CAN TIME SERIES FOUNDATION MODELS SCALE AT TEST TIME?

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time scaling aims to improve forecasts through additional inference computation, yet its benefits beyond strong raw-input baselines remain unclear. We study four frozen time series foundation models (TSFMs) on univariate point-forecasting tasks from eight datasets, distinguishing fixed-input sampling, multi-view inference, and history-based selection and aggregation. Repeated sampling improves forecasts with diminishing returns: on eight designated series, doubling the sample count from 64 to 128 reduces average relative MAE by less than 0.6% for each stochastic model. Multi-view aggregation can provide further gains. Across 48 additional series from three datasets, `Moirai`'s historically selected multi-view ensemble achieves a dataset-balanced mean MAE reduction of 7.42% against a historically selected raw-input ensemble. However, gains vary across models and tasks, and stronger raw-input controls reduce some apparent improvements. Historical improvements can reverse at test time, while candidate-pool expansion can worsen selection and incur additional calibration costs. Our findings distinguish gains from sampling, views, and aggregation, highlighting the gap between candidate potential and forecasting improvements realizable using observed history. Code is available at https://anonymous.4open.science/r/forecasting-tts-fb18b6a3/.

1 INTRODUCTION

Once a time series foundation model (TSFM) is frozen, can additional inference computation improve its forecasts? Pretrained TSFMs support drawing forecast samples or querying alternative input representations without updating model weights (Ansari et al., 2024; Woo et al., 2024). These strategies resemble test-time scaling in large language models, where candidate generation and evaluation can improve reasoning performance (Wang et al., 2023; Snell et al., 2025). Recent forecasting studies investigate repeated sampling with self-consistency (Liu et al., 2025a) and diversified inputs with aggregation and oracle selection (Hua et al., 2026). Which gains persist beyond well-configured raw-input forecasts remains an open empirical question.

In forecasting, generating more samples, changing the input, and choosing a forecast are distinct operations. At a fixed input, more samples can refine a point estimate without improving the predictive distribution. Alternative views change the forecasting task, but need not improve its candidates. Future observations are unavailable for selection, and configurations favored by history may fail at test time. Raw-input context selection and ensembling can also improve the reference forecast. These distinctions motivate a more specific question:

> *Which test-time gains survive sufficiently sampled raw-input baselines, history-based configuration selection, and calibration costs?*

Our central finding is that **repeated sampling improves point forecasts with diminishing returns; further gains from input views depend on aggregation, the raw-input reference, and the transfer of historical selection to future data**. We therefore assess both the improvements over calibrated raw-input controls and the computation required to obtain them.

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.

We organize a controlled empirical study around *sampling* and *aggregation* (Figure 1). Single-view inference samples a fixed input; multi-view inference uses raw contexts, cycle–residual decompositions, trend–residual decompositions, or interleaved temporal subseries. These views change how temporal structure is presented to the model, complementing input-adaptation approaches such as TATO (Qiu et al., 2026). We combine forecasts by pointwise means or medians, with or without history-based selection. Our historical verifiers rank configurations by forecast errors on observed windows; they do not verify the correctness of a current future trajectory.

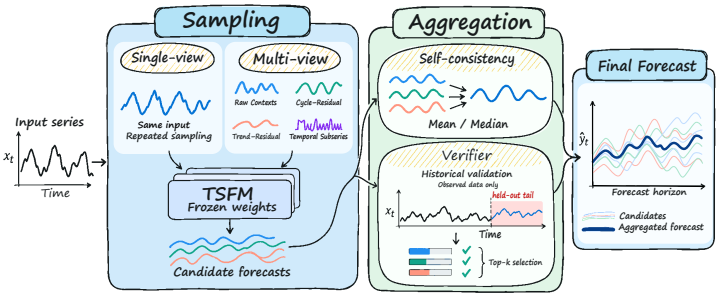

Figure 1: **Sampling and aggregation with a frozen TSFM.** Single- or multi-view inference produces candidates for Mean / Median aggregation or history-based selection. View forecasts are reconstructed before cross-view aggregation; historical verification scores configurations on observed targets, not current future outcomes.

We evaluate `Chronos-T5-Base` (Ansari et al., 2024), `Sundial-Base-128M` (Liu et al., 2025b), `Moirai-1.0-R-Base` (Woo et al., 2024), and `TimesFM 3.0` (Das et al., 2024) on univariate point-forecasting tasks from eight datasets. We abbreviate these checkpoints as `Chronos`, `Sundial`, `Moirai`, and `TimesFM`. Weights remain frozen; stochastic sampling is evaluated where supported, while multi-view inference also applies to deterministic forecasts. MAE is primary and MSE complementary. Raw-input controls apply the same historical scoring criterion to raw configurations, separating the value of transformations from raw selection and combination. Matched residual-predictor replacements test the TSFM's contribution within a decomposition. Future-informed selection is diagnostic only, and costs include historical calibration and online probes.

Three findings answer our central question. First, repeated sampling has diminishing returns: on the eight designated OT series, doubling the sample count from 64 to 128 improves average relative MAE by less than 0.6% for each stochastic model. Second, selected multi-view aggregation can improve well-sampled raw ensembles, even though transformed view families are not uniformly better on their own. On sixteen additional Electricity series, `Moirai`'s historical top-3 Median reduces MAE by 10.04% against historical raw top-k Median, with fifteen series improving. Yet on four prespecified Traffic series, its improvement is 22.48% against all-raw Median but only 0.28% against historical raw top-3 Median. Third, historical selection and its costs limit realizable gains. For `TimesFM`, 13 of 21 historical single-view improvements reverse on the 48-series test evaluation (Appendix D.1); expanding `Moirai`'s candidate pool worsens Traffic top-3 MAE under the same historical rule. Thus, neither a larger candidate pool nor a better historical score guarantees a better deployed forecast.

Our main contributions are:

- We distinguish fixed-input sampling gains from multi-view inference gains through a controlled evaluation of four frozen TSFMs across eight datasets, with configuration selection restricted to observed history.
- We use raw-input and component controls to test which improvements survive raw selection and ensembling, and whether a TSFM's contribution within a decomposition improves the complete forecast.
- We diagnose the practical limits of additional inference through historical-selection transfer, candidate-pool expansion, and calibration and probing costs, separating candidate potential from realizable forecasting gains.

2 Related Work

Test-time scaling in LLMs. Self-consistency (Wang et al., 2023) improves LLM reasoning by sampling multiple reasoning paths and aggregating their answers. Repeated-sampling studies show

that candidate coverage can continue to increase with the sampling budget even when practical selection methods plateau (Brown et al., 2024). Snell et al. (2025) analyze verifier-guided search and test-time response refinement, showing that effective compute allocation depends on problem difficulty. These findings motivate treating candidate generation and candidate evaluation as distinct uses of test-time computation.

Test-time scaling in TSFMs. ReC4TS (Liu et al., 2025a) evaluates reasoning strategies for zero-shot forecasting and reports an empirical test-time scaling law based on self-consistency. Hua et al. (2026) study standard and diversified sampling in TSFMs, using input perturbations to examine how the diversity–fidelity trade-off affects inference scaling. Kayaalp et al. (2026) explore a complementary approach based on selecting and ensembling pretrained forecasting models under computational constraints. Our comparisons hold one checkpoint fixed at a time and vary its input configurations, sample counts, and aggregation rules. We examine when these inference strategies yield gains over strong raw-input configurations, with selection restricted to observed history and compute accounting that includes calibration and online probing.

Input adaptation and multi-view forecasting. TATO (Qiu et al., 2026) optimizes an input transformation pipeline for a frozen TSFM, including context slicing, normalization, and outlier correction. Structured forecasting representations provide further motivation for alternative views: CycleNet (Lin et al., 2024) forecasts residuals around learned periodic cycles, while SutraNets (Bergsma et al., 2023) reformulate long sequences as lower-frequency subseries and model dependencies across them. In our setting, cycle-residual and interleaved-subseries views expose temporal structure through inference-time transformations of a frozen checkpoint's input. We evaluate the incremental value of these views against raw-input configurations selected using the same historical criterion.

3 Experimental Setup

We study how sampling and aggregation affect forecasting accuracy and inference cost with four frozen TSFMs on eight datasets. Following Figure 1, we first define how repeated sampling and alternative input views produce candidate forecasts (Section 3.1), then how self-consistency and historical verification turn these candidates into a point forecast (Section 3.2). The controlled evaluation in Section 3.3 places these choices under a shared forecasting protocol, using raw-input and component controls to identify the sources of any gains.

3.1 Sampling

3.1.1 Single-view sampling

We fix the input configuration and examine how forecasting error changes with the number of samples. `Chronos`, `Sundial`, and `Moirai` generate $R = 128$ samples per prediction branch; we evaluate nested prefixes $N \in \{1, 8, 20, 32, 64, 128\}$ using pointwise means and medians. `TimesFM` returns a deterministic P50 forecast and participates in the multi-view comparisons. For the fixed-input diagnostic, we select the raw configuration by calibration MAE at $N = 128$ with median aggregation, then hold it unchanged across all sample counts and both aggregators.

The native sampling interfaces differ across models. In particular, `Chronos` shares a median continuation history beyond its first 64 predicted steps, so its 96-step sample paths are not independent complete trajectories. We therefore include a decoding-matched control in which each trajectory continues from its own sampled history.

3.1.2 Multi-view sampling

A view specifies the input history, its transformation, and any model-specific configuration. Motivated by classical decomposition methods such as STL (Cleveland, 1990) and multiscale forecasting in TimeMixer (Wang et al., 2024), we query the same frozen model with raw inputs and three families of transformed inputs:

Raw contexts. Each raw view directly forecasts the last L observations. We vary the context length and, for `Moirai`, cross each length with patch settings 128 and auto. These configurations test whether changing the raw input alone improves forecasting.

Cycle–residual views. Inspired by residual cycle forecasting in CycleNet (Lin et al., 2024), we estimate a periodic template by averaging historical observations at each phase, subtract it, and forecast the residual with the TSFM. We then add the repeated template to the residual forecast. Each dataset uses two prespecified periods: 24 and 168 for hourly data, 96 and 192 for 15-minute data, 144 and 288 for Weather, and 5 and 20 for Exchange. The template uses only observed history and introduces no learned model parameters.

Trend–residual views. We use a causal 24-point moving average to separate trend and residual components, following the decomposition principle in Autoformer (Wu et al., 2021). The first observation is replicated on the left to preserve context length. The frozen TSFM forecasts the two components separately, and their point forecasts are added to recover the full prediction.

Temporal subseries. Following the subseries representation in SutraNets (Bergsma et al., 2023), we present the model with lower-frequency sequences. We split the history into two interleaved phases, forecast each independently for $H/2$ steps, and interleave their predictions in future-phase order. This changes the temporal resolution seen by the model while retaining all observations.

Each context/patch configuration thus contributes five views: one raw, two cycle–residual, one trend–residual, and one temporal-subseries view. Table 4 in Appendix A.2 lists the context lengths and resulting candidate-pool sizes.

3.2 Aggregation

Self-consistency. Following the use of multiple samples in self-consistency (Wang et al., 2023; Liu et al., 2025a), we form point forecasts using pointwise means or medians. For stochastic models, we first aggregate samples within each prediction branch, then reconstruct a complete view forecast by restoring the cycle, adding trend and residual, or interleaving subseries. We subsequently combine complete view forecasts using a pointwise mean or median. This order matters: for example, adding the two branch medians in a trend–residual view need not equal the median of sample-wise sums.

Historical verification. The fixed verifier ranks views by their average forecasting MAE on calibration windows and either selects the best view or combines the three highest-ranked views by mean or median. Except in the fixed-input sampling diagnostic, these rankings are computed separately for each sample count and within-branch aggregator. Local variants update the choice using the two most recent completed H-step windows, $[t-2H, t-H)$ and $[t-H, t)$, at forecast origin t. All selection targets are therefore observed before the current prediction.

Diagnostic oracle. We also use the realized test horizon to select one complete view forecast per test window. This oracle measures the potential of the available candidates; it is unavailable at inference time and does not bound arbitrary forecast combinations.

3.3 Controlled Evaluation

Models and datasets. We evaluate `Chronos-T5-Base` (Ansari et al., 2024), `Sundial-Base-128M` (Liu et al., 2025b), `Moirai-1.0-R-Base` (Woo et al., 2024), and `TimesFM 3.0` (Das et al., 2024) on ETTh1, ETTh2, ETTm1, ETTm2, Electricity, Exchange, Traffic, and Weather. The main horizon is $H = 96$ steps. The common evaluation set comprises each dataset's designated OT series and four prespecified additional series each from Electricity, Traffic, and Weather, for 20 series in total. For these three datasets, expanded evaluations of `Moirai` and `TimesFM` use 16 additional series each. We report OT and additional-series groups separately and align cross-model comparisons to the same series and forecast origins.

Forecasting protocol. We use frozen-weight, history-calibrated univariate forecasting: weights remain fixed, while configuration selection and simple residual-model fitting may use observed

historical targets. Each series is standardized using its training-split mean and standard deviation, shared across all methods. We use fixed chronological ETT splits and 70%/10%/20% train/calibration/test splits for the remaining datasets, with at most 16 calibration and 32 test windows per series. Forecast origins lie on an H-step grid and are uniformly subsampled when the grid exceeds the window cap. Calibration precedes testing; sequential prediction allows local validation to use earlier test observations once they become available.

