TARGET FIDELITY IN TIME-SERIES FORECAST EVALUATION: A CONTROLLED STUDY OF FORECAST COMBINATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Persistent recording disturbances can make forecasting benchmarks reward agreement with altered observations. Evaluating robustness therefore requires distinguishing this agreement from fidelity to the original recorded trajectory. The distinction is difficult to measure when input repair, predictor training, and forecast combination change together. We introduce a controlled protocol that applies disturbances across the forecast boundary while retaining the pre-injection future as an observed reference. Forecasting heads are trained with equal individual losses and frozen before combination fitting, allowing the same prediction to be scored against retained-source and altered-observation targets under matched repair conditions. In a post-hoc analysis with causal repair, changing only the scoring target reverses comparisons between fixed forecasts, including reversals whose directions persist across training seeds. Oracle access to retained source inputs improves source-target accuracy while increasing altered-target error. These results motivate pairing robustness claims with an explicit scoring target and the available repair information.

1 INTRODUCTION

Forecasting benchmarks compare predictions with recorded future values. Those values may contain a physical signal, meaningful operating phases, and persistent recording disturbances. A sensor that stays at zero across the forecast boundary illustrates the ambiguity: forecasting zero matches the recorded future, while recovering the trajectory retained before a controlled disturbance was introduced answers a different question. Both errors are measurable when an intervention preserves the original observations. Their interpretation depends on the intended target.

For robustness evaluation, the target specifies what a successful forecast should preserve. Improving fidelity to a retained source and reproducing an altered observation can call for different predictions. Interpreting the observation pattern also requires context: Figure 1 contrasts flat and zero-valued ETT intervals flagged by an audit with Solar-Energy's recurring night-zero operating phase. These examples motivate stating both the scoring reference and the observation process when interpreting an accuracy gain.

Separating these objectives poses a measurement and attribution problem. Natural series provide recorded observations; accuracy against a latent clean trajectory requires an independently verified reference. Quality-aware predictors can also change input repair, gate features, and forecast weights together, so their combined error change reflects several operations. Prior work shows that imperfect volatility proxies and test-label errors can affect model rankings (Patton, 2011; Northcutt et al., 2021), while forecasting studies document sensitivity to evaluation choices (Hewamalage et al., 2023; Brigato et al., 2025). Recent fault benchmarks examine corrupted inputs with an unchanged future target (Kim et al., 2026; Zhao et al., 2026; Windmann et al., 2026). Studying a disturbance that persists into the future requires retaining alternative scoring references and separating target changes from changes to the predictor's inputs and weights.

We address this problem with a controlled dual-target protocol. Prescribed transformations affect both the observed input and the future, while the retained pre-injection series supplies an observed reference that includes its original natural irregularities. Scoring each prediction against both futures

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

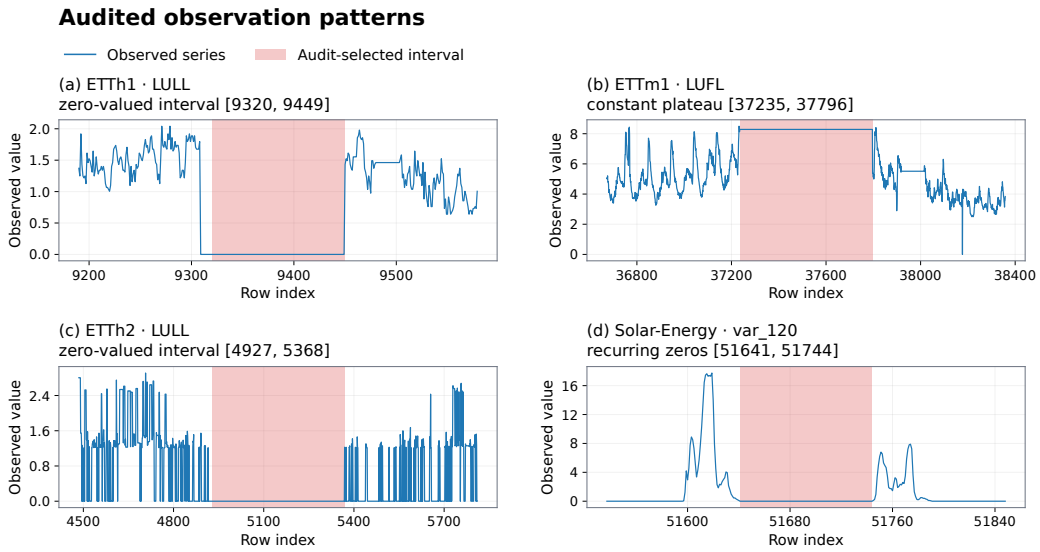


Figure 1: Observation patterns motivating explicit evaluation targets. The three ETT panels show zero or near-constant intervals flagged by the archived offline audit. Solar-Energy shows a meaningful night-zero phase. Blue traces are observations and red spans are detector-selected intervals; the horizontal axes are row indices. An audit flag identifies a pattern for examination, while its interpretation depends on the measurement process and operating regime.

isolates the effect of the target choice. Varying repair access with the predictor fixed measures how recovered inputs affect those scores. To separate forecast combination from predictor training, we train a shared seven-head bank with equal individual losses, freeze its parameters, and fit combination rules on a later time block. Single-head selection, averaging, learned static weights, and matched quality-conditioned gates then combine the same head forecasts within each repair condition. Adaptive Information Filtering (AIF) supplies the head architecture and reliability controller; Figure 2 shows the resulting comparison.

