Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Auditing Frozen Time-Series Foundation Models Under Informative Context Missingness

Muhammetalp Erdem

muhammetalperdem@trabzon.edu.tr

Trabzon University

Research Article

Keywords: time-series forecasting, missing data, foundation models, forecast evaluation, uncertainty quantification

Posted Date: September 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10871922/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Auditing Frozen Time-Series Foundation Models Under Informative Context Missingness

Muhammetalp Erdem[1*]

[1*]Department of Computer Engineering, Trabzon University, Trabzon, Türkiye.

Corresponding author(s). E-mail(s): muhammetalperdem@trabzon.edu.tr;

Abstract

Time-series foundation models are usually evaluated on complete contexts, although operational histories are incomplete, structured, and mechanism dependent. We present a pre-registered audit protocol for forecasting under informative context missingness and apply it to two frozen backbones, Chronos-T5-small and TimesFM 2.5, across 12 datasets, 48 series, three missingness rates, and five seeds. The protocol holds out the test mask generators from adapter fitting, fixes the primary rate, contrast, aggregation, and correction family before analysis, and carries the dataset–series–origin hierarchy into every interval and test. Its pre-registered contrast between mechanism-balanced and average-risk adaptation favours balancing for both backbones ($\mathbf{-0.74}$, exact $\boldsymbol{p = 0.065}$ for Chronos-T5-small; $\mathbf{-2.41}$, Holm-adjusted $\boldsymbol{p = 0.0020}$ for TimesFM 2.5). The audit's more consequential finding is what that contrast cannot show. Against the no-adaptation anchor – a method the protocol registered but the confirmatory family never tested – all seven learned adaptation objectives score worse than leaving the backbone alone, on 11 of the 12 datasets individually for every objective and both backbones (sign test $\boldsymbol{p = 0.006}$). We further show that the twelfth dataset violates the normalisation the protocol assumes, carrying roughly forty times the weight of any other in an unweighted mean; excluding it on that data-side criterion makes the anchor gap unambiguous ($\mathbf{+7.4}$ and $\mathbf{+7.7}$, both $\boldsymbol{p = 0.001}$). Mechanism balancing is therefore the least harmful member of a set of harmful interventions rather than an improvement, and comparisons between adaptation objectives are uninterpretable without an explicit no-adaptation anchor.

Keywords: time-series forecasting, missing data, foundation models, forecast evaluation, uncertainty quantification

1 Introduction

Large pretrained forecasting models have broadened zero-shot and lightly adapted time-series prediction Ansari et al. (2024); Das et al. (2024). Their practical deployment, however, often starts from a partially observed context: sensor outages, delayed reports, and event-driven recording may make missingness informative rather than a benign random deletion. Missing-data theory distinguishes mechanisms because the observation process can change what is learnable from an observed history Rubin (1976). Yet forecasts under incomplete contexts are frequently assessed with a single synthetic mechanism or with point accuracy alone.

This paper began from a deliberately narrow question: when a frozen time-series foundation model is equipped with a small global adapter, does a mechanism-balanced adaptation curriculum improve probabilistic forecasts under informative missingness relative to ordinary average-risk adaptation? We fixed the primary rate, contrast, aggregation, and correction family before looking at any result, and registered ten methods, of which the confirmatory family tested two.

That design produced a clean answer to the question we asked and a more useful answer to one we had not. Balanced adaptation does score better than average-risk adaptation. But the protocol also registered a `frozen` method – the backbone applied directly to the incomplete context, with no adapter at all – and once that anchor is analysed, both adaptation objectives are worse than doing nothing, as are the five secondary objectives. The confirmatory contrast is a comparison between two ways of making the forecast worse. It is not visible as such from inside the confirmatory family, because that family compares the two objectives only with each other.

We report this because it generalises past our own study. A contrast between two variants of an intervention is uninformative about whether the intervention helps, yet it is the contrast that pre-registration protects and the one a reader is most likely to see. An audit protocol that does not force the no-intervention anchor into the reported family will keep producing well-powered, correctly-corrected, and misleading comparisons.

Our contributions are the following.

1. **A mechanism-aware audit protocol** (specified in full in Appendix A) for incomplete-context forecasting: exact-count mask generators with a hidden-value causal canary, a test mask bank whose generators never enter adapter fitting, chronologically disjoint origins, and a dataset–series–origin inferential hierarchy carried into every reported interval (Sections 3 and 4).
2. **A no-adaptation anchor as a required reported quantity.** We show that all seven learned adaptation objectives are worse than the frozen backbone on 11 of 12 datasets for both backbones, and we make the anchor a build-time requirement of the released artifact rather than an optional courtesy (Section 5.2; the claim-to-evidence mapping is Appendix C).
3. **An exact cluster-level test and a direction count.** Twelve dataset clusters admit only 2^{12} sign assignments, so we enumerate them rather than sampling, and

we pair the mean-based test with a cluster direction count that a single dominant cluster cannot overturn (Section 4.4).

4. **A normalisation-validity audit** that recomputes every series' scale denominator from source and identifies where the pre-registered normalisation fails. One dataset carries roughly forty times the weight of any other in an unweighted mean, which is enough to reverse the apparent ranking of every baseline (Section 5.3).

These address four recurrent weaknesses in incomplete-context evaluation. A fixed missing rate does not fix the information removed: eight isolated observations, a periodic phase, and an event-aligned segment may all carry different predictive content. A large collection of forecast horizons does not create an equally large number of independent experimental units: horizons share a context, origins overlap in time, and series share a dataset. An attractive point estimate does not establish transfer: the uncertainty procedure must follow the data hierarchy and the confirmatory family must be fixed in advance. Finally, calibrated and uncalibrated intervals answer related but different questions, so we report the raw probabilistic score as the primary endpoint and validation-only conformal calibration as a secondary diagnostic.

The study was built to stay informative if its headline hypothesis failed, and that is what happened. It identifies where mechanism balancing does not transfer, whether explicit mask features account for an apparent gain, whether robust objectives behave similarly, whether the result persists outside synthetic masks, and – through the anchor – whether the whole family of interventions is worth deploying. We therefore keep confirmatory, anchor, robustness, diagnostic, task-compatible, and natural-missingness families separate throughout rather than pooling them into one leaderboard.

2 Related work

Four literatures meet in this study: general-purpose forecasting backbones, missingness-aware sequence modelling, robust training objectives, and the evaluation machinery – proper scoring, distribution-free calibration, and clustered inference – that decides what a comparison between them is allowed to claim.

2.1 Time-series foundation models

Chronos tokenises scaled and quantised series and trains a language-model backbone over them, while TimesFM is a decoder-only patched forecaster with a native quantile head Ansari et al. (2024); Das et al. (2024). Moirai adds any-variate attention and multiple patch sizes for cross-frequency transfer, and MOMENT provides an openly released family spanning several analysis tasks Woo et al. (2024); Goswami et al. (2024). These models descend from the long-horizon transformer line that Informer represents Zhou et al. (2021), but their defining property for our purposes is that a single pretrained set of weights is applied across domains without per-dataset fitting.

That property is what makes the evaluation question sharp. GIFT-Eval was introduced precisely because heterogeneous ad-hoc protocols made cross-model claims hard to compare Aksu et al. (2024), and our panel is drawn from its source collection. What such benchmarks assume, and what we stress here, is a complete context: the model

is handed an uninterrupted history and asked for the future. We keep the backbones frozen and vary only what surrounds them, so that any difference is attributable to the intervention rather than to the forecaster.

2.2 Missingness as a mechanism, not a rate

Rubin's taxonomy separates missing completely at random, missing at random, and missing not at random by how the observation process relates to observed and unobserved values, and the distinction matters because it changes what is identifiable from an observed history Rubin (1976); Little and Rubin (2019). Deep sequence models have taken the same point empirically. GRU-D feeds masks and elapsed-time gaps directly into the recurrence on the grounds that missingness patterns are themselves informative, BRITS imputes bidirectionally while treating imputed values as trainable, and mTAN learns continuous-time embeddings for irregular sampling Che et al. (2018); Cao et al. (2018); Shukla and Marlin (2021).

A separate line treats reconstruction as the goal. GAIN adapts adversarial training to imputation, SAITS uses self-attention with a joint reconstruction and imputation objective, and TSI-Bench standardises the comparison across many such methods Yoon et al. (2018); Du et al. (2023, 2024). More recently the foundation-model idea has been carried into imputation itself: MoTM builds a continuous-time imputation model and a follow-up study asks directly whether time-indexed foundation models are the right vehicle for it Le Naour et al. (2025, 2026). Forecasting architectures designed for incomplete multivariate input, such as S4M and CoIFNet, occupy a third position by refusing to separate the two stages Peng et al. (2025); Tang et al. (2025).

Our setting sits deliberately outside all three. We neither reconstruct the context as a goal in itself nor train a missingness-aware forecaster; we hold a frozen backbone fixed and vary only the objective of a small global adapter placed around it. That makes the comparison narrow, and – as our anchor result shows – it also makes it easy to lose track of whether the surrounding intervention is worth anything at all.

2.3 Incomplete contexts as a distribution-shift problem

Classical missing-data terminology is useful but does not by itself specify a forecasting stress test. In forecasting, the target lies after the incomplete context, so a mask alters the effective recency, seasonal coverage, and representation supplied to the forecaster. Two masks satisfying the same rate and the same broad label can remove very different lag structures: eight isolated deletions, a deleted periodic phase, and an event-aligned burst are not interchangeable. We therefore operationalise mechanisms through explicit frozen generators and pair every informative construction with a count-matched control. The labels describe generators in this benchmark and are not claims that a real-world observation process has been causally identified.

