether internally reconstructed values influence the decoder.

D.7.1 Attention summaries and input interventions

Coarse attention summaries can remain stable while representations change. Attention entropy measures diffuseness, and register-token mass measures the share assigned to the dedicated summary token. Under 70% missingness, both overlap across clean and three missingness conditions (Figure 6(a,b)), even as hidden states drift by mechanism-dependent amounts (panel c). To locate the affected computation, we therefore intervene on input channels and attention connections rather than infer it from these aggregate summaries.

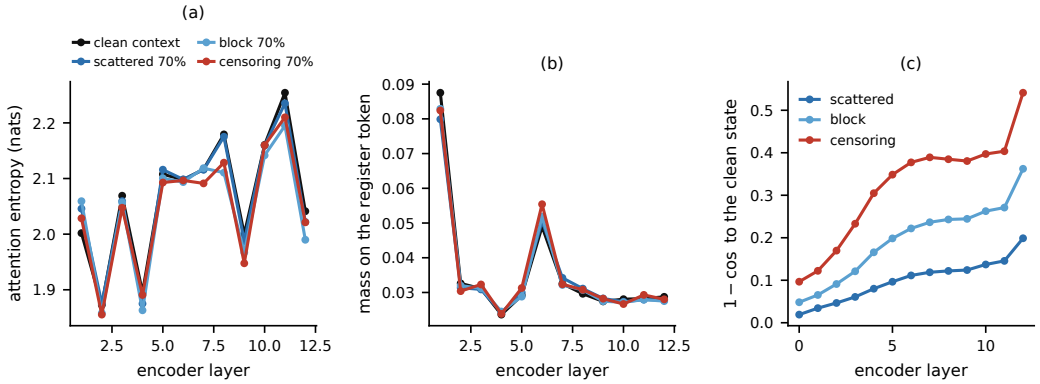

Figure 6: Attention statistics under 70% missingness versus a clean context, per layer, averaged over three datasets. The four conditions overlap in (a) and (b).

A separate input intervention crosses a corrupted-versus-oracle normalisation scale with corrupted-versus-oracle patch content to locate the damage of a zero fill in a statistics channel and a value channel (Table 12). The two interventions are crossed rather than assumed additive, so the table also reports their interaction. At rate 0.5, the statistics channel accounts for 71.5% of the damage and the value channel for 17.9%; at rate 0.7, the value channel accounts for 74.0% and the interaction is −28.8%. The statistics intervention acts on the two normalisation quantities identified in Corollary 2. The dominant contribution changes with the missing rate, and the interaction quantifies why the two channel effects cannot be assumed additive.

Table 12: Share of zero-fill damage by channel (Chronos-Bolt, dataset-averaged). **Bold** marks the larger channel contribution in each row: normalisation statistics at rate 0.5, patch values at 0.7. The share recovered by rescaling falls from 82.1% to 26.0%.

rate	statistics channel	value channel	interaction	recovered by rescaling
0.5	**71.5%**	17.9%	10.6%	82.1%
0.7	54.8%	**74.0%**	−28.8%	26.0%

D.7.2 Reconstructed hidden states can remain unused by the decoder

A linear probe predicts true values from hidden states, with R^2 measuring explained target variance. It recovers fully-missing patch values at mid-depth (Figure 7(a)), with R^2 rising from 0 at the embedding layer. Under censoring, near-zero R^2 indicates little linear reconstruction.

Attention knockout blocks selected query–key connections. Blocking all patch keys for fully-missing queries reduces reconstruction to $R^2 = 0$ under full knockout (Figure 7(b)). Knocking layers 0–3 removes the early rise, after which blocks 4–5 rebuild most of the signal. Knocking layers 4–7 leaves no subsequent rebound, while knocking layers 8–11 leaves earlier R^2 unchanged (Figure 8). The segment controls locate where the linearly decodable signal is built: early removal can be repaired by later layers, whereas removing the middle segment prevents that rebound.

Despite these changes in reconstruction, all 24 knockout cells leave the forecast float64-identical to the no-knockout value (Figure 7(c)). The decoder uses the patch mask to ex-

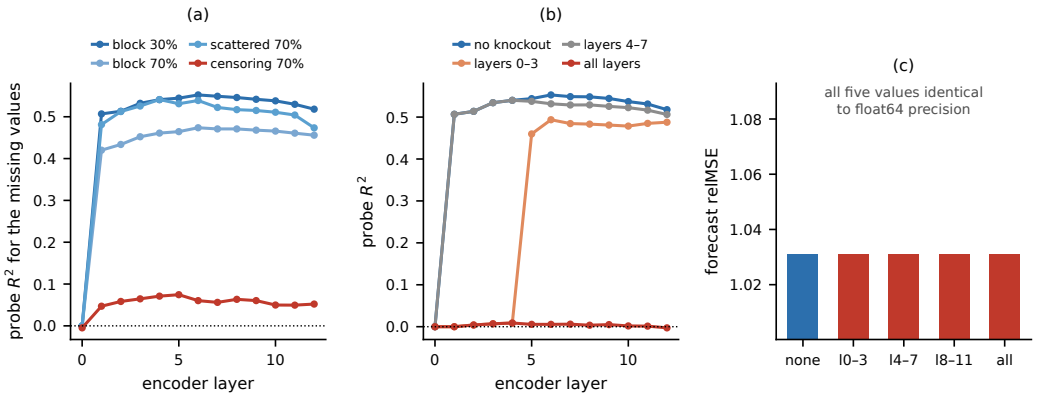

Figure 7: Linear probes on hidden states (a), the same probes under attention knockout (b), and the forecast error under every knockout (c).

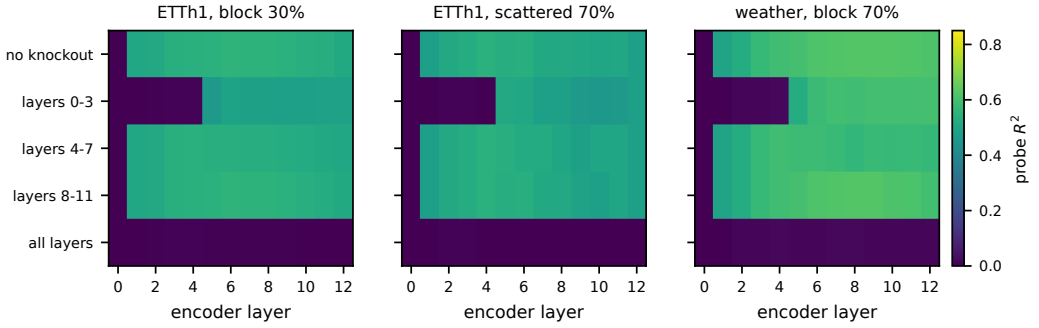

Figure 8: Probe R^2 across encoder layers (columns) under each knockout segment (rows).

clude fully-missing positions from cross-attention. This suggests a second restriction: a value may be represented inside the encoder yet be unavailable at its direct decoder connection. The paired open-mask intervention below tests that connection.

D.7.3 Reconstruction quality alone does not predict forecast improvement

Splitting the probe targets by type (Figure 9(a)) shows different levels of linear decodability. Observed values inside a heavily-holed window are recovered almost perfectly ($R^2 \approx 0.85$–0.95), fully-missing patches under scattered or block missingness are recovered at $R^2 \approx 0.55$–0.67, and fully-missing patches under value censoring have the lowest reconstruction score (0.12). These scores measure linear accessibility rather than all information in the hidden states.

Adding an auxiliary reconstruction objective changes probe R^2 from 0.478 to 0.487 and forecast relMSE from 0.800 to 0.794 (Figure 9(b)). Improving the auxiliary reconstruction score therefore produces little additional forecast change in this configuration. The decoder-access intervention tests whether making those representations directly available changes their influence.

D.7.4 Opening the decoder mask exposes the knockout effect

To test whether the decoder's cross-attention mask blocks the reconstructed signal, we compare the same encoder knockout with the mask closed and open. With the mask left native, knocking out all twelve encoder blocks changes the forecast by $\max|\Delta| = 0.0$ exactly, replicating Figure 7(c) under a fine-tuned model. With the same checkpoint and the same knockout, but the decoder allowed to attend to fully-missing positions, the intervention changes the forecast by $\max|\Delta| = 7.89$. Thus, decoder access determines whether this knockout affects the forecast.

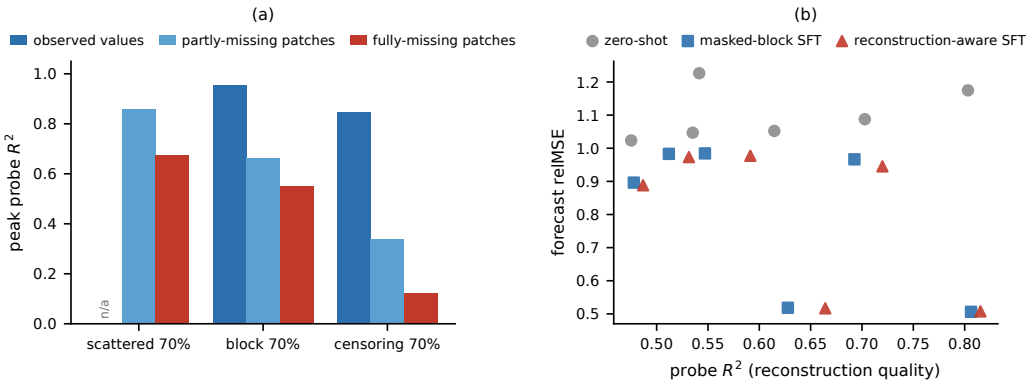

Figure 9: Linear reconstruction by target type (a); reconstruction and forecasting after fine-tuning (b).

Opening the mask yields a paired relMSE gain relative to the same checkpoint with the native mask on heavily censored windows: +5.7% on ETTh1 (paired $t = -15.5$, winning on 98% of windows), +6.9% on ETTm1, and a flat result on weather. The affected cells remain roughly 4× degraded relative to clean-context relMSE. The intervention establishes a useful decoder contribution on ETTh1 and ETTm1 while leaving substantial forecast degradation. Encoder reconstruction and decoder access are distinct steps toward using missing-position information.

D.7.5 Input-content and decoder-access restrictions are distinct

Two separate gates act on fully-missing patches. At the encoder input, the declared path zeroes fill content before embedding, leaving the normalisation-statistics route; at the decoder, the patch mask excludes their hidden states from cross-attention. The encoder knockout result and the paired open-mask intervention establish the decoder gate, while restoring input content enables the fill response measured in Section 4.4. Restoring an external fill and exposing an internally reconstructed state therefore change different operations. The paired interventions separate their effects on forecasting.

D.8 Does repair access support substitution between sources?

Section 5 establishes a response to declared fills. If that response also supports *source selection*, reliable covariates should help most when the target repair is poor, and reliable repairs should reduce dependence on corrupted covariates. The repair/context (R/C) probe tests these paired effects on the covariate-capable Chronos-2 interface.

On ETTh1/ETTm1/weather, the probe varies two sources independently: the target repair q_r (zero / linear / oracle fill) and the context quality q_c (clean / corrupted). We hole a target channel with scattered or block outages and supply its four most correlated channels as covariates, with correlation computed on the first 80% of the timeline only. At nominal rates 0.3 and 0.7, the vanilla checkpoint receives silent fills, whereas the +5k fair-budget control and CPT retrofit use the restored endpoint, declaring both target fills and corrupted covariates. We use 150 windows per cell; errors are relMSE against each model's own clean univariate forecast, paired per window, with cell medians pooled over strata (Table 13). These ratios compare within-model degradation, not shared absolute risk. The extension tests nominal rates 0.9 and 1.0: nominal block rate 0.9 masks about 57% of positions because blocks overlap, while the full-outage arm explicitly masks every position.

Table 13: The R/C probe at nominal rates 0.3 and 0.7. Entries pool cell-median relMSE, using each model's own clean univariate forecast as denominator.

model	zero, clean ctx	linear, clean ctx	oracle, clean ctx	zero, corrupt ctx
vanilla Chronos-2	1.913	1.084	0.986	1.984
+5k control	1.629	1.044	1.029	1.580
CPT retrofit (ours)	1.063	1.021	1.009	1.058

For vanilla Chronos-2, pooled relMSE changes from 1.913 with zero fill to 0.986 with oracle-reference fill. Under block outages at nominal rate 0.7, corrupting covariates costs 0.67 relMSE points with zero fill and 0.02 with oracle-reference fill. Covariate quality matters more when the target repair is poor in this condition.

The CPT retrofit lowers zero-fill degradation relative to its own clean level, but its covariate effects remain near zero: pooled corrupt-minus-clean differences are −0.005, +0.005 and +0.007 for zero, linear and oracle-reference fills, respectively, and −0.016 for zero fill at nominal rate 0.9. The paired manipulation therefore separates improved univariate robustness from beneficial covariate substitution, for which this checkpoint provides little evidence. Its univariate training did not directly train that capability. The conditional-Gaussian task in Appendix G.3 tests source selection under a separate training setup.

Complete outage changes the input regime. When every position is declared missing, zero and linear fills coincide and produce an input-independent near-zero forecast (relMSE 8.35). Oracle-reference content with the same all-missing flags yields median relMSE $\sim 10^9$ on $O(10)$ data. The control remains finite through the same restored endpoint, and vanilla remains finite through its native path. The paired comparison therefore localises the failure to the CPT weights interacting with the restored full-outage input.

Mechanism of the blow-up. At this corner the declaration mask removes all 32 context patches from the attention key set, so every model's forecast reduces to a per-window constant $y = \sinh(v)\,\sigma + \mu$: the normalized readout v is a single weight-dependent vector (input-independent to 2×10^{-8} across windows, identical for oracle-reference and zero fills, for the restored and native endpoints, and with or without covariates), and the data enter only through the instance-norm statistics (μ, σ). Vanilla and the control emit $|v| \approx 0.4$, a benign predict-the-mean constant after un-scaling, while the retrofit emits $|v| \approx 12.2$, which the arcsinh inverse of the instance norm exponentiates into $|\hat{y}| \approx 10^5\sigma$; a zero fill instead hits the 10^{-5} scale floor, turning the same v into the complementary near-zero degenerate forecast. The locus is the CPT weights' final-layer-norm and output-head readout direction: encoder activations never leave the vanilla envelope (max ratio 0.96), attention entropy is statistically identical to vanilla's on the same input, and the instance-norm statistics are ordinary, ruling out an attention collapse or a normalization degeneracy. WiSE-FT cuts the v excess only $\sim$ 20% (12.24 → 9.85), which the exponential un-scaling still converts into a $\sim$ 80× relMSE reduction. This diagnosis motivates testing full-outage declared windows in the CPT corpus, a regime none of the shipped or retrofitted recipes currently trains. Complete outage is a distinct all-missing configuration beyond the training maximum of 0.7, rather than an estimated failure threshold for intermediate rates.

E How missingness and evaluation choices change measured performance

Section 4.2 shows that an incomplete benchmark does not define a unique forecasting comparison. This appendix separates the effects of the missingness mechanism, the supplied fill and the windows included in the score. It then asks whether the same patterns persist across datasets and metrics, and whether mechanism-aware calibration can repair the observed interval-coverage failures.

E.1 Missing rate alone does not specify forecast damage

Deleting the same fraction of a history can remove different predictive information. We first compare point errors under matched rates, then examine whether prediction intervals reveal the resulting damage. The two deployment comparisons test the practical implication: a fill that works for one pattern need not work for another. Appendix C specifies the datasets.

E.1.1 Mechanisms differ at the same missing rate

Figure 3, left, sweeps four mechanisms across four rates with a linear fill. Under scattered dropout, error stays near clean through 50% missing and reaches 1.26× clean at 70%. Value censoring at 70% reaches 2.52×. On ETTh1 at that rate, error on future values above the context's 90th percentile is

Table 14: **Interface probes and missingness damage across nine datasets.** Chronos-Bolt base, $L = 512$, $H = 64$, 300 windows. The probe columns compare plain inputs with the finite-fill mask path at rate 0.3. The latter has $\rho \leq 10^{-5}$, displayed as 0.00; the separate NaN path has zero permutation and redraw responses, so its ratio is undefined. Damage cells report paired per-window relMSE/relMAE under linear fills at rate 0.7. **Bold** marks the worst mechanism in each row: value censoring on eight datasets and extreme censoring on exchange.

		ρ (rate 0.3)		relMSE / relMAE at rate 0.7			
dataset	series	plain	declared	scattered	block	censoring	extreme
ETT (electricity transformer)							
ETTh1	2,100	0.92	0.00	1.28 / 1.18	1.31 / 1.20	**3.29 / 2.05**	1.90 / 1.48
ETTh2	2,100	0.79	0.00	1.12 / 1.08	1.16 / 1.10	**2.88 / 1.80**	1.49 / 1.26
ETTm1	2,100	0.92	0.00	1.12 / 1.08	1.37 / 1.21	**3.93 / 2.20**	2.38 / 1.64
ETTm2	2,100	0.93	0.00	1.10 / 1.06	1.19 / 1.10	**2.69 / 1.74**	1.56 / 1.28
other sources							
weather	6,166	0.99	0.00	1.04 / 1.02	1.00 / 1.00	**1.42 / 1.19**	1.21 / 1.12
electricity	6,900	0.92	0.00	3.07 / 1.85	3.54 / 2.01	**22.35 / 5.03**	8.10 / 3.02
traffic	7,200	0.92	0.00	2.51 / 1.90	3.15 / 2.31	**11.20 / 4.29**	5.84 / 3.17
exchange	2,400	0.89	0.00	1.04 / 1.03	0.97 / 0.98	1.34 / 1.19	**2.86 / 1.84**
illness	917	0.95	0.00	1.03 / 1.03	1.69 / 1.38	**4.34 / 2.22**	2.40 / 1.64

2.37× the error on the rest, compared with 0.70× under scattered dropout. Thus the two mechanisms differ both in overall error and in its concentration on extreme targets.