Raw-input and component controls. To isolate the added value of transformed views, we compare a historically selected raw configuration, an ensemble of all raw configurations, a historically selected raw top-k ensemble with $k = \min(3, V_{\text{raw}})$, and multi-view top-3 aggregation. The historical selectors share the same MAE criterion. To identify the TSFM's contribution within a cycle view, we hold the historical template and its configuration fixed while replacing the residual predictor with zero, the last residual, or a fitted autoregression. Naive, seasonal-naive, and history-selected simple forecasts provide additional references. Raw-only and component attribution analyses were added after inspecting the main results.

Evaluation metrics. MAE is the primary point-forecast metric, with MSE as a complementary measure; both use training-standardized units. We additionally specify MASE (Hyndman & Koehler, 2006) for scale-independent evaluation, using a seasonal-difference scale computed once from each series' training segment and shared across all methods and views. Historical selection continues to use the fixed MAE-based rules.

Cross-dataset summaries. Let $E_s(m)$ denote the mean test-window MAE of method m on series s. Against an explicitly named raw-input reference r, the relative change is $100[E_s(m)/E_s(r) - 1]$; negative values indicate improvement. We average these changes within each dataset and then equally across datasets, retaining separate OT and additional-series groups. To assess sensitivity to this summary, we additionally specify a dataset-balanced geometric mean of error ratios (Aksu et al., 2024; Qiao et al., 2026) and the proportions of series that improve, degrade, or tie.

Uncertainty and computation. We compute 95% intervals within each dataset and series group using 2,000 paired time-block bootstrap replicates, conditional on the generated sample pools and calibrated configurations. Runs use FP32 on an NVIDIA GeForce RTX 5080 GPU. Cost estimates separate current prediction, online historical validation, and amortized calibration, using synchronized prediction-API times. Sample and view counts are not common units of computation across models, and cached sample prefixes provide no measured runtime curves at smaller N.

Appendix A collects exact split boundaries, decoding settings, transformation mappings, selection variants, and statistical and timing conventions, together with data-quality limitations and supplementary controls.

4 Results

We evaluate sampling-based test-time scaling through seven research questions. **(RQ1)** How much does repeated sampling improve forecasting? (Section 4.1) **(RQ2)** Which input views improve forecasts, and on which tasks? (Section 4.2) **(RQ3)** Do multi-view gains persist beyond raw-input selection and aggregation? (Section 4.3) **(RQ4)** What does the TSFM contribute within a transformed view? (Section 4.4) **(RQ5)** When does aggregation improve on its constituent forecasts? (Section 4.5) **(RQ6)** Does expanding the candidate pool improve forecasting? (Section 4.6) **(RQ7)** When are the accuracy gains worth the inference cost? (Section 4.7) Unless noted, stochastic branches use Median of $N = 128$ samples and `TimesFM` uses P50. Relative changes average per-series error ratios; negative values indicate improvement.

4.1 How Much Does Repeated Sampling Improve Forecasting?

Initial gains and diminishing returns. Increasing N from 1 to 128 reduces average per-series MAE by 12.08% for `Chronos`, 13.47% for `Sundial`, and 32.87% for `Moirai` (Figure 2). The corresponding reductions from 64 to 128 are 0.05%, 0.22%, and 0.59%. The curves need not be monotone: `Sundial`'s average change from 32 to 128 samples is $+0.27\%$.

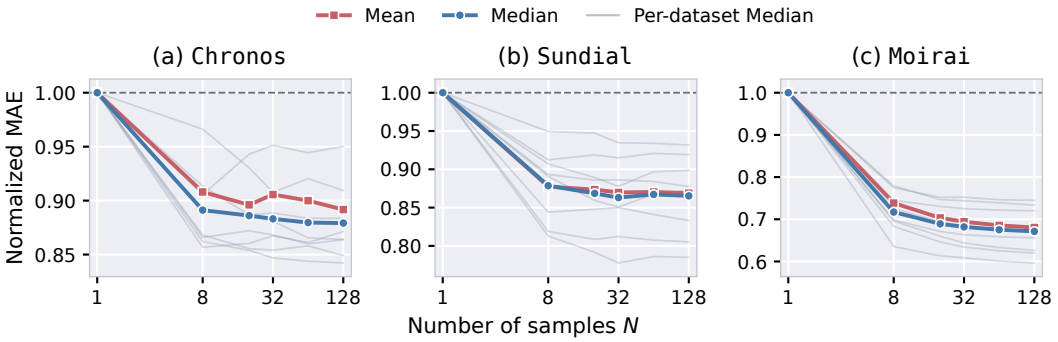

Figure 2: **Single-view sampling on eight OT series.** Historically selected raw configurations are fixed across N. Colored lines average per-series MAE ratios to $N = 1$; gray lines show individual Median curves. Panel ranges differ. Points are nested prefixes of one sample pool, without independent-seed intervals.

Table 1: **Illustrative cycle–residual improvements.** Changes and conditional 95% intervals compare Cycle + TSFM with raw top-k Median; the last column uses matched Cycle + AR. L/P gives context/period, with auto patch for `Moirai`. These post-hoc examples use $H = 96$ and historical selection. Dashed lines mark zero on the common $[-30, +10]\%$ interval scale.

Model	Series	L/P	vs raw top-k (%)	95% interval	Interval	vs AR (%)
`Moirai`	Elec. / 270	512/24	−10.22	[−18.00, −2.67]		−15.82
`Moirai`	Elec. / 282	512/24	−19.67	[−23.87, −14.38]		−1.80
`Sundial`	Traffic / 335	2048/168	−7.88	[−17.15, +0.16]		−8.21
`Chronos`	ETTm2 / OT	512/96	−13.70	[−28.73, +7.09]		−6.41

A reference for evaluating views. We assess views against sufficiently sampled raw forecasts at $N = 128$. For `Moirai`'s sixteen Electricity series, increasing raw samples from 64 to 128 improves MAE by only 0.57%; historical multi-view aggregation at $N = 128$ still improves over raw top-k by 10.04% (Table 2). These comparisons use different references and costs, but illustrate improvements beyond increasing the raw sample count. `Chronos`'s native paths share continuation history; an independent-trajectory control retains the qualitative sampling pattern (Appendix B.2).

4.2 Which Input Views Improve Forecasts?

Cycle–residual forecasts. Cycle views give several clear improvements over direct raw inference. Across `Moirai`'s eight OT tasks, historical selection within the cycle family improves average MAE by 4.05%, with six tasks improving; trend and subseries family selection instead increase average MAE by 13.48% and 18.69%. On four additional Electricity series, `Moirai`'s historically selected cycle view reduces MAE by 9.98% relative to selected raw, with a conditional 95% interval of $[-20.23, -1.26]\%$. Table 1 illustrates individual cycle forecasts that also improve over raw top-k. For Electricity column 270, `Moirai`'s MAE falls from 0.17776 for raw top-k to 0.15960 for the cycle forecast, a 10.22% reduction. The same forecast improves on a matched AR residual predictor by 15.82%. Configurations use historical selection; these post-hoc examples accompany the complete-group comparisons.

Other views and complete family comparisons. Trend and subseries views also have useful settings. `Moirai`'s historical single-view selector chooses a trend view on ETTh2 OT, improving MAE over selected raw by 4.15%, although its interval includes zero. In the separate `TimesFM` subseries experiment, historical top-3 Median improves over raw-only Median by 5.06% on ETTm2 at $H = 192$ ($[-9.65, -0.59]\%$). That pool has only two raw configurations, so raw-only Median equals raw top-k Median. Appendix B.4 reports complete family comparisons using calibration-selected configurations. The subseries horizon and pool remain separate from the common $H = 96$ evaluation.

Table 2: **Historical multi-view top-3 Median versus raw top-k Median.** Entries give relative MAE changes and conditional 95% intervals. Wins counts improvements over raw top-k; Beats all counts improvements over every constituent's test MAE. Four-series groups are complete for every model and are nested in the sixteen-series groups where available. Miniature intervals share a $[-25, +15]\%$ scale and dashed zero line; their colors denote direction, not significance.

Model	Dataset	n	ΔMAE (%)	95% interval	Interval	Wins	Beats all
Chronos	Electricity	4	−1.53	[−4.66, +1.65]		4/4	4/4
	Traffic	4	−0.54	[−5.18, +5.35]		2/4	2/4
	Weather	4	−1.11	[−6.20, +4.79]		2/4	1/4
Sundial	Electricity	4	+2.26	[−0.37, +5.11]		0/4	0/4
	Traffic	4	−6.92	[−13.04, −1.16]		4/4	4/4
	Weather	4	−0.06	[−4.14, +4.12]		2/4	2/4
Moirai	Electricity	4	−11.29	[−19.15, −3.72]		4/4	1/4
	Traffic	4	−0.28	[−4.94, +5.11]		1/4	2/4
	Weather	4	+1.09	[−2.72, +6.00]		1/4	2/4
TimesFM	Electricity	4	+0.81	[−3.18, +5.89]		1/4	1/4
	Traffic	4	+2.85	[−1.70, +9.93]		1/4	0/4
	Weather	4	−0.22	[−6.38, +6.83]		2/4	2/4
Expanded evaluation ($n = 16$ per dataset)							
Moirai	Electricity	16	−10.04	[−15.09, −5.17]		15/16	8/16
	Traffic	16	−4.20	[−8.83, +0.58]		13/16	9/16
	Weather	16	−8.04	[−16.46, −0.09]		10/16	8/16
TimesFM	Electricity	16	−0.65	[−2.26, +1.23]		11/16	12/16
	Traffic	16	+2.25	[−1.01, +6.14]		4/16	1/16
	Weather	16	−4.17	[−9.56, +1.26]		12/16	7/16

4.3 Do Multi-View Gains Persist Beyond Raw-Input Controls?

Gains beyond calibrated raw ensembles. Moirai improves over raw top-k by 10.04% on sixteen Electricity series ($[-15.09, -5.17]\%$), with fifteen series improving. Sundial improves by 6.92% on four Traffic series ($[-13.04, -1.16]\%$), with all four improving (Table 2). Across Moirai's 48 additional series, the dataset-balanced improvement is 7.42%, with 38 improvements, one tie, and nine degradations. The corresponding TimesFM average is only 0.86%, with 27 improvements and 21 degradations.

Why the raw reference matters. On four Moirai Traffic series, the apparent improvement is 22.48% against all-raw Median, but only 0.28% against historically selected raw top-k Median, with an interval spanning zero. On Electricity, the four-series improvement against raw top-k remains 11.29%. On the eight OT series, the same multi-view comparison yields changes of -2.90%, $+3.41\%$, -3.99%, and $+0.02\%$ for Chronos, Sundial, Moirai, and TimesFM, respectively. On twelve additional series, Chronos improves by 1.06%, with eight series improving. Complete macro summaries, absolute errors, geometric averages, and MSE/MASE comparisons appear in Appendix B.3.

4.4 What Does the TSFM Contribute Within a View?

Component improvement and complete forecasts. Figure 3 replaces only the residual predictor of each historically selected cycle view. On four Traffic series, TimesFM's cycle forecast improves MAE by 22.10% over Cycle + AR, yet remains 16.74% worse than selected raw. This distinction persists across sixteen Traffic series: the corresponding changes are -28.79% and $+17.85\%$.

Interpreting cycle gains. On four Electricity series, Moirai improves over raw by 9.98%, while its 3.54% improvement over AR has an interval spanning zero. On sixteen series, the changes become -5.80% against raw and $+1.04\%$ against AR, both with intervals spanning zero. The stronger 10.04% gain in Table 2 belongs to the full historical top-3 strategy and cannot be attributed to cycle alone. The individual cases also differ: Electricity column 282 improves by 19.67% over raw top-k, but only 1.80% over AR (Table 1). Appendix B.5 reports all direct paired component comparisons.

4.5 When Does Aggregation Improve on Its Constituents?

Combination can exceed each candidate. For `Moirai`, historical top-3 Median has lower test MAE than each of its three constituents on 8/16 Electricity series and 9/16 Traffic series (Table 2, Beats all). The three candidates are fixed using history; the best constituent's whole-test-segment MAE is a retrospective reference. Thus, combination provides gains beyond selecting any one of these three fixed candidates.

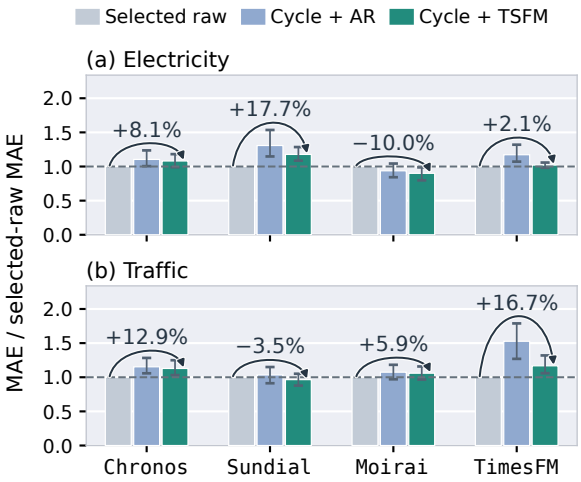

Figure 3: **Cycle forecasts with matched residual predictors.** Bars average per-series MAE relative to selected raw on four additional series per panel; whiskers give paired conditional 95% intervals. AR and TSFM share the historically selected cycle settings. Arrows give the TSFM forecast's percentage change from raw; negative is better.