Separating the training mask bank from the test mask bank makes this an out-of-distribution evaluation rather than an interpolation. The adapter is fitted on isolated deletions, strict observed-history MAR, and geometric bursts; it is evaluated on clean contexts and on periodic or event-focused masks whose generators never entered fitting. This is the same concern that motivates test-time adaptation work, where the

question is whether a correction estimated under one regime survives into another, and where evaluation protocol turns out to matter as much as method: TAFAS adapts a forecaster online under non-stationarity, and a systematic evaluation of test-time adaptation under compute constraints found that reported rankings change substantially once the protocol is held fixed Kim et al. (2025); Alfarra et al. (2024). Our adapters are fitted offline rather than at test time, but the failure mode we report – a correction that helps in one regime and hurts across a panel – is of that family.

2.4 Robust objectives

Mechanism balancing is one member of a family of reweighting objectives. Equal weighting changes a known sampling measure; group DRO optimises worst-group risk over predefined groups; CVaR targets the high-loss tail irrespective of group identity Sagawa et al. (2020); Rockafellar and Uryasev (2000). The choice of robustness set is not a detail: uniform performance guarantees depend on which distributions the objective is asked to be robust against Duchi and Namkoong (2021). We therefore treat group DRO and CVaR as distinct registered controls rather than as alternative implementations of the same idea, so that a gain specific to the balanced curriculum can be told apart from a generic consequence of emphasising difficult episodes.

2.5 What the evaluation machinery has to supply

We report quantile forecasts under the interval score because proper scoring rules assess sharpness and calibration jointly rather than separately Gneiting and Raftery (2007), and because the interval score has an operational decomposition into width and miss penalties that makes a regression in one component visible Bracher et al. (2021). The quantile levels themselves follow the pinball-loss convention Koenker and Bassett (1978). MASE is reported as a complementary scale-free point diagnostic Hyndman and Koehler (2006), with its seasonal-naive denominator estimated per series as in the M-competition protocols Makridakis et al. (2020); it never substitutes for the primary probabilistic estimand.

Coverage is a separate question from score. Conformal prediction supplies finite-sample distribution-free guarantees under exchangeability Vovk et al. (2005); Angelopoulos and Bates (2023), which time series violate, and a body of work addresses that violation directly – through adaptive online correction under distribution shift, through conformal procedures built for multi-horizon forecasting, and through sequential re-estimation of the residual quantile Gibbs and Candès (2021); Stankevičiūtė et al. (2021); Xu and Xie (2023). Our calibration step is deliberately the simplest member of this family: an additive offset from validation residuals only, per method, backbone, rate, and horizon. It is reported to show whether coverage can be repaired without target leakage, never to replace the raw primary result or to select the primary method.

Finally, the inference has to respect the design. Horizons share a context, origins overlap in time, and series share a dataset, so treating them as independent units inflates precision. We resample datasets, then series, then contiguous origin blocks,

following the block-bootstrap treatment of dependent sequences and the cluster-bootstrap literature for nested data Politis and Romano (1994); Field and Welsh (2007), and we sign-flip at the highest independent level, where cluster-robust practice warns that a small number of clusters is the binding constraint Cameron et al. (2008). Multiplicity across the two confirmatory tests is controlled by Holm's step-down procedure Holm (1979). Section 5.2 shows that with twelve clusters this machinery, correctly applied, can still be the wrong instrument for the question a reader will actually ask.

3 Problem formulation and methods

For an observed context $x_{1:T}$ with observation mask $m_{1:T}$ and a future target $y_{T+1:T+H}$, a frozen backbone produces quantiles $(q_{0.1}, q_{0.5}, q_{0.9})$. Chronos-T5-small uses retained forecast trajectories to estimate these quantiles, while TimesFM 2.5 supplies them through its continuous quantile head. Point comparisons use $q_{0.5}$.

3.1 Episodes and prediction contract

An episode is indexed by dataset d, series s, chronological origin o, split v, mechanism z, missing rate r, and mask identifier k. The context length is $L = 80$ and the forecast horizon is $H = 24$. The clean context precedes the origin, while the target is the next 24 observations. A binary vector $m \in \{0, 1\}^L$ records missing context positions. Targets are never masked in the synthetic experiment, and natural-missingness episodes are admitted only when the complete target horizon is observed.

All forecasts follow the `q1-predictions-v2` contract, whose fields are listed in Supplementary Table S1. In addition to episode indices, every row records backbone, method, seed, horizon, target, comparable point forecast, optional native point, $q_{0.1}$, $q_{0.5}$, $q_{0.9}$, calibration status, interval source, MASE and interval-score denominators, seasonal lag, model and adapter revisions, runtime, run identifier, and data and panel hashes. Quantile validity is enforced through $q_{0.1} \leq q_{0.5} \leq q_{0.9}$. This explicit contract makes it possible to verify that two methods used the same target, mask, denominator, and model revision.

Chronos generates 100 retained trajectories for every eligible context. Empirical trajectory quantiles define the reported distribution; the median is used for comparable point error. TimesFM supplies the same three quantiles through its continuous quantile head, and its native point is retained separately. Adapter methods reuse the appropriate frozen output cache, preventing their comparison from being contaminated by different stochastic backbone draws.

3.2 Missingness representation

The global adapter receives a representation derived from the incomplete context, never from a hidden value. Available components include deterministic context imputation, binary mask, age since last observation, missing rate, and summary features. The `context_no_explicit_mask` control removes mask, age, and missing-rate features. It is not strictly mask-independent because interpolation remains a function of the observation pattern. The mask/age-only control takes the complementary view:

it removes direct context values while retaining missingness geometry and the cached forecast. Together, the ablations distinguish reliance on value context from reliance on explicit observation-process features without changing the target episodes.

3.3 Adaptation objectives

Each global adapter is trained at the level *backbone* × *missing-rate* × *method* × *seed* and is then applied to every evaluation series. This avoids fitting a distinct adapter to each test series. Average adaptation minimizes the pooled training loss. Mechanism-balanced adaptation gives equal mechanism-level contribution to the curriculum. We additionally evaluate MCAR-only adaptation, canonical exponentiated-gradient group DRO, Rockafellar–Uryasev CVaR, mask/age-only features, and a context-only ablation. The latter omits explicit mask, age, and rate features, although deterministic imputation can still convey indirect missingness information.

Let $\ell_i(\theta)$ be the adapter loss on episode i, and let G_z contain training episodes generated by $z \in \{\text{MCAR}, \text{MAR}, \text{burst}\}$. Average adaptation minimizes

$$\mathcal{J}_{\text{avg}}(\theta) = N^{-1} \sum_{i=1}^{N} \ell_i(\theta), \tag{1}$$

whereas balanced adaptation minimizes

$$\mathcal{J}_{\text{bal}}(\theta) = \frac{1}{3} \sum_{z} |G_z|^{-1} \sum_{i \in G_z} \ell_i(\theta). \tag{2}$$

Equations 1 and 2 therefore differ only in the measure used to aggregate the same mechanism-labelled training bank. It is not a comparison between different adapter capacities or between per-series and global fitting.

For group DRO, mechanism weights are updated by canonical exponentiated-gradient steps so that mechanisms with larger current loss receive more mass. CVaR uses the Rockafellar–Uryasev construction to emphasize the high-loss tail irrespective of mechanism identity. MCAR-only adaptation restricts training to MCAR episodes. These remain secondary controls and are never substituted for the balanced method after results are observed.

3.4 Baselines and the no-adaptation anchor

Three of the ten registered methods fit no global adapter, and what each one controls for is different.

`frozen` is the *no-adaptation anchor*: the backbone is given the incomplete context directly and its output is scored unchanged. It shares the episodes, masks, targets, seasonal lags, and denominators of every other method, and because backbone outputs are cached, it consumes the identical forecast draws. It is therefore the reference against which any adaptation must justify its existence, and the quantity that tells a reader whether the intervention helps at all rather than which variant of it helps most. An earlier draft of this manuscript described `frozen` in its method list and reported no

number for it anywhere; the released artifact now fails its own verification if that recurs.

`linear_interpolation` controls for the possibility that a trivial deterministic repair of the context is sufficient, separating the value of a learned intervention from the value of any intervention.

`saits_imputation` is the learned-imputation comparator. SAITS is fitted per dataset, rate, and seed on training masks only, with epoch selection on validation imputation loss, and its imputed test contexts are passed to both backbones. We fit it univariately because the common protocol is univariate, which is not the multivariate regime SAITS was designed for; its results should be read as a competent baseline under our task definition rather than as a tuned upper bound for the method.

S4M is not forced into the common interface at all. It is trained from pinned official source on the multivariate ETT1 and Jena Weather subsets with seven and 21 channels, and because it emits point forecasts its $q_{0.1}$ and $q_{0.9}$ come from validation-residual conformal calibration. It is accordingly labelled secondary and task-compatible rather than ranked against the univariate family.

3.5 Metrics

The primary outcome is the normalized interval score at 20% missingness. Its denominator and the MASE denominator are estimated from each series' training prefix using the official seasonal-naive lag, or pre-registered lag 1 when no seasonal definition is available. Lower normalized interval score is better.

For nominal miscoverage $\alpha = 0.2$, endpoints l, u, and target y, the interval score is

$$S_\alpha(l, u; y) = (u - l) + \frac{2}{\alpha}(l - y)\mathbb{1}\{y < l\} + \frac{2}{\alpha}(y - u)\mathbb{1}\{y > u\}. \tag{3}$$

It rewards sharp intervals while penalizing the distance of misses. Each series is normalized by a scale estimated only from its training prefix. Official-frequency lags are 1,440 for minute, 24 for hourly, seven for daily, 52 for weekly, 12 for monthly, and four for quarterly data. Undefined frequencies use the pre-registered lag-1 fallback. Non-finite or degenerate denominators fall back to one under the frozen validation rule.