The primary study covers eight datasets at horizon 96, with a seven-dataset extension at horizon 336. Four prespecified comparisons measure the incremental effects of quality coordinates and explicit reliability control. A post-hoc analysis of the saved errors shows that, under causal repair, switching the scoring target changes which combination has the lowest MAE; we report the reversal counts together with error margins and consistency across seeds. Oracle access to retained source inputs reduces source-target error while increasing altered-target error. On unmodified inputs, static weighting has the lowest aggregate MAE among the studied rules at both horizons.

The study contributes a paired-target formulation that links changes in forecast comparisons to changes in the scoring reference (Section 3); a controlled protocol that separates target choice, repair access, and forecast combination (Section 4); and measurements of target-dependent rankings and repair-conditioned errors across horizons, seeds, and corruption types (Section 5). A complete eight-dataset, four-horizon comparison of end-to-end forecasters and an analysis of natural audit views provide comparisons on the original recorded targets (Section 6).

2 Related Work

Targets, faults, and evaluation. Patton studies how imperfect volatility proxies affect continuous-forecast comparisons and characterizes losses that preserve rankings under conditionally unbiased proxies (Patton, 2011). RobustTSF analyzes anomalous inputs and training targets, with additional experiments on anomalous test inputs and noisy-validation model selection (Cheng et al., 2024). TSRBench, TS-Fault, and SensorFault-Bench test structured input faults while retaining the future scoring target (Kim et al., 2026; Zhao et al., 2026; Windmann et al., 2026). Our intervention retains both the pre-injection and altered futures, so the same prediction can be evaluated under either

objective. Our controlled panel measures paired target differences under structured transformations across the forecast boundary, with input repair separated from frozen-head combination. TFB, GIFT-Eval, and forecasting evaluation guidance motivate consistent pipelines, chronological separation, and coverage reporting (Qiu et al., 2024; Aksu et al., 2025; Hewamalage et al., 2023).

Forecast combination and reliability information. FFORMA learns forecast weights from time-series features (Montero-Manso et al., 2020), while mixture-of-linear-experts specializes predictors across temporal patterns (Ni et al., 2024). FAME trains a statistics-based sparse router over pretrained frozen experts, separating validation-based suitability mining from later router calibration (Li et al., 2026). Our shared-bank comparison varies scoring targets, repair access, gate quality coordinates, and explicit reliability control. Simple averaging provides a reference motivated by the forecast combination puzzle (Smith & Wallis, 2009). The bank draws on direct linear, decomposition, patch, convolutional, variate, and spectral representations (Zeng et al., 2023; Nie et al., 2023; Liu et al., 2024; Luo & Wang, 2024; Xu et al., 2024). These heads supply complementary forecasts to the shared comparison.

3 Evaluation Targets and Controlled Comparisons

Let $X \in \mathbb{R}^{L\times C}$ and $Y^s \in \mathbb{R}^{H\times C}$ be a retained observed input and future. A prescribed transformation produces $(\widetilde{X}, Y^o)$, where $Y^o = Y^s + \delta$ is the altered observation. A forecasting rule f receives $\widetilde{X}$, input-side audit summaries, and the repair tensor permitted by condition r. Its prediction $\hat{Y}_{f,r}$ has two empirical risks:

$$R_s(f,r) = \mathbb{E}\,\ell(\hat{Y}_{f,r}, Y^s), \qquad R_o(f,r) = \mathbb{E}\,\ell(\hat{Y}_{f,r}, Y^o). \tag{1}$$

The expectation denotes the specified origins and intervention groups. Neither target enters prediction or combination selection. R_s measures fidelity to the retained recorded trajectory; R_o measures agreement with the transformed observation. Both are operational evaluation objectives. Natural benchmarks supply no independently verified latent clean target.

Why comparisons can change. For two fixed forecasts f, g, define the squared-loss difference $D_t = R_t(f,r) - R_t(g,r)$, $t \in \{s, o\}$. Expanding the two squares gives

$$D_o - D_s = -2\mathbb{E}\langle \hat{Y}_{f,r} - \hat{Y}_{g,r}, \delta\rangle. \tag{2}$$

The inner product uses the same component averaging as the squared loss. Target changes affect comparisons through their alignment with the difference between forecasts. If this expected inner product is zero, the squared-loss comparison is unchanged; our prescribed transformations allow either alignment. For component-averaged absolute loss, the triangle inequality gives $|D_o - D_s| \le 2\mathbb{E}\|\delta\|_1$, with the norm also averaged over components. The ranking analyses measure target sensitivity directly using MAE, MSE, and SMAPE.