Historical rules and candidate opportunities. On the eight OT series, `Moirai`'s all-view Median changes MAE by $+1.80\%$ relative to selected raw, historical single selection by -4.09%, and historical top-3 Median by -5.21% (Figure 4a). Selection is not uniformly beneficial: `Sundial`'s top-3 is 1.52% worse than selected raw. The future-informed single-view oracle improves average OT MAE by 20.31–30.52%, substantially more than the historical rules. It selects a complete view forecast separately at each origin using future observations; it is unavailable at inference time and does not bound arbitrary forecast combinations. Appendices B.6 and B.3 include all historical rules and the Mean/Median sensitivity comparison.

4.6 Does Expanding the Candidate Pool Improve Forecasting?

Additional Moirai patch sizes. Adding eight raw configurations with patch sizes 8/16/32/64 expands the full pool from 20 to 28 while retaining the historical selection rule (Figure 4b). On sixteen Traffic series, expanded top-3 Median is 5.28% worse than the original top-3 ($[+0.88, +11.14]\%$). The corresponding changes are -0.21% on Electricity and -0.02% on Weather. The original candidates remain available, yet the expanded search worsens Traffic forecasts under this historical rule.

4.7 When Are the Gains Worth the Inference Cost?

Useful gains and unchanged forecasts. On four `Moirai` Electricity series, historical top-3 improves MAE by 11.72% over selected raw, while total estimated cost rises from 0.25 to 1.18 seconds per origin (Figure 4c). On four `Chronos` Traffic series, historical single selection returns raw for every series, preserving its error but raising cost from 15.52 to 56.73 seconds once calibration is charged. Top-3 costs 88.93 seconds for a 0.54% improvement whose interval spans zero (Figure 4d).

Timing scope. These estimates amortize calibration over 32 origins and measure prediction-API throughput, including online probes but excluding external transformations. Thermal conditions were uncontrolled; the timings do not establish an intrinsic efficiency ranking across models.

5 Discussion and Limitations

Practical implications. Repeated sampling shows diminishing returns, while suitable views and aggregation can improve forecasts beyond strong raw baselines. These gains depend on the model, dataset, and historical selector: larger candidate pools can worsen selection, and better residual prediction need not improve the complete forecast. Aggregation can outperform every constituent, so the best single candidate does not describe all available gains. Practical comparisons should

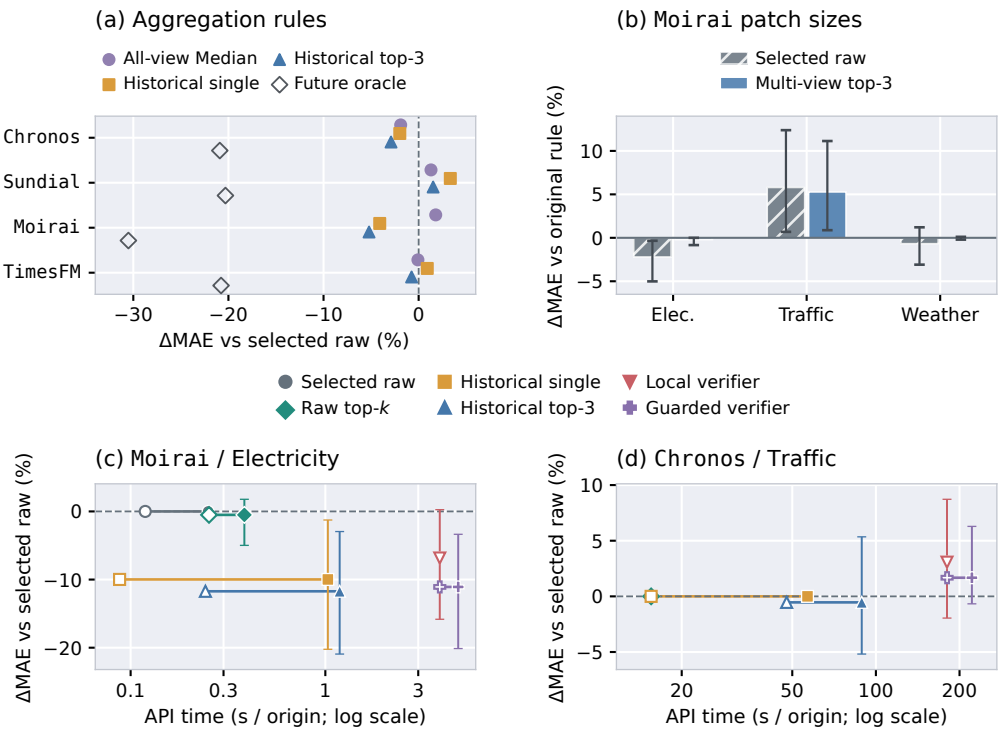

Figure 4: **Aggregation, candidate-pool expansion, and inference cost.** (a) Eight OT series; hollow diamonds denote the future-informed oracle. (b) `Moirai`, sixteen series per dataset: raw/full pools expand from 4/20 to 12/28. (c,d) Four series per dataset. Hollow markers charge prediction and online probes; filled markers add calibration over 32 origins. Lines connect identical forecasts; intervals describe conditional accuracy uncertainty.

account for prediction, calibration, and probing costs and use sufficiently sampled, historically calibrated raw baselines. Reliable historical selection, adaptive budget allocation, and reuse of calibrated choices remain directions for further study.

Selection transfer. Oracle improvements reveal useful forecasts within the candidate pool, but do not show that observed history can identify them. The gap between oracle and historical selection motivates evaluating whether configuration rankings transfer to later windows, rather than equating candidate availability with deployable gains.

Limitations. Our evidence covers four frozen checkpoints, mainly $H = 96$ univariate point forecasts, and finite sets of views and selection rules. It does not establish probabilistic calibration, universal scaling laws, or a causal benefit of pretraining. Sampling interfaces differ, uncertainty intervals condition on generated pools and calibrated choices, and some controls are post hoc. Pretraining-data overlap is unaudited. Weather normalization retains finite sentinel values and wind direction uses linear loss; prediction-API timings exclude external transformations and reflect uncontrolled thermal conditions. Appendices A and B.1 provide details and evaluation coverage.

6 Conclusion

We studied sampling-based test-time scaling for frozen TSFMs through single- and multi-view sampling, self-consistency, and historical verification. Repeated sampling improves point forecasts with diminishing returns; suitable views and historical aggregation can produce additional gains beyond calibrated raw-input ensembles. Cycle–residual cases and ensembles that outperform their constituents demonstrate two ways these gains arise. Their effectiveness depends on the model and task: improving a residual predictor need not improve the complete forecast, and expanding a candidate pool can worsen historical selection. Test-time scaling should therefore be assessed through both the quality of its candidates and the accuracy and total cost of the forecast produced using observed history.

AI USE STATEMENT

We used generative AI tools to help formulate research drafts and experimental comparisons, explore related literature, and develop and debug inference, evaluation, and plotting code. They also assisted in summarizing experimental outputs, drafting and revising the manuscript, and preparing figures, tables, and LaTeX source. AI-assisted code and numerical summaries were checked through executable tests, recomputation from cached forecasts, and consistency checks against experiment records. We take responsibility for the final text, claims, code, and figures, including material developed with AI assistance.

REPRODUCIBILITY STATEMENT

Section 3 specifies the inference choices, controls, and evaluation protocol. Appendix A provides data splits, model interfaces, decoding settings, view construction, historical selection rules, and statistical and timing conventions. Appendix B reports the evaluation coverage and supplementary comparisons. Quantitative figures and tables are generated from cached forecasts using shared analysis and plotting scripts. Source hashes, coverage, and plotted values are recorded with the analysis snapshot described in Appendix B.1. Code and numerical artifacts are available at https://anonymous.4open.science/r/forecasting-tts-fb18b6a3/.

REFERENCES

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, 2024. URL https://arxiv.org/abs/2410.10393.

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

Shane Bergsma, Tim Zeyl, and Lei Guo. Sutranets: Sub-series autoregressive networks for long-sequence, probabilistic forecasting. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=sC4RbbVKbu.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling, 2024. URL https://arxiv.org/abs/2407.21787.

R. B. Cleveland. Stl : A seasonal-trend decomposition procedure based on loess. 1990. URL https://api.semanticscholar.org/CorpusID:64570714.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10148–10167. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/das24c.html.

Ruijin Hua, Zichuan Liu, Kun Zhang, and Yiyuan Yang. Diversified scaling inference in time series foundation models, 2026. URL https://arxiv.org/abs/2601.17376.

Rob J Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22:679–688, 2006. URL https://api.semanticscholar.org/CorpusID:15947215.

Mert Kayaalp, Ali Caner Turkmen, Oleksandr Shchur, Pedro Mercado, Abdul Fatir Ansari, Michael Bohlke-Schneider, and Bernie Wang. Test-time efficient pretrained model portfolios for time series forecasting. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=iqUMjxfDNH.

Shengsheng Lin, Weiwei Lin, Xinyi Hu, Wentai Wu, Ruichao Mo, and Haocheng Zhong. Cyclenet: Enhancing time series forecasting through modeling periodic patterns. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 106315–106345. Curran Associates, Inc., 2024. doi: 10.52202/079017-3373. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/bfe7998398779dde03cad7a73b1f81b6-Paper-Conference.pdf.

Haoxin Liu, Zhiyuan Zhao, Shiduo Li, and B. Aditya Prakash. Evaluating system 1 vs. 2 reasoning approaches for zero-shot time series forecasting: A benchmark and insights, 2025a. URL https://arxiv.org/abs/2503.01895.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39295–39317. PMLR, 13–19 Jul 2025b. URL https://proceedings.mlr.press/v267/liu25be.html.

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 *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=79TgfXHbsK.

Yunzhong Qiu, Zhiyao Cen, Zhongyi Pei, Chen Wang, and Jianmin Wang. Adapt data to model: Adaptive transformation optimization for domain-shared time series foundation models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=uTK1SNgi1N.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and JUN ZHOU. Timemixer: Decomposable multiscale mixing for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7oLshfEIC2.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53140–53164. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/woo24a.html.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=J4gRj6d5Qm.

A Experimental Details

A.1 Data and Forecasting Protocol

Splits and forecast origins. We use half-open, zero-based train, calibration, and test intervals. Table 3 records the main $H = 96$ boundaries and window counts. For non-ETT datasets of length T, the training endpoint is $\lfloor 0.7T \rfloor$ and the calibration endpoint is $T - \lfloor 0.2T \rfloor$. The fixed ETT boundaries leave the remaining CSV tail unused. Within an interval $[a, b)$, eligible origins are $a, a + H, \ldots$ with the full forecast horizon contained in the interval. If the grid exceeds the window cap, we select uniformly spaced grid indices, including its endpoints, using integer-valued `linspace`. The caps are 16 calibration windows and 32 test windows.

Table 3: Main forecasting splits and numbers of selected windows. All endpoints are exclusive.

Dataset	Train end	Calibration end	Test end	Cal. windows	Test windows
ETTh1 / ETTh2	8640	11520	14400	16	30
ETTm1 / ETTm2	34560	46080	57600	16	32
Electricity	18412	21044	26304	16	32
Exchange	5311	6071	7588	7	15
Traffic	12280	14036	17544	16	32
Weather	36887	42157	52696	16	32

Available history. At origin t, local validation forecasts the two completed windows $[t - 2H, t - H)$ and $[t - H, t)$ from their respective preceding histories. Both targets have been observed by t. Later forecast origins may therefore use outcomes from earlier parts of the test interval. This is sequential forecasting, rather than predicting the entire test segment before receiving any new observations.

Series groups. Each dataset includes its designated OT series. For Electricity, Traffic, and Weather, a permutation of the non-OT columns using `default_rng(20260924)` specifies 16 additional series per dataset. `Moirai` and `TimesFM` use all 16; `Chronos` and `Sundial` use the first four from the same list. The common set therefore consists of eight OT series and 12 additional series, with models aligned by series, horizon, and origin. The shared non-OT columns are Electricity `270, 157, 242, 266`, Traffic `539, 335, 471, 206`, and Weather `max. wv (m/s), wd (deg), wv (m/s), Tdew (degC)`. Results from the four-series and 16-series groups are distinguished explicitly.

Preprocessing and limitations. Each series uses its training-split mean and population standard deviation, with the latter floored at 10^{-6}. Non-finite observations are forward-filled, with leading missing values set to zero; calibration, local-validation, and test targets are required to be finite. Finite entries equal to -9999 remain in the training normalization statistics of three selected Weather series: wind speed, maximum PAR, and OT. The checked generation contexts and targets do not directly intersect these entries, but the resulting normalization can still affect predictions. No cleaned-data rerun has been performed. Wind direction uses ordinary linear-coordinate errors, without a circular angular-loss sensitivity analysis. Pretraining-data overlap has not been audited.