MASE uses $q_{0.5}$ as the common point forecast. Coverage is the fraction of targets between $q_{0.1}$ and $q_{0.9}$, and width is $q_{0.9} - q_{0.1}$. These are secondary diagnostics. An interval-score improvement need not imply a MASE improvement; in fact, at 20% the balanced-minus-average MASE differences are slightly positive for both backbones.

3.6 Validation-only calibration

For each method, backbone, rate, and horizon, validation nonconformity is

$$a_i = \max\{q_{0.1,i} - y_i,\ y_i - q_{0.9,i},\ 0\}. \tag{4}$$

The higher empirical 0.8 quantile of these residuals supplies an additive offset. The lower endpoint is decreased and the upper endpoint increased by that amount. Test

targets never enter offset estimation. Since calibration changes width and coverage after a raw interval exists, calibrated scores are secondary and cannot overwrite the raw primary endpoint.

4 Experimental design

The synthetic panel contains 12 datasets, 48 series, and seven application-domain labels (economic/financial, energy, healthcare, nature, sales, transport, and web/cloud operations). Each series records five training, three validation, and eight chronological test origins. Training and validation masks use MCAR, strict observed-history MAR, and exact-k truncated-geometric burst masks. Test masks are held out from this bank and comprise clean, periodic informative/control, and event informative/control families. Every mask removes exactly round($80r$) observations at rate $r \in \{0.1, 0.2, 0.4\}$.

4.1 Panel composition and chronology

Table 1 summarizes the immutable panel. Four series were selected from each dataset, creating equal dataset-level representation without treating series as independent datasets. Lengths range from 208 to 17,420 observations and frequencies range from hourly to monthly. The primary analysis does not weight a long series more heavily because it contains more raw observations: every series contributes the same eight test origins.

Table 1 Immutable synthetic-panel inventory. Every dataset contributes four series and the same 5/3/8 train/validation/test origin structure.

Dataset	Domain	Freq.	Series	Min. len.	Max. len.
M4 Daily	Econ./financial	daily	4	220	4,211
M4 Hourly	Econ./financial	hourly	4	748	1,008
Electricity	Energy	weekly	4	208	208
ETT1	Energy	hourly	4	17,420	17,420
COVID deaths	Healthcare	daily	4	212	212
Bitbrains fast storage	Web/cloud ops	hourly	4	721	721
Jena Weather	Nature	daily	4	366	366
Temperature/rain	Nature	daily	4	725	725
Restaurant	Sales	monthly	4	296	478
Hierarchical sales	Sales	weekly	4	260	260
LOOP Seattle	Transport	daily	4	365	365
M-DENSE	Transport	daily	4	730	730

Origins are recorded in increasing chronological order after reserving an 80-step context and 24-step target. Training, validation, and test origin identities are disjoint. Normalization and denominators are fit on a training prefix only. The manifest retains

native identifier, selected dimension, frequency, length, split origins, scale, source path, and data hash, permitting reconstruction without relying on incidental row order.

4.2 Mask generators and leakage controls

For MCAR, exactly k positions are sampled without replacement. The MAR generator proceeds sequentially. At each time it calculates a decision probability from earlier values that remained observed, accounts for the number of positions still required to reach k, and appends the current value to history only if it remains observed. A hidden-value canary generates a mask, drastically perturbs exactly the hidden values, and regenerates the mask with the same seed. Equality is required. This operational check is stronger than merely labeling a generator "MAR."

The burst generator repeatedly draws a start and a geometric run length with probability 0.25, adds currently available forward positions, and truncates its union at exactly k. Periodic-informative masks derive a phase modulo eight from maximal strictly observed-history activity; periodic controls select the phase randomly. Event-informative masks rank the observed-history activity signal; event controls rank random scores. Both informative/control pairs have the same missing count. Their geometry comparisons are diagnostic rather than causal identification because equal counts do not make every geometric feature identical.

Table 2 summarises the role and principal safeguard of each bank. The panel manifest contains 9,216 mask rows. Clean masks are repeated by contractual rate but hide zero values. At rates 0.1, 0.2, and 0.4, every non-clean mask hides exactly eight, 16, and 32 context positions. Train/validation and test mask identities are disjoint, and no periodic or event generator is admitted to adapter fitting.

Table 2 Mask-bank roles and safeguards.

Bank	Mechanisms	Principal safeguard
Training	MCAR; observed-history MAR; geometric burst	exact-k count and causal MAR canary
Validation	same three training families	distinct IDs and validation-only selection
Test	clean; periodic informative/control; event informative/control	generators absent from fitting bank
Natural	realized incomplete context	complete future target; chronological isolation

SAITS is trained only with training masks at each dataset–rate–seed and is used to impute test contexts before forecasting by each frozen backbone. S4M is trained from its fixed official implementation on the task-compatible multivariate subsets of ETT1 and Jena Weather; its conformal intervals are therefore labelled secondary and

task-compatible. Natural-missingness experiments use chronologically later, complete-target episodes with naturally incomplete context from UCI Household Power and *temperature_rain_with_missing*; no synthetic-method retuning occurs.

4.3 Campaign organization and compute discipline

The Supplementary Information records the failure and repair ledger behind these choices. The campaign separates cached backbone inference, global adapter fitting, adapter evaluation, SAITS fit/evaluation, S4M fit/evaluation, natural-missingness evaluation, calibration, and analysis. Shards are append-only. A job moves through pending, running, complete, or failed states, and restart logic skips a completed shard only after hash validation. This design allowed interrupted Colab sessions to resume without rerunning completed scientific work.

The predeclared budget was at most 40 Colab Pro+ GPU-hours, with 34 hours for new jobs and six reserved for infrastructure or out-of-memory retry. Profiling separated backbone episodes, adapter fitting/evaluation, SAITS, and S4M. Batching and caching could be optimized if projections exceeded budget, but the panel, five seeds, backbones, and methods could not be narrowed after observing results. The finished campaign contains 711 completed jobs, 494 prediction shards, and 26,105,469 raw prediction rows. The analysis source contains 978,864 episode-metric rows and is identified by SHA-256 `2f9d1e384a28ba2bb493c31c60178ec4e916b5910e00ee8a1141eb899cddf5e0`.

For the primary family, seeds are averaged within cells, periodic and event informative conditions are combined at the series–origin level, and inference preserves nested dataset → series → contiguous-origin blocks. We use 10,000 hierarchical bootstrap replicates and a 10,000-repetition dataset-cluster sign-flip permutation test. Holm correction is applied across the two backbone-specific primary tests. Horizon-wise additive conformal offsets learned on validation traces yield a pre-specified calibrated secondary view; the primary interval score is raw.

4.4 Inferential hierarchy

The hierarchy is part of the estimand; Appendix B restates it as a checklist and the Supplementary Information lists the products each step writes. First, five seeds are averaged within each dataset–series–origin–backbone–method–mechanism–rate–calibration cell. Seed repetitions therefore cannot inflate sample size. Periodic-informative and event-informative scores are then averaged equally within each series–origin, and the balanced-minus-average difference is formed on matched cells. At the primary rate, each backbone contributes $12 \times 4 \times 8 = 384$ paired origin contrasts.

The bootstrap samples datasets with replacement, series with replacement within a sampled dataset, and circular moving blocks of ordered origins within a sampled series. With eight origins, the frozen rule $\mathrm{round}(\sqrt{8})$ gives block length three. The 95% interval is the percentile interval over 10,000 replicates. The sign-flip permutation is conservative at the highest independent level: one Rademacher sign is assigned per dataset, preserving all nested series and origin dependence. An earlier analysis that flipped origins independently produced an overly small Chronos p-value and is explicitly superseded by this v3 analysis.

Holm correction covers exactly the two backbone-specific primary tests. Method controls, missing-rate sensitivity, geometry, calibration, natural missingness, and S4M remain separate secondary families. We do not move a favorable secondary comparison into the confirmatory family after inspection.

5 Results

We report five families and never merge them: the pre-registered confirmatory contrast, the no-adaptation anchor, a normalisation-validity analysis with its sensitivity view, robustness analyses at the two non-primary rates, and diagnostics. Only the first carries the registered Holm correction.

5.1 Confirmatory contrast

Table 3 reports the pre-registered balanced-minus-average contrast at 20% missingness on raw intervals. Negative values favour mechanism-balanced adaptation. Twelve dataset clusters admit only $2^{12} = 4096$ distinct sign assignments, so we enumerate all of them instead of sampling; the exact p-values agree with the 10,000-replicate Monte-Carlo values of the sealed analysis to within 0.001.

Table 3 Pre-registered balanced-minus-average contrast at 20% missingness, raw intervals. Negative favours mechanism-balanced adaptation. The sign test counts how many of the twelve dataset clusters point the same way.

Backbone	Effect	95% CI	Exact p	Holm p	Sign test
Chronos-T5-small	−0.743	[−1.573, 0.563]	0.0649	0.0649	10/12, 0.039
TimesFM 2.5	−2.409	[−5.559, −0.835]	0.0010	0.0020	11/12, 0.006

Figure 1a shows the contrast against the anchor of Fig. 1b. The direction favours balanced adaptation for both backbones and the statistical support differs. The Chronos-T5-small interval crosses zero; the TimesFM 2.5 interval does not and its Holm-adjusted result is $p = 0.0020$. A leave-one-dataset-out jackknife shows neither estimate rests on a single cluster: the Chronos effect ranges over $[-0.99, -0.53]$ and the TimesFM effect over $[-2.63, -1.46]$ across the twelve deletions.