The same comparison extends beyond the three datasets swept in detail. Table 14 repeats the measurement on nine standard benchmark datasets spanning 7 to 862 channels and four domains. Scattered dropout at 70% missing costs 1.03–3.07× the clean error; value censoring at the same rate costs 1.34–22.35×, and the ordering scattered < block < extreme < censoring holds for eight of the nine evaluated datasets.

On exchange, extreme censoring is the worst cell (2.86 against 1.34 for ordinary censoring), reversing the ordering seen on the other eight datasets. Censoring therefore produces the dominant high-error pattern across this grid, with a different ordering on exchange.

The main pattern persists under absolute error. The second number in each cell of Table 14 uses relMAE on the same forecasts and windows. Magnitudes shrink most under censoring, but censoring remains worst on eight of nine datasets, with exchange the exception. The mechanism ranking is therefore not created by squared-error weighting, although the magnitude of the relative excess remains metric-dependent. The separate closure check is reported in Appendix E.5.

E.1.2 Coverage can fall while intervals narrow

Predictive interval width can fail to reveal deteriorating coverage. Figure 3, middle, reports empirical coverage and width of Chronos-Bolt's nominal 80% interval. Under scattered dropout, ETTh1 coverage changes from 0.742 to 0.702 as the interval widens. Under value censoring, coverage falls to 0.217 while width decreases from 4.51 to 3.13.

The missed coverage is concentrated on high targets: top-decile coverage is 0.000 at rates above 0.3 in this panel. The same censored cells have narrower intervals, so width does not signal this tail-specific failure. Under censoring, coverage should be inspected separately for high and ordinary targets.

E.1.3 Coverage loss can also occur with wider intervals

The paired windows show why under-coverage and narrowing should be measured separately. Figure 10 extends the example of Figure 23 to three datasets using vanilla Chronos-Bolt, the declared path, linear fill, rate 0.7, L=512, H=64, and channel 0. We select the first window in protocol order whose censored 80% band under-covers and is narrower than clean, while scattered-dropout coverage is at least 0.9 with a wider band. The search starts at K=6 seeded windows and doubles K when none qualifies. ETTh1 uses the same rule without the narrowing condition.

Weather and electricity illustrate joint narrowing and under-coverage: censored widths are 0.72× and 0.21× clean, with coverage 0.59 and 0.27; scattered-dropout coverage is 1.00 and 0.91. ETTh1 instead has coverage 0.33 with width 1.86× clean. All 24 scanned ETTh1 windows widen, averaging 1.34× clean with mean coverage 0.40. Coverage loss therefore need not accompany narrowing.

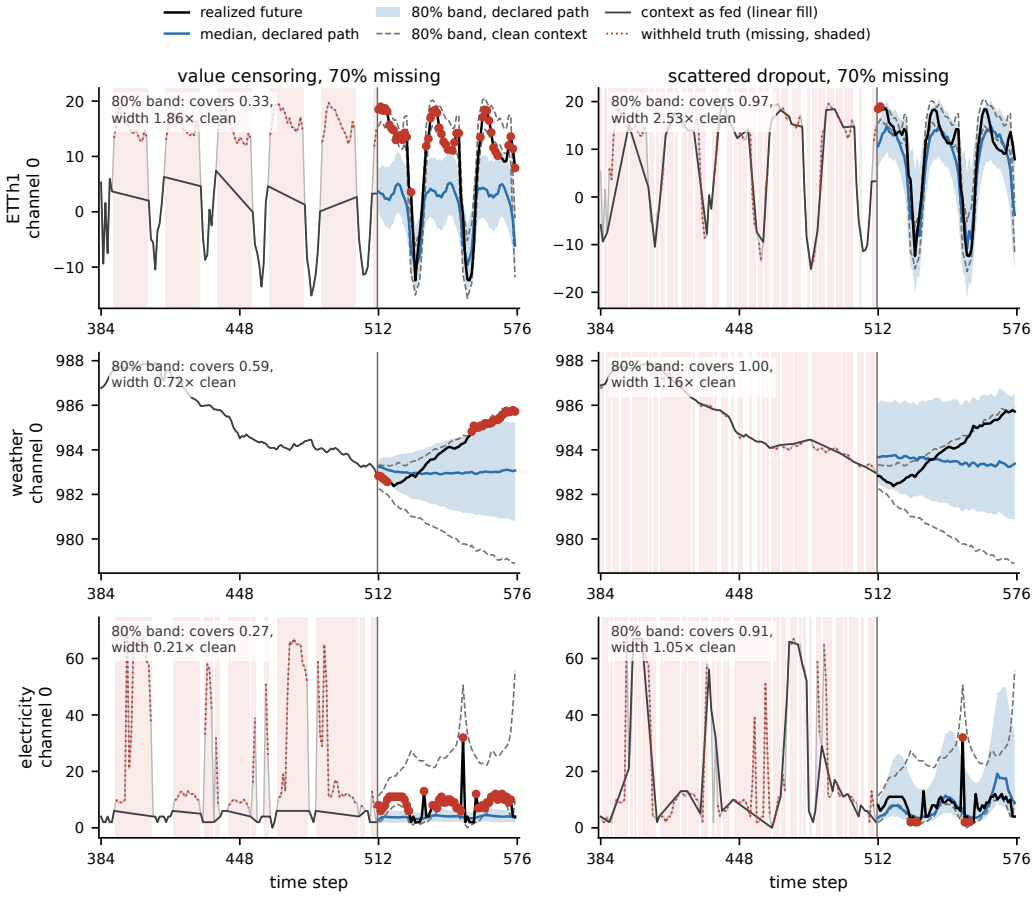

Figure 10: Case-study grid behind Figure 23, same protocol (vanilla Chronos-Bolt, declared path, linear fill, rate 0.7, L=512, H=64, channel 0, last 128 of the 512 context steps shown). One row per dataset at the window chosen by the mechanical selection rule (K=12 for ETTh1 and weather, K=24 for electricity; ETTh1 without the narrowing condition, which no scanned window of its channel 0 satisfies). Red dots mark future observations outside the declared 80% interval.

E.1.4 The best fill reverses between two real deployments

Figure 3, right, compares fixed fills on two real deployments. On Penmanshiel wind-farm curtailment, zero-fill has NMSE 28.43, compared with 60.16 for linear interpolation. On METR-LA sensor outages, the ordering reverses: zero-fill has NMSE 21.19, compared with 2.75 for linear interpolation.

Zero-fill beats linear interpolation by about a factor of two on Penmanshiel and loses by about a factor of eight on METR-LA. The useful fill choice therefore depends on deployment. Explaining this reversal requires more than ranking reconstruction quality: a model may receive the fill's temporal content, only its normalisation statistics, or neither. The fixed-input probes in Section 3.3 identify which route is available before the quality sweep tests its forecasting effect.

E.1.5 The mask-and-scale representation carries predictive signal

A Penmanshiel predictor given the binary missingness pattern together with two scale scalars reaches median NMSE 2.13, compared with 5.29 for a context-mean baseline. It also beats a founda-

tion model with access to the observed sequence on 43% of windows. Both comparisons use the joint mask-and-scale representation. The result motivates keeping the observation flag available alongside reconstructed values, while leaving the flag's individual contribution unresolved. In the interface tests, fixing that flag while varying fills isolates the response to the reconstruction without also changing the declared missingness pattern.

E.1.6 Convention sensitivity persists under a second window protocol

Figure 24, right, compares the random-window audit with a last-window-per-series protocol. Under the latter, convention spreads are 22%, 35%, and 10% for Chronos-Bolt, Chronos-2, and Moirai 2.0 in the highest stratum, compared with 0.2–0.7% in the lowest. Larger high-stratum spreads therefore occur under both tested window protocols.

E.2 The fill ranking changes with the missingness mechanism

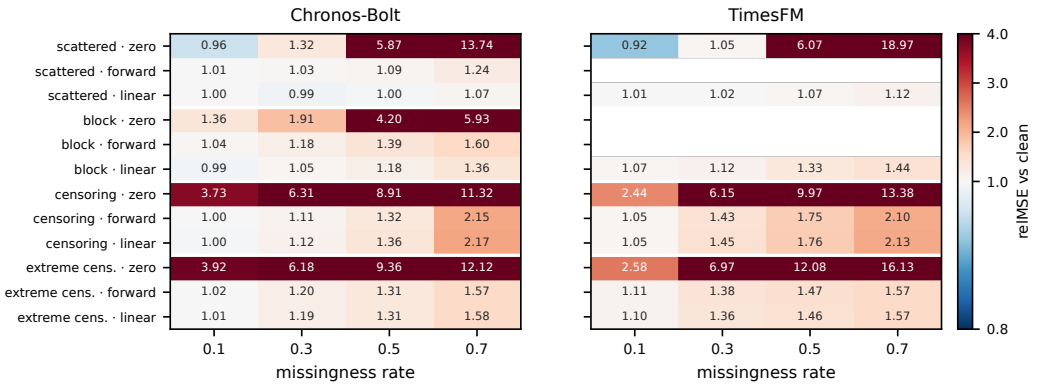

Figure 11: relMSE against a clean context, dataset-averaged. Blank cells were not run.

Figure 11 expands the synthetic grid of Figure 3 for two families. The grid shows an interaction between fill convention and missingness mechanism, not a single fill ranking: zero-fill reaches 13.7× and 19.0× clean error at 70% scattered missingness, whereas forward-fill and linear interpolation stay closer to clean under scattered dropout and separate more under block outages. Even these fills leave errors above twice clean in high-rate censoring cells. The grid establishes the consequence of choosing a convention under each mechanism. The paired permutation measurement below then distinguishes this error response from access to position-specific fill content.

E.3 Forecast damage varies while the access restriction persists

Table 14 summarises the nine-dataset grid in Figure 12(a). Panel (a) measures relative forecast damage, whereas panel (b) measures whether rearranging filled values changes the output. This pairing separates forecast consequence from content access: plain-fill ratios range from 0.79 to 0.99, while the declared-path output response to permutation is zero at the reported precision on all nine datasets. On the audited statistics-only path, permutation preserves the multiset and therefore the mean and variance delivered to the model. The zero response is thus a property of that statistical route. A displayed ratio of 0.0000 represents zero permutation sensitivity when redraw produces a nonzero response; zero-over-zero is undefined. Across these datasets, the same restriction coexists with different forecast damage. This motivates checking whether access measurements predict the score changes caused by switching conventions.

E.4 Within-path access does not determine between-convention score spread

Permutation sensitivity changes fills within one input path. Convention spread can also change the path itself and its treatment of the observation flag. Figure 13 tests the relation between these two measurements across eleven checkpoints, using synthetic masks for the probe and recorded gaps for the score spread (Sections 3.3 and 4.2).

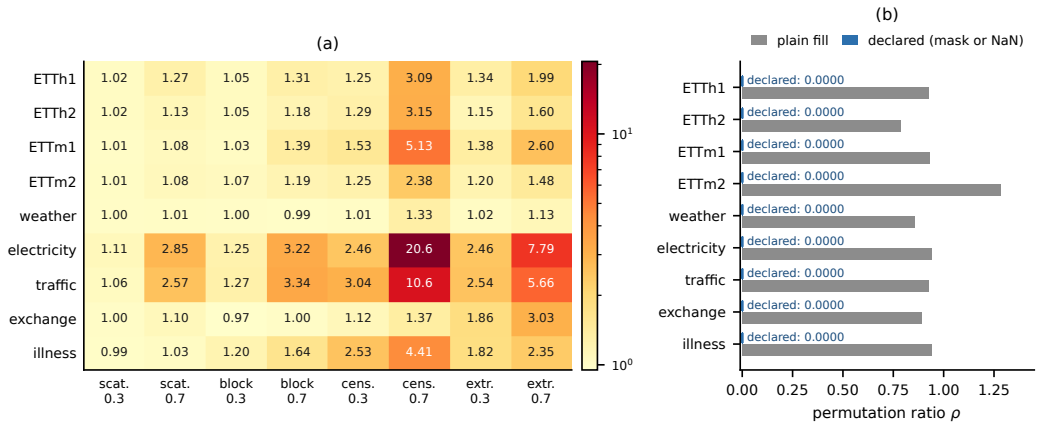

Figure 12: The damage grid (a) and the permutation probe (b) on nine datasets.

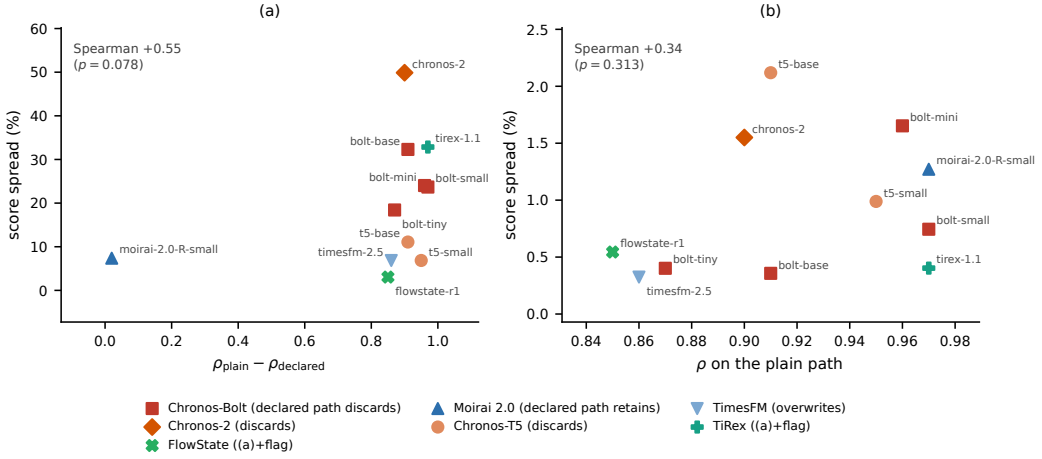

Figure 13: Score spread across fill conventions against the two probe quantities.

We fix the dataset with the most missing context among the sampled windows (12.15% of context points; 17.1% of its raw series), so every checkpoint is compared on the same panel. The two percentages refer to sampled context points and raw-series points, respectively; for some checkpoints, the largest spread instead occurs on a nearly complete dataset with just 63 windows. The Timer checkpoint is excluded before computing the association because raw NaN yields a non-finite forecast and undefined spread. On the fixed dataset, the difference between path probes has Spearman correlation +0.55 ($p = 0.078$) and Pearson correlation +0.33 ($p = 0.32$) with convention spread across eleven checkpoints, including TiRex and FlowState. The point estimates suggest an association on this common panel, with substantial uncertainty and measure-dependent strengths.

Changing the probe definition while keeping the eleven-checkpoint set gives a plain-path Spearman correlation of +0.08. Changing instead the sample pool, from the fixed dataset to the full corpus, also weakens the association: five of the seven datasets in this checkpoint comparison have under 1.5% missing context. The two results vary different axes, so the relationship depends on both the measured path difference and the amount of missing-context support in the evaluation windows.

FlowState has the smallest spread in this comparison, 3.0%, with observed-point statistics and a trained flag channel. Chronos-2 and TiRex have spreads of 49.9% and 32.8% despite rank-zero declared paths. Rank describes fill sensitivity within a fixed path; changing the convention can also change the path and its treatment of the flag. Rank within one path cannot predict the score spread across different input conventions.

As a within-family size control, we compare the Chronos-Bolt ladder on this same dataset. The spreads are 18.4%, 24.0%, 23.7%, and 32.3% for 8.7M, 21M, 48M, and 205M checkpoints. The largest checkpoint has the largest spread, with a small reversal between the middle two sizes. Size is therefore associated with, but does not determine, convention spread in this panel.

The coarse path label also leaves substantial variation. The three largest spreads belong to discarders (Chronos-2 49.9%, TiRex 32.8%, and Chronos-Bolt base 32.3%), whereas the measured retainer has spread 7.4%. Yet FlowState is a low-spread discarder on this same panel. The access class identifies which intervention a path can transmit; evaluating the trained response remains necessary to determine its convention spread. This eleven-checkpoint correlation sample is smaller than the full census, which includes six retaining checkpoints across three Moirai generations (Table 10).

E.5 Absolute error preserves the main pattern but changes its magnitude

Squared error gives greater weight to extreme residuals, which are concentrated under censoring. Recomputing damage and closure under relMAE on the same forecasts and windows tests whether that weighting creates the mechanism ranking. Each arm uses its own clean MAE.