A.2 Model Interfaces and Sample Pools

Checkpoints and configurations. Table 4 specifies the loaded model variants and common pools. `Moirai` crosses each context length with patch settings 128 and auto. V counts configurations: two cycle periods contribute two views, and a view may require multiple prediction branches. Prediction batch sizes are 2, 4, 4, and 16 for `Chronos`, `Sundial`, `Moirai`, and `TimesFM`, respectively.

Checkpoint release dates. `Chronos-T5-Base` weights released in March 2024; `Sundial-Base-128M` weights released in May 2025; `Moirai-1.0-R-Base` weights released in March 2024; `TimesFM 3.0` weights released in August 2026.

Table 4: Model checkpoints and common candidate pools. V_{raw} counts raw configurations and V counts all views; a view may require multiple prediction branches.

Checkpoint	Context lengths	V_{raw}	V	Output per branch
`Chronos-T5-Base`	512	1	5	128 samples
`Sundial-Base-128M`	512, 2048	2	10	128 samples
`Moirai-1.0-R-Base`	512, 2048	4	20	128 samples
`TimesFM 3.0`	512, 2048	2	10	Deterministic P50

Sample counts and decoding. Stochastic models generate $R = 128$ samples per branch, with nested prefixes $N \in \{1, 8, 20, 32, 64, 128\}$ evaluated under both mean and median aggregation. `Chronos` uses temperature 1.0, top-k 50, and top-p 1.0. `Moirai` samples from a distribution produced by its backbone; auto patch selection additionally compares historical likelihoods internally. `Sundial` uses 50 flow-sampling steps and a 720-step output block. Thus, 128 samples do not imply 128 complete backbone executions for every model. `TimesFM` uses independent one-dimensional queries and returns deterministic P50 forecasts. Generation uses one predetermined seed scheme indexed by task, view, branch, and batch; it does not provide independent seed replications.

`Chronos` continuation control. For $H = 96$, `Chronos`'s native interface continues beyond 64 steps using a median history shared across the $R = 128$ first-block samples. Its sample prefixes are therefore neither independent full-horizon trajectories nor equivalent to fresh calls with `num_samples`$= N$. The decoding-matched independent control uses ETTh1, ETTm2, Electricity, and Traffic OT series, with eight uniformly selected test origins per series. Each of 128 trajectories is generated with its own continuation, using batches of 16 copied contexts and one sample per context. Raw and historically selected view configurations are inherited from native calibration. Native and independent forecasts are compared on the same eight origins; smaller-N prefixes still do not constitute separately timed runs.

Fixed-input sampling diagnostic. We select the raw configuration using calibration MAE at $N = 128$ with median aggregation, then hold it constant across all N and both aggregators. Its $N = 1$ result therefore conditions on this calibration. Other selection comparisons recalibrate separately for each sample count and within-branch aggregation rule.

A.3 View Construction and Mapping to the Forecast Horizon

Common candidate pool. Each context/patch configuration contributes raw, cycle(P_1), cycle(P_2), trend24, and subseries2 views. The period pairs are 24/168 for hourly datasets, 96/192 for 15-minute datasets, 144/288 for Weather, and 5/20 for Exchange. These values are fixed without test-set search. The full pools require 7, 14, 28, and 14 prediction branches per origin for `Chronos`, `Sundial`, `Moirai`, and `TimesFM`, respectively. Branch counts are interface-level counts and do not include all internal work such as `Moirai`'s auto patch evaluation.

Cycle–residual mapping. For a context $x_{0:L-1}$ and period P, define $c_r = \text{mean}\{x_i : 0 \le i < L,\ i \bmod P = r\}$. The TSFM receives $x_i - c_{i \bmod P}$. At future offset h, its residual prediction is added to $c_{(L+h) \bmod P}$. CycleNet (Lin et al., 2024) learns recurrent cycle parameters; our control uses the directly estimated historical phase template.

Trend–residual mapping. The trend is a causal 24-point moving average, with the first observation replicated on the left to preserve context length. We predict the trend and the residual $x - \text{trend}$ separately and add the two point forecasts. This is an input decomposition, not smoothing of the raw forecast. STL (Cleveland, 1990) provides a classical seasonal–trend decomposition reference, while moving-average neural decomposition is exemplified by Autoformer (Wu et al., 2021); our implementation uses the causal moving average described here.

Subseries mapping. For factor k, branch r contains $x_r, x_{r+k}, \ldots$ and forecasts H/k steps. Its predictions fill future positions satisfying $(L + h) \bmod k = r$. This alignment is necessary when L is not divisible by k. SutraNets (Bergsma et al., 2023) model dependencies across subseries

autoregressively; our frozen-model queries predict the branches independently. TimeMixer (Wang et al., 2024) motivates the use of temporal resolutions, but its learned multiscale mixing blocks are not part of our inference procedure.

Order of operations. For stochastic models, take the mean or median of the first N samples separately within each branch. Next, restore a cycle, sum trend and residual, or interleave subseries to obtain one complete view point forecast. Only then combine forecasts across views. In particular, the sum of branch medians is generally different from the median of sample-wise sums.

A.4 Historical Selection and Diagnostic Oracle

Historical scores. For a specified sample count and within-branch aggregator, let $\ell(u, v)$ denote view v's complete-horizon MAE at origin u. With calibration origins $\mathcal{C}$, define

$$a_v = \frac{1}{|\mathcal{C}|} \sum_{u \in \mathcal{C}} \ell(u, v), \qquad b_{t,v} = \frac{\ell(t - 2H, v) + \ell(t - H, v)}{2}. \tag{1}$$

Fixed raw, view, and cycle selection minimize a_v within the raw, full, and cycle pools, respectively. Top-3 selection takes the three lowest-scoring views and combines their current forecasts by mean or median. Raw top-k applies the same criterion within the raw pool, with $k = \min(3, V_{\text{raw}})$. Ties in historical scores follow the fixed candidate order.

Family-specific reporting. To compare view families, we additionally minimize the same calibration score a_v separately within the cycle, trend, and subseries pools. The chosen configuration is then held fixed throughout the test segment. This reporting extension is computed from existing prediction caches after inspecting the main results; it does not tune on test outcomes or introduce additional model calls. Cycle selection reproduces the existing fixed-cycle rule. We report every family and retain the original OT, first-four, and sixteen-series groups, including unfavorable results.

Aggregation versus its constituents. For each series, the three top-ranked configurations are fixed using calibration MAE. We compare the test MAE of their pointwise Median with the minimum of their three individual test MAEs, each averaged over all test origins. The Beats all count requires a reduction greater than 10^{-12}. The best constituent is identified retrospectively for this diagnostic; it is not used to generate the aggregate or to select an inference rule. This comparison differs from the per-origin, future-informed oracle below.

Local and guarded rules. Local selection minimizes $b_{t,v}$; the prior–local variant minimizes $0.75a_v + 0.25b_{t,v}$. The guard accepts the latter candidate only if it strictly beats the selected raw forecast on each local window and has mean local MAE below 0.98 times raw. Otherwise it returns raw, while retaining the cost of all completed probes. The coefficients and threshold are fixed. Unfiltered all-view and all-raw ensembles use no external historical selection.

Future-informed oracle. The oracle uses the realized test horizon to select one complete, already aggregated view forecast per origin. It neither stitches together time points from different views nor selects from the union of all raw stochastic samples. It is a diagnostic of the available view forecasts, rather than a deployable rule or an error bound for arbitrary ensembles.

A.5 Controls and Supplementary Comparisons

Simple and component controls. Naive repeats the last observation; seasonal naive repeats the last P observations. For the historically selected cycle configuration, the component controls retain its context, period, patch setting, and template. The residual predictor is replaced by zero, the last observed residual, or a ridge autoregression with an unpenalized intercept, penalty 1, and lag $\min(P, 24, \lfloor L/4 \rfloor)$, recursively forecast for H steps. The half-raw/half-template blend retains a TSFM call. A separate historical selector chooses among the fixed simple-control pool. Component replacements identify the residual predictor's contribution under the matched configuration; they do not independently optimize each replacement or isolate the causal effect of pretraining.

Expanded candidate pools. The `Moirai` patch control adds eight raw configurations: context lengths 512/2048 crossed with patches 8/16/32/64. The raw pool expands from 4 to 12 and the full pool from 20 to 28, with the original historical selection rules retained. Only calibration and test predictions are added; this is not a complete view-by-patch grid or an expansion of local-probe candidates. The `TimesFM` OT-specific subseries comparison uses raw and subseries factors 2, 3, and 4 at both context lengths, at horizons 96 and 192. Its eight-view pool is distinct from the common ten-view pool and has no cycle or trend candidates.

Adapted TATO comparison. Our finite-budget adaptation of TATO (Qiu et al., 2026) uses the authors' code with `Moirai` (patch 128) and the median of 128 samples. It evaluates nested prefixes of 20 and 100 TPE trials, with up to 16 training and 16 calibration windows, and retains both the weighted criterion and MAE reranking of the same shortlist. The maximum context is 1440, with a vanilla context of 672. Its inputs use filled original units, while the common-pool experiments use training-standardized inputs. Differences in candidate search, history use, and preprocessing require separate reporting; this adaptation is not a full reproduction of the original study or a matched-cost common-pool comparison.

A.6 Metric Definitions and Cross-Dataset Summaries

Point errors and MASE. Let $E_s(m)$ be method m's MAE averaged over selected test windows on series s, in training-standardized units. MSE is computed on the same predictions and windows. For supplementary scale-independent evaluation, define MASE (Hyndman & Koehler, 2006) as

$$q_s = \frac{1}{T_s - p_s} \sum_{i=p_s+1}^{T_s} |z_{s,i} - z_{s,i-p_s}|, \qquad \mathrm{MASE}_s(m) = \frac{E_s(m)}{q_s}, \tag{2}$$

where $z_{s,1:T_s}$ is the training segment in the same standardized units. The fixed lag p_s is 24 for hourly data, 96 for 15-minute data, 144 for Weather, and 5 for Exchange. The scale uses the full training segment before any view transformation, is shared across methods and models, and remains fixed across forecast origins. The existing preprocessing is retained, including the Weather limitation above. MASE values are averaged within datasets and then equally across datasets, separately for each series group.

Arithmetic relative change. Against a named reference r, define $R_s = E_s(m)/E_s(r)$ and $\Delta_s = 100(R_s - 1)$. We first average Δ_s across series within a dataset and then equally across datasets in the group. Negative values indicate improvement. The primary references are the historically selected raw configuration and raw top-k ensemble; other controls are named explicitly in each comparison.

Geometric relative change. For each model and prespecified series group, define

$$G(m; r) = \exp\left[\frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \frac{1}{|\mathcal{S}_d|} \sum_{s \in \mathcal{S}_d} \log R_s\right], \tag{3}$$

where $\mathcal{D}$ contains the datasets and $\mathcal{S}_d$ the selected series in dataset d for that group. Thus, datasets receive equal total weight regardless of series count. $G < 1$ indicates improvement; $100(G - 1)$ expresses the relative change as a percentage. This is a paired, raw-reference summary motivated by relative evaluation in GIFT-Eval (Aksu et al., 2024) and TIME (Qiao et al., 2026); it is not a leaderboard score for either benchmark. Because q_s is fixed, the per-series MASE ratio equals R_s whenever defined and supplies no independent evidence of relative improvement.

Improvement, degradation, and ties. We compare each series' mean test MAE with its reference. A tie satisfies $|E_s(m) - E_s(r)| \le 10^{-12} + 10^{-6} E_s(r)$; differences outside this fixed tolerance determine improvement or degradation by their sign. Each proportion is computed within a dataset and then averaged equally across datasets. The rates are descriptive and do not imply statistical significance.

Boundary cases and reporting scope. A zero seasonal scale leaves MASE undefined. A zero reference error leaves the corresponding error ratio and full-group relative summaries undefined; affected-series counts and absolute errors are retained. When every reference error is positive, a zero method error gives a zero geometric mean. No epsilon is added, and undefined cases are not silently removed to create a complete-group score. All summaries retain OT/additional-series separation and explicit four-series/16-series group sizes. The supplementary metrics evaluate the existing forecasts without changing the MAE-based historical selectors.

A.7 Conditional Uncertainty

Paired resampling. Dataset/group-level 95% intervals use 2,000 paired bootstrap replicates with `default_rng(20260924)`. We resample circular blocks of $\min(4, n_{\text{test}})$ selected forecast origins, using the same origin indices across series in a group, and then resample series with replacement. Method and reference share all sampled indices. The training scales, calibration decisions, and generated samples remain fixed. For an OT group with one series, uncertainty comes only from origin-block resampling.

Interpretation. Adjacent elements in a block refer to the selected-origin list, which may be a subsampled grid; they need not cover consecutive $4H$ time steps. The intervals condition on the current sample pools and selected series, do not cover independent generation seeds, and are not corrected for multiple comparisons. An interval spanning zero does not establish equivalence. Dataset-level intervals do not constitute a confidence interval for an across-dataset macro summary.