The aggregate effect hides offsetting changes in the score components. For Chronos, balanced adaptation raises coverage by 0.032 and narrows intervals by 14.96 scale units relative to average adaptation, but raises MASE by 0.111. For TimesFM, coverage rises by 0.050 and width falls by 4.75 while MASE rises by 0.241. The contrast is therefore not a point-accuracy gain; it reflects the interval score's joint treatment of width and misses, which is why that score was fixed as primary rather than chosen after the fact.

Table 4 separates the two held-out informative families. Under Chronos, balanced adaptation moves raw interval score from 48.95 to 48.21 on event masks and from 43.96 to 43.21 on periodic masks; under TimesFM the corresponding changes are 46.95 to 44.41 and 41.44 to 39.16. The direction is consistent across both held-out families,

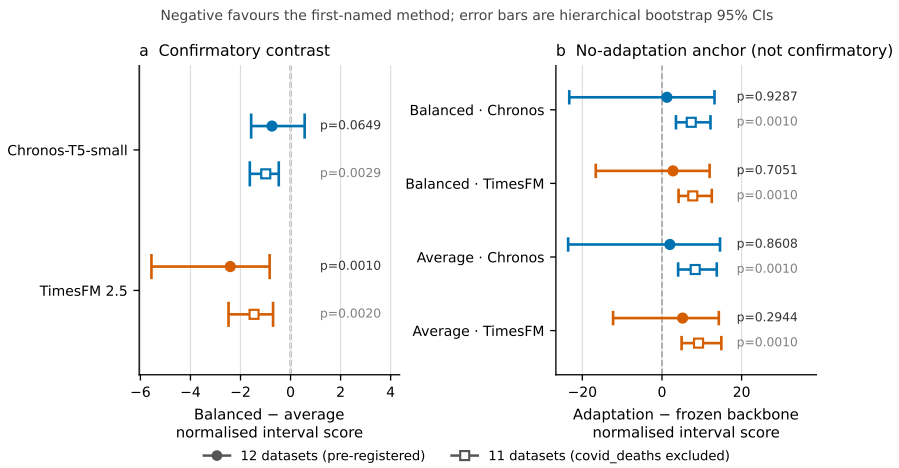

Fig. 1 Left: the pre-registered balanced-minus-average contrast. Right: the no-adaptation anchor of Section 5.2, on its own scale. Filled circles are the pre-registered twelve-dataset estimates and open squares the eleven-dataset sensitivity view of Section 5.3; error bars are hierarchical bootstrap 95% confidence intervals and p-values are exact.

but the inferential unit remains the dataset-level hierarchy rather than these four aggregate entries.

Table 4 Raw performance at 20% on held-out informative masks, all twelve datasets. Values are means over seed-averaged series–origin cells. Lower score and MASE are better; nominal coverage is 0.80.

Backbone	Method	Mechanism	Interval score	MASE	Coverage	Width
Chronos	Average	Event	48.948	7.428	0.579	3563.6
Chronos	Balanced	Event	48.212	7.534	0.611	3543.8
Chronos	Average	Periodic	43.961	6.581	0.585	3572.2
Chronos	Balanced	Periodic	43.212	6.698	0.617	3562.1
TimesFM	Average	Event	46.951	9.591	0.658	3533.8
TimesFM	Balanced	Event	44.414	9.822	0.708	3520.1
TimesFM	Average	Periodic	41.438	8.552	0.665	3539.3
TimesFM	Balanced	Periodic	39.159	8.802	0.716	3543.5

5.2 The no-adaptation anchor

The confirmatory family compares two adaptation objectives with each other. It cannot say whether adapting is better than not adapting, because that comparison is not in it. The protocol registered `frozen` – the backbone applied directly to the incomplete context, with no adapter (Section 3.4) – as one of its ten methods, on exactly the

same episodes, masks, targets, and denominators. We analyse it here with the identical hierarchical bootstrap and exact cluster sign-flip procedure, and label the family *anchor*: these contrasts were pre-registered as methods but not as confirmatory tests, so they carry their own Holm correction and never enter the registered one.

Table 5 places every registered method on the primary estimand's own cells. Reading the first column, the ordering is not the one the confirmatory contrast implies. The frozen backbone is the best method for both backbones; linear interpolation and SAITS imputation follow; the seven learned adaptation objectives occupy the bottom of the table. Balanced adaptation is the best of those seven, which is what the confirmatory contrast detected. Figure 2 shows the same ordering: the three methods that fit nothing are the three best methods.

Table 5 Every pre-registered method at 20% missingness on raw intervals, over the held-out informative families. The first numeric column is the pre-registered twelve-dataset mean; the second excludes `covid_deaths` on the data-side criterion of Section 5.3; the third is the gap to the frozen backbone in that view. Lower is better throughout.

Backbone	Method	12 datasets	11 datasets	− frozen
Chronos	Frozen backbone	44.47	6.22	—
Chronos	Linear interpolation	47.22	7.94	+1.72
Chronos	SAITS imputation	87.88	8.72	+2.50
Chronos	Mask/age-only	54.01	9.17	+2.95
Chronos	Balanced adaptation	45.71	13.57	+7.35
Chronos	CVaR	47.87	14.46	+8.24
Chronos	Average adaptation	46.45	14.56	+8.35
Chronos	No explicit mask	48.43	15.36	+9.14
Chronos	MCAR-only adaptation	48.99	15.65	+9.44
Chronos	Group DRO	50.22	16.34	+10.12
TimesFM	Frozen backbone	39.02	7.09	—
TimesFM	Linear interpolation	38.99	7.09	+0.01
TimesFM	SAITS imputation	62.17	8.64	+1.56
TimesFM	Mask/age-only	52.46	10.63	+3.54
TimesFM	Balanced adaptation	41.79	14.82	+7.73
TimesFM	Average adaptation	44.19	16.28	+9.19
TimesFM	No explicit mask	47.86	16.85	+9.77
TimesFM	Group DRO	45.52	17.03	+9.94
TimesFM	CVaR	47.34	17.36	+10.27
TimesFM	MCAR-only adaptation	43.58	17.41	+10.32

Table 6 gives the anchor contrasts with full uncertainty. On the pre-registered twelve datasets the mean-based evidence is inconclusive: every interval is wide and no exact p approaches significance. That width has a single cause. The per-cluster means

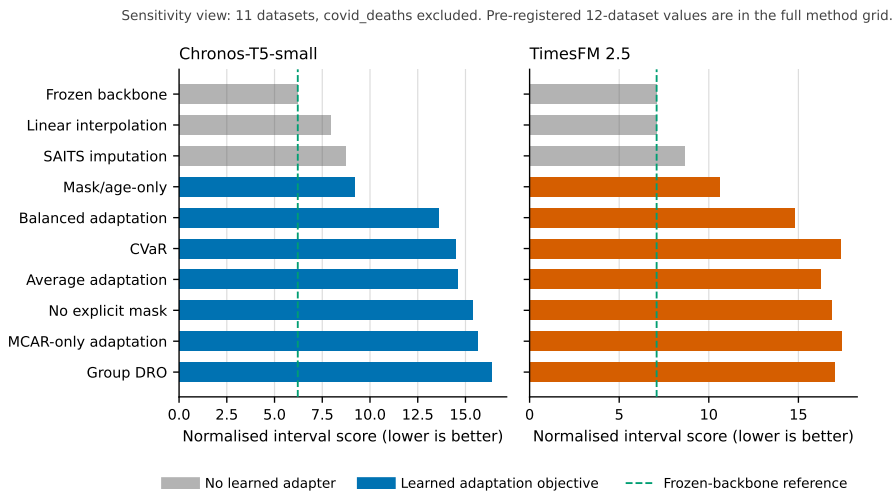

Fig. 2 Every pre-registered method against the frozen-backbone reference at 20% missingness, in the eleven-dataset view of Section 5.3. Grey bars fit no adapter. The three methods that fit nothing are the three best methods.

are consistent in sign but not in magnitude, and one cluster is two orders of magnitude larger than the rest, so the mean-based test spends nearly all its power on that one cluster. The cluster direction count is not vulnerable in that way, and it is unambiguous (`anchor_direction_summary.csv` in the released artifact): for every one of the seven learned objectives, on both backbones, adaptation is worse than the frozen backbone on at least 11 of the 12 datasets, each with sign-test $p \leq 0.006$. In twelve of these fourteen combinations the one dataset that favours adaptation is `covid_deaths`, and Section 5.3 shows why that dataset behaves differently from the rest of the panel.

Table 6 Anchor contrasts at 20% missingness, raw intervals. Positive values mean the adaptation method is *worse* than the reference. Holm correction within this family leaves all four adapter-minus-frozen tests at $p = 1.000$ and is omitted from the table; the confirmatory family of Table 3 is untouched. Sign test counts clusters favouring the reference.

Contrast	Backbone	Effect	95% CI	Exact p	Sign, p
Balanced − frozen	Chronos	+1.25	[−23.29, 13.19]	0.929	11/12, 0.006
Balanced − frozen	TimesFM	+2.77	[−16.63, 11.97]	0.705	11/12, 0.006
Average − frozen	Chronos	+1.99	[−23.57, 14.58]	0.861	11/12, 0.006
Average − frozen	TimesFM	+5.17	[−12.28, 14.28]	0.294	11/12, 0.006
Balanced − linear	Chronos	−1.51	[−25.72, 9.06]	0.999	10/12, 0.039
Balanced − linear	TimesFM	+2.80	[−16.61, 11.98]	0.696	11/12, 0.006
Balanced − SAITS	Chronos	−42.17	[−176.36, 7.28]	1.000	11/12, 0.006
Balanced − SAITS	TimesFM	−20.38	[−104.65, 8.49]	1.000	11/12, 0.006

Two rows deserve comment because their means and their direction counts disagree. Balanced adaptation appears to beat linear interpolation for Chronos (−1.51) and to beat SAITS by a wide margin on both backbones (−42.17 and −20.38), yet loses to both on 10 to 11 of the 12 clusters. The means are carried by one cluster on which the imputation baselines fail badly. Reporting only the mean would have produced the claim that the adapter beats a strong imputation baseline, which the cluster-level evidence does not support.