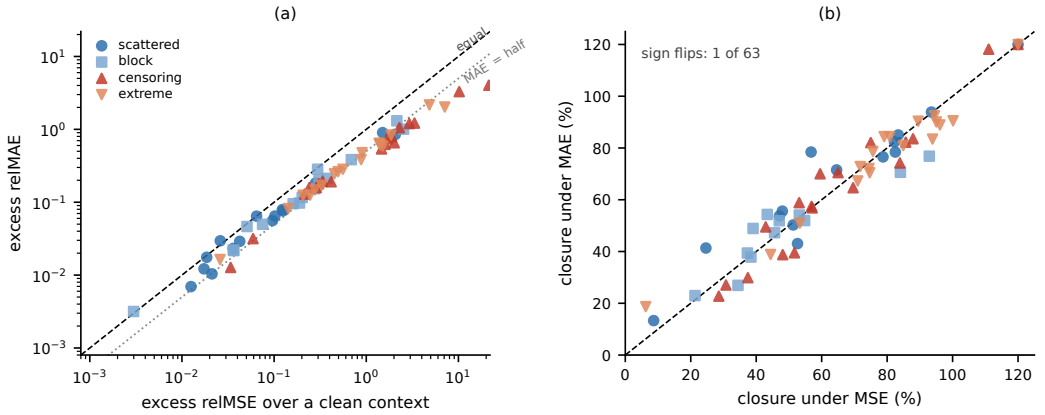

Figure 14: Paired measurements under both metrics; diagonal points have equal values.

The magnitudes shrink, by a factor that depends on the mechanism. Figure 14(a) plots every damage cell of Table 14 under both metrics. On the relative-error scale, all 72 cells lie between the equal line and the half line. On the separate excess-over-clean scale, the median ratio of relMAE excess to relMSE excess is 0.42 under value censoring and 0.47 under extreme censoring, against 0.57 under block outages and 0.60 under scattered dropout. The larger reduction under censoring is consistent with error concentrated on extremes; at rate 0.7, its median excess falls from 2.29 to 1.05.

Closure is similar in aggregate, with one sign reversal. The mechanism ranking is unchanged under relMAE: censoring remains the worst cell on the same eight of nine datasets, and exchange remains the one exception. A separate analysis in Figure 14(b) tests relative-degradation closure under MAE and MSE, using each arm's own-clean denominator on cells with sufficient declared-path excess. The in-sample median moves from 53.1% to 57.2% and the held-out median moves from 73.2% to 71.5%. The sign changes in exactly one of 63 cells: weather under block outages at rate 0.7, where closure was +0.3% under MSE and is −13.0% under MAE. The aggregate closure result is therefore similar under the two metrics, while a cell near zero under MSE changes sign. As in the MSE scorecard, positive MAE closure compares own-clean relative degradation rather than absolute error on a common scale.

The probe is similar under two output norms. Replacing the RMS output-change norm with an L_1 norm changes the measured ρ by at most 0.015 on the nine datasets. Both declared paths retain zero permutation response, reported as 0.0000 where the ratio is defined. This zero response follows from the implemented permutation invariance; the similarity of nonzero ratios is an empirical result.

E.6 TARGET-COMPLETENESS SELECTION ON PENMANSHIEL

A target-completeness filter changes the population of windows behind an accuracy score. The real-data protocol with fully observed targets is compared here with scoring all available target points on Penmanshiel.

On Penmanshiel we rebuild the window set without the completeness requirement and score observed target points. The filter removes 26% of windows. Their median NMSE is 17.65, compared with 1.66 for retained windows, a 10.6× difference. Applying the filter lowers the reported median by 15.4%, with a larger gap in mean error. Both estimates score observed targets only.

The filter preferentially removes difficult windows in this comparison. This is separate from input handling: the GIFT-Eval audit keeps its own observed-target window set fixed while changing the convention supplied to the model.

E.7 RELEASED-MODEL ACCURACY ON RECORDED-GAP STRATA

Table 15 compares models on the GIFT-Eval audit protocol: median MASE against an in-window naive-1 baseline, observed target points, and random windows. The standard rows use linear fill. Overall pooled median MASE is 5% lower for TiRex than vanilla Chronos-Bolt and 7% higher for FlowState. Model errors differ even on nearly complete windows, so the table reflects baseline forecasting accuracy as well as performance on missing inputs.

Stratifying these windows locates performance differences within the real corpus. TiRex's NaN row ranges from 0.621 in the lowest missing-fraction stratum to 0.928 in the highest, with larger values in some intermediate strata. These are comparisons between different window groups, so their ordering combines group composition and model response. The fixed-input probes of Section 3.3 separately isolate fill-content sensitivity without changing the evaluated window.

Table 15: The field on GIFT-Eval's missing-containing subsets, Table 4's protocol. Median MASE per missing-fraction stratum; linear fill except for TiRex's separate declared (NaN) row. The window sets differ across missingness strata.

model	overall	< 1%	1–5%	5–15%	15–30%	> 30%
Chronos-Bolt base (vanilla)	0.782	0.661	0.854	1.367	1.738	1.730
Chronos-2 (vanilla)	0.761	0.629	0.860	1.445	2.020	1.822
Moirai 2.0 (vanilla)	0.781	0.633	0.892	1.829	2.406	1.938
FlowState	0.839	0.703	0.908	1.924	2.285	1.924
TiRex	0.742	0.621	0.827	1.137	1.453	1.468
TiRex, NaN path	0.732	0.621	0.826	1.037	1.018	0.928
TimesFM 2.5	0.793	0.655	0.877	1.867	2.175	1.843

E.8 MECHANISM DETECTION AND GROUP-WISE CALIBRATION

The interval results show that pooling mechanisms can hide concentrated under-coverage. This suggests calibrating errors within groups that distinguish those conditions. Because the generating mechanism is not given at deployment, a detector assigns groups from the mask and observed context. Synthetic masks test the resulting coverage against known mechanisms; recorded gaps test transfer of the frozen grouping rule.

E.8.1 CONSTRUCTING GROUPS FROM OBSERVABLE FEATURES

The detector combines mask-geometry features, including missing rate and run length, with observed-context summaries and a chi-square statistic. The summaries include observed-zero fraction, a scale-fallback indicator, and mask-value correlation. All features are computed from the mask and observed context, so the detector does not use the hidden missing values. The chi-square feature alone performs near chance on the two censoring classes in Figure 15(a), whereas the multi-feature classifier improves separation. Value censoring and long blocks remain hardest to distinguish.

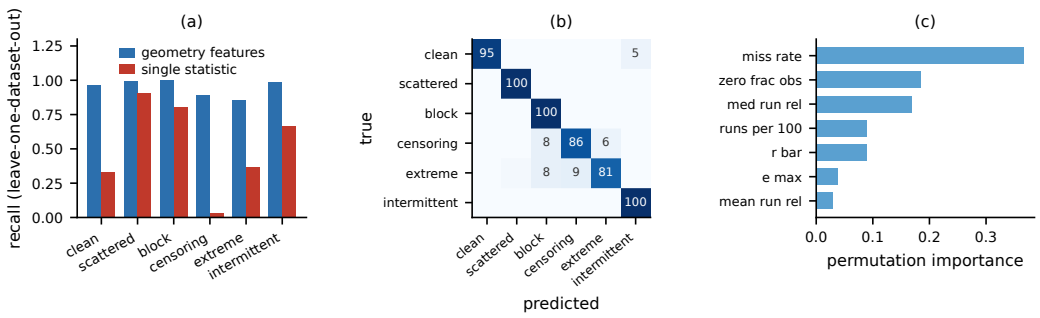

Figure 15: Window-level mechanism classification, leave-one-dataset-out.

E.8.2 Group-wise calibration improves censoring coverage

Split conformal calibration adjusts intervals using quantiles of held-out errors (Lei et al., 2018; Romano et al., 2019). A pooled adjustment uses all calibration windows; Mondrian calibration computes an adjustment within each predicted mechanism group. Missing-pattern coverage can differ from marginal coverage (Zaffran et al., 2023). Pooled split conformal reaches the 0.90 target at the dataset level while leaving synthetic censoring cells at 0.808 mean coverage and 0.732 at worst. With the detector fitted on three datasets and frozen for evaluation on nine, group-wise calibration raises mean coverage across eighteen censoring cells from 0.442 native and 0.808 pooled to 0.901. Sixteen of eighteen cells are within 0.03 of nominal. Held-out and in-sample dataset summaries are 0.901 ± 0.014 and 0.901 ± 0.018, respectively; coverage under predicted grouping differs by at most 0.029 from true-mechanism grouping. In this same synthetic evaluation, median interval-width multipliers are $1.4\times$ for clean or scattered groups and $3.7\times$ for censored groups.

On real curtailment, coverage rises from 0.710 to 0.870; on METR-LA outages it rises from 0.710 to 0.910. Figure 16(a,b) separates these synthetic and real evaluations, so the synthetic width multipliers above are not width estimates for the two deployments.

Transfer to GIFT-Eval is partial. On its nine missing-containing datasets, adding mechanism and rate groups raises coverage in the highest missing-fraction stratum from 0.829 under pooled calibration to 0.855, still below the nominal 0.90. These interval adjustments leave the point forecast unchanged. Appendix E.8.3 details the stratum results.

E.8.3 Mechanism and rate groups on recorded GIFT-Eval gaps

The corpus is the nine missing-containing GIFT-Eval datasets, read with the protocol of Section 4.2. Contexts retain the recorded gaps, no synthetic masks are applied, and only observed target points are scored. Windows are randomly split into a calibration half and a test half. The detector is the same frozen classifier, fitted on ETTh1, ETTm1 and weather; it is applied to GIFT-Eval without refitting.

Nearly complete windows establish the native reference. Windows with under 1% missing context have native coverage 0.822 over 1569 windows, below the nominal 0.90. Thus the native model already under-covers on nearly complete inputs in this corpus. We retain 0.90 as the calibration target and use 0.822 to describe coverage before the missingness strata are compared.

Pooled calibration leaves a high-missingness shortfall. Native coverage is 0.089 lower in the highest than the lowest missing-fraction stratum (Figure 16c). A single conformal pool reaches 0.899 overall but 0.829 in the highest stratum. Overall coverage thus conceals a group-specific shortfall. The intermediate strata are non-monotonic and contain different windows.

Rate groups refine the mechanism groups. Mechanism grouping raises coverage in the highest stratum from 0.829 to 0.840. This stratum contains 83 test windows, and its coverage point estimate remains below nominal. Detected classes span missing fractions from 2% to 60%, so a single mechanism group combines materially different rates; this motivates adding rate groups. Across

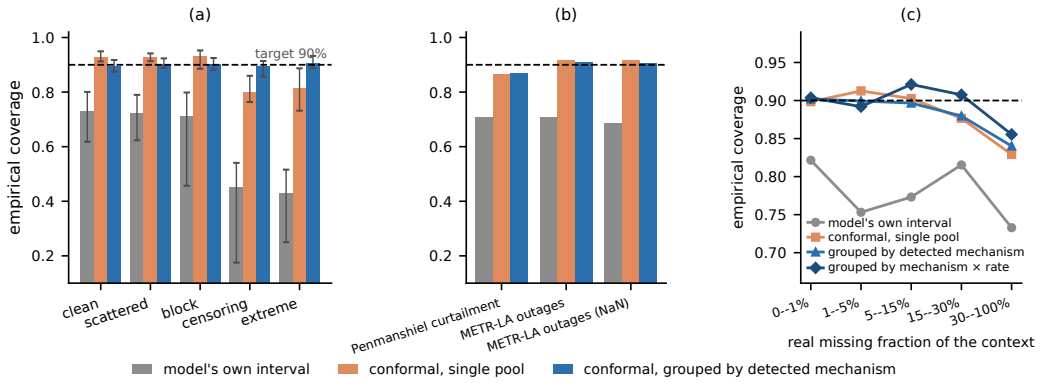

Figure 16: **Coverage under mechanism-grouped calibration.** The detector is fitted on three datasets and frozen, then evaluated on nine datasets with synthetic masks (a) and two real deployments (b); bars in (a) span the nine datasets. Panel (c) groups GIFT-Eval windows by their real missing fraction. $L = 512$, $H = 96$, nominal 0.90 interval.

Table 16: Coverage on recorded GIFT-Eval gaps, grouped by each window's missing fraction. **Bold** marks the value closest to nominal 0.90 in each row. Grouped calibration comes within ±0.03 of nominal in four strata; the highest remains below target.

missing fraction	n	native	single pool	Mondrian	Mondrian × rate
0–1%	1569	0.822	0.898	**0.902**	0.904
1–5%	459	0.753	0.913	**0.899**	0.892
5–15%	125	0.773	0.903	**0.897**	0.921
15–30%	48	0.815	0.877	0.880	**0.907**
> 30%	83	0.733	0.829	0.840	**0.855**

detected classes, coverage ranges from 0.868 to 0.916. Adding rate groups brings the 15–30% stratum to 0.907 and the highest to 0.855, still 0.045 below target.

The frozen detector assigns clean 2184 times, scattered 796, extreme 614, censoring 559, intermittent 421, and block 82. These are predicted classes under transfer from synthetic masks: more windows resemble the synthetic censoring patterns than the block patterns. The predicted labels supply calibration groups; they are not observations of the real generating mechanisms.

Coverage repair and interval width remain distinct. Intervals widen across the real missing-fraction strata in GIFT-Eval. Together with the ETTh1 case in Appendix E.1.3, this separates coverage shortfalls from narrowing: low coverage can accompany wider intervals as well as narrower ones.

F Harmful fills and auxiliary-channel corrections

The fill-quality sweep asks whether accessible content can help. The paired perturbation experiments here test the other direction: whether a poor fill can damage the forecast through that same route. Auxiliary-channel experiments then ask whether a different information source helps when the target history is incomplete. These are distinct interventions: covariates enter outside the bound on fill access (Section 3.2).

F.1 An accessible fill channel can transmit harmful content

We keep observations fixed and perturb only the fill, with every chosen value constrained to the window's observed range. The search therefore tests harmful content without using values outside the observed amplitude range. Its untargeted objective maximises error; its targeted objective drives

the forecast mean to half its clean value. Finite search reports attained damage, which lower-bounds the optimum over the constrained fill set.

On METR-LA, the same intervention produces different damage through the three paths. The untargeted choice reaches 3.06× at the median through plain fill, 1.18× through the rank-two declared path and 1.00× through NaN; the targeted variant reaches 2.43×, 1.04× and 1.00×. The NaN column is exactly 1.00 at the median, the 90th percentile and the maximum. This agrees with the constant candidate-fill mapping of the discarded-fill path.

The Penmanshiel rows show why accessibility does not fix damage magnitude. At 14% missingness with curtailment geometry, the same procedure achieves 1.01×, and the targeted variant slightly *reduces* error. The window, missingness pattern and input path jointly determine the attained forecast damage.

Chronos-2 preserves the contrast between content and discard. The gradient-based search uses Chronos-Bolt's replicated forward path. We test Chronos-2 with a gradient-free hill-climbing search under the same protocol and the same 412 METR-LA windows: 8 rounds of 32 range-clamped candidates, keeping the incumbent with the highest true NMSE (the same oracle untargeted objective). This search uses forecast queries and ground-truth errors; its attained damage lower-bounds the range-constrained optimum.

The best fixed fill for Chronos-2 is NaN (NMSE 1.158 against 1.220 linear). The search instead exposes the content-reading path, starting from linear rather than the discarded-fill NaN arm. It attains 1.40× at the median, 10.5× at the 90th percentile and 172× at the maximum; error more than doubles in 32% of windows. Under NaN, the candidate fills are discarded and the forecasts are unchanged. Thus Chronos-2 also exposes harmful fill changes on its content-reading path while remaining invariant on its discarded-fill path. The within-model contrast identifies this restriction; the different search methods prevent using the cross-model multipliers as a robustness ranking.

Declared content retention transmits damage too. We test fill perturbations through Moirai 2.0's declared path. The same gradient-free search, run on the ETTh1/ETTm1/weather grid at nominal rate 0.5, attains 1.29–2.49× at the median and 4.4–23.9× at the 90th percentile, against 1.02–1.26× for Chronos-Bolt's rank-two declared path on identical windows. Combined with the quality sweep in Section 4.4, these results establish the two roles of retained content: it permits a useful correction and also admits a harmful one. The interface determines access, while the supplied values and trained response determine the measured error change.

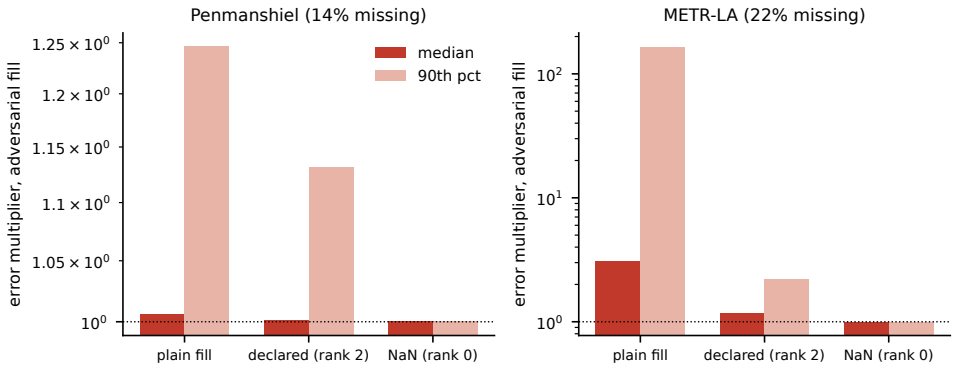

Figure 17: Attained damage from range-constrained fill search across interfaces. The discarded-fill NaN path remains at 1.00.