A.8 Computation Accounting

Measured prediction time. Runs use FP32 on an NVIDIA GeForce RTX 5080 Laptop GPU. For each view v, synchronized prediction-API wall time T_v over its unique generated origins $\mathcal{O}_v$ gives $c_v = T_v/|\mathcal{O}_v|$. This is an estimate under the recorded batching conditions, including operations inside the API. It excludes external decomposition and aggregation, AR fitting, initial weight loading, and file I/O.

Deployment estimates. Estimated per-origin cost is current prediction cost plus online historical-probe cost and necessary calibration amortized over the evaluated test origins. Historical raw selection charges all searched raw configurations; full-pool selection and top-3 charge the entire calibration pool. A single predetermined raw configuration incurs no unnecessary external selection calibration. Local rules charge two full-pool historical probes per origin, including probes performed before a guarded fallback. Internal model work, such as `Moirai`'s auto patch evaluation, remains part of c_v.

Timing limitations. The timings describe observed API throughput, not end-to-end request latency, FLOPs, or total project compute including interrupted and discarded work. Hardware thermal conditions were not controlled, and thermal throttling was observed during some `Chronos` runs. Times are therefore unsuitable for attributing small differences to intrinsic model efficiency. Cached $R = 128$ sample prefixes supply no separately measured runtime at smaller N; we do not extrapolate such curves.

B Additional Results

B.1 Evaluation Coverage

The results use a 26 September 2026 snapshot of completed, independently checked forecasts. All eight OT series are available for all four models. `Moirai` and `TimesFM` have sixteen additional series in each of Electricity, Traffic, and Weather. `Chronos` and `Sundial` have all four prespecified additional series in each of the three datasets. The four-model common evaluation set is complete: eight OT and twelve additional series per model, with the two groups reported separately. The snapshot includes 160/160 main execution tasks and 152/152 common-family tasks after excluding the sixteen `TimesFM` subseries-specific tasks and including its eight common-family OT tasks. The `Moirai` small-patch controls, `Chronos` independent-trajectory controls, and adapted TATO

comparison are complete at 56, 4, and 8 tasks, respectively. These counts describe executions, not independent datasets.

All newly reported MASE values use the fixed training scales in Appendix A.6. The historical selectors, predictions, preprocessing, and original MAE/MSE values are unchanged. All 56 training scales are positive, and the reported reference errors are nonzero. Numerical figure points and source hashes accompany the plotting scripts, including all available series rather than only the illustrated examples. Weather retains the preprocessing and angular-error limitations described in Appendix A.1.

B.2 Sampling and the Chronos Interface Control

MSE and within-view aggregation. Figure 5 repeats the fixed-input OT comparison with MSE. It uses the same generated pools, raw configurations, and test origins as the MAE curves. Mean and Median are reported separately; their finite-sample behavior need not be monotone or identical across error metrics. The faint dataset curves show variation across tasks, rather than repeated-seed uncertainty.

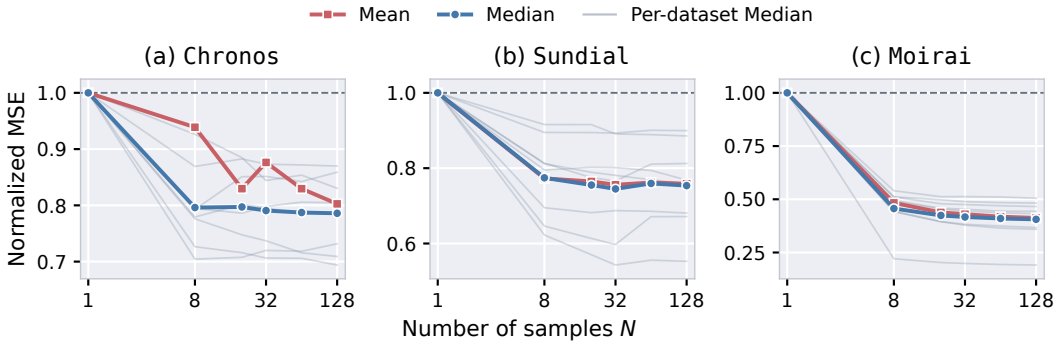

Figure 5: **Fixed-input sampling under MSE.** Thick curves average per-series error ratios over the eight OT datasets; faint curves show individual Median trajectories. Vertical ranges differ across panels. The configuration is selected at $N = 128$ by historical Median MAE and then held fixed. Thus, even $N = 1$ conditions on the same historical selection.

Independent full-horizon trajectories. Figure 6 shows all five unique configurations from the four completed `Chronos` controls. The raw and historically selected configurations coincide on ETTh1, Electricity, and Traffic, so each is shown once. On ETTm2 they differ and both are retained. Each comparison uses the same eight selected test windows, with configuration choices inherited from native calibration. For the four raw configurations, independent Median aggregation reduces MAE by 10.90–26.50% from $N = 1$ to 128 and by 0.006–1.47% from 64 to 128. Relative to the native $N = 128$ forecasts on those same windows, independent trajectories change MAE by -2.95% on ETTh1, $+4.42\%$ on ETTm2, -2.05% on Electricity, and -0.49% on Traffic. Thus, independent continuation preserves the qualitative sampling pattern while affecting the absolute forecast errors in both directions. The runs use one generation-seed scheme, and their prefixes are not separately timed budgets.

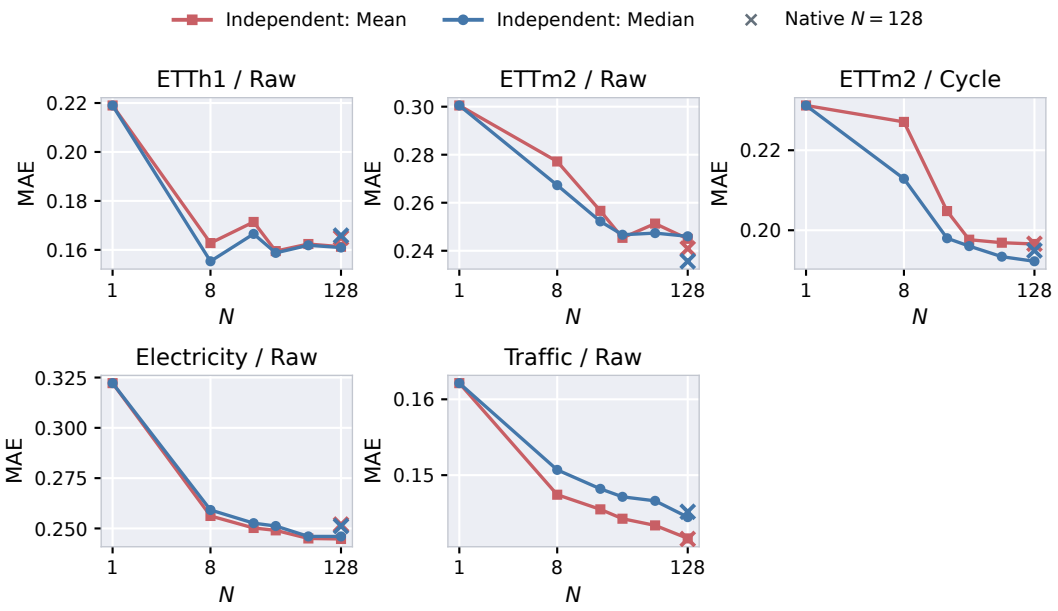

Figure 6: **`Chronos` independent-trajectory controls.** Lines use independent full-horizon sample prefixes; crosses show native $N = 128$ on the same eight origins. Each panel has its own MAE scale. Configurations are inherited from historical calibration, with no test-based reselection and no duplicated raw/selected-view panels.

B.3 Absolute Metrics, References, and Additional Series

Complete macro summaries and raw controls. Table 5 retains all complete OT and additional-series macro comparisons, including the 48-series `Moirai` and `TimesFM` groups. All-raw and raw top-k separate combination from historical configuration selection. The arithmetic and geometric summaries use the same paired per-series ratios and dataset weights; no across-dataset confidence interval is inferred from the dataset-level intervals.

Absolute errors and training-scaled evaluation. Table 6 reports absolute MAE and MSE in training-standardized units, together with MASE, on all eight OT series. The scale for MASE is shared by methods within each series; consequently, per-series relative MAE and MASE ratios agree. Absolute MASE aggregates answer a different question from macro averages of paired relative errors, since they retain each series' error magnitude after seasonal scaling. We do not interpret a MASE value below one as proof of beating a seasonal-naive forecast on the test windows: its denominator is the training seasonal-difference error.

Sensitivity to the metric and aggregation rule. Table 7 retains both cross-view Mean and Median and both MAE and MSE. Within-view aggregation remains Median throughout this table, separating this comparison from the within-view Mean/Median sampling curves. `Moirai`'s advantage over raw top-k is larger with cross-view Mean than with Median on both complete groups; this does not retrospectively select a different aggregator for each dataset. `Sundial`'s additional-series MAE gain reverses under MSE, and `TimesFM`'s small OT MAE changes do not translate into MSE gains. MASE is supplementary and does not change any calibrated decisions.

Dataset-level effects and larger groups. Table 2 reports all available additional-series groups and conditional intervals. The first-four and sixteen-series groups overlap and are not independent replications. For example, `Moirai`'s Traffic effect against raw top-k is -0.28% on four series and -4.20% on sixteen; `TimesFM`'s corresponding changes are $+2.85\%$ and $+2.25\%$. `Moirai`'s sixteen-series groups improve in 38 of 48 individual series, tie in one, and degrade in nine; the dataset-balanced arithmetic and geometric changes are -7.42% and -7.96%. For `TimesFM`

Table 5: **Complete macro comparisons against raw-input controls.** The first three columns use selected raw as the reference; the next two compare multi-view top-3 Median directly with raw top-k Median. I/T/D counts improvements, ties, and degradations against raw top-k. First-four groups are nested in sixteen-series groups. All four models have complete OT and twelve-series additional groups.

	Against selected raw (%)			Multi-view vs raw top-k (%)		
Model	**All raw**	**Raw top-k**	**Multi-view**	**Arithmetic**	**Geometric**	**I/T/D**
Eight OT series; eight datasets						
Chronos	0.00	0.00	-2.90	-2.90	-2.98	6/0/2
Sundial	-1.81	-1.81	+1.52	+3.41	+3.34	1/0/7
Moirai	+6.61	-1.18	-5.21	-3.99	-4.13	7/0/1
TimesFM	-0.79	-0.79	-0.74	+0.02	-0.01	4/0/4
Twelve additional series; four per dataset						
Chronos	0.00	0.00	-1.06	-1.06	-1.09	8/0/4
Sundial	+0.28	+0.28	-1.26	-1.57	-1.67	6/0/6
Moirai	+16.20	-0.86	-4.35	-3.50	-3.78	6/0/6
TimesFM	+1.90	+1.90	+3.10	+1.15	+1.10	4/0/8
48 additional series; sixteen per dataset						
Moirai	+17.89	-0.74	-8.13	-7.42	-7.96	38/1/9
TimesFM	+1.92	+1.92	+0.99	-0.86	-0.95	27/0/21

these changes are -0.86% and -0.95%, with 27 improved and 21 degraded series. These group-dependent results qualify the scope of any general claim about multi-view inference.

Table 6: **Absolute OT metrics.** Raw is the historically selected raw configuration; raw top-k and multi-view top-3 use Median across configurations. Stochastic branches use Median of 128 samples, and `TimesFM` uses P50. All comparisons have $H = 96$ and identical test origins within a dataset. Model interfaces differ, so this is not an equal-compute ranking.