5.3 Normalisation validity

Every score in this paper is divided by a per-series scale estimated from that series' training prefix. That is only meaningful when the prefix represents the period being scored. We recomputed the denominator for all 48 series directly from source (Table 7, Fig. 3). Forty-four series behave as intended, with test-period level between 0.4× and 4.5× their prefix level and a median of 1.02. The four `covid_deaths` series do not. Two have a training prefix that is identically zero, so their denominator falls back to the pre-registered value of one and their scores are not normalised at all; the other two have test-period levels 254× and 224× their prefix level. The prefix for these series is pre-pandemic and the test origins are at the epidemic peak, so the assumption the normalisation rests on fails outright.

Table 7 Normalisation validity by dataset. The ratio is the test-period mean absolute level divided by the training-prefix mean absolute level, per series; a value near one means the prefix represents the test period. Frozen score is the dataset's mean normalised interval score under the frozen backbone at 20%.

Dataset	Fallback denominators	Level ratio	Frozen score
`covid_deaths`	2 of 4	224–254 (2 undefined)	465.21
All other datasets	0 of 44	0.38–4.47	2.73–11.30

The consequence is quantitative, not cosmetic. The frozen backbone scores 465.21 on `covid_deaths` and between 2.73 and 11.30 on every other dataset, so an unweighted mean over twelve datasets gives this one roughly forty times the weight of any other. That is enough to reverse the apparent ordering of the baselines in Table 5: SAITS imputation looks catastrophic at 87.88 over twelve datasets and unremarkable at 8.72 over eleven.

We therefore report a sensitivity analysis that excludes `covid_deaths` (Table 8). Three properties keep it from being a selection on the outcome. The criterion is computed from the input series alone by a separate published script, with no reference to any method's score. It is not uniformly favourable: it strengthens the Chronos confirmatory effect and weakens the TimesFM one, and a test in the released artifact fails if that ever stops being true. And the pre-registered twelve-dataset estimate is reported beside it everywhere, including in both panels of Fig. 1. The pre-registered

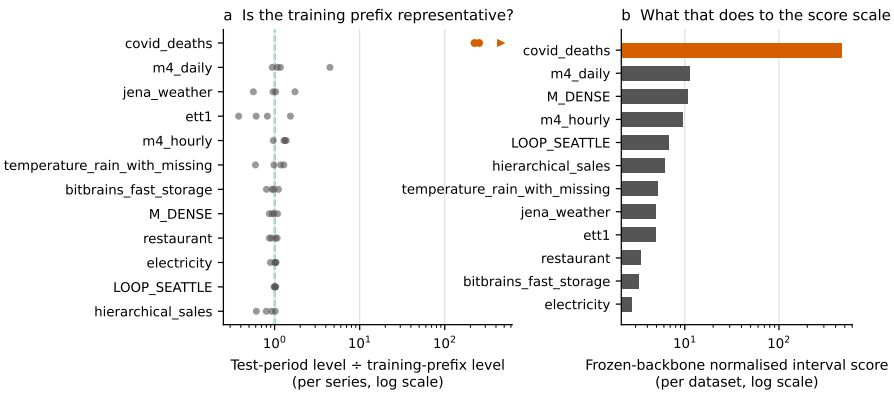

Fig. 3 Left: test-period level divided by training-prefix level for each of the 48 series; triangles mark the two series whose prefix is identically zero. Right: mean normalised interval score per dataset under the frozen backbone. One dataset sits roughly forty times above the rest on the axis that every unweighted mean in this paper runs over.

result remains the primary one. We do not claim the eleven-dataset numbers as a confirmatory finding; we claim that the twelve-dataset means are dominated by a dataset whose normalisation is invalid, and that both views should be read together.

Table 8 Sensitivity analysis excluding `covid_deaths` at 20% missingness, raw intervals. The pre-registered twelve-dataset effect is repeated for comparison. Negative favours the first-named method for the confirmatory contrast; positive means adaptation is worse for the anchors.

Contrast	Backbone	12 ds	11 ds	95% CI (11 ds)	Exact p
Balanced − average	Chronos	−0.74	−0.99	[−1.63, −0.47]	0.0029
Balanced − average	TimesFM	−2.41	−1.46	[−2.48, −0.70]	0.0020
Balanced − frozen	Chronos	+1.25	+7.35	[3.51, 12.19]	0.0010
Balanced − frozen	TimesFM	+2.77	+7.73	[4.16, 12.52]	0.0010
Average − frozen	Chronos	+1.99	+8.35	[4.06, 13.76]	0.0010
Average − frozen	TimesFM	+5.17	+9.19	[4.93, 14.91]	0.0010
Balanced − linear	Chronos	−1.51	+5.64	[2.14, 10.02]	0.0020
Balanced − linear	TimesFM	+2.80	+7.73	[4.16, 12.51]	0.0010
Balanced − SAITS	Chronos	−42.17	+4.85	[1.04, 8.39]	0.0010
Balanced − SAITS	TimesFM	−20.38	+6.18	[3.37, 9.65]	0.0010

In that view the anchor result is no longer ambiguous. Adaptation is worse than the frozen backbone by 7.35 and 7.73 normalised interval-score units, worse than linear interpolation, and worse than SAITS imputation, with every exact p at or near the smallest value eleven clusters can produce. The confirmatory contrast survives in both views and becomes significant for Chronos, which is precisely why we do not promote it.

5.4 Secondary analyses

The 10% and 40% analyses are pre-declared robustness rates, not replacements for the primary estimand. Figure 4 shows the confirmatory contrast and the anchor across all three rates.

5.4.1 Rate sensitivity

At 10%, balanced adaptation has lower interval score, lower MASE, higher coverage, and narrower intervals than average adaptation for both backbones. At the primary rate, interval score and coverage improve but MASE worsens slightly. At 40%, TimesFM retains a smaller advantage (−1.04) while Chronos changes sign (+0.66). Table 9 gives the full set. The anchor tells a different story across rates: the adapter's deficit against the frozen backbone shrinks as missingness grows, and for TimesFM at 40% the adapter finally overtakes it (−3.0). Whatever the adapters learn is of some use only when the context is severely degraded, and on this panel that regime is outside the pre-registered primary rate.

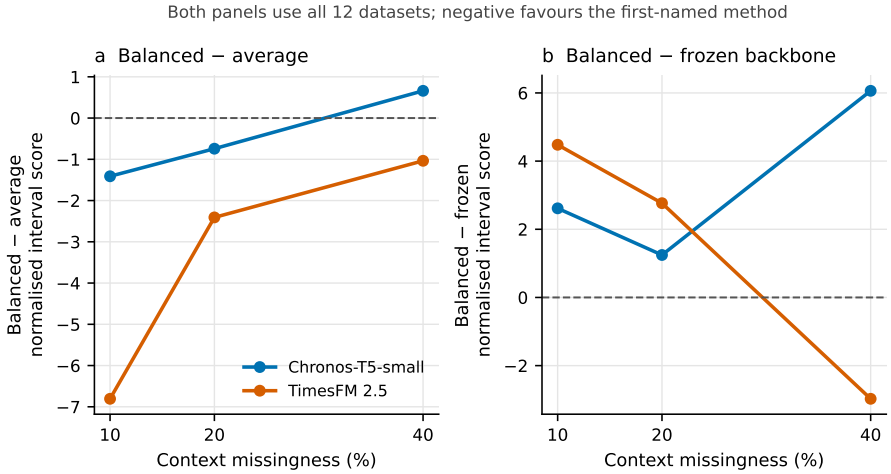

Fig. 4 Rate sensitivity, all twelve datasets. Left: the confirmatory balanced-minus-average contrast. Right: the balanced-minus-frozen anchor. Negative favours the first-named method. Both are secondary at 10% and 40%.

5.4.2 Objective and feature controls

Supplementary Table S4 compares the five secondary objectives with average adaptation at 20% on the same informative cells. Group DRO and CVaR do not reproduce the focal improvement: their interval scores are worse for both backbones. MCAR-only adaptation is worse for Chronos and modestly better for TimesFM. Removing explicit mask features is worse for both. The mask/age-only adapter sharply narrows intervals and improves MASE but loses enough coverage to receive a worse proper score.

Table 9 Balanced-minus-average robustness contrasts across rates, all twelve datasets. Negative interval-score and MASE effects are better.

Backbone	Rate	Score	MASE	Coverage	Width
Chronos	0.10	−1.411	−0.176	+0.0117	−15.0
Chronos	0.20	−0.743	+0.111	+0.0319	−15.0
Chronos	0.40	+0.660	+0.094	−0.0153	−36.1
TimesFM	0.10	−6.806	−0.309	+0.0727	−133.3
TimesFM	0.20	−2.409	+0.241	+0.0503	−4.8
TimesFM	0.40	−1.036	−0.226	+0.0041	−93.1

These comparisons are all within the learned family; Table 5 shows that the family as a whole sits below every method that fits nothing.

5.4.3 Informative versus control geometry

At 20%, event-informative masks are harder than event controls for both focal methods and backbones, with raw interval-score gaps from 3.59 to 5.69. Periodic-informative masks are *easier* than their random-phase controls, with gaps from −3.67 to −4.84. The informative label therefore does not imply uniform difficulty; it means the mask construction depends on the observed-history activity feature. The opposing signs also justify combining the two informative families with equal weight in the primary estimand rather than selecting whichever favours a method. For balanced adaptation the event gap is 3.59 under Chronos and 5.69 under TimesFM, and the periodic gaps are −4.84 and −3.67; average adaptation shows nearly the same geometry, so the balanced benefit is not explained by turning both held-out families into uniformly easier cases.