F.2 Neighbour benefit depends on the mechanism and available observations

Auxiliary channels provide information not present in the target's reconstructed history. We test their contribution with Chronos-2 under synthetic masks over real multivariate data, then under METR-LA's recorded outages (Figure 18, left and right). The synthetic setting masks the target channel

and supplies its four most-correlated neighbours as fully observed covariates, making neighbour availability explicit.

With the target kept as a filled channel, adding neighbours leaves median relMSE at 1.357 under block outages at nominal rate 0.7. It reduces the scattered-dropout ratio from 1.211 to 1.084 and makes the small value-censoring improvement 3.059 → 2.967. These within-path differences measure the neighbour increment. The declared path gives 1.023 under block outages, 1.096 under scattered dropout, and 3.503 under value censoring. Because the corresponding declared-without-neighbour arms are not reported here, these values compare endpoints under a different interface. The filled-path paired contrasts establish a mechanism-dependent neighbour increment. The synthetic 95% recovery uses a separate own-clean relative-degradation summary, not that neighbour increment.

Real METR-LA outages restrict neighbour availability: of 618 windows, only 38 (6.1%) have two fully healthy top-correlated neighbours. Allowing neighbours with at most 5% missingness yields the 98 windows evaluated here. In this selected set, declaring plus neighbours improves the median window by 2% and the mean by 40% (2.078 → 1.238). The larger mean reduction locates the benefit in the high-error part of this selected distribution. In deployment, the constraint is therefore not just whether a model can use covariates, but whether sufficiently observed neighbours are available. This result is conditional on that availability, whereas the synthetic comparison supplies fully observed neighbours and uses an own-clean reference.

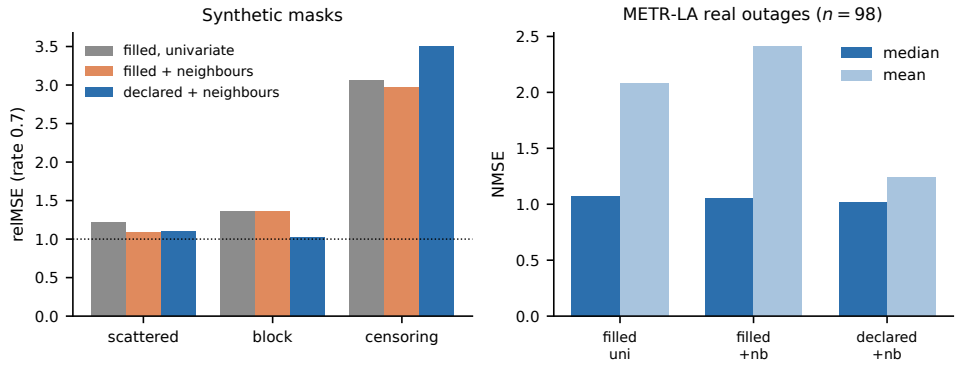

Figure 18: Cross-channel borrowing with Chronos-2: synthetic masks (left) and real METR-LA outages (right). Only 6.1% of real outage windows have two healthy top-correlated neighbours.

G Learning robustness and learning to use reconstructed content

The released-model comparisons leave interface design and pretraining entangled. The within-family factorial first separates these factors at fixed architecture and budget. Recipe comparisons then ask what exposure makes retained repairs useful, and a two-source task tests substitution between repairs and covariates on a common risk scale. Finally, CPT applies the resulting design to released checkpoints, with matched-budget controls and native-path comparisons identifying which gains each adaptation achieves.

G.1 A within-family factorial separates access from training exposure

The decisive contrast is between learning to predict despite gaps and learning to use their reconstructed values. We cross vanilla and restored interfaces with filtered and mechanism-diverse training, holding the corpus, architecture and training budget fixed. The filtered arms share weights, making their evaluation difference an interface intervention.

Corpus. We use 23 Chronos corpus subsets (`autogluon/chronos_datasets`), excluding all evaluation sources: the nine benchmark datasets (ETTh/m 1/2, weather, electricity, traffic, exchange,

illness; the corpus contains no ETT), the GIFT-Eval missing-containing sources (`kdd_cup_2018`, `temperature_rain`, `car_parts`, `covid_deaths`), Penmanshiel and METR-LA. The corpus contains 386,618 real series (4.7B points) plus 1M KernelSynth series sampled at 15%, with per-subset caps so no source dominates. Windows are sampled uniformly with complete targets.

Arms. All five arms are Chronos-Bolt tiny (8.7M), randomly initialised with the same initial seed for paired contrasts, trained 15,000 steps at batch 1024 (AdamW, lr 3×10^{-4} cosine, fp32, NaN-gradient guard) with the bolt-native next-block quantile loss:

Table 17: The pretraining arms; the filtered-corpus control is bit-identical to baseline and is therefore not a row (see the shared-weights control below). Interface: vanilla = content zeroed at masked positions; restored = content + flag. Regime: filtered = no augmentation; mechdiv = mechanism-diverse missingness (50% of windows; mechanism ~ U{scattered, block, censoring, extreme}, rate $\sim U(0.05, 0.7)$, half declared, half α-blend filled, $\alpha \sim U(0, 1)$); moirai2 = single scattered patch-mask at 50% with causal-mean fill (a Moirai-2.0-style control).

arm	interface	regime	reads as
baseline	vanilla	filtered	filtered baseline
restored + mech-diverse	restored	mechdiv	interface and augmentation
baseline + mech-diverse	vanilla	mechdiv	augmentation alone
Moirai-2.0 recipe	restored	moirai2	scattered patch-mask control

Common reference computation. The evaluation harness reproduces the stored vanilla Chronos-Bolt tiny probe values over nine cells to a maximum deviation of 0.001. With randomly initialised weights, the interface-switchable encoder matches the vanilla plain-input forward exactly ($\max |\Delta| = 0.0$). These comparisons establish the numerical reference before interface and training interventions are applied. The factorial then trains its own arms from scratch on the corpus above, separating that controlled comparison from the released-checkpoint probe.

Shared weights in the filtered arms. On fully observed training inputs, the vanilla and restored interfaces perform the same operation. The two filtered arms therefore share trained weights and clean predictions. They remain distinct evaluation cells: declaring holes through the restored path opens content access that was not exercised during training. Table 2's negative closure measures the effect of this changed evaluation path.

Access and robustness have different signatures. Vanilla augmentation leaves the fill-quality slope near zero, whereas the trained restored arm has a substantial permutation response and closes relative degradation (Table 18). Reporting response, closure and clean error together distinguishes use of repaired content from improved prediction despite gaps.

Paired seeds and uncertainty. The full factorial and fill-diverse arms use three paired training seeds, 20260824, 20260901 and 20260902, varying initialisation, window sampling and augmentation draws together. Each metric is aggregated within a seed before computing the cross-seed mean and sample standard deviation ($n - 1$ denominator). All arm-to-baseline contrasts pair the same training seed and evaluation grid. Table 19 separates variation across training seeds from the window-paired bootstrap intervals computed within each seed.

The paired records are stored in `s48_eval_seeds.json` for baseline, restored-plus-mechanism-diverse and fill-diverse, and in `s48_eval_bdseeds.json` for filtered-restored and vanilla-augmented arms. Together they contain the three-seed factorial of Table 2.

Changing the recipe improves consistency as well as closure. Across three seeds, fill-diverse reaches 93.8% ± 2.5 closure, compared with 72.7% ± 16.1 for restored-plus-mechanism-diverse. The mean difference is 21.1 percentage points. Fill-diverse also has more $\alpha = 0$ cells no worse than baseline in every seed and clean ratios 0.975, 0.963 and 0.910.

For restored-plus-mechanism-diverse, clean ratios are 1.055, 1.036 and 1.009; two of the three meet the 5% clean-cost check. These clean costs belong to the same arm as the closure comparison

Table 18: Access, quality response and accuracy over three paired training seeds (mean ± sample s.d.). Slopes and ρ use scattered and block masks; closure and floor use all four mechanisms. The Moirai-2.0 recipe comparison uses the separate nine-dataset grid.

Comparison	Measured response
The baseline retains the access restriction ($\rho \approx 0$, flat sweep)	$\rho < 0.001$ in all seeds; |slope| = 0.011 ± 0.003
restored + mech-diverse reads content ($\rho \geq 0.5$) and closes ≥ 50% of the baseline's relative degradation at α=1	All seeds; ρ = 0.852 ± 0.034, closure 72.7% ± 16.1
At α=0, restored + mech-diverse is no worse than baseline in ≥ 60% of cells	All seeds; (18.3 ± 1.2)/24
baseline + mech-diverse (augmentation, vanilla interface) stays flat in α	All seeds; |slope| = 0.006 ± 0.005
restored + mech-diverse's clean accuracy within 5% of the baseline's	Two of three seeds meet the check; ratio 1.033 ± 0.023 (Table 19)
The restored path reopens sensitivity to bad fills relative to baseline	All seeds; 1.18 ± 0.03 vs 1.06 ± 0.02
Recipe contrast: Moirai-2.0-style versus mechanism-diverse augmentation	Nine-dataset closure 36.2% ± 28.1 versus 73.3% ± 19.3 (Table 20)

Table 19: Mechanism-diverse and fill-diverse training at three seeds. Closure is the median relative-degradation contrast against the same-seed reference at α=1, with a 95% interval from a window-paired bootstrap; floor counts cells of 24 where the arm is no worse than the status quo at α=0; clean is the median clean-MSE ratio over the nine datasets.

arm	seed	closure α=1	bootstrap 95%	floor	clean
restored + mech-diverse	1	55.2%	[+27, +67]%	17/24	1.055
restored + mech-diverse	2	87.1%	[+85, +95]%	19/24	1.036
restored + mech-diverse	3	75.8%	[+63, +83]%	19/24	1.009
fill-diverse	1	91.7%	[+87, +97]%	20/24	0.975
fill-diverse	2	96.5%	[+96, +101]%	22/24	0.963
fill-diverse	3	93.1%	[+89, +98]%	20/24	0.910
restored + mech-diverse, mean ± s.d.		72.7% ± 16.1		18.3 ± 1.2	1.033 ± 0.023
fill-diverse, mean ± s.d.		93.8% ± 2.5		20.7 ± 1.2	0.949 ± 0.035

above. The fill-diverse arm's corresponding bad-fill damage is 1.20×, 1.28× and 1.34×. The recipe comparison thus needs both useful-fill recovery and harmful-fill response: more consistent closure does not remove the channel through which poor repairs can act.

Clean effects vary across datasets. Across three paired seeds, the per-dataset clean-context median MSE ratios of restored + mech-diverse to baseline are ETTh1 0.918 ± 0.018, ETTm1 1.045 ± 0.032, weather 1.199 ± 0.113, ETTh2 0.986 ± 0.028, ETTm2 1.002 ± 0.020, electricity 0.990 ± 0.011, traffic 1.043±0.010, exchange 1.252±0.113 and illness 1.147±0.067. The aggregate 1.033±0.023 first takes the median over datasets within each seed; it summarises the heterogeneous costs across datasets.

Numerical and loading conventions. Training windows are finite before augmentation; the sampler introduces missingness only after window selection. Training uses TF32 matrix multiplication, while evaluation uses strict fp32; these input and precision conventions define the reproducibility boundary. Loading the compressed corpus arrays into RAM once avoids repeated archive reads during sampling, an I/O optimization that changes throughput rather than the model or its comparisons.

G.2 Training on repair quality changes the use of retained content

After fixing a restored interface, the next question is which training inputs make it useful. The recipe comparisons use Chronos-Bolt tiny with a fixed budget (15,000 steps, batch 1024), restored interface, corpus and sampler. The augmentation recipe varies across arms.

Contiguous-patch masking (CPM) removes runs of one to five 16-value patches, replacing their content with zeros and retaining the observation flag. Declaration-only training uses four missingness mechanisms and zero fills. Fill-diverse training uses block outages with uniformly sampled interpolation quality. The full-dosage mixture uses the same interpolation over all four mechanisms: every augmented window receives a blended fill, with no separate zero-fill endpoint.

Table 20: **Nine-dataset training-recipe comparison at three seeds.** All entries are means ± sample s.d. of per-seed statistics for Chronos-Bolt tiny. The grid has nine datasets, four mechanisms and two rates, with 150 windows per cell. Closure at $\alpha = 1$ uses 62, 71 and 64 cells with same-seed baseline excess > 0.02; floor counts cells of 72 no worse than baseline at $\alpha = 0$. The endpoint columns are median own-clean relMSE across all 72 cells within each seed. This grid uses a different loader from Table 2.

arm	closure (%)	floor (/72)	relMSE $\alpha = 0$	relMSE $\alpha = 1$
baseline (status quo)	–	–	1.075 ± 0.036	1.127 ± 0.076
restored + mech-diverse	73.3 ± 19.3	57.0 ± 3.0	1.023 ± 0.006	1.023 ± 0.006
Moirai-2.0 recipe	36.2 ± 28.1	37.7 ± 13.1	1.111 ± 0.074	1.105 ± 0.087
CPM block-mask	40.5 ± 54.0	46.7 ± 7.8	1.058 ± 0.006	1.063 ± 0.031
declare-only	2.7 ± 63.0	44.7 ± 8.1	1.064 ± 0.017	1.137 ± 0.045
fill-diverse	94.4 ± 5.5	61.7 ± 5.5	1.012 ± 0.003	1.004 ± 0.003
mech-diverse + full dosage	97.6 ± 2.6	63.3 ± 5.5	1.007 ± 0.006	1.001 ± 0.002

Repair exposure outperforms declaration alone on this grid. Within the ladder's fixed Chronos-Bolt-tiny architecture, restored interface, corpus, sampler and 15,000-step budget, fill-diverse reaches 94.4% ± 5.5 closure on the nine-dataset grid, compared with 73.3% ± 19.3 for the mechanism-diverse mixture and 2.7% ± 63.0 for declaration-only training. The full-dosage mixture reaches 97.6% ± 2.6. The corresponding mean floor counts are 61.7, 57.0, 44.7 and 63.3 of 72. The contrast supports training on repaired values rather than merely marking holes. These are recipe-level comparisons: exposure to fills, declarations and mechanisms changes together, so their differences measure the full input mixtures rather than one independent augmentation ingredient.

The larger closure must be read alongside clean and bad-fill responses. The three-seed clean and attack summaries for the mixture and fill-diverse arms are reported in Table 2. The other recipes' separate initial-seed clean ratios and attack damage are 1.023 and 1.22× for Moirai-2.0-style, 1.095 and 1.26× for CPM, 1.036 and 1.36× for declaration-only, and 1.011 and 1.71× for full dosage. These readouts use the scorecard protocol, not the nine-dataset sweep endpoints.

The fill-diverse recipe also works at larger tested sizes. The single-seed fill-diverse runs reach 76.3% closure at 48M and 99.5% at 205M on the three-dataset grid, with clean ratios 0.929 and 1.050. The window-bootstrap closure intervals are $[71, 89]$% and $[98, 101]$%; floor counts are $9/24$ and $15/24$, and attack damage is 1.30× and 1.36×. The mechanism-diverse 48M run has −4.8% closure. These results extend fill-diverse training to two larger sizes, with size-dependent effects.

The three-seed recipe sweep is stored in `s48_eval_9ds_lite.json`. Each seed is compared with its own baseline under the same loader and window set.

G.3 Reliable repairs and covariates as alternative information sources

Useful target repairs need not establish that a model can instead rely on covariates. The bivariate task in Figure 25 tests this stronger property by varying the two sources independently and supplying a common statistical reference. Conditional on each series' sampled generating parameters, the process is Gaussian and its reference forecast follows from Gaussian conditioning. The target combines an AR(1) process ($\phi \sim U(0.7, 0.95)$) with three Gaussian harmonics, with periods drawn from $\{192, 384, 768\}$ and energy share $U(0.6, 0.85)$. The context is $c = \rho x + \sqrt{1 - \rho^2}\eta$, where the entries satisfy $\eta_t \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ and $\rho \sim U(0.5, 0.9)$.

The four arms cross interface access and corpus augmentation at 8.7M parameters and 15,000 training steps, using three seeds. Evaluation varies target-repair quality q_r and context qual-

ity q_c independently on a 5×5 grid over 256 held-out series. Here q_r blends the filled target entries toward the truth and q_c blends the whole repaired context toward the truth: at $q = 0$ the corresponding repaired values use the separately generated standard-normal noise draws, at $q = 1$ they use the true target or context values, and intermediate values use the implemented linear blend. The masks include scattered, block and tail outages at rates 0.3 and 0.7. Bayes conditioning and naive trust provide reference predictions.

A common risk scale. Every panel divides mean squared error by the same clean Bayes mean risk, 0.3047024699418246. Model panels average realised test MSE over the three seeds; the references are conditional expected risks under the known generator. At tail-outage rate 0.7, the trained retaining arm has values 3.17 ± 0.10, 1.36 ± 0.04, 1.48 ± 0.04 and 1.41 ± 0.07 at $(q_r, q_c) = (0,0), (0,1), (1,0), (1,1)$ (mean ± sample s.d., three seeds). The corresponding Bayes references are 2.38, 1.06, 1.00 and 1.00. The paired corners show that either reliable source lowers error for the trained retaining arm. Both remain above the conditional reference, separating useful substitution from attainment of the generator's optimal risk.