	MAE			MASE		MSE	
Dataset	**Raw**	**Raw top-k**	**Multi-view**	**Raw**	**Multi-view**	**Raw**	**Multi-view**
`Chronos`							
ETTh1	0.1712	0.1712	0.1742	0.6154	0.6261	0.0497	0.0524
ETTh2	0.2775	0.2775	0.2725	1.0580	1.0390	0.1315	0.1284
ETTm1	0.1148	0.1148	0.1126	0.4113	0.4035	0.0263	0.0241
ETTm2	0.2104	0.2104	0.1910	0.8039	0.7298	0.0879	0.0687
Electricity	0.2908	0.2908	0.2866	0.5454	0.5376	0.1645	0.1621
Exchange	0.2407	0.2407	0.2447	3.2686	3.3226	0.0966	0.0951
Traffic	0.1650	0.1650	0.1585	0.6495	0.6241	0.1132	0.1009
Weather	0.0271	0.0271	0.0248	0.2815	0.2581	0.0013	0.0012
`Sundial`							
ETTh1	0.1686	0.1682	0.1685	0.6061	0.6056	0.0528	0.0518
ETTh2	0.2608	0.2566	0.2664	0.9943	1.0158	0.1214	0.1235
ETTm1	0.1170	0.1135	0.1193	0.4194	0.4276	0.0263	0.0265
ETTm2	0.1820	0.1725	0.1728	0.6955	0.6600	0.0647	0.0572
Electricity	0.2243	0.2239	0.2476	0.4207	0.4645	0.0949	0.1196
Exchange	0.2850	0.2661	0.2853	3.8694	3.8733	0.1254	0.1259
Traffic	0.1720	0.1668	0.1645	0.6773	0.6475	0.0997	0.0957
Weather	0.0206	0.0217	0.0221	0.2141	0.2294	0.0009	0.0010
`Moirai`							
ETTh1	0.2002	0.1911	0.1881	0.7196	0.6760	0.0657	0.0572
ETTh2	0.2841	0.2852	0.2697	1.0833	1.0283	0.1350	0.1242
ETTm1	0.1200	0.1182	0.1142	0.4300	0.4093	0.0261	0.0232
ETTm2	0.2082	0.2123	0.2088	0.7952	0.7977	0.0787	0.0781
Electricity	0.2747	0.2787	0.2514	0.5153	0.4716	0.1443	0.1251
Exchange	0.2371	0.2220	0.2370	3.2191	3.2183	0.0921	0.0921
Traffic	0.1916	0.1897	0.1681	0.7545	0.6616	0.1251	0.1000
Weather	0.0274	0.0274	0.0259	0.2844	0.2695	0.0014	0.0013
`TimesFM`							
ETTh1	0.1770	0.1731	0.1677	0.6360	0.6029	0.0541	0.0508
ETTh2	0.2533	0.2593	0.2583	0.9657	0.9848	0.1163	0.1187
ETTm1	0.1013	0.1012	0.1025	0.3630	0.3673	0.0204	0.0204
ETTm2	0.1882	0.1804	0.1863	0.7190	0.7117	0.0766	0.0730
Electricity	0.2739	0.2580	0.2502	0.5137	0.4692	0.1475	0.1251
Exchange	0.2511	0.2500	0.2514	3.4092	3.4137	0.0995	0.0999
Traffic	0.1258	0.1292	0.1346	0.4955	0.5300	0.0672	0.0757
Weather	0.0215	0.0218	0.0212	0.2234	0.2204	0.0010	0.0009

Table 7: **Metric and aggregation sensitivity.** Relative changes compare multi-view top-3 to raw top-k using the same cross-view aggregation. "Extra 4" contains four additional series per dataset, twelve in total. The final two columns give dataset-balanced absolute MASE for selected raw and multi-view Median.

		MAE change (%)		MSE change (%)		MASE	
Group	Model	Mean	Median	Mean	Median	Raw	Multi-view
OT	Chronos	-2.84	-2.90	-6.75	-6.63	0.954	0.943
	Sundial	+2.74	+3.41	+4.29	+6.08	0.987	0.990
	Moirai	-7.14	-3.99	-7.43	-5.22	0.975	0.942
	TimesFM	-0.04	+0.02	+0.31	+1.58	0.916	0.913
Extra 4	Chronos	+0.16	-1.06	-8.04	-6.52	0.804	0.792
	Sundial	-1.34	-1.57	+1.00	+2.03	0.776	0.764
	Moirai	-12.03	-3.50	-14.31	-5.89	0.878	0.828
	TimesFM	+0.77	+1.15	-0.83	+0.81	0.754	0.774

B.4 Historically Selected View Families

Selection within each family. Tables 8 and 9 compare calibration-selected cycle, trend, and subseries configurations with selected raw. We reconstruct the choices and test errors from all 152 completed common-family tasks, at $N = 128$ or deterministic P50. The raw and cycle choices reproduce the existing selections, and all candidate test errors agree with the original caches. The family comparisons were added during result analysis; all configurations are chosen using calibration outcomes alone. They differ from an oracle that reports the best test configuration within each family.

Table 8: **Complete view-family macro summaries.** Entries give relative MAE changes (%) against selected raw, followed by improved/total series. Each family uses its own historical calibration winner. These descriptive macro summaries do not have across-dataset uncertainty intervals. Additional groups contain three datasets; the twelve-series groups are nested in the 48-series groups where available.

Group	Model	Cycle	Trend	Subseries
Eight OT	Chronos	+3.81 (2/8)	+14.19 (2/8)	+7.85 (1/8)
	Sundial	+6.50 (1/8)	+9.54 (0/8)	+12.72 (2/8)
	Moirai	−4.05 (6/8)	+13.48 (2/8)	+18.69 (4/8)
	TimesFM	+4.24 (3/8)	+2.62 (3/8)	+6.60 (3/8)
Twelve additional	Chronos	+9.52 (1/12)	+15.12 (1/12)	+16.25 (1/12)
	Sundial	+6.24 (5/12)	+10.05 (0/12)	+14.13 (0/12)
	Moirai	+0.11 (4/12)	+15.18 (1/12)	+28.02 (0/12)
	TimesFM	+7.68 (1/12)	+7.42 (0/12)	+11.22 (1/12)
48 additional	Moirai	+0.13 (26/48)	+18.97 (6/48)	+31.13 (5/48)
	TimesFM	+9.15 (6/48)	+6.90 (8/48)	+11.26 (3/48)

Scope of the cycle examples. Moirai's cycle family improves six of eight OT tasks, with a mean relative change of -4.05%, but its mean change across 48 additional series is $+0.13\%$. On the sixteen-series groups, cycle changes MAE by -5.80% on Electricity, $+4.22\%$ on Traffic, and $+1.96\%$ on Weather. The full top-3 strategy improves all three dataset averages against raw top-k, illustrating that the value of the selected aggregate is not described by any one family average. Trend and subseries have individual useful settings but predominantly unfavorable family averages under this common-pool protocol. The separate TimesFM subseries experiment changes both the available factors and, at $H = 192$, the horizon; its positive ETTm2 result should not be merged with this table.

Table 9: **View-family effects on all additional-series groups.** Entries are relative MAE changes (%) and conditional 95% intervals against selected raw. Each family's configuration is chosen by historical MAE. All first-four groups are complete; the sixteen-series groups extend the same pre-specified sequence of columns.

Model	Dataset	Cycle	Trend	Subseries
Shared evaluation ($n = 4$ per dataset)				
Chronos	Electricity	+8.09 [−1.01, +18.10]	+4.84 [−0.47, +11.05]	+14.56 [+6.03, +24.31]
	Traffic	+12.87 [+3.09, +24.94]	+36.52 [+23.68, +50.28]	+29.87 [+19.90, +39.81]
	Weather	+7.60 [−5.74, +22.40]	+4.00 [−2.37, +13.72]	+4.33 [−3.59, +15.38]
Sundial	Electricity	+17.71 [+8.63, +28.49]	+15.83 [+8.16, +25.99]	+24.18 [+14.83, +33.58]
	Traffic	−3.49 [−12.22, +5.13]	+8.29 [+4.43, +13.08]	+14.36 [+9.69, +18.60]
	Weather	+4.51 [−1.98, +13.13]	+6.05 [−0.09, +14.80]	+3.87 [+0.16, +8.85]
Moirai	Electricity	−9.98 [−20.23, −1.26]	+2.19 [−7.29, +10.28]	+24.66 [+14.98, +34.30]
	Traffic	+5.93 [−3.35, +15.80]	+37.23 [+23.52, +54.59]	+57.78 [+35.93, +88.96]
	Weather	+4.39 [−1.39, +11.30]	+6.12 [−1.82, +16.40]	+1.61 [−3.24, +6.76]
TimesFM	Electricity	+2.08 [−1.94, +5.81]	+5.26 [+0.46, +13.89]	+4.33 [−1.52, +10.56]
	Traffic	+16.74 [+5.93, +31.95]	+15.26 [+8.12, +24.74]	+24.90 [+13.48, +40.42]
	Weather	+4.22 [−1.02, +10.53]	+1.74 [−2.26, +6.89]	+4.41 [+0.36, +11.59]
Expanded evaluation ($n = 16$ per dataset)				
Moirai	Electricity	−5.80 [−12.87, +2.65]	+9.20 [+2.37, +19.77]	+29.74 [+23.31, +36.68]
	Traffic	+4.22 [−3.60, +13.19]	+37.81 [+27.72, +51.06]	+64.23 [+46.42, +90.06]
	Weather	+1.96 [−18.11, +52.84]	+9.92 [+1.50, +31.64]	−0.59 [−6.50, +5.24]
TimesFM	Electricity	+1.97 [−0.45, +4.62]	+3.87 [+1.31, +7.94]	+4.48 [+1.55, +7.58]
	Traffic	+17.85 [+7.42, +30.97]	+14.66 [+9.35, +21.65]	+26.47 [+15.53, +41.23]
	Weather	+7.64 [+0.98, +18.46]	+2.16 [−4.22, +18.66]	+2.83 [−1.51, +7.86]

B.5 Matched Cycle–Residual Controls

Table 10 gives the direct paired comparisons behind the cycle analysis, including the larger groups. Both columns evaluate the same historically selected cycle forecast, but change the reference: replacing its residual predictor with AR, or forecasting directly from selected raw input. Their percentage changes have different denominators and are not additive contributions. In particular, the intervals for the two normalized bar heights in Figure 3 cannot be subtracted to obtain an interval for Cycle + TSFM versus Cycle + AR. The direct paired intervals are reported here.

Table 10: **Cycle + TSFM against matched AR and direct raw forecasts.** Entries give relative MAE changes (%) and paired conditional 95% intervals, with the column heading naming the reference. All cycle settings are selected by historical MAE and retained for the AR replacement. First-four and sixteen-series groups overlap and are not independent replications.

Model	Dataset	n	vs Cycle + AR	vs selected raw
Chronos	Electricity	4	−1.71 [−14.10, +6.82]	+8.09 [−1.01, +18.10]
	Traffic	4	−2.11 [−3.89, −0.56]	+12.87 [+3.09, +24.94]
	Weather	4	−1.13 [−7.14, +8.26]	+7.60 [−5.74, +22.40]
Sundial	Electricity	4	−9.36 [−19.65, −2.82]	+17.71 [+8.63, +28.49]
	Traffic	4	−6.73 [−11.01, −2.50]	−3.49 [−12.22, +5.13]
	Weather	4	−3.82 [−8.21, +0.40]	+4.51 [−1.98, +13.13]
Moirai	Electricity	4	−3.54 [−17.49, +4.81]	−9.98 [−20.23, −1.26]
	Traffic	4	−1.37 [−3.76, +0.48]	+5.93 [−3.35, +15.80]
	Weather	4	+0.93 [−5.76, +7.96]	+4.39 [−1.39, +11.30]
TimesFM	Electricity	4	−12.57 [−21.75, −6.01]	+2.08 [−1.94, +5.81]
	Traffic	4	−22.10 [−34.96, −8.40]	+16.74 [+5.93, +31.95]
	Weather	4	−1.03 [−8.45, +5.55]	+4.22 [−1.02, +10.53]
Expanded evaluation ($n = 16$ per dataset)				
Moirai	Electricity	16	+1.04 [−4.44, +7.86]	−5.80 [−12.87, +2.65]
	Traffic	16	−1.42 [−4.13, +0.57]	+4.22 [−3.60, +13.19]
	Weather	16	+1.18 [−6.30, +11.85]	+1.96 [−18.11, +52.84]
TimesFM	Electricity	16	−17.47 [−25.87, −10.17]	+1.97 [−0.45, +4.62]
	Traffic	16	−28.79 [−37.92, −19.80]	+17.85 [+7.42, +30.97]
	Weather	16	−6.05 [−14.79, +1.16]	+7.64 [+0.98, +18.46]

B.6 Historical Rules and Candidate Expansion

All frozen historical rules. Table 11 reports the complete OT summaries for the historical rules compared in the main text. The local verifier uses only the last two completed forecast windows; prior–local combines those scores with calibration scores, and the guarded rule may return raw input. Moirai's top-3 is the best of these listed deployable rules by average OT MAE, whereas Sundial's guarded rule improves on the other listed rules. Neither a more local score nor an additional safeguard consistently dominates across models. The oracle uses future observations and is kept separate in interpretation.

Aggregation beyond the constituent forecasts. The Beats all column in Table 2 compares top-3 Median against the best of its three fixed constituents by whole-test-segment MAE. On the sixteen-series groups, Moirai exceeds every constituent on 8 Electricity, 9 Traffic, and 8 Weather series. The comparison is descriptive and retrospective, with configuration choices fixed by history. It neither supplies a deployable best-constituent selector nor establishes that any particular error-correlation pattern causes the improvement.

Raw selection with additional patch sizes. The small-patch experiment expands Moirai's raw pool from 4 to 12 and the full pool from 20 to 28, while leaving the historical scoring rule fixed.

Table 11: **Aggregation on the eight OT series.** Entries are dataset-balanced arithmetic MAE changes (%) against selected raw. All-view and top-3 use Median; the last column is the future-informed single-view diagnostic. These are descriptive macro summaries, without an across-dataset confidence interval.

Model	All-view	Single	Top-3	Local	Prior–local	Guarded	Oracle
Chronos	-1.89	-1.98	-2.90	+4.53	-1.74	-2.74	-20.91
Sundial	+1.30	+3.33	+1.52	+2.45	+0.53	-1.14	-20.31
Moirai	+1.80	-4.09	-5.21	-0.50	-2.88	-2.13	-30.52
TimesFM	-0.08	+0.91	-0.74	+5.31	+1.51	+1.65	-20.77

It adds raw forecasts only, rather than a complete transformation-by-patch grid. Figure 4(b) shows that raw selection itself can deteriorate on Traffic alongside expanded multi-view top-3. Both the raw and multi-view references therefore depend on the patch configurations included in their pools. The corresponding changes on Electricity and Weather are small under the same selection rules.