5.4.4 Calibration

Validation-derived additive conformal offsets raise empirical coverage close to nominal without touching test targets. For the two focal methods at 20%, raw coverage across informative mechanisms ranges from 0.579 to 0.716 and calibrated coverage from 0.799 to 0.812. Chronos average adaptation moves from 0.579 to 0.805 on event masks and 0.585 to 0.804 on periodic masks; balanced adaptation reaches 0.809 in both families; TimesFM behaves analogously. For the focal methods the calibrated interval score barely moves (−0.15 for Chronos and −0.02 for TimesFM in the eleven-dataset view), as expected when calibration targets coverage rather than the score itself.

Figure 5 shows why this has to be reported per method. Averaged over all ten methods the calibration cost appears to be +6.1 and +9.0 interval-score units, but essentially all of that comes from the mask/age-only control, whose raw intervals are so narrow that the offsets required to reach nominal coverage are enormous. The focal methods are unaffected on that scale.

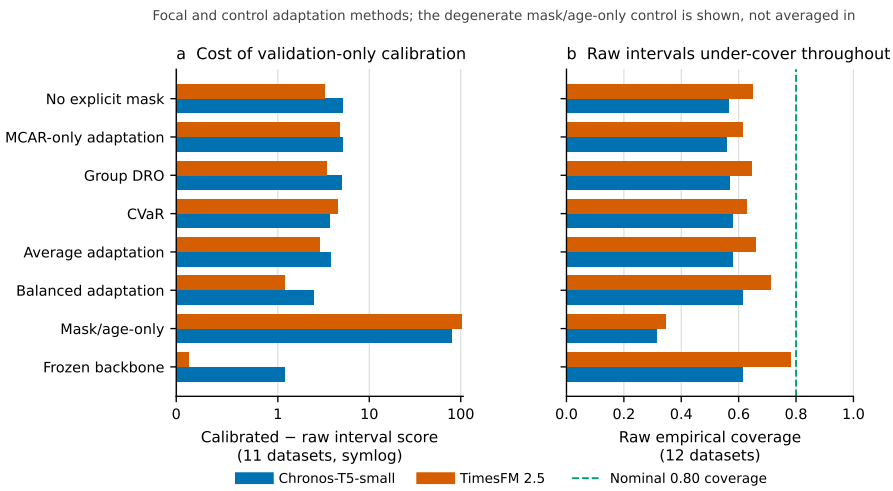

Fig. 5 Validation-only conformal calibration, by method rather than pooled. Left: change in normalised interval score on a symlog axis; the focal methods move by one to five units while the degenerate mask/age-only control moves by 80 to 101. Right: raw empirical coverage against the nominal 0.80. Pooling these methods into one average, as an earlier version of this figure did, reports the control's behaviour as if it were the method's.

5.4.5 Natural missingness

Natural-missingness evaluation is an external stress test on two datasets, not a pooled extension of the panel. At synthetic training rate 0.2, balanced adaptation has lower raw interval score than average adaptation on both natural datasets and both backbones: 5.686 versus 5.974 for Chronos and 8.904 versus 9.368 for TimesFM on the temperature/rain data, and 14.848 versus 15.129 and 12.170 versus 12.720 on UCI Household Power. These are descriptive means over 24 and 32 seed-averaged episodes. MASE changes are small and slightly adverse and coverage is generally higher under balance. No retuning uses natural test traces. Two datasets covering only nature and energy cannot support a confirmatory claim, and we attach no synthetic-panel p-values to them.

5.4.6 S4M task-compatible subset

S4M provides a reference point for an architecture built for multivariate incomplete-input forecasting. On ETT1, calibrated interval scores are 4.02, 4.09, and 4.56 at rates 0.1, 0.2, and 0.4; MASE rises from 0.995 to 1.066 and coverage declines from 0.656 to 0.611. On Jena Weather, interval scores are 9.20, 9.44, and 9.59, MASE is approximately 1.93–1.97, and coverage approximately 0.56–0.57. These cannot be ranked directly against the univariate table because S4M uses multivariate channels and residual-conformal intervals. Their role is to show that the campaign covers a dedicated missingness-aware architecture where task compatibility permits.

6 Discussion

The result we set out to measure is real and small: mechanism-balanced adaptation scores better than average-risk adaptation, reliably for TimesFM 2.5 and directionally for Chronos-T5-small. The result that matters is the one the confirmatory family structurally could not see: neither objective, and none of the five secondary objectives, reaches the frozen backbone. Balanced adaptation is the best member of a set of interventions that all make the forecast worse.

6.1 Why a well-designed confirmatory family can still mislead

Nothing went wrong with the pre-registration. The rate, contrast, aggregation, hierarchy, and correction family were fixed in advance; the test is exact; the seeds are averaged before inference; the intervals respect the data hierarchy. The confirmatory contrast is a correct answer to the question it encodes. The difficulty is that the question encodes a comparison *within* an intervention, and a reader who sees a significant, correctly-corrected, hierarchy-aware effect will reasonably conclude that the intervention is worth using.

This is not a hazard specific to our design. Any protocol whose registered family compares variants of a method – objectives, curricula, schedules, prompt strategies – inherits it, and pre-registration does nothing to help, because the discipline of pre-registration is about not changing the question after seeing the answer, not about whether the question was the right one. The remedy is cheap and mechanical: register the no-intervention condition as a reported quantity, not merely as a method in a list. Our own protocol registered `frozen` and still failed to report it, which is precisely why we now enforce it as a build-time check on the released artifact rather than as an intention.

6.2 Why the mean-based test was the wrong instrument here

The anchor contrasts are inconclusive on the pre-registered twelve datasets by any mean-based criterion: intervals from $[-23, 13]$, exact p from 0.29 to 0.93. The same data are decisive when read at cluster level, where every learned objective loses on at least 11 of 12 clusters. Both readings are correct, and they differ because the estimand differs. A dataset-clustered sign flip on a mean is a test about the mean, and when one cluster contributes two orders of magnitude more than any other, that mean is close to being a statement about one cluster.

We report both rather than choosing. A cluster count discards magnitude and would be the wrong summary if the effects were small and mixed in sign; a cluster mean discards the fact that eleven independent panels agree. Where a design has few top-level units and heterogeneous scales – which describes most multi-dataset forecasting benchmarks – neither alone is adequate.

6.3 Why normalisation deserves its own audit

Normalised scores are usually treated as a solved problem: divide by a seasonal-naive scale and datasets become comparable. Section 5.3 shows the assumption underneath

that habit. The scale is estimated from a training prefix, so it transfers only when the prefix represents the scored period. For a strongly non-stationary series it does not, and the pre-registered fallback for a degenerate denominator – substituting one – silently returns an unnormalised score rather than flagging a failure.

The effect was large enough to invert conclusions. Over twelve datasets SAITS imputation looks catastrophic at 87.88 and the adapter appears to beat it by 42 units; over the eleven datasets whose normalisation holds, SAITS scores 8.72 and beats the adapter by 4.85. We would have reported the first as a finding. The general lesson is that a normalisation constant estimated per series is a modelling assumption with its own validity conditions, and a benchmark that does not check them can be dominated by whichever series violates them most.

6.4 What the backbone difference does and does not show

The two backbones differ in pretraining, forecast representation, and probabilistic interface: Chronos quantiles are empirical summaries of stochastic trajectories, while TimesFM emits quantiles directly. A global adapter can therefore meet different error geometry under the same context, mask, target, and loss, and the confirmatory effect is roughly three times larger for TimesFM. But the anchor result is nearly identical across the two (+7.35 and +7.73 in the eleven-dataset view), so the backbone difference is a difference in how much the adapters hurt, not in whether they hurt. The Chronos result should not be described as a failed replication of the TimesFM one; the two are parallel pre-registered tests whose intervals overlap substantially.

6.5 Why the robust objectives did not rescue the family

Equal mechanism weighting, group DRO, and CVaR encode different priorities. Equal weighting changes a known sampling measure; group DRO shifts mass toward the currently high-loss predefined group; CVaR targets the high-loss tail regardless of group identity. Their empirical behaviour need not agree, and it does not: both are worse than average adaptation at the primary rate for Chronos, and group DRO is worse for TimesFM. MCAR-only adaptation modestly improves TimesFM's interval score while reducing its coverage.

Read within the learned family these are informative contrasts about robustness-set choice. Read against the anchor they are variations among seven ways of losing to the frozen backbone by between 7.4 and 10.3 units. This does not disprove distributionally robust optimisation; it shows that when the intervention itself is net harmful on a panel, comparisons of its robustness set are answering a question that has not yet been earned.

6.6 What we think the adapters actually did

We did not design the study to explain the anchor gap, so what follows is a hypothesis the data constrain rather than a result. The adapters operate on the backbone's output distribution using features derived from the incomplete context, and they are fitted globally at the backbone–rate–method–seed level and then applied to every series. A single global correction has to be right on average across twelve domains with very

different forecastability. The per-dataset table shows it is not: the correction that helps the one pathological, high-variance dataset is the same correction applied to eleven well-behaved ones, where the frozen backbone was already close to its own best achievable interval score.

Two observations support this reading. The deficit shrinks monotonically with missingness rate and reverses for TimesFM at 40%, which is where the frozen backbone's own context becomes badly degraded and there is genuinely something to correct. And the mask/age-only ablation, which removes context values and keeps only missingness geometry, produces by far the narrowest intervals and the worst coverage, indicating that the adapters are learning a scale correction rather than a mechanism-specific one. A per-dataset or online adapter is a different intervention and this study says nothing about it.