Information loss within the generator. Keeping the generator and rate at 0.7, the scattered-mask conditional reference is 1.03–1.07 times its clean risk. This is distinct from the illustrated tail-outage corner with both sources at quality zero, whose conditional reference is 2.38; the two comparisons change mask geometry and source quality, respectively. They describe different available-information settings. Naive trust reaches 2.94 at that corner, providing a strategy comparison against the conditional reference.

Interface and training responses. At $(q_r, q_c) = (0,0)$, the native augmented arm has shared-scale error 7.42 ± 0.77, compared with 4.81 ± 0.18 for the native filtered arm. This is the augmentation contrast at a blocked interface: neither native arm exposes a declared content channel, so it does not test learned use of a flag. The restored filtered arm has errors 20.50 ± 8.07, 20.21 ± 8.41, 20.10 ± 5.00 and 19.52 ± 5.22 at the four corners in the same order; this is access without missingness training. Changing a declaration flag while fixing content also changes the trained retaining arm's forecast. Opening a path permits a response to the supplied information; the trained retaining arm demonstrates how exposure to content and declarations changes that response.

On scattered-mask grids, learned and reference repair-weight magnitudes have three-seed Spearman correlations 0.922 ± 0.031 at rate 0.3 and 0.905 ± 0.019 at rate 0.7. The correlations compare magnitudes, leaving agreement in signs unresolved.

Relative to the corresponding native filtered arm at each seed, the trained retaining arm's clean-context MSE increases by $5.4\% \pm 4.1$ across three seeds. We report this clean-cost comparison separately from the arena's shared-scale missingness errors, whose denominator is the common Bayes risk.

G.4 Adapting released weights to reconstructed inputs

The from-scratch comparisons motivate retaining repairs and training across their quality. CPT tests whether released weights can acquire this response with a short adaptation. The tiny released-checkpoint scorecard below and the 205M base result are separate comparisons. For the tiny arms, each cell has 150 evaluation windows, and before training the interface replica matches the vanilla forward exactly. The reported CPT seed schedule is 20260826, 20260901 and 20260902; seed summaries are marked by spreads or ranges. The Chronos-Bolt base retrofit takes 43 minutes on one GPU. Appendices G.4–G.6 give the family-specific configurations and scorecards.

G.4.1 Separating relative degradation from absolute error

The training scorecards use *paired per-window relMSE*: the error under a condition divided by the same model's error on the same window with a clean context, summarised by the median over windows. This measures degradation relative to each arm's clean performance. Cross-arm differences in these ratios are not absolute risk reductions.

Closure as a relative-degradation contrast. For one cell (dataset × mechanism × rate) at $\alpha = 1$,

$$\gamma = \frac{e_{\text{ref}} - e_{\text{mod}}}{e_{\text{ref}} - 1},$$

where each e is that arm's own-clean relMSE, and the reference is vanilla on its declared path. The denominator is the reference's relative degradation at the oracle endpoint. For ETTm1 value censoring at rate 0.7, values $e_{\text{ref}} = 5.15$ and $e_{\text{mod}} = 3.17$ give $(5.15 - 3.17)/4.15 = 48\%$. This percentage contrasts normalised degradation; it is not the fraction of a common absolute error removed. The aggregate is the median over cells with $e_{\text{ref}} - 1 > 0.02$, excluding small denominators.

Absolute-scale win rates. Absolute errors complement own-clean ratios by comparing forecasts on a common scale. Against the vanilla checkpoint the retrofit wins 74.0% of windows under zero fill, 62.1% under linear and 61.1% under a perfect fill (66,668 series-rows per fill, rate 0.7, declared path); every dataset×fill cell wins more than half its windows (minimum 50.5%, weather at perfect fill). Weather, whose clean error rises 13.1%, returns a positive mean delta on two fills, but 61–65% of that mass sits in the top 10 of its 12,600 rows; under zero fill the retrofit improves every statistic. Per-window win rates on GIFT-Eval's real missingness appear in Section 5.

The permutation probe ρ. In equation 6, permuting fills among missing positions preserves their multiset, including mean and variance, while changing their arrangement. The independent-redraw denominator changes both content and potentially its statistics. A small ratio with a nonzero denominator indicates little response to the tested arrangement changes. If both responses vanish, the ratio is undefined; a model without a declaration path has no declared-path ρ.

G.4.2 Access without adaptation differs from trained repair use

Table 21: CPT-from-vanilla arms at Chronos-Bolt tiny. Spreads and ranges summarise three seeds; unqualified entries are point estimates (median closure of the reference arm's relative degradation at $\alpha = 1$; floor = cells of 24 where the arm is no worse than vanilla at $\alpha = 0$; clean = median MSE ratio over the nine benchmarks; worst fill = median degradation under an adversarially chosen fill).

arm	ρ declared	closure α=1	floor α=0	clean	worst fill
vanilla (reference)	0.001	–	–	1.000	1.01×
interface restored, untrained	0.807	8.4%	12/24	1.000	1.55×
adapter only (600 steps)	0.747	37.4% ± 8.7	15/24	1.026	1.25×
adapter + CPT	0.789	82.0% ± 4.0	18/24	1.062	1.40×
CPT control (no missing)	0.000	−18.9% ± 9.4	9/24	1.038	1.02×
one-stage CPT (fill-diverse)	0.808	75.7% ± 2.1	18/24	1.048	1.43×
CPT (mech-diverse)	0.724	71.7%	17/24	1.254	1.14×
CPT retrofit (WiSE-FT)	0.748	61.7% ± 1.6	17–18/24	1.058	1.20×

The ladder separates opening access from learning how to use it. Restoring the untrained interface gives $\rho = 0.807$ but 8.4% closure and bad-fill damage 1.55×. The adapter reaches 37.4% ± 8.7 closure; one-stage fill-diverse CPT reaches 75.7%±2.1. The no-missingness control has −18.9%±9.4 closure, separating the tested missingness-augmentation package from clean-only training at the same additional budget. The gain belongs to the tested training package, whose fill and mechanism exposure change together.

Across the paired seeds, adapter-plus-CPT exceeds one-stage CPT by 0.9, 5.9 and 12.1 closure points. Applying weight interpolation to the retrofit yields 61.7% ± 1.6 closure with clean ratio 1.058 and bad-fill damage 1.20×. The two-stage recipe changes closure; weight interpolation also changes clean accuracy and fill sensitivity.

At Chronos-Bolt base (205M), the reported one-stage fill-diverse CPT run reaches 78.4% closure, with window-bootstrap interval $[71, 82]$% and clean ratio 1.041. This is a single-run result with a window-bootstrap interval, not a three-seed estimate. The base result and the tiny scorecard are separate comparisons: the tiny scorecard's clean ratios 1.038 and 1.058 correspond to +3.8% for the control and +5.8% for the interpolated retrofit.

G.4.3 Training must cover the supplied fill convention

The any-fill matrix compares the CPT retrofit (WiSE-FT) with vanilla using each model's own-clean relMSE. It covers nine benchmarks × four fills (zero, forward, linear and oracle), aggre-

gated over the four mechanisms at rate 0.7. Zero fill is the vanilla model's native values-only convention; forward and linear are supplied numeric fills, while oracle is the true-value endpoint. The retrofit receives the numeric fills together with observation flags. At a ±2% band, the retrofit has 29 wins, 4 ties and 3 losses (ETTh2 zero +13%, ETTm2 zero +3%, ETTh1 ffill +5%). Both oracle and linear fill win throughout.

Table 22: Zero fill (the hardest column; the vanilla model's native convention). Median relMSE.

	ETTh1	ETTm1	weather	ETTh2	ETTm2	electricity	traffic	exchange	illness
vanilla	1.78	2.14	1.82	1.80	2.13	8.43	4.34	14.67	2.19
CPT retrofit	1.81	2.01	**1.43**	2.04	2.19	**7.13**	**3.81**	**6.79**	**2.03**

In the matched blend-only ablation, zero-fill relMSE is 3–19 times the vanilla value across datasets. The matched full recipe is the CPT retrofit row in Table 22. Training on its separate zero-content declaration endpoint lowers zero-fill degradation relative to blend-only. The ablation connects a specific deployment input to the exposure needed during adaptation.

G.4.4 Synthetic-gap gains transfer unevenly to recorded outages

Recorded outages change both the data distribution and the gap patterns encountered by the retrofit. Penmanshiel curtailment and METR-LA outages use held-out test halves and fixed fills under Appendix H.5's protocol. On these real-missingness tests, the CPT and vanilla point estimates are close: zero-fill NMSE is 20.2 versus 19.9 on Penmanshiel; linear-fill NMSE is 2.11 versus 2.14 on METR-LA. The clean-context NMSE on METR-LA is 2.06 versus 2.99, a separate clean-reference comparison. The clean and missing-input results therefore describe different effects of adaptation on this deployment.

The tested from-scratch arms have Penmanshiel clean NMSE 16.9–76.1, compared with 5.85 for vanilla. Separately, for the learned fill in Appendix B.2, CPT yields Penmanshiel NMSE 21.4, compared with 16.2 through vanilla. These are different tests: clean performance of from-scratch arms versus a learned-fill CPT comparison. Together they show that the synthetic-missingness benefit does not transfer uniformly across the tested real-missingness settings.

G.4.5 Weight interpolation changes, but does not uniformly remove, clean costs

Weight interpolation controls how far the adapted weights move from the released model. Its clean-accuracy effect remains dataset-dependent. Using 30% CPT weights gives a 30% increase on m4_hourly, compared with 53% for the 50% interpolation, so interpolation changes but does not remove the dataset-specific regression. A separate clean-anchor distillation test with $\lambda = 10$ gives 33.5% closure and clean ratio 1.110 at tiny, and 45.9% closure with ratio 1.038 at base; these are not direct rankings against interpolation because the objective and size change. The no-missingness controls also incur clean increases of 3.8–12.9%. These observations locate the clean-retention issue in the tested objectives, weights and datasets rather than in additional optimisation alone.

G.4.6 Attention responds when fill quality changes

On the same ETTh1 windows, vanilla attention received by masked patches changes from 0.804 to 0.808 as α increases from 0 to 1, while the CPT retrofit changes from 0.832 to 0.859 (Figure 26). Fill quality is manipulated through the supplied values, not provided as an explicit quality label. The attention response is an internal signature of content sensitivity. The paired permutation forecasts in Appendix I test whether the changed content also affects the output.

G.5 Chronos-2: restored fills and native missing-value handling

Chronos-2 tests whether the same adaptation principle transfers to a second family, while retaining its native missing-value path as a separate reference. On its native path, the interface-switchable replica matches the vanilla forward and loss bit-exactly ($\max|\Delta| = 0.0$ on all 18 clean and holed gate cells). The dual path removes the content-zeroing line, which defines the input change being

tested. For the CPT comparison, the run uses the bolt recipe: 5,000 steps from the vanilla checkpoint, mechanism-diverse augmentation (declaration endpoint + α-blend), and effective batch 1024. The initial run uses seed 20260901; the scorecard reports three-seed spreads where indicated. Because Chronos-2's group attention is $O(B^2)$ in the batch, the effective batch is accumulated over four microbatches of 256, with group attention computed within each microbatch (62 minutes on one GPU for the initial run). Thus the effective batch controls optimizer aggregation, while attention sees each microbatch separately. The retrofit uses WiSE-FT 0.5 interpolation with the vanilla weights.

Table 23: The Chronos-2 CPT retrofit, same columns as Table 21: ρ on the dual path; closure at α=1 (median over the 70 cells with vanilla excess > 0.02); floor at α=0 (cells of 72); clean is the three-seed mean of per-seed median ratios over nine benchmarks; worst fill on the same 412 METR-LA windows as Appendix F.1. Each cell uses 150 windows. Spreads are standard deviations over three seeds; worst-fill results use one seed.

arm	ρ dual	closure α=1	floor α=0	clean	worst fill
vanilla (reference)	0.000	–	–	1.000	1.40×
CPT	0.825 ± 0.030	80.8% ± 1.1	65–67/72	1.043	1.08×
CPT retrofit (WiSE-FT)	0.880 ± 0.023	55.5% ± 2.2	59–61/72	1.007	1.11×

Numeric-fill improvement and native declaration are different comparisons. The grid uses $\alpha \in \{0, 0.5, 1\}$. For the dual arms, the NaN-declared cell is the trained declaration endpoint, zero content plus a flag; NaN and zero coincide by construction. Vanilla instead uses observed-only statistics on its native NaN path. In the 36-cell declaration comparison, NativeNaN wins 26 cells and the retrofit wins 10. The retrofit improves the tested zero and linear inputs, while the native observed-only path remains the stronger choice in most of this declaration grid.

Interpolation balances clean accuracy and oracle closure. For the initial seed, interpolation changes the clean ratio from 1.039 to 0.993 and closure from 79.5% to 53.9%. At $\alpha = 0.5$, closure is 82.4% before interpolation and 77.6% after it. The average comparison is therefore a trade-off: interpolation lowers the clean ratio while also lowering closure at the reported endpoints. Two difficult cells describe separate limits. Illness retains a clean regression (1.75 before interpolation, 1.48 after), while on electricity under 70% value censoring with zero fill, the retrofit has relMSE 79.3 versus vanilla's 31.0. The electricity cell uses the dual path's full-context statistics, distinct from NativeNaN's observed-only statistics, so it is a missing-context path result rather than another summary of the clean/closure trade-off. Retaining content and choosing the normalisation statistics are therefore separate design decisions for this family.

Bad-fill sensitivity. Under the evaluated attack budget, damage is 1.08–1.11× for CPT and its interpolation, compared with 1.40× for vanilla. Opening a content path determines which perturbations can reach the forecast; it does not determine their error magnitude.

Omitting the zero-content endpoint leaves that input untrained. The blend-only arm uses the same 5,000 steps, seed and corpus, but fills every holed window with an α blend, so it does not train on the standalone zero-content input. At the separately tabulated zero-fill column, its median relMSE is 20.13, compared with vanilla's 2.03, with no wins in 36 cells; the worst exchange cell is 3,327.8 versus 3.59. This zero-fill column is distinct from the sweep's $\alpha = 0$ endpoint, which uses the linear fill for the closure and floor summaries. On that trained sweep, the same arm has linear-fill relMSE 1.24 and closure 84.6%, with 65/72 cells no worse than vanilla at $\alpha = 0$; its clean ratio is 1.020 and its permutation response is $\rho = 0.864$. The arm learns to use the interpolation-quality sweep but responds poorly to zero content, which that sweep does not supply. Including the standalone endpoint addresses this mismatch in the tested recipe; the comparison does not isolate the numerical operation behind these errors.

G.6 MOIRAI 2.0: ADAPTING USE WITHOUT CHANGING CONTENT ACCESS

Moirai 2.0 already passes fills and flags into the content path. Keeping this interface unchanged tests the learned-use component directly: CPT can change how repairs affect forecasting without opening a new input route.

Forward computation and loss. At the evaluation batch size, our call path matches the library's own GluonTS inference path exactly ($\max|\Delta| = 0.0$ on 60 clean windows, fp32). This exact match is conditional on that batch size: at batch 512 cuBLAS selects different kernels and the gap is 1.9×10^{-5}; all evaluation uses one batch size across arms. Separately, the training loss is the released quantile loss (`PackedQuantileMAELoss`), not a re-implementation. The probe reproduces the census value before any training: the vanilla model declared $\rho = 0.949$.

Training. We use 5,000 steps from the vanilla checkpoint with the same corpus and mechanism-diverse augmentation, mapped to Moirai 2.0's (target, observed_mask) input. The initial seed is 20260901 (18 minutes on one GPU); the scorecard reports three-seed spreads where indicated. To limit extreme normalised targets, we skip training windows whose future lies more than 100 observed-context standard deviations away (1.7% of draws). This filter is part of the training distribution, not an evaluation cleanup. Moirai 2.0's scaler floors at 10^{-5}, and unfiltered, 1.3% of near-constant windows carried over 99% of batch loss; the same low-quality filter exists in the model's own pretraining. The filter therefore changes training exposure, while the evaluation windows remain unchanged. After CPT, validation clean loss is 0.1841 against the vanilla model's 0.1831.

Table 24: The Moirai 2.0 CPT retrofit, same columns as Table 21: ρ on the declared path; closure at α=1 against the vanilla model's *declared* path (which already moves with α (vanilla endpoints 1.500 → 1.141 overall; 1.374 → 1.176 under scattered dropout at rate 0.7)), median over the 68 cells with vanilla excess > 0.02; floor at α=0 (cells of 72); clean is the three-seed mean of per-seed median ratios over nine benchmarks; worst fill on the same 412 METR-LA windows as Appendix F.1. Three seeds (spread = ± sd; worst fill single-seed), 150 windows per cell.

arm	ρ declared	closure α=1	floor α=0	clean	worst fill
vanilla (reference)	0.949	–	–	1.000	1.41×
CPT	0.856 ± 0.019	36.3% ± 0.4	65/72	1.033	1.23×
CPT retrofit (WiSE-FT)	0.895 ± 0.018	54.3% ± 1.3	66/72	1.016	1.30×

An unchanged interface supports a different learned response. Content sensitivity remains (ρ 0.949 → 0.895). At rate 0.7, within each checkpoint the fill-quality sweep moves from $\alpha = 0$ to $\alpha = 1$: vanilla changes 1.682 → 1.183, while the interpolated CPT checkpoint changes 1.499 → 1.123. These arrows are within- checkpoint quality endpoints, not a vanilla-to-CPT comparison. The leaderboard comparison across checkpoints also improves (zero fill 11.56 → 2.96, linear 1.60 → 1.49, 36 of 36 and 32 of 36 cells won; NaN and zero coincide for this interface).