B.7 Temporal Subseries and Adapted Input Optimization

`TimesFM` subseries at two horizons. Table 12 uses the separate eight-view pool: raw and phase-aligned factors 2, 3, and 4 at two context lengths. It contains neither cycle nor trend candidates, and therefore is not interchangeable with the common-family `TimesFM` results. Historical selection and top-3 aggregation produce mixed changes at both horizons. In particular, Traffic top-3 degrades at both $H = 96$ and $H = 192$, while ETTm2 improves. The test-origin grids differ across horizons, so the differences between columns do not isolate a causal horizon effect.

Table 12: **`TimesFM`'s separate temporal-subseries comparison.** Relative MAE changes (%) are measured against the raw context selected within the same horizon and pool. Selection uses history; no test-optimal factor is reported as an inference rule.

	$H = 96$		$H = 192$	
Dataset	Single view	Top-3 Median	Single view	Top-3 Median
ETTh1	0.00	-5.11	0.00	+7.64
ETTh2	0.00	+3.97	0.00	-2.94
ETTm1	+19.01	+4.65	-3.89	-2.62
ETTm2	-2.15	-2.58	-6.48	-9.74
Electricity	0.00	-7.14	0.00	+1.14
Exchange	+1.11	+1.17	0.00	+1.36
Traffic	0.00	+8.66	0.00	+16.03
Weather	0.00	-3.51	+4.47	-0.68

Adapted TATO comparison. Table 13 reports the finite-budget adaptation of TATO (Qiu et al., 2026) alongside patch-128 controls. The two TATO columns reuse the same 100-trial shortlist with its weighted criterion or MAE reranking. On Electricity, the patch-128 view forecast has MAE 0.3179 versus 0.4528 for the adapted TATO variants, while on ETTh1 adapted TATO's MAE reranking reaches 0.1814 versus 0.1915 for the patch-128 view forecast. The additional auto-patch option further changes the reference, most clearly on Electricity and Traffic. These outcomes motivate reporting input and configuration controls, but do not establish superiority over the original TATO study: search budgets, history use, and preprocessing differ as specified in Appendix A.5.

Table 13: **MAE for the adapted TATO comparison and patch controls.** Raw and view configurations are historically selected. "p128" fixes patch size 128; "p128/auto" additionally permits the model's auto-patch interface. The different search and preprocessing protocols preclude an equal-budget ranking.

	Raw	View	Adapted TATO, $B = 100$		View
Dataset	p128	p128	Weighted	MAE rerank	p128/auto
ETTh1	0.2002	0.1915	0.1831	0.1814	0.1861
ETTh2	0.3249	0.2856	0.3140	0.3140	0.2723
ETTm1	0.1200	0.1161	0.1184	0.1184	0.1143
ETTm2	0.2230	0.2056	0.2100	0.2100	0.2056
Electricity	0.4510	0.3179	0.4528	0.4528	0.2672
Exchange	0.2371	0.2371	0.2536	0.2796	0.2371
Traffic	0.5073	0.1996	0.4901	0.4901	0.1916
Weather	0.0275	0.0257	0.0268	0.0268	0.0239

C Qualitative Forecasting Examples

These cases visualize frozen forecasts for retrospectively chosen model–series pairs. Raw and view configurations are selected separately by calibration MAE. For each case, the displayed window has the signed test MAE difference (view minus selected raw) closest to the median across that series; ties use the earlier origin. This retrospective window choice leaves model configurations unchanged. All plots show 96 observed steps and 96 future steps in training-standardized units.

C.1 Following a Cycle–Residual Forecast

Figure 7 follows `Moirai` on Electricity column 270 through decomposition and reconstruction. The forecast improves over selected raw on this window, although both residual predictors miss part of the late increase. Across all 32 test windows, Cycle + TSFM reduces MAE by 14.01% against selected raw and by 15.82% against matched Cycle + AR (Table 1).

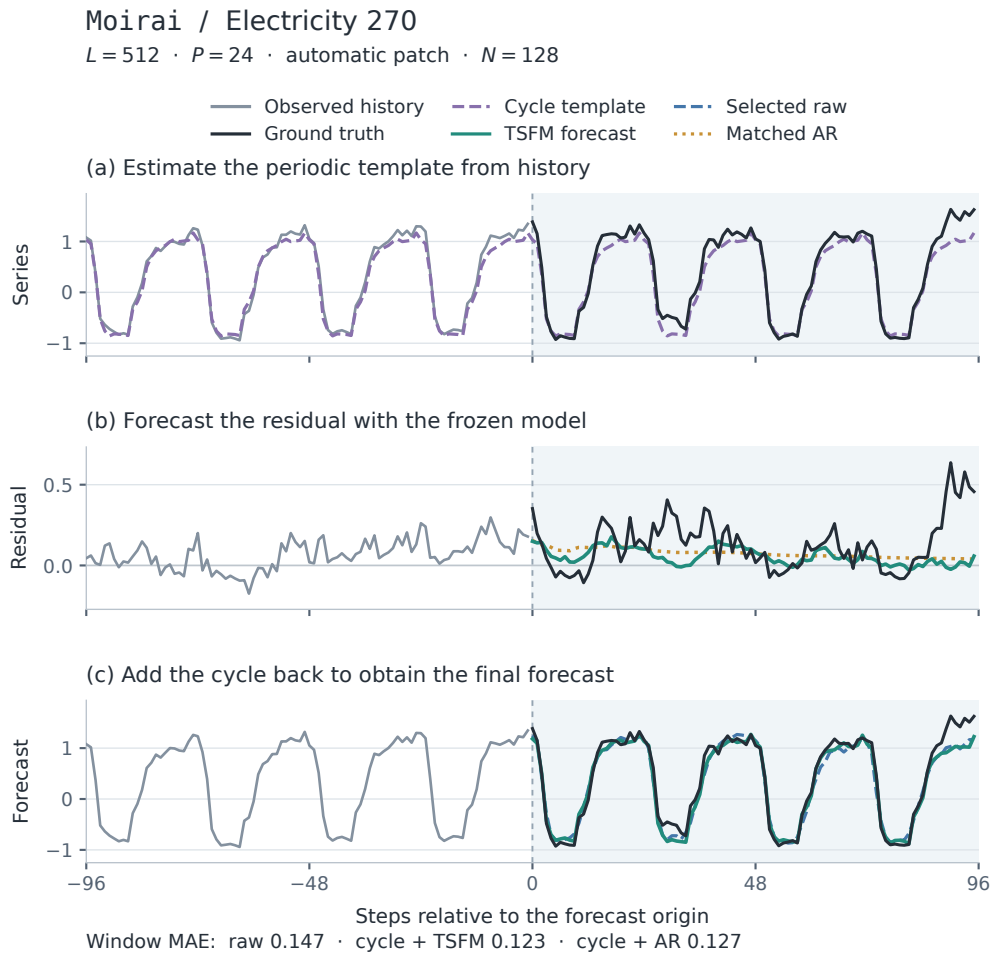

Figure 7: **A cycle view from input decomposition to final forecast.** Electricity column 270, starting at zero-based row 21,332. (a) The period-24 template uses the preceding 512 observations. (b) `Moirai` and matched AR use the same residual history; the future residual in black is ground truth only. (c) Restoring the cycle gives the forecast compared with selected raw. Values are training-standardized, stochastic forecasts use Median of 128 samples, and shading marks the forecast horizon.

C.2 Contrasting View Families and Local Outcomes

Figure 8 contrasts cycle, trend, and subseries views across the four checkpoints. In panel (a), `Chronos`'s cycle view improves series-average MAE but loses on the displayed window, illustrating that an average benefit need not hold at every origin. In panel (b), `Moirai`'s trend view improves only slightly and still misses the downward movement in the target. Panel (c) gives a useful subseries setting for `Sundial`; panel (d) shows a historically selected cycle configuration for `TimesFM` that is worse than selected raw both locally and over the test series. The complete family comparisons in Tables 8 and 9 establish the broader scope of these behaviors.

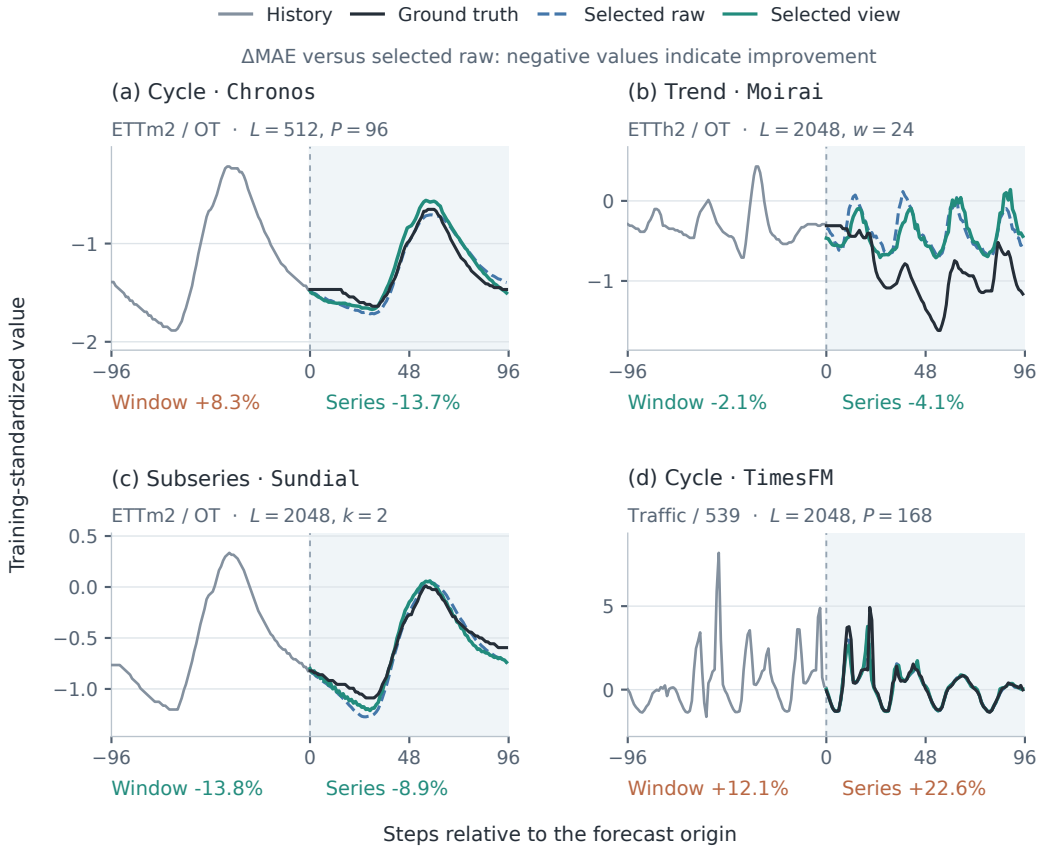

Figure 8: **View effects differ across families, series, and forecast origins.** Each green curve uses the calibration-selected configuration within the named family; the blue curve uses selected raw. The panel labels give the transformed view's context length L and period P, moving-average window w, or subseries factor k. "Window" is the relative MAE change on the displayed horizon; "Series" compares MAEs averaged over all test windows of that series. Negative values indicate improvement, with no confidence interval inferred from these illustrations. The zero-based forecast origins are (a) 50,784, (b) 12,096, (c) 49,344, and (d) 15,284. Stochastic branches use Median of 128 samples; `TimesFM` uses deterministic P50. Curves retain their original values, and vertical scales differ across panels.

D Further Experimental Diagnostics

D.1 Transfer of Historical Gains to the Test Period

Figure 9 compares the calibration and test MAE changes of the historically selected single view against selected raw. The common evaluation set contains the same eight OT and twelve additional series for every model. Selection uses only calibration outcomes; a positive test change therefore identifies a historical improvement that did not transfer. The figure retains every transformed selection in this set, including unfavorable outcomes. Selections that return raw coincide with the reference and are counted but omitted from the scatter.