6.7 Calibration as a separate engineering decision

Raw intervals substantially under-cover their nominal 80% level in every focal condition. The validation-only additive correction raises coverage close to nominal for both methods and backbones without touching test targets, and for the focal methods it changes the interval score by less than 0.2 units. That is operationally useful: an application may need an explicit calibration layer regardless of its adaptation objective, and an interval-score gain must not be read as calibrated uncertainty.

Calibration does not remove the need for the raw comparison, and reporting it pooled across methods actively misleads. Figure 5 shows the pooled average is dominated by the degenerate mask/age-only control. We recommend a two-stage deployment audit: compare raw models under a proper score against a no-intervention anchor, then evaluate a frozen validation-only calibration procedure under the same chronological protocol, per method.

6.8 Implications for evaluation practice

Four recommendations follow, in the order we would apply them. First, report the no-intervention anchor in the same table as the confirmatory contrast; if the intervention cannot beat doing nothing, the comparison among its variants is not yet meaningful. Second, evaluate at least one held-out mask family: training and testing only MCAR measures within-generator performance and hides dependence on missingness geometry. Third, carry the data hierarchy into uncertainty estimation and pair the mean-based cluster test with a cluster direction count, because with a dozen top-level units a single cluster can dominate the mean. Fourth, audit the normalisation itself, from the input series and before looking at any score.

6.9 What can and cannot be claimed

The strongest supported statement is this. At 20% informative context missingness, on the pre-registered panel, mechanism-balanced adaptation lowers mean normalised interval score relative to average-risk adaptation, reliably for TimesFM 2.5 (Holm-adjusted $p = 0.0020$) and directionally but inconclusively for Chronos-T5-small. Against the frozen backbone, all seven learned objectives are worse on at least 11 of 12

datasets for both backbones, and significantly worse in the eleven-dataset view whose normalisation the audit validates.

Several tempting statements are unsupported. The study does not show that balanced adaptation is useful in deployment; on this panel it is not. It does not show that the anchor gap is significant on the pre-registered twelve datasets by a mean-based test; it is not. It does not establish that no lightweight adapter could help, only that these seven, fitted globally at this scale, do not. It does not establish a causal missingness mechanism for the natural datasets, and it does not show that every informative mask is harder than its control.

6.10 Limitations

The evidence is bounded by the specified panel, horizons, adapters, and missingness constructions, and the following boundaries are the ones we consider most likely to matter.

6.10.1 Construct validity

The names periodic-informative and event-informative denote algorithmic generators. They do not imply that real outages follow the same process or that a mask is missing not at random in a causal sense. The informative generators use strictly observed-history activity, which prevents hidden-value leakage but narrows the class of dependence studied. Control masks match missing count and broad family but not every statistic of run length, lag coverage, or information content. The fixed context and horizon simplify comparison but limit generality: an 80-step context may contain fewer than one official seasonal cycle for minute data and many cycles for monthly data, and a 24-step horizon has different semantic duration across frequencies.

6.10.2 Internal and statistical validity

Five seeds characterise adapter variation but do not create five independent datasets, and we average them before inference. Eight origins give chronological coverage but remain short enough that block-length choice can influence uncertainty; the fixed square-root rule avoids tuning it to the desired result. Twelve top-level datasets bound the resolution of the exact sign-flip test at $1/4096$ and, more importantly, make the mean-based anchor tests underpowered against a heterogeneous alternative. The exclusion of `covid_deaths` is defined on the input series and reported alongside the full panel throughout, but it remains a post-hoc analysis and we do not present its p-values as confirmatory.

6.10.3 Model and implementation validity

Only two backbones are tested, and results need not transfer to larger Chronos variants, other TimesFM revisions, decoder-only alternatives, or fine-tuned models. The adapters are global at a backbone–rate–method–seed level; per-dataset, per-series, and online adapters answer different questions and are not evaluated here. SAITS depends on its training configuration and is fitted univariately, which is not the regime it was

designed for; its eleven-dataset performance is respectable but should not be read as a tuned upper bound. S4M is evaluated only on a task-compatible multivariate subset with residual-conformal intervals. The context-only control remains indirectly mask-dependent through deterministic imputation.

6.10.4 External validity and release constraints

Natural missingness is observed in only UCI Household Power and one temperature/rain dataset, covering two domains. Operational systems may have censoring, sensor replacement, reporting delay, multivariate dependence, or distribution shift absent from these traces. The public repository excludes large raw shards, pretrained weights, and source datasets, so the strongest public reproducibility claim applies to the claim-bearing tables, manifests, figures, and analysis code, all of which regenerate from the released artifact. A fresh Ubuntu/CUDA smoke reproduction of model inference remains an open release gate.

7 Conclusion

Across a pre-registered multi-domain audit, mechanism-balanced adaptation improves the primary normalised interval score relative to average-risk adaptation, reliably for TimesFM 2.5 and directionally for Chronos-T5-small. Against the no-adaptation anchor that the same protocol registered but its confirmatory family never tested, all seven learned adaptation objectives are worse than leaving the backbone alone on at least 11 of 12 datasets for both backbones, and significantly worse once the one dataset whose normalisation the audit invalidates is set aside. Mechanism balancing is the least harmful member of a harmful family, not an improvement.

Three methodological points generalise beyond this panel. A confirmatory family that compares variants of an intervention cannot tell a reader whether the intervention is worth deploying, and pre-registration does not protect against that, so the no-intervention anchor belongs in the reported family and, in our artifact, in a build-time check. With a dozen top-level clusters and heterogeneous scales, a cluster-level mean test and a cluster direction count answer different questions and both should be reported. And a per-series normalisation constant is a modelling assumption with validity conditions that deserve auditing from the input data, before any score is computed; here one dataset that violated them carried roughly forty times the weight of any other and inverted the apparent ranking of every baseline.

Future work should test whether the anchor gap persists for per-dataset or online adapters, for larger backbones, and under real operational missingness, and should establish at what degree of context degradation a lightweight correction begins to pay for itself. Our rate sensitivity suggests that point exists but lies beyond the missingness levels this panel was designed around.

A Detailed protocol specification

A.1 Frozen constants

This appendix records the frozen constants and generator contracts referenced from Sections 3 and 4. The core constants are $L = 80$, $H = 24$, missing rates $\{0.10, 0.20, 0.40\}$, seeds $\{0, 1, 2, 3, 4\}$, and origin counts 5/3/8. The two synthetic-panel backbones are `chronos_t5_small` and `timesfm_2p5`. The methods are frozen, linear interpolation, average adaptation, MCAR adaptation, balanced adaptation, group DRO, CVaR, mask/age-only, context without explicit mask, and SAITS imputation. S4M is outside that common method list because it uses task-compatible multivariate subsets.

The primary rate is 0.20, primary calibration status is raw, and the two primary mechanisms are periodic-informative and event-informative. The two are averaged after seed averaging and before matched method differencing. The confirmatory treatment is balanced adaptation and the reference is average adaptation. Any table generated under a different rate, method pair, calibration status, or mechanism aggregation is secondary.

A.2 Exact-k contract

For each non-clean mask, $k = \text{round}(Lr)$. The generator must return a Boolean vector of shape 80 whose sum is exactly k. This is checked for every stored mask rather than only in unit fixtures. Stable seeds are derived from canonical serialized job components through SHA-256, making the same job identity reproducible across restarts. Mask identifiers and mask-content hashes are stored separately: two job descriptions cannot silently point to different masks, and identical clean masks can be recognized even when contractual episode identifiers differ.

A.3 Seasonality and normalization boundary

Seasonal lag is determined from the recorded frequency token before a hyphen suffix. Minute and legacy "T" frequencies map to 1,440, hourly to 24, daily to seven, weekly to 52, monthly to 12, and quarterly to four. Any unrecognized token maps to lag one with an explicit fallback label. The denominator is calculated from absolute lagged differences in the training prefix. Future perturbation must not change normalization or denominator fields.

B Reproducibility checklist

1. Confirm that the panel has 12 datasets, 48 series, context 80, horizon 24, and 9,216 mask records.
2. Confirm exact missing counts 8, 16, and 32 and zero for clean masks.
3. Run the hidden-value MAR canary and chronology/normalization boundary tests.
4. Verify that train/validation and test mask IDs are disjoint and test generators do not enter training.

5. Verify all prediction shards for duplicates, finite values, target/mask alignment, and quantile monotonicity.
6. Average five seeds before any comparison; do not treat seed or horizon rows as independent samples.
7. Combine periodic- and event-informative conditions equally inside each series–origin.
8. Resample dataset, series, and contiguous origin blocks in that order; use one sign per dataset for permutation.
9. Apply Holm correction only to the two primary backbone tests; correct the anchor family separately and never merge the two.
10. Report the no-adaptation anchor in the same table as the confirmatory contrast, and pair every cluster-level mean test with a cluster direction count.
11. Recompute every series' interval-score denominator from source and confirm that the training prefix represents the test period before interpreting any normalised mean.
12. Recompute MASE and normalized interval score from raw predictions and training-prefix denominators.
13. Regenerate every table from raw episode metrics in a CPU clean room and compare hashes; regenerate every figure from those tables alone.
14. Produce a fresh Ubuntu/CUDA smoke shard before final archival release.

C Claim ledger

Table 10 maps each claim to the evidence that licenses it. The ledger is enforced by a repository check that requires the manuscript to describe the Chronos confirmatory result as inconclusive, retain the corrected adjusted p-values, report the no-adaptation anchor, and state the eleven-dataset analysis as a sensitivity view rather than a confirmatory one. It rejects text asserting significance for both backbones, universal improvement, or a deployment recommendation for the adapters. This guard cannot replace human review, but it prevents an accidental edit from reverting to the superseded optimistic interpretation.