Three distinct operations. Changing the target in Eq. (1) keeps predictions fixed. Changing repair access keeps the checkpoint and combination rule fixed but can change its predictions. Changing the combination rule holds the bank's seven forecasts fixed within an input condition. Earlier latent-artifact and route-cap calculations that motivated AIF are provided separately in Appendix G.

4 Experimental Design

4.1 Frozen Forecasts and Seven Combination Rules

The bank has seven internal heads: direct linear, trend-seasonal decomposition, patch encoding, temporal convolution, variate mixing, spectral-periodic, and audit residual. All are supervised through complete predictions after inverting reversible instance normalization (RevIN) (Kim et al., 2022). We minimize the equally weighted mean of their individual $0.7\mathrm{MAE} + 0.3\mathrm{MSE}$ losses. No mixture loss or routing prior enters bank training. Shared feature modules remain part of the bank. Its parameters are frozen before fitting any combination rule, and hashes verify that subsequent fitting leaves them unchanged.

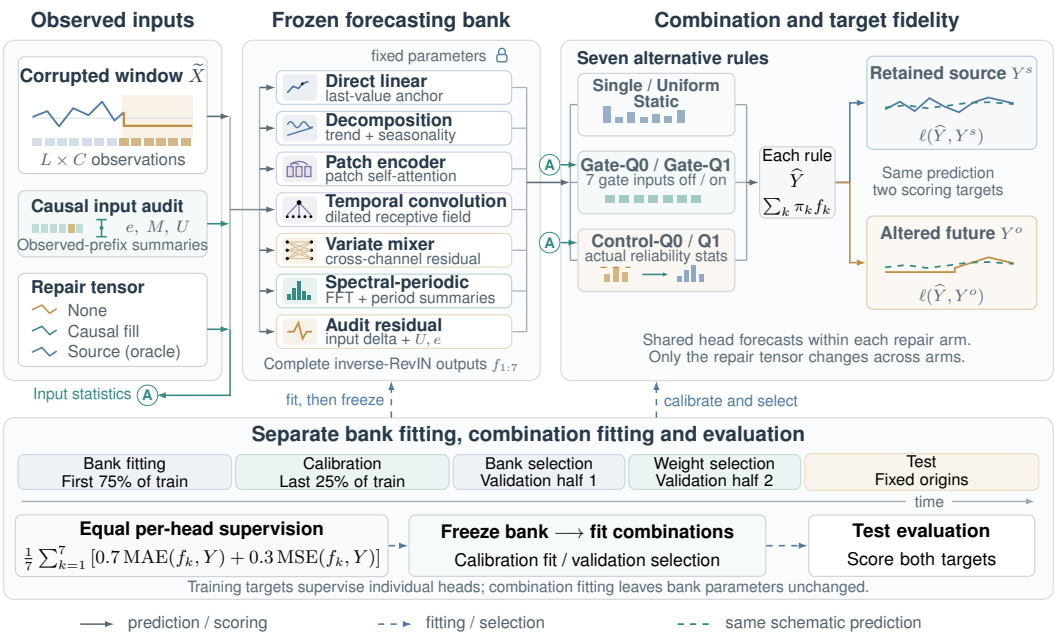


Figure 2: Controlled study design. Top: observed inputs and one repair tensor feed seven frozen heads; seven alternative rules combine their complete forecasts, and each prediction is scored against both the retained source and altered observation. Raw input, audit metadata and uncertainty stay fixed across repair arms; source-input repair is an oracle reference. Ports A supply input statistics to gates and controllers; Q0 masks seven normalized gate coordinates, while controllers use actual statistics. Bottom: chronological blocks separate supervised fitting and selection, with shared train-fitted preprocessing. Dashed arrows denote fitting or selection; curves, tiles and icons are schematic.

For complete head forecasts f_k, a combination outputs $\hat{Y} = \sum_{k=1}^{7} \pi_k f_k$, with nonnegative weights summing to one. **Single** selects one head using validation MAE. **Uniform** uses $\pi_k = 1/7$. **Static** learns one softmax weight vector. **Gate-Q0** and **Gate-Q1** are identical $18 \to 32 \to 7$ MLPs with GELU and dense softmax; Q0 zeros seven reliability coordinates after calibration normalization, whereas Q1 retains them. Initialization, batches, objective, and selection are matched. Both gates apply the same direct-route floor.

Control-Q0 and **Control-Q1** apply AIF's fixed reliability controller to the corresponding gate weights. The controller caps audit-residual and variate mass, redistributes it to direct and decomposition heads, and reserves spectral mass under its activation rule. It always reads the actual input statistics. Q0 therefore denotes removal of specified gate coordinates; the underlying heads and the explicit controller still have their documented inputs. The controlled comparison measures the controller as a package. Appendix A specifies the coordinates and settings; Appendix I.4 gives its equations.

4.2 Coverage, Temporal Separation, and Audit

Eight datasets are included: ETTh1, ETTh2, ETTm1, ETTm2, Exchange-Rate, Solar-Energy, Electricity, and Weather. Lookback is 96. Horizon 96 is the prespecified primary study, with seeds 17, 23, and 29. Horizon 336 is an extension using the same protocol. All 24 primary banks and 21 extension banks completed. Exchange-Rate's