Under the tested attack, damage decreases from 1.41× to 1.30× at the median and from 14.3× to 5.4× at the 90th percentile. This interface has no content-free path on which to test fill immunity.

Exact fills test a distinct endpoint. Raw CPT is non-monotone at $\alpha = 1$ on ETTh1, ETTh2 and electricity at rate 0.7. On ETTh1 under scattered dropout, relMSE is 1.21 at $\alpha = 0$, 1.16 at $\alpha = 0.5$, and 1.73 at $\alpha = 1$. The last value exceeds vanilla's 1.18, while the same perfect fill *without* the flag scores exactly 1.000. The overshoot is therefore specific to the raw CPT flagged endpoint in these cells. The paired unflagged input shows that supplying the true values is insufficient when their declaration changes the trained response. WiSE-FT reduces negative-closure cells from 31% to 9% and changes the worst closure from −6.05 to −0.52. A mismatch between training blends and the exact $\alpha = 1$ endpoint is a candidate explanation, but these interventions do not isolate it. The established design implication is to evaluate exact flagged fills separately, even when intermediate-quality fills improve forecasting.

G.7 Cross-family results under explicit fill conventions

Table 25 is the common fill/path comparison for every evaluated model and dataset: zero and linear fill, plus an oracle reference, at rate 0.7. The columns align the supplied fill while retaining each model's available input path: declared paths are used where available, and values-only models are identified separately. This makes the role of the input convention explicit in the cross-family score comparison. Mechanism-level cells are retained in the result JSONs.

Table 25: **Relative-degradation profiles across checkpoints.** Cells use the own-clean relMSE and hierarchical aggregation of Table 3 on nine benchmarks at 70% missing. Zero and linear fills represent practical choices; oracle fills are diagnostic references. Bold marks the lowest relative-degradation value, not the lowest absolute forecast error. On their declared paths, TimesFM, TiRex and FlowState replace or discard the external fill, so these columns are invariant to its value. The tested Timer-base checkpoint has no declaration path; raw NaN input produces non-finite forecasts.

Model	ETTh1	ETTm1	weath.	ETTh2	ETTm2	electr.	traffic	exch.	illness	*med.*
perfect fill										
Chronos-Bolt base (vanilla)	1.51	1.85	1.41	1.38	1.41	4.88	3.31	1.30	2.02	1.51
+ 5k steps, no missingness	1.40	1.94	1.31	1.34	1.32	4.19	3.47	1.83	2.38	1.83
Moirai 2.0 (vanilla)	**1.13**	1.28	1.04	1.12	1.32	**1.37**	1.84	1.08	1.18	1.18
TimesFM 2.5	1.55	1.86	1.13	1.37	1.38	5.77	5.63	1.23	19.29	1.55
TiRex	1.59	1.91	1.49	1.34	1.45	6.82	5.03	1.41	1.80	1.59
FlowState	1.54	1.16	1.28	1.30	1.15	5.13	4.20	1.35	1.09	1.30
Timer-base	no declaration endpoint; raw NaN non-finite									
Chronos-Bolt base + CPT (ours)	1.28	1.27	1.18	1.22	1.16	1.73	1.61	1.17	1.08	1.22
Chronos-2 (vanilla)	1.53	1.76	1.54	1.37	1.40	4.73	3.97	1.40	2.27	1.54
Chronos-2 + CPT (ours)	1.31	1.52	1.23	1.25	1.29	1.70	1.74	1.09	**0.99**	1.29
Moirai 2.0 + CPT (ours)	1.17	**1.14**	**1.03**	**1.11**	**1.11**	1.43	**1.55**	**1.02**	1.10	**1.11**
linear fill										
Chronos-Bolt base (vanilla)	1.57	1.84	1.31	1.36	1.40	5.75	4.09	1.32	2.09	1.57
+ 5k steps, no missingness	1.63	1.94	1.23	1.37	1.32	6.30	4.56	1.64	2.45	1.64
Moirai 2.0 (vanilla)	1.60	1.90	1.28	1.44	1.58	5.18	5.42	1.20	1.77	1.60
TimesFM 2.5	1.55	1.86	1.13	1.37	1.38	5.77	5.63	1.23	19.29	1.55
TiRex	1.59	1.91	1.49	1.34	1.45	6.82	5.03	1.41	1.80	1.59
FlowState	1.54	**1.16**	1.28	**1.30**	**1.15**	5.13	4.20	1.35	**1.09**	**1.30**
Timer-base	no declaration endpoint; raw NaN non-finite									
Chronos-Bolt base + CPT (ours)	1.51	1.69	**1.12**	1.33	1.32	4.53	3.68	1.11	1.75	1.51
Chronos-2 (vanilla)	1.52	1.75	1.38	1.36	1.38	5.12	4.18	1.46	2.34	1.52
Chronos-2 + CPT (ours)	**1.44**	1.48	1.23	1.34	1.36	**3.50**	**3.42**	1.11	1.46	1.44
Moirai 2.0 + CPT (ours)	1.49	1.66	1.14	1.34	1.29	4.08	4.53	**1.09**	1.90	1.49
zero fill										
Chronos-Bolt base (vanilla)	2.21	2.54	3.51	3.34	3.79	14.39	5.30	46.48	2.88	3.51
+ 5k steps, no missingness	4.74	6.76	9.32	13.95	16.88	54.41	9.80	393	4.81	9.80
Moirai 2.0 (vanilla)	6.87	10.79	5.45	11.56	12.58	82.20	22.26	13.96	4.81	11.56
TimesFM 2.5	1.55	1.86	**1.13**	1.37	1.38	5.77	5.63	**1.23**	19.29	1.55
TiRex	1.59	1.91	1.49	1.34	1.45	6.82	5.03	1.41	1.80	1.59
FlowState	**1.54**	**1.16**	1.28	**1.30**	**1.15**	**5.13**	4.20	1.35	**1.09**	**1.30**
Timer-base	no declaration endpoint; raw NaN non-finite									
Chronos-Bolt base + CPT (ours)	1.70	1.96	1.56	2.03	2.26	5.74	**3.84**	7.18	2.29	2.26
Chronos-2 (vanilla)	1.67	1.86	2.10	1.68	1.73	5.99	4.36	5.02	2.53	2.10
Chronos-2 + CPT (ours)	1.68	1.57	1.66	1.58	1.51	8.89	4.25	2.70	1.61	1.66
Moirai 2.0 + CPT (ours)	2.45	3.35	1.86	2.92	2.64	21.72	10.57	2.99	2.96	2.96

G.8 Additional clean training does not reproduce the repair recipe

To distinguish the recipe from additional optimisation, each family receives the same 5,000 steps on the same corpus without missingness (Section 5). The controls preserve the seed and sampler stream, changing the training regime to complete inputs. Their vanilla reference cells reproduce those of the corresponding recipe runs exactly. Table 26 compares both oracle closure and weak-fill response, because a gain on one endpoint need not extend to the other.

Table 26: The no-missingness control and CPT retrofit. Closure contrasts relative degradation at $\alpha = 1$; floor counts cells no worse than vanilla at $\alpha = 0$. For Chronos-Bolt, closure, floor and clean use tiny, while zero-fill relMSE uses base with dataset-hierarchical aggregation. The other families' zero-fill values use flat-36 medians at rate 0.7, unlike the hierarchical medians in Table 3. Floor denominators are 24 for tiny and 72 otherwise.

family	arm	closure α=1	floor α=0	clean	zero fill
Chronos-Bolt	control	−18.9%	9/24	1.038	3.51 → 9.80
	CPT retrofit	+61.7%	17–18/24	1.058	3.51 → 2.26
Chronos-2	control	+70.0%[a]	51/72	1.052	2.03 → 10.78
	CPT retrofit	+55.5%	59–61/72	1.007	2.03 → 1.81
Moirai 2.0	control	+13.0%	48/72	1.025	8.97 → 8.64
	CPT retrofit	+54.3%	65–66/72	1.016	8.97 → 2.80

[a] On the unflagged plain path, a perfect fill reproduces the clean input and hence relMSE 1.000. The restored declared path has a different input flag, so its oracle closure also depends on that response. The control's 70.0% oracle closure coexists with zero wins in 36 zero-fill cells. At $\alpha = 0$, the reported closure is 15% for the control and 68% for the recipe.

The no-missingness controls worsen zero-fill relMSE for Chronos-Bolt and Chronos-2, while Moirai 2.0 changes from 8.97 to 8.64. Moirai's control reaches 13.0% closure, compared with 54.3% for the retrofit. Additional clean training does not reproduce both the oracle response and zero-fill gain.

G.9 Native-path training cannot create an absent fill channel

TimesFM, TempoPFN and Timer-base (84M; `timer-base-84m`) receive 5,000 steps of mechanism-diverse CPT and a no-missingness control. Each uses its native input path; no new declaration or content channel is added. Table 27 reports means and sample standard deviations over two CPT training seeds (20260903 and 20260904); the no-missingness control has one seed. The comparison holds the input mapping fixed, so it tests whether training changes performance without changing access to externally supplied repairs.

Table 27: **Native-path CPT on boundary conventions.** TimesFM and TempoPFN have zero permutation and redraw responses on the declared path before and after CPT, so their ratio ρ is undefined. The quality-sweep gain is zero because externally supplied fills are discarded; this differs from cross-arm closure in Appendix G.4.1. Leaderboard medians aggregate 36 dataset–mechanism cells at rate 0.7. Clean costs use the three probe datasets. CPT entries report two-seed means ± sample s.d.; stock values refer to one released checkpoint. Timer-base has no declared path; its leaderboard and clean readouts use plain inputs. The Chronos-Bolt reference combines base sensitivity and tiny closure from separate runs, as labelled.

model	convention	declared ρ: stock → CPT	sweep gain	leaderboard zero fill	clean cost
TimesFM	overwriting	– → –	0% ± 0	2.02 → 1.43 ± 0.10	(10.6 ± 9.6)%
TempoPFN	learned token	– → –	0% ± 0	1.48 → 1.31 ± 0.01	(7.1 ± 2.9)%
Timer-base	no declared path	non-finite (NaN input)		1.39 → 1.14 ± 0.06	(188.0 ± 40.9)%
Chronos-Bolt (reference)	zeroing → restored	0.000 → 0.832 (base)	76.7% closure (tiny)	fill-responsive	+4% (base)

Declared and plain paths. CPT changes performance on these native paths without exposing discarded fills. TempoPFN's declared-path zero-fill relMSE changes from 1.48 to 1.31 ± 0.01, compared with 1.48 for its control, while declared fills remain inaccessible. TimesFM retains plain-path sensitivity ($\rho = 0.862$ before CPT and 0.959 ± 0.045 after), but its native declaration overwrites external repairs. Improved prediction from gaps and learned use of external repairs therefore remain distinguishable even after continued training.

Timer-base training convention. For optimization, Timer-base trains with observed-context normalisation and a Huberised next-patch loss with threshold 20; raw-scale training produced non-finite gradient norms on batches with $|x| \gtrsim 10^5$. For evaluation, the original raw-input convention is retained, under which both CPT and the control show large clean-error changes: 188.0% ± 40.9

for the two CPT seeds and 344.8% for the control. This drift cannot be attributed to CPT alone. Independently, plain-path ρ changes from 0.843 to 0.788 ± 0.005 after CPT; there is no declared-path statistic, and NaN inputs produce non-finite logits. Thus the stabilization choice governs training feasibility, while the absent declared path governs the boundary interpretation. The model identities, input mappings and per-cell results are recorded in `experiments/s71_boundary/`.

H From reconstruction quality to forecasting benefit

Section 4.4 varies reconstruction quality while holding the model and missing positions fixed. Here we extend that comparison from interpolation-to-truth blends to learned imputers. The blends identify the response available through each interface; real residuals test whether imputation accuracy predicts that response. Residual structure and recorded-gap transfer then explain why the two comparisons can differ, followed by interventions on values, missingness metadata and internal reconstruction.

H.1 A controlled quality sweep separates access from adaptation

The sweep changes the content of a repair without changing its declared missing positions. Figure 19 complements the adapted result in Section 4.4 with zero-shot inference and a fourth interface that both restores content and computes statistics from observed points only. The zero-shot row uses three datasets and the adapted row uses all nine, with each interface measured relative to its own clean reference.

The statistics-only declared path changes little along the blend, both before and after adaptation. Zero-shot endpoints are $1.21 \rightarrow 1.22$ under scattered dropout and $1.10 \rightarrow 1.09$ under block outages at rate 0.7. Training the input projection does not make the discarded temporal content available on that path.

On the zero-shot top row, which uses three datasets, restoring content at $\alpha = 0$ raises relative degradation under scattered dropout from 1.21 for the declared path to 1.32 for the restored path. The adapted bottom row uses nine datasets rather than the same three paired windows: its own-clean ratios are close after 600 steps of input-embedding adaptation (1.148 against 1.150), but the restored path remains higher under block outages (1.129 against 1.118). These comparisons show the starting cost of exposing imperfect repairs within each panel. Their different datasets and own-clean references prevent interpreting the row difference as a paired adaptation gain.

At $\alpha = 1$, plain fill gives exactly 1.00 because the input is the clean context; it is therefore the oracle-reference endpoint, not a deployable imputer. At $\alpha = 0$, plain linear fill is the deployable baseline, and block outages give its highest relative degradation.

H.2 Transfer of the adapted interface and its clean-reference change

The adapter's transfer test asks whether the learned response extends beyond its training datasets. It adapts the input embedding for 600 steps at batch 48 and learning rate 10^{-4} with the backbone frozen. It is trained once on the three datasets above the horizontal rule in Figure 21(a), then frozen for the remaining held-out datasets. Cells where the declared path's own excess over clean is below 0.02 are marked n/a because the closure denominator is too small for a stable comparison. Displayed values are clipped to ±100% for visualisation only.

Panel (b) reports the clean-context change caused by adaptation. Clean error improves by 0.5–3% on the ETT and weather family and degrades by 5–9% on electricity, traffic and illness. Panel (a) normalises each interface by its own clean error. The largest closures occur on datasets where the clean denominator also worsens. Reading the two panels together therefore separates reduced relative degradation from the change in complete-input accuracy that supplies its reference.

H.3 Learned imputers do not follow a single blend-quality ordering

The blend sweep changes errors along one controlled path. Learned imputers can produce different residual structures at similar reconstruction accuracy. Table 28 tests these actual outputs through the same interfaces. For each dataset, SAITS and BRITS are trained in the two variants listed in the table: mechanism-matched all, and mb, trained only on scattered and block missingness (Appendix H.7).

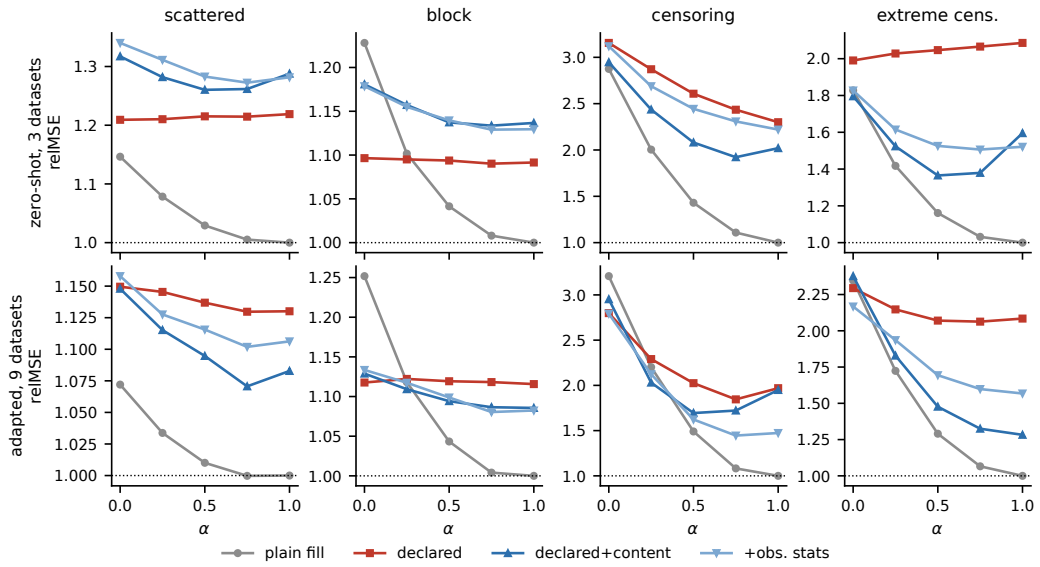

Figure 19: Zero-shot (top, three datasets) and adapted (bottom, nine) sweeps, rate 0.7.

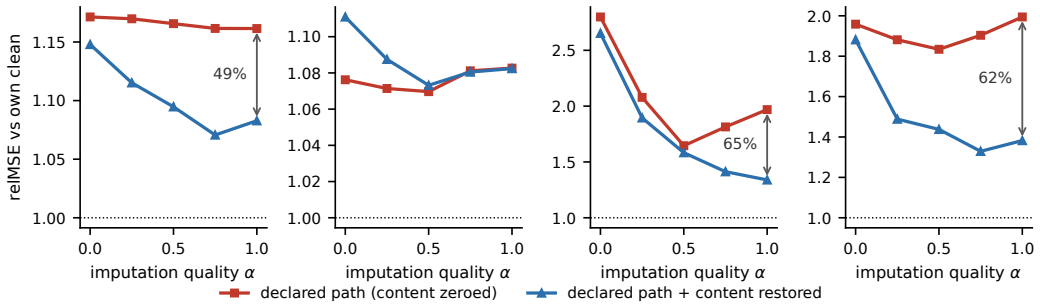

Figure 20: The in-sample row of Figure 4: the adapter's own three datasets, same layout. Columns (left to right): scattered dropout, block outages, value censoring, extreme censoring.

Table 28: **Real imputers evaluated through declared and restored interfaces.** SAITS and BRITS use the per-dataset mechanism-matched all and scattered/block-trained mb variants, on 300 windows, three datasets, four mechanisms and two rates. Cells won counts the restored path's lower own-clean relMSE in 24 cells; win share and closure use paired windows. Oracle denotes a ground-truth fill. Full results are in Appendix H.3.

fill	cells won (of 24)	median win share	median rel.-degrad. closure
linear	11	50.8%	−1.7%
SAITS mb	13	51.7%	+6.8%
SAITS all	13	53.8%	+10.6%
BRITS mb	11	50.9%	−7.2%
BRITS all	11	50.7%	−6.7%
oracle	21	58.3%	+40.2%

The table keeps these variants as separate rows rather than pooling them. Each imputed context is then evaluated through the declared path and the two content-restored variants, using the adapted input projections. Windows, masks and baselines match the main grid (300 windows, three datasets, four mechanisms, rates 0.3 and 0.7, own-clean normalisation). The linear-fill rows reproduce the sweep of Appendix H.1 to within 0.5%.

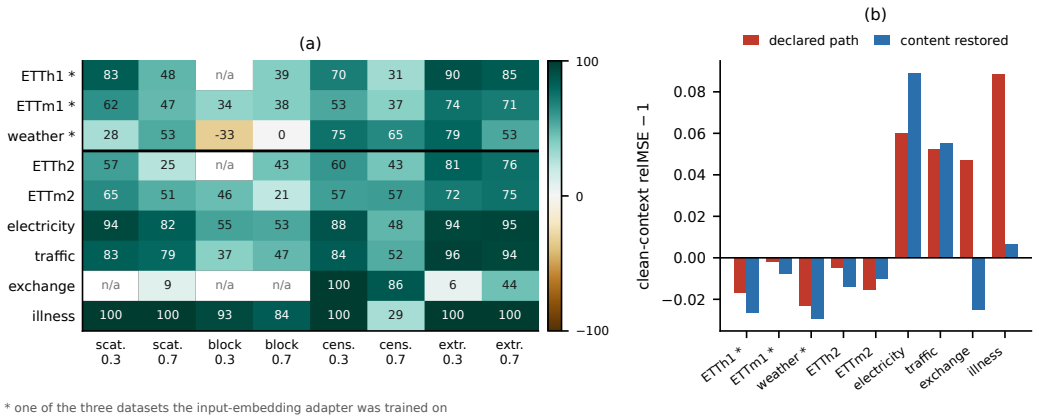

Figure 21: Closure by dataset and mechanism (a), and the clean-context control (b).

The oracle-reference comparison first isolates the interface response to a perfect fill. On ETTh1 under scattered dropout at rate 0.7, replacing linear fill with the oracle changes the declared-path relMSE from 1.171 to 1.162, whereas the same linear-to-oracle replacement changes the restored path from 1.221 to 1.084. At $\alpha = 1$, the restored path has lower own-clean relMSE in all 16 ETT cells; under heavy value censoring, the declared-to-restored interface comparison is $1.420 \rightarrow 1.291$. This comparison establishes the response to exact temporal content. The learned-imputer rows then test whether estimated content can realise a similar benefit.

With SAITS, the restored path has lower own-clean relMSE in 13 of the 16 ETT cells, but none of the 8 weather cells, where linear interpolation is already near-perfect. Under heavy value censoring, replacing linear fill with SAITS on the fixed declared path changes ETTh1 relMSE from 2.799 to 1.660 at rate 0.7; this arrow changes the imputed content, not the interface. A real imputer can therefore improve the retained statistics route even when pointwise content is zeroed, while the additional content path still gives a dataset-dependent response. The exact-fill comparison isolates access to true temporal content; SAITS shows that real repairs can also act through retained statistics, with additional content access helping only where the trained response uses it.

BRITS reaches an effective α up to 0.8 by imputation error, yet the restored path has higher median relative degradation than the declared path. Reconstruction error alone therefore does not place its forecasts on the blend sweep. The next analysis tests whether patch-scale residual structure distinguishes these responses.

H.4 Patch-scale residual structure helps distinguish imputer responses

The same residual energy can be spread across a patch or concentrated in a common offset. Because Chronos-Bolt projects patches rather than individual values, this distinction provides a concrete diagnostic of the imputer mismatch. We test whether patch-level offsets track the measured closure.

Patch-level residual energy. Chronos-Bolt uses non-overlapping patches of 16 steps and a learned patch projection. To describe residual structure at this scale, we separate the energy of per-patch offsets from total residual energy. For a nonzero residual r on the missing positions of one window, define the *patch-DC share*, where DC denotes the constant component

$$\psi = \frac{\sum_p n_p \bar{r}_p^2}{\sum_p \sum_{t \in p} r_t^2} \in [0, 1], \tag{17}$$

where p runs over patches with missing positions, n_p counts those positions, and $\bar{r}_p$ is their mean residual. A constant residual within each patch gives $\psi = 1$. The statistic describes residuals at the model's patch scale. It does not assume that the learned projection averages its inputs: zero-mean residuals can still change the embedding, so ψ is not the embedding's response itself.

We compute ψ on the residuals that produced Section 4.4's forecasts, using the windows, masks and SAITS/BRITS checkpoints of Appendix H.3. The grid covers three datasets × four mechanisms × two rates × five fills. We retain the 115 cells whose declared-path excess exceeds 0.02, keeping the relative-degradation closure denominator away from zero.

Pooled and within-mechanism associations. Across the 115 cells, ψ has Spearman correlation -0.272 ($p = 0.003$) and Pearson correlation -0.056 ($p = 0.55$) with relative-degradation closure, indicating a weak monotone association and little linear association. Imputation RMSE has correlation $+0.233$ ($p = 0.012$) in the same pooled grid. Pooling mixes datasets, rates and imputers; the following comparisons progressively hold some of these factors fixed to examine the association more closely.

Because ψ is near 0.9 under censoring and 0.3 under scattered dropout, we also split cells at each mechanism's own median ψ. Higher-ψ groups have lower closure in all four mechanisms:

mechanism	$\psi \leq$ median	$\psi >$ median	gap
scattered dropout	+14.3%	−1.1%	15.4 pts
block outage	+12.7%	−0.5%	13.2 pts
value censoring	−3.1%	−18.2%	15.1 pts
extreme censoring	+13.7%	−3.5%	17.3 pts
pooled	+12.8%	−10.7%	Mann–Whitney $p = 0.0004$

Within-mechanism Spearman coefficients are -0.33, -0.48, -0.38 and -0.59, with median -0.43, so the ψ–closure association remains when the comparison is restricted to a single missingness mechanism. Mechanism differences alone therefore do not account for the ordering. Dataset and rate still vary within these groups, motivating the matched imputer comparison below.

Paired imputer comparison. Pairing BRITS and SAITS on dataset, mechanism and rate, higher ψ accompanies lower closure in 30 of 44 pairs (binomial $p = 0.011$). Median BRITS-minus-SAITS differences are $+0.026$ in ψ and $+0.008$ in imputation RMSE. The aggregate RMSE difference is small, but individual pairs are not matched on accuracy. Holding dataset, mechanism and rate fixed preserves the ordering in a majority of pairs, strengthening the residual-structure association without establishing it as the complete cause of BRITS's forecast response.

What the diagnostic makes available. Computing ψ requires residuals against the missing true values, so it is an offline diagnostic of a completed forecast rather than a deployment-time routing signal. Within this grid, the exploratory split at $\psi = 0.86$ separates closure groups of +7.8% and −15.0% ($p = 0.002$). The split summarises this measured grid. Using it as a threshold on new datasets would require a separate transfer test and a way to estimate residuals without hidden truth.

Low ψ also does not ensure positive closure. Weather under scattered dropout at rate 0.7 has $\psi = 0.30$ and imputation RMSE 0.04, yet closure is −188%. Linear interpolation is already near-perfect in these cells, and adaptation changes the clean reference. Residual structure, initial excess degradation and the clean-context effect must therefore be considered together when interpreting this diagnostic.

H.5 Real curtailment tests the learned fill against a feasible fixed baseline

The real-data comparison removes the synthetic masking intervention and evaluates the repairs on recorded curtailment. The learned fill of Appendix B.2 is fitted on one half of the real curtailment windows and evaluated on the other. The admissible class is centred on the selected best fixed fill, so the held-out comparison includes that baseline as a feasible candidate. Centring instead on linear interpolation excluded that fixed fill in 66.3% of windows. The selected centre therefore makes the baseline available within the learned fill's feasible class (App. H.6). Figure 22 reports the paired outcomes for each window.

On real curtailment, the learned fill wins on 62% of held-out windows with a paired median ratio of 0.938 and a two-sided sign test at $p = 1.7 \times 10^{-5}$. This is a 6% reduction relative to the selected fixed fill at the median of the per-window ratios. The ratio distribution has a long right tail, with 9.2% of windows degrading by more than 2× and 3.6% by more than 5×.

A plateau-shaped augmentation does not reproduce the real benefit. We augment training with clean windows clamped to the k-th largest value to imitate a curtailment plateau. This augmentation worsened every evaluated cell: test median relMSE changed from 1.499 to 2.163 and win rate from 62% to 45%, with degradation also on the training half. This clamp-based augmentation recipe therefore fails on the tested real windows; the comparison does not isolate the responsible mismatch. Reproducing the plateau's shape is insufficient for this augmentation to improve the recorded-gap task.

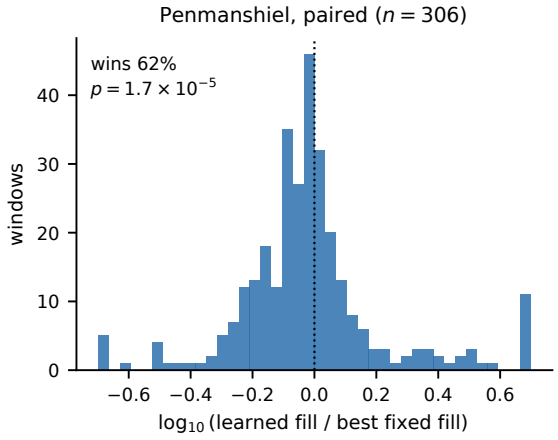

Figure 22: Stored per-window error ratios on real curtailment (306 windows), learned fill relative to the selected fixed fill. Displayed ratios are clipped to $[0.2, 5]$ before the logarithmic transform, so endpoint bins include values beyond that range.

H.6 Metric and optimisation choices needed to interpret the comparison

Own-clean normalisation measures relative degradation. Own-clean relMSE is the measure used for the synthetic sweeps and interface comparisons: it divides each configuration's corrupted-context error by that configuration's clean-forecast error. A lower relMSE therefore means less degradation within that configuration, not lower absolute error; aggregated clean errors cannot be multiplied by aggregated relMSE to recover corrupted-context absolute error, as the patch-size comparison illustrates (Appendix D.5). Closure inherits the own-clean reference. The real-curtailment migration instead reports a paired per-window ratio of learned-fill error to selected-fixed-fill error, so its 6% median describes a paired relative error reduction, not own-clean closure. An advantage in pooled absolute MSE requires aggregating errors on a shared scale.

Typical-window gain and aggregate error can differ. The per-window error distribution is heavy-tailed: on ETTh1 the clean per-window MSE has mean 10.36 and median 1.12. High-error windows can dominate a ratio of means. In the real-curtailment experiment of Appendix B.2, the median per-window ratio shows a 6% improvement, while the aggregate mean error worsens.

Finite forecasts do not guarantee valid gradients. Chronos-Bolt's instance normalisation computes $\sigma = \sqrt{\overline{(x-\mu)^2}}$ and then guards the value with $\texttt{where}(\sigma = 0, \epsilon, \sigma)$. The guard fixes the forward value, but backpropagation still encounters the infinite derivative at $\sqrt{0}$. A window with a constant *filled* context can therefore turn an upstream module's weights to NaN. Those weights produce an all-NaN context, which the model masks entirely before emitting its no-information constant. Forecasts thus remain finite but become input-independent. When training an input-side module through frozen Chronos-Bolt, we skip steps whose gradient norm is non-finite.

H.7 Interventions on values, metadata and internal reconstruction

The remaining comparisons act on different sources of information. Changing a fill modifies estimated values; a missingness channel adds metadata; reconstruction training changes

internal representations. Their distinct effects help identify which operation is needed rather than treating all modifications as one repair.

Fill-value interventions. These interventions test whether changing the filled values can repair the forecast through the existing interface. Rescaling zero fill by observed statistics recovers ∼82% of zero-fill damage under scattered dropout but worsens censoring. Tobit imputation, based on a censored regression model, improves synthetic censoring (2.52 → 1.97) and loses on real curtailment. Within the tested end-to-end candidates in Table 28, the in-domain learned-imputer rows give the strongest block-missingness results among non-oracle fills and partially improve censoring (2.17 → 1.36). The common pattern is that a useful value intervention depends on the mechanism and data source. Its role is to supply a correction through the route identified by the interface tests, not to change the interface itself.

Missingness-channel interventions. These interventions add information about the missingness pattern, not the missing values. Under the same tested grid, a learned mask token improves block-missingness results but does not transfer to censoring; a shared token trained across domains wins none of the evaluated comparisons. GRU-D-style structured features yield no measured gain in this setup, while natural-language mechanism descriptions produce a content-independent constant shift. These channels can label where information is absent, but they do not carry the fill itself. Adding a missingness projection alongside the adapted input projection improves the worst cells, yet still does not beat plain linear fill under censoring. The shared conclusion is that metadata can modulate the response, but using position-specific fill information requires content access and a path trained to use it; retained statistics can separately transmit a correction.

Internal reconstruction. Internal reconstruction tests a different bottleneck from the input interface (App. D.7). Auxiliary reconstruction training changes the probe and forecast metrics little, while opening the decoder mask yields modest gains on censored ETTh1 and ETTm1; forecasts nevertheless remain far above clean-context error. The paired knockout tests identify the decoder's use of those states, while the remaining error shows that making an internal reconstruction accessible is not enough to recover the complete-context forecast. This intervention addresses a different operation from retaining externally supplied fills.

H.7.1 Tokenisation retains the distinction between declaration and fill

Chronos-T5 maps values to discrete tokens rather than embedding patches, and has its own declared path: NaN positions become a pad token with attention mask zero, and the scale statistic is computed over observed points only. Table 29 completes the grid for that family.

Table 29: Chronos-T5, declared path versus linear fill (dataset-averaged relMSE; lower is better). Declaration gives lower relMSE at three of four rates under scattered and block missingness, at the two lower rates under value censoring, and at no tested rate under extreme censoring.

Mechanism	path	p=0.1	p=0.3	p=0.5	p=0.7
scattered dropout	declared	1.013	1.060	1.138	1.328
	linear	1.037	1.066	1.156	1.306
block outage	declared	1.108	1.154	1.325	1.426
	linear	1.099	1.182	1.360	1.560
value censoring	declared	0.988	1.233	1.442	1.845
	linear	1.008	1.244	1.438	1.709
extreme censoring	declared	0.995	1.232	1.308	1.380
	linear	0.882	1.082	1.167	1.248

Chronos-T5's declared path has lower relMSE at three of four rates for scattered dropout and block outage, at the two lower rates for value censoring, and at none for extreme censoring (Table 29). This rate-dependent comparison also occurs with tokenisation, so it is not specific to patching.

H.7.2 Scale fallback on sparse inputs

A quantile-binning tokeniser substitutes scale 1.0 for non-positive scales and clamps token indices. On all-zero streams, forecasts remain near zero regardless of scale, and changing the fallback worsens performance. For these sparse inputs, modifying the scale fallback does not replace the missing predictive information.

H.7.3 Extreme censoring leaves a distinct reconstruction problem

The tested interventions improve other cells while leaving substantial relative degradation under extreme censoring (Section 4.4). This pattern separates opening an information route from supplying and using an adequate reconstruction of the hidden tail. The former is established by the interface interventions; the latter remains unresolved in these experiments, rather than being ruled out by an error lower bound.

I Visual comparisons linking missingness, access and learning

The panels below connect the main argument at the level of individual forecasts and paired comparisons. They show how gaps change the evaluated problem, how retained repairs change forecasting, and where the trained model responds to those repairs.

I.1 What changes when incomplete windows are evaluated

The same-window cases compare interval coverage and width under different missingness mechanisms. The audit plots then hold the windows fixed while changing the input convention, and separately compare window-selection protocols (Section 4.2). Together they distinguish a forecasting response from a change in which cases contribute to the score.

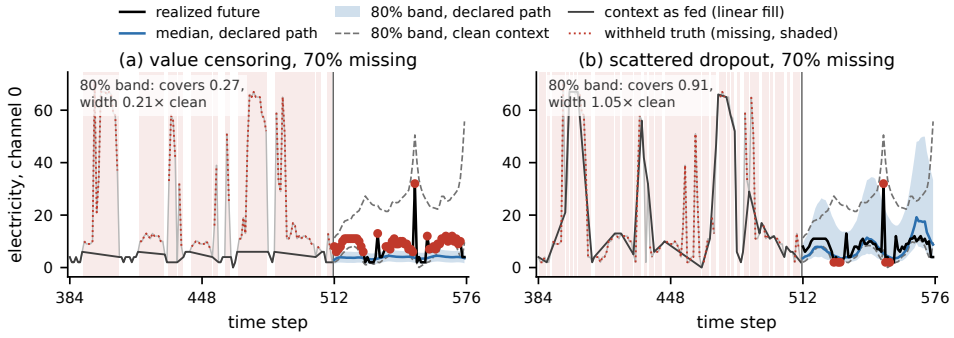

Figure 23: One window, two mechanisms, same protocol: a single electricity window (channel 0) forecast by vanilla Chronos-Bolt through its declared path (linear fill plus observation mask, rate 0.7, L=512, H=64; the last 128 of the 512 context steps are shown, missing positions shaded, the withheld values dotted). (a) Value censoring: the declared 80% band narrows to 0.21× the clean-context band (dashed) while the realized future exits it at 47 of 64 steps (red dots). (b) Scattered dropout on the same window: the band widens to 1.05× clean and covers 58 of 64 steps. The window is the first in protocol order in which the declared band under-covers its nominal 0.8 under censoring while narrower than the clean band and covers at least 0.9 of the horizon under scattered dropout while wider; the scan pool starts at the first K=6 windows of the seeded protocol and had to double to K=24. The dataset is chosen mechanically too (lowest censored coverage among qualifying windows); all three datasets appear in Figure 10. Under censoring the interval narrows around the wrong answer; under scattered dropout it widens and covers.

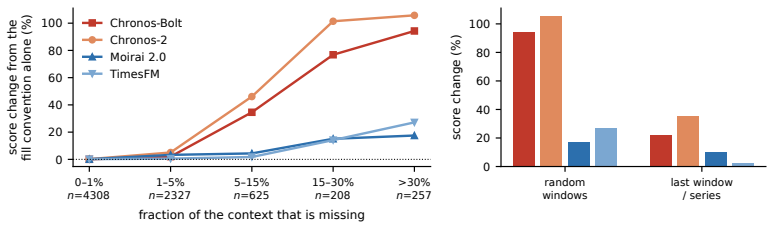

Figure 24: Left: the spread between the best and worst of four fill conventions, as a percentage of the best, from per-stratum median MASE on pooled GIFT-Eval windows, stratified by the window's own missing fraction; n is the stratum size. Right: the highest stratum under a second window protocol (last window per series), same quantity, colours as on the left.

I.2 From fill response to learned use

The released-model sweeps vary fill quality within each checkpoint, showing that retained content can change the forecast. Differences across families motivate the within-family controls in Section 4.5, where access and training exposure are separated. The two-source task extends the question to reliable repairs and auxiliary observations on a common risk scale; the CPT table reports how the recipe transfers to released weights.

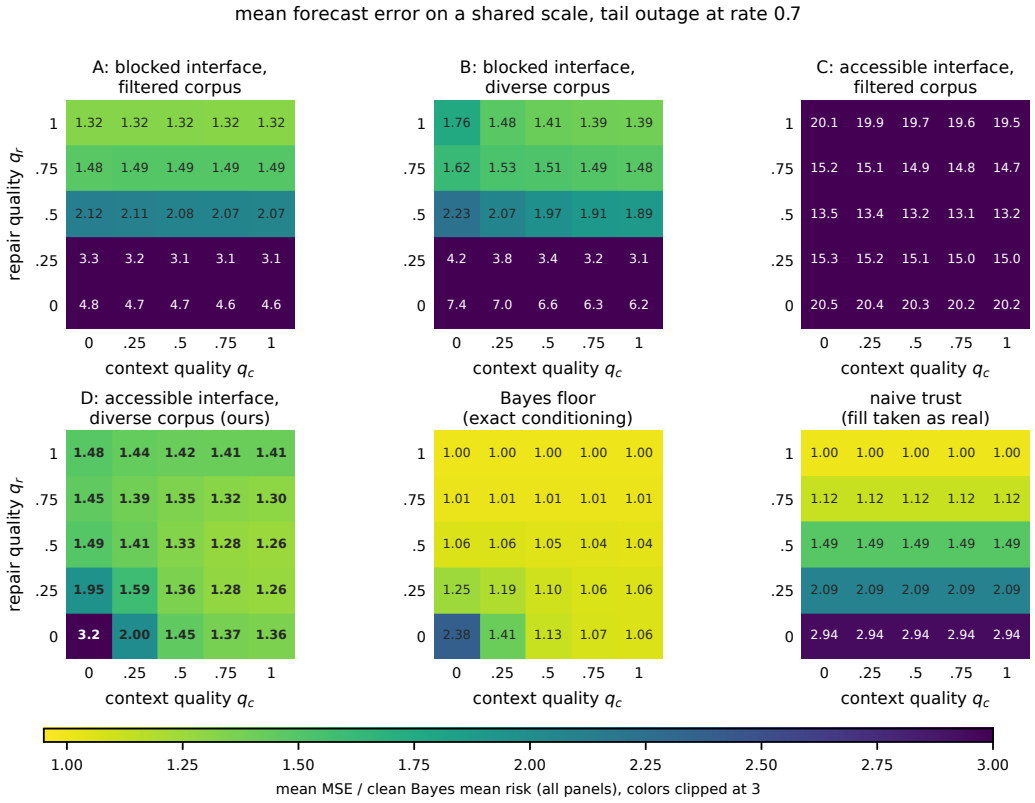

Figure 25: **Two-source responses on a shared error scale.** Mean MSE divided by clean Bayes risk on 256 synthetic two-series tasks with a 70% tail outage. Axes independently vary repair quality q_r and auxiliary-context quality q_c. Model cells average test errors over three seeds; Bayes and naive-trust references use conditional expected risks. The trained retaining arm uses both sources but remains above the conditional optimum. Colours saturate at 3; larger values are printed. Three-seed corner standard deviations, generator and reference definitions appear in Appendix G.3.

Table 30: **Fill-quality sweeps on two declared interfaces.** Both receive identical filled windows; Moirai 2.0 retains the content and Chronos-Bolt masks it before patch embedding. $\alpha = 1$ supplies true missing values. Cells use own-clean relative errors, averaged over ETTh1, ETTm1 and weather, with $L = 512$, $H = 64$ and 150 windows per cell. **Bold** marks oracle endpoints.

		Chronos-Bolt (declared)		Moirai 2.0 (declared)	
Mechanism	rate	$\alpha=0$	$\alpha=1$	$\alpha=0$	$\alpha=1$
scattered dropout	0.3	1.029	1.029	1.359	**1.312**
	0.7	1.215	1.219	1.402	**1.291**
block outage	0.3	1.020	1.024	1.071	**1.030**
	0.7	1.094	1.086	1.244	**1.072**
value censoring	0.3	1.197	1.182	1.388	1.105
	0.7	3.067	2.213	3.255	1.190

Table 31: **Relative degradation of the CPT retrofit and its controls.** Each cell is median own-clean relMSE at 70% missing: first over windows, then over four mechanism medians for each dataset; *med.* aggregates the nine datasets. Zero and linear fills are deployable inputs; oracle fills are diagnostic references. The *clean* column uses the dataset-median clean-MSE ratios defined in Table 3, from the same saved leaderboard runs; n/r marks missing matching control summaries. These differ from the separately scoped scorecards in Table 26. Arms are the native checkpoint, equal-budget missingness-free CPT (+5k-ctrl), and the retrofit (+CPT). Bold marks the lowest relative-degradation value within a family and column, not a cross-configuration absolute-error ranking. Table 4 gives MASE results on the real datasets.

model	arm	clean	ETTh1	ETTm1	weath.	ETTh2	ETTm2	electr.	traffic	exch.	illness	*med.*
perfect fill												
	vanilla	–	1.51	1.85	1.41	1.38	1.41	4.88	3.31	1.30	2.02	1.51
Chronos-Bolt	+5k-ctrl	+4.3%	1.40	1.94	1.31	1.34	1.32	4.19	3.47	1.83	2.38	1.83
	+CPT (ours)	+4.2%	**1.28**	**1.27**	**1.18**	**1.22**	**1.16**	**1.73**	**1.61**	**1.17**	**1.08**	**1.22**
	vanilla	–	1.53	1.76	1.54	1.37	1.40	4.73	3.97	1.40	2.27	1.54
Chronos-2	+5k-ctrl[†]	n/r	**1.11**	**1.24**	1.25	**1.11**	**1.17**	**1.45**	**1.46**	1.11	1.35	**1.24**
	+CPT (ours)	−0.7%	1.31	1.52	**1.23**	1.25	1.29	1.70	1.74	**1.09**	**0.99**	1.29
	vanilla	–	1.13	1.28	1.04	1.12	1.32	1.37	1.84	1.08	1.18	1.18
Moirai 2.0	+5k-ctrl	n/r	**1.09**	1.25	1.07	**1.04**	1.18	**1.34**	1.86	1.08	1.14	1.14
	+CPT (ours)	+1.1%	1.17	**1.14**	**1.03**	1.11	**1.11**	1.43	**1.55**	**1.02**	**1.10**	**1.11**
linear fill												
	vanilla	–	1.57	1.84	1.31	1.36	1.40	5.75	4.09	1.32	2.09	1.57
Chronos-Bolt	+5k-ctrl	+4.3%	1.63	1.94	1.23	1.37	**1.32**	6.30	4.56	1.64	2.45	1.64
	+CPT (ours)	+4.2%	**1.51**	**1.69**	**1.12**	**1.33**	**1.32**	**4.53**	**3.68**	**1.11**	**1.75**	**1.51**
	vanilla	–	1.52	1.75	1.38	1.36	1.38	5.12	4.18	1.46	2.34	1.52
Chronos-2	+5k-ctrl	n/r	1.50	1.81	1.28	**1.33**	1.38	4.78	4.38	1.14	2.35	1.50
	+CPT (ours)	−0.7%	**1.44**	**1.48**	**1.23**	1.34	**1.36**	**3.50**	**3.42**	**1.11**	**1.46**	**1.44**
	vanilla	–	1.60	1.90	1.28	1.44	1.58	5.18	5.42	1.20	1.77	1.60
Moirai 2.0	+5k-ctrl	n/r	1.61	1.76	1.34	1.38	1.35	5.02	5.22	1.16	**1.69**	1.61
	+CPT (ours)	+1.1%	**1.49**	**1.66**	**1.14**	**1.34**	**1.29**	**4.08**	**4.53**	**1.09**	1.90	**1.49**
zero fill												
	vanilla	–	2.21	2.54	3.51	3.34	3.79	14.39	5.30	46.48	2.88	3.51
Chronos-Bolt	+5k-ctrl	+4.3%	4.74	6.76	9.32	13.95	16.88	54.41	9.80	393	4.81	9.80
	+CPT (ours)	+4.2%	**1.70**	**1.96**	**1.56**	**2.03**	**2.26**	**5.74**	**3.84**	**7.18**	**2.29**	**2.26**
	vanilla	–	**1.67**	1.86	2.10	1.68	1.73	**5.99**	4.36	5.02	2.53	2.10
Chronos-2	+5k-ctrl	n/r	7.38	10.78	8.92	29.37	25.57	130	20.16	182	7.07	20.16
	+CPT (ours)	−0.7%	1.68	**1.57**	**1.66**	**1.58**	**1.51**	8.89	**4.25**	**2.70**	**1.61**	**1.66**
	vanilla	–	6.87	10.79	5.45	11.56	12.58	82.20	22.26	13.96	4.81	11.56
Moirai 2.0	+5k-ctrl	n/r	6.92	10.10	5.20	11.94	12.53	79.11	20.39	12.87	4.60	11.94
	+CPT (ours)	+1.1%	**2.45**	**3.35**	**1.86**	**2.92**	**2.64**	**21.72**	**10.57**	**2.99**	**2.96**	**2.96**

[†]Chronos-2's control reads a perfect fill through an untrained content-passing interface: the fill is the ground truth, so the excess closes mechanically, not by learning (Table 26).

I.3 Attention and permutation responses after CPT

Permuting fills preserves their multiset while changing their temporal arrangement. The forecast panels therefore test position-specific response after CPT, while the attention panels locate accompanying changes in selected attention shares (Section 5). Attention summaries locate internal changes accompanying the position-specific forecast sensitivity established by the intervention.

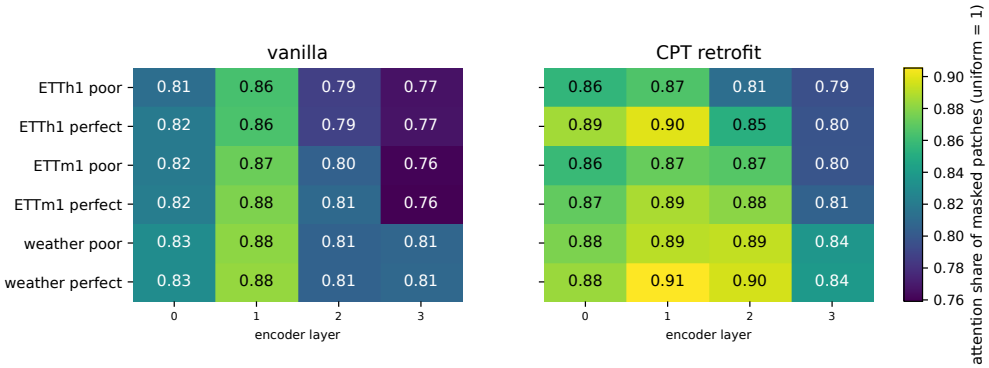

Figure 26: Attention share received by patches containing masked points, relative to uniform, per encoder layer, vanilla Chronos-Bolt tiny (left) vs the CPT retrofit (right), at the two ends of the fill-quality axis; median over windows, block outages at rate 0.5.

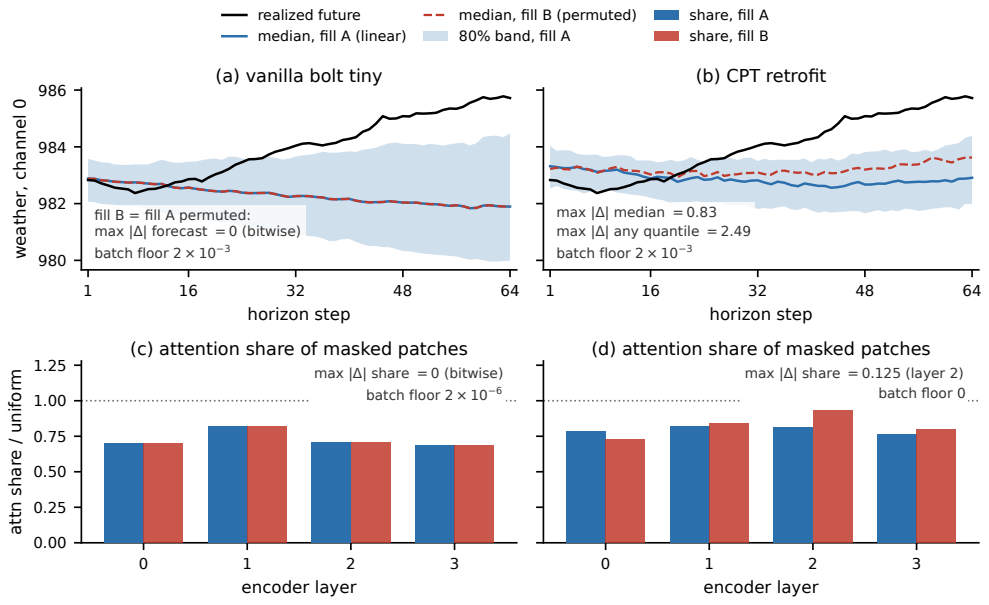

Figure 27: The permutation test on a single window (weather, channel 0; block outages at rate 0.5; L=512, H=64; the declared path: linear fill plus observation mask; vanilla Chronos-Bolt tiny vs the CPT retrofit). Fill B permutes the 198 fill values among the missing positions, preserving their mean and variance bitwise. (a) The vanilla model's forecast is bitwise identical under the two fills (max $|\Delta| = 0$ across all nine quantiles; the batch-composition float floor is 2×10^{-3}), and (c) its per-layer attention share of the masked patches is unchanged. (b) The CPT retrofit reads the fill: the same permutation moves its median by up to 0.83 (any quantile 2.49) and (d) shifts the attention shares by up to 0.125. The retrofit breaks permutation invariance on this window despite unchanged statistics.