Supplementary information. A separate Supplementary Information document contains the prediction-contract schema (Table S1), the inventory of claim-bearing analysis products (Table S2), the failure and repair ledger (Table S3), and the secondary objective and feature controls (Table S4). Verification manifests and reproducibility instructions accompany the software release.

Author contributions. The author confirms sole responsibility for conceptualization, methodology, software, validation, formal analysis, investigation, data curation, visualization, and the writing of the original draft and its revision.

Funding. The author received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability. The reproduction and verification artifact accompanies this submission as a related file and contains every claim-bearing result table, all figures,

Table 10 Claim-to-evidence ledger used during manuscript verification.

Claim	Evidence selector	Allowed interpretation
TimesFM benefits at 20%	primary contrast, `timesfm_2p5`	strong, Holm-significant evidence
Chronos direction favors balance	primary contrast, `chronos_t5_small`	directional and inconclusive
Rate dependence	all six rate-sensitivity rows	robustness only
Calibration raises coverage	focal 20% paired calibration rows	validation-derived secondary result
Natural results favor balance descriptively	two datasets, two backbones, rate 0.2 raw	external descriptive stress test
S4M evaluated	six S4M rows	task-compatible multivariate subset only
Adaptation loses to the frozen backbone	anchor direction summary, all seven objectives	11 of 12 clusters per objective, both backbones
Anchor gap is not significant on the full panel	anchor contrasts, twelve datasets	mean-based tests are underpowered here
Anchor gap is significant excluding `covid_deaths`	robust aggregation, eleven datasets	sensitivity analysis, not confirmatory
One dataset is not validly normalised	normalisation audit, 48 series	data-side criterion, computed before any score
Analytical closure passes	automated closure-check report	CPU analysis and hash-chain acceptance

the evidence manifests, and the frozen analysis code. It will be released publicly with an archival DOI upon acceptance. Source datasets and pretrained model assets are obtained from their original providers under their respective terms and are not redistributed.

Code availability. The analysis code, verification tooling, and the frozen snapshot that produced the sealed evidence are included in the accompanying artifact and will be released publicly with an archival DOI upon acceptance.

Declarations. The author declares no competing interests.

Use of generative AI. Generative AI tools (OpenAI Codex and Anthropic Claude) assisted with drafting and software-engineering support under the author's supervision. The author takes responsibility for the final manuscript, analyses, and claims.

References

Aksu T, Woo G, Liu J, et al (2024) GIFT-Eval: A benchmark for general time series forecasting model evaluation. arXiv preprint arXiv:241010393 URL https://arxiv.org/abs/2410.10393

Alfarra M, Itani H, Pardo A, et al (2024) Evaluation of test-time adaptation under computational time constraints. In: Proceedings of the 41st International Conference on Machine Learning, vol 235. PMLR, URL https://proceedings.mlr.press/v235/alfarra24a.html

Angelopoulos AN, Bates S (2023) Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning 16(4):494–591. https://doi.org/10.1561/2200000101

Ansari AF, Stella L, Turkmen C, et al (2024) Chronos: Learning the language of time series. Transactions on Machine Learning Research URL https://openreview.net/forum?id=gerNCVqqtR

Bracher J, Ray EL, Gneiting T, et al (2021) Evaluating epidemic forecasts in an interval format. PLOS Computational Biology 17(2):e1008618. https://doi.org/10.1371/journal.pcbi.1008618

Cameron AC, Gelbach JB, Miller DL (2008) Bootstrap-based improvements for inference with clustered errors. The Review of Economics and Statistics 90(3):414–427. https://doi.org/10.1162/rest.90.3.414

Cao W, Wang D, Li J, et al (2018) BRITS: Bidirectional recurrent imputation for time series. In: Advances in Neural Information Processing Systems, pp 6776–6786, URL https://proceedings.neurips.cc/paper/2018/hash/734e6bfcd358e25ac1db0a4241b95651-Abstract.html

Che Z, Purushotham S, Cho K, et al (2018) Recurrent neural networks for multivariate time series with missing values. Scientific Reports 8(1):6085. https://doi.org/10.1038/s41598-018-24271-9

Das A, Kong W, Sen R, et al (2024) A decoder-only foundation model for time-series forecasting. In: Proceedings of the 41st International Conference on Machine Learning, vol 235. PMLR, pp 10148–10167, URL https://proceedings.mlr.press/v235/das24c.html

Du W, Côté D, Liu Y (2023) SAITS: Self-attention-based imputation for time series. Expert Systems with Applications 219:119619. https://doi.org/10.1016/j.eswa.2023.119619

Du W, Wang J, Qian L, et al (2024) TSI-Bench: Benchmarking time series imputation. arXiv preprint arXiv:240612747 URL https://arxiv.org/abs/2406.12747

Duchi JC, Namkoong H (2021) Learning models with uniform performance via distributionally robust optimization. The Annals of Statistics 49(3):1378–1406. https://doi.org/10.1214/20-AOS2004

Field CA, Welsh AH (2007) Bootstrapping clustered data. Journal of the Royal Statistical Society Series B: Statistical Methodology 69(3):369–390. https://doi.org/10.1111/j.1467-9868.2007.00593.x

Gibbs I, Candès EJ (2021) Adaptive conformal inference under distribution shift. In: Advances in Neural Information Processing Systems, pp 1660–1672, URL https://proceedings.neurips.cc/paper/2021/hash/0d441de75945e5acbc865406fc9a2559-Abstract.html

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102(477):359–378. https://doi.org/10.1198/016214506000001437

Goswami M, Szafer K, Choudhry A, et al (2024) MOMENT: A family of open time-series foundation models. In: Proceedings of the 41st International Conference on Machine Learning, vol 235. PMLR, pp 16115–16152, URL https://proceedings.mlr.press/v235/goswami24a.html

Holm S (1979) A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6(2):65–70

Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. International Journal of Forecasting 22(4):679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Kim H, Kim S, Mok J, et al (2025) Battling the non-stationarity in time series forecasting via test-time adaptation. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 17868–17876, https://doi.org/10.1609/aaai.v39i17.33965

Koenker R, Bassett G (1978) Regression quantiles. Econometrica 46(1):33–50. https://doi.org/10.2307/1913643

Le Naour E, Nabil T, Agoua G (2025) MoTM: Towards a foundation model for time series imputation based on continuous modeling. arXiv preprint arXiv:250713207 URL https://arxiv.org/abs/2507.13207

Le Naour E, Nabil T, Petralia A, et al (2026) Are time-indexed foundation models the future of time series imputation? Transactions on Machine Learning Research URL https://arxiv.org/abs/2511.05980

Little RJA, Rubin DB (2019) Statistical Analysis with Missing Data, 3rd edn. John Wiley & Sons, Hoboken, NJ, https://doi.org/10.1002/9781119482260

Makridakis S, Spiliotis E, Assimakopoulos V (2020) The M4 competition: 100,000 time series and 61 forecasting methods. International Journal of Forecasting 36(1):54–74. https://doi.org/10.1016/j.ijforecast.2019.04.014

Peng J, Yang M, Zhang Q, et al (2025) S4M: S4 for multivariate time series forecasting with missing values. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=zJT8ZE3rRl

Politis DN, Romano JP (1994) The stationary bootstrap. Journal of the American Statistical Association 89(428):1303–1313. https://doi.org/10.1080/01621459.1994.10476870

Rockafellar RT, Uryasev S (2000) Optimization of conditional value-at-risk. The Journal of Risk 2(3):21–41. https://doi.org/10.21314/JOR.2000.038

Rubin DB (1976) Inference and missing data. Biometrika 63(3):581–592. https://doi.org/10.1093/biomet/63.3.581

Sagawa S, Koh PW, Hashimoto TB, et al (2020) Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=ryxGuJrFvS

Shukla SN, Marlin BM (2021) Multi-time attention networks for irregularly sampled time series. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=4c0J6lwQ4_

Stankevičiūtė K, Alaa AM, van der Schaar M (2021) Conformal time-series forecasting. In: Advances in Neural Information Processing Systems, URL https://proceedings.neurips.cc/paper/2021/hash/312f1ba2a72318edaaa995a67835fad5-Abstract.html

Tang K, Zhang J, Meng H, et al (2025) CoIFNet: A unified framework for multivariate time series forecasting with missing values. arXiv preprint arXiv:250613064 URL https://arxiv.org/abs/2506.13064

Vovk V, Gammerman A, Shafer G (2005) Algorithmic Learning in a Random World. Springer, New York, https://doi.org/10.1007/b106715

Woo G, Liu C, Kumar A, et al (2024) Unified training of universal time series forecasting transformers. In: Proceedings of the 41st International Conference on Machine Learning, vol 235. PMLR, pp 53140–53164, URL https://proceedings.mlr.press/v235/woo24a.html

Xu C, Xie Y (2023) Sequential predictive conformal inference for time series. In: Proceedings of the 40th International Conference on Machine Learning, vol 202. PMLR, pp 38707–38727, URL https://proceedings.mlr.press/v202/xu23r.html

Yoon J, Jordon J, van der Schaar M (2018) GAIN: Missing data imputation using generative adversarial nets. In: Proceedings of the 35th International Conference on Machine Learning, vol 80. PMLR, pp 5689–5698, URL https://proceedings.mlr.press/v80/yoon18a.html

Zhou H, Zhang S, Peng J, et al (2021) Informer: Beyond efficient transformer for long sequence time-series forecasting. In: Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, pp 11106–11115, https://doi.org/10.1609/aaai.v35i12.17325

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ijmlcsupplementarylatex.zip