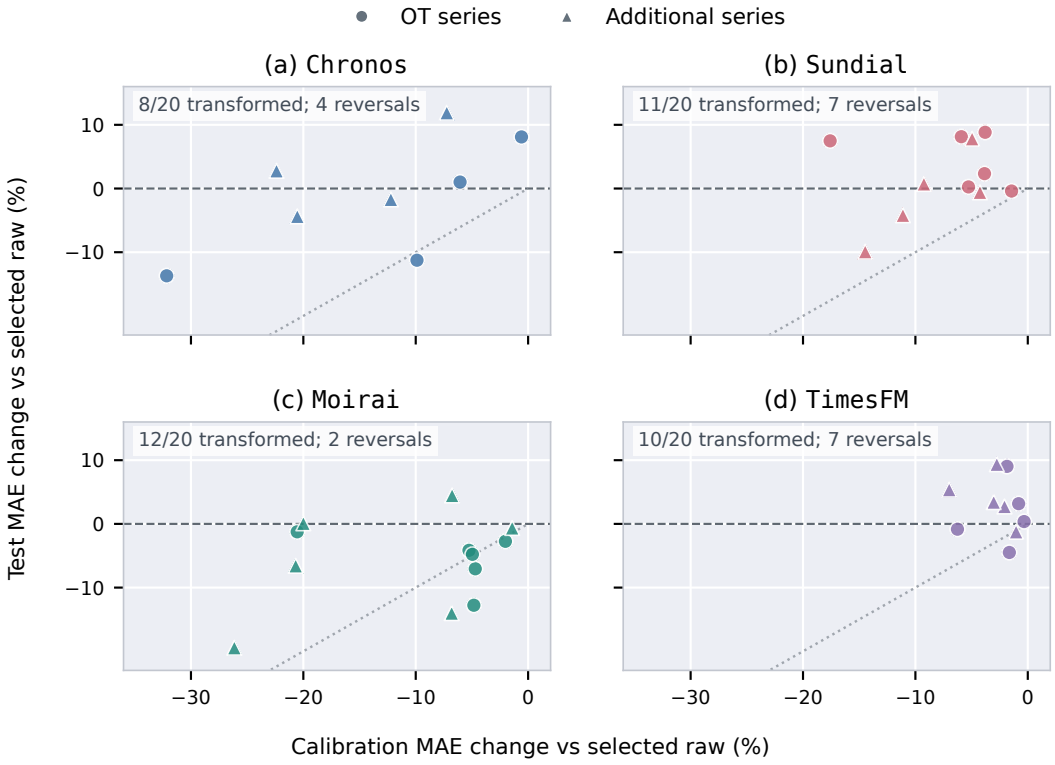

Figure 9: **Historical gains can reverse on the test segment.** Each point is one historically selected transformed view, evaluated relative to the raw configuration selected on the same calibration windows. Circles denote OT series and triangles the twelve additional series; all panels use the same axes. Negative values indicate improvement. Points above the dashed horizontal line lose to raw on the test segment, and the dotted diagonal marks equal calibration and test changes. Panel counts include all twenty series, including omitted raw selections at $(0, 0)$. These are descriptive per-series changes, without independent-seed intervals.

Table 14 extends the diagnosis to every complete series group. The rank correlation is Spearman's ρ between candidates' mean calibration and test MAE within each series, averaged within datasets and then across datasets. The reversal count uses the fixed error tolerance from Appendix A.6 on both calibration and test errors. On the eight OT series, five of `Sundial`'s six historical improvements reverse; none of `Moirai`'s six do. On 48 additional series, `TimesFM` reverses 13 of 21 historical improvements despite a mean candidate-rank correlation of 0.768. Thus, positive agreement across the full candidate ranking does not guarantee that its historical winner improves on raw at test time.

Table 14: **Historical selection transfer and aggregation.** Transformed counts non-raw single-view selections out of all series. Reversed / gained counts test degradations among selections that improve calibration MAE. Rank ρ summarizes agreement across the complete candidate pool. Beats all counts top-3 Median forecasts whose whole-test MAE is lower than each of their three fixed constituents. OT and additional groups are separate; the twelve-series groups are nested in the 48-series groups. The rank and constituent comparisons use test outcomes only for diagnosis.

Group	Model	Transformed	Reversed / gained	Rank ρ	Beats all
Eight OT	Chronos	4/8	2/4	0.662	3/8
	Sundial	6/8	5/6	0.406	2/8
	Moirai	6/8	0/6	0.592	4/8
	TimesFM	5/8	3/5	0.473	1/8
Twelve additional	Chronos	4/12	2/4	0.750	7/12
	Sundial	5/12	2/5	0.817	6/12
	Moirai	6/12	2/6	0.803	5/12
	TimesFM	5/12	4/5	0.746	3/12
48 additional	Moirai	31/48	8/31	0.791	25/48
	TimesFM	21/48	13/21	0.768	20/48

D.2 Sensitivity to Simple Forecasting References

Table 15 compares the same multi-view top-3 Median forecast against raw top-k and against the historically selected simple-control pool. The latter includes naive and seasonal-naive forecasts, cycle templates with simple residual predictors, and a half-raw/half-template blend. Its choice uses calibration MAE and may retain a TSFM call through the blend. The reference pool consequently contains both forecasts with and without TSFM calls.

The two references answer different questions. Across Moirai's 48 additional series, the mean improvement changes from 7.42% against raw top-k to 2.49% against selected simple controls. Each of its three dataset-level intervals against the simple reference spans zero. For TimesFM, top-3 improves on simple controls in 47 of 48 series, but improves on raw top-k in only 27. On Traffic, its mean change is -6.82% against simple controls and $+2.25\%$ against raw top-k. Exceeding a simple reference consequently provides different evidence from improving an already calibrated TSFM forecast.

Table 15: **The same multi-view forecast under two reference choices.** Entries are relative MAE changes (%) for historical multi-view top-3 Median. Macro rows give datasets equal weight; dashes indicate that no across-dataset confidence interval is estimated. The final six rows give paired conditional 95% intervals for sixteen additional series per dataset. I/T/D counts improvements, ties, and degradations against the historically selected simple reference. All configurations and both references are fixed using observed history.

Group	Model	n	vs raw top-k	vs simple	95% interval (simple)	I/T/D
Eight OT	Chronos	8	−2.90	−3.91	—	5/0/3
	Sundial	8	+3.41	−9.27	—	6/0/2
	Moirai	8	−3.99	−6.36	—	5/0/3
	TimesFM	8	+0.02	−11.26	—	7/0/1
Twelve additional	Chronos	12	−1.06	−3.75	—	10/0/2
	Sundial	12	−1.57	−12.97	—	12/0/0
	Moirai	12	−3.50	−5.69	—	7/0/5
	TimesFM	12	+1.15	−6.54	—	12/0/0
48 additional	Moirai	48	−7.42	−2.49	—	27/0/21
	TimesFM	48	−0.86	−8.12	—	47/0/1
Expanded evaluation ($n = 16$ per dataset)						
Electricity	Moirai	16	−10.04	−2.14	[−7.63, +3.03]	9/0/7
Traffic		16	−4.20	−1.75	[−5.50, +1.96]	8/0/8
Weather		16	−8.04	−3.57	[−13.55, +5.91]	10/0/6
Electricity	TimesFM	16	−0.65	−7.93	[−12.62, −3.74]	16/0/0
Traffic		16	+2.25	−6.82	[−11.49, −2.70]	15/0/1
Weather		16	−4.17	−9.61	[−15.98, −4.09]	16/0/0

D.3 Complete Patch-Expansion Comparisons

Table 16 gives all eight OT comparisons and the three sixteen-series groups for Moirai's expanded raw pool. The added patches expand raw candidates from 4 to 12 and the full pool from 20 to 28; the historical scoring rule is retained. All columns preserve an original-pool reference so that a weaker expanded raw forecast cannot inflate the apparent value of expanded top-3. For the sixteen Traffic series, expanded raw increases MAE by 5.79% relative to original raw, and expanded top-3 increases MAE by 5.28% relative to original top-3. By contrast, ETTh1 OT improves by 7.86% under the latter comparison. The complete table therefore retains both improvements and degradations from the added candidates. Its comparisons are paired within a series and horizon; the columns have different denominators and cannot be subtracted to decompose a gain.

Table 16: **Adding raw patch configurations under fixed historical selection.** Entries give relative MAE changes (%) and paired conditional 95% intervals. "Raw" in the first two column denominators is the original historically selected raw configuration. The last column directly compares expanded and original top-3 Median, with Median of 128 samples within each view. OT rows use one series; each additional-series row uses sixteen. The added candidates are raw inputs, not a complete transformation-by-patch grid.

Dataset	Group	Expanded raw / raw	Expanded top-3 / raw	Expanded / original top-3
ETTh1	OT	-6.45 $[-12.85, -0.32]$	-13.45 $[-19.57, -7.01]$	-7.86 $[-14.28, -1.55]$
ETTh2	OT	-1.90 $[-10.49, +8.05]$	-6.47 $[-12.14, +0.32]$	-1.48 $[-7.31, +5.40]$
ETTm1	OT	$+0.00$ $[+0.00, +0.00]$	-4.81 $[-16.42, +5.65]$	$+0.00$ $[+0.00, +0.00]$
ETTm2	OT	$+0.00$ $[+0.00, +0.00]$	$+0.31$ $[-13.50, +14.88]$	$+0.00$ $[+0.00, +0.00]$
Electricity	OT	$+0.00$ $[+0.00, +0.00]$	-7.16 $[-12.25, -2.25]$	$+1.45$ $[-2.62, +5.88]$
Exchange	OT	$+0.00$ $[+0.00, +0.00]$	-0.02 $[-0.83, +0.55]$	$+0.00$ $[+0.00, +0.00]$
Traffic	OT	$+2.93$ $[-6.99, +14.33]$	-8.94 $[-18.76, -1.29]$	$+3.84$ $[+1.93, +6.13]$
Weather	OT	$+0.00$ $[+0.00, +0.00]$	-5.25 $[-21.88, +14.84]$	$+0.00$ $[+0.00, +0.00]$
Electricity	16 extra	-2.20 $[-5.01, -0.34]$	-10.04 $[-15.79, -4.22]$	-0.21 $[-0.84, +0.00]$
Traffic	16 extra	$+5.79$ $[+0.68, +12.39]$	$+1.41$ $[-3.51, +7.39]$	$+5.28$ $[+0.88, +11.14]$
Weather	16 extra	-0.67 $[-3.09, +1.21]$	-11.00 $[-19.81, -2.26]$	-0.02 $[-0.21, +0.11]$

D.4 Prediction, Verification, and Calibration Costs

Table 17 extends the cost comparison to all four models on the same twelve additional series. Each series has 32 evaluated test origins, allowing the calibration charge to be shown on a common amortization horizon. We average per-origin costs within each series, within each dataset, and finally across the three datasets. The total is the sum of current prediction, online historical probes, and one calibration cost divided by 32.

The decomposition separates different sources of overhead. For `Moirai`, top-3 requires 0.348 seconds of current prediction versus 0.339 for raw top-k, but total estimated cost rises from 0.534 to 1.694 seconds after calibration. For `Chronos`, the two full-pool historical probes add 170.854 seconds per origin to both local and guarded verification. Returning to raw after a failed guard does not recover the already incurred probe cost. These costs accompany modest or unfavorable accuracy changes in several rows.

Raw top-k uses all available raw configurations for `Chronos`, `Sundial`, and `TimesFM`, so it needs no external calibration to determine membership. For `Moirai` it selects three of four raw configurations and incurs that selection cost. All timings retain the measured batching and thermal conditions and exclude external transformations, control fitting, weight loading, and I/O. They estimate prediction-API throughput under these measurement conditions.

Table 17: **Cost decomposition on the common twelve additional series.** Times are seconds per forecast origin; Calibration/32 amortizes the complete required calibration over 32 origins. Total is computed from unrounded components. The final column reports dataset-balanced MAE change relative to raw top-k Median within the same model. Historical top-3 uses cross-view Median; all stochastic branches use Median of 128 samples. No smaller-N runtimes or independent timing intervals are inferred.

Model	Rule	Prediction	Probes	Calibration /32	Total	△MAE (%)
`Chronos`	Selected raw	15.662	0.000	0.000	15.662	+0.00
	Raw top-k	15.662	0.000	0.000	15.662	+0.00
	Historical single	18.658	0.000	42.714	61.372	+0.72
	Historical top-3	56.819	0.000	42.714	99.532	−1.06
	Local verifier	16.912	170.854	0.000	187.766	+5.26
	Guarded verifier	16.431	170.854	42.714	229.998	+1.19
`Sundial`	Selected raw	0.064	0.000	0.067	0.132	−0.25
	Raw top-k	0.135	0.000	0.000	0.135	+0.00
	Historical single	0.069	0.000	0.469	0.537	−0.82
	Historical top-3	0.243	0.000	0.469	0.712	−1.57
	Local verifier	0.083	1.875	0.000	1.958	+1.13
	Guarded verifier	0.068	1.875	0.469	2.412	−1.52
`Moirai`	Selected raw	0.141	0.000	0.196	0.337	+0.90
	Raw top-k	0.339	0.000	0.196	0.534	+0.00
	Historical single	0.140	0.000	1.346	1.487	−2.13
	Historical top-3	0.348	0.000	1.346	1.694	−3.50
	Local verifier	0.147	5.385	0.000	5.532	−1.21
	Guarded verifier	0.134	5.385	1.346	6.866	−3.20
`TimesFM`	Selected raw	0.027	0.000	0.021	0.049	−1.82
	Raw top-k	0.043	0.000	0.000	0.043	+0.00
	Historical single	0.030	0.000	0.137	0.167	−0.19
	Historical top-3	0.093	0.000	0.137	0.230	+1.15
	Local verifier	0.029	0.549	0.000	0.578	+2.33
	Guarded verifier	0.029	0.549	0.137	0.715	−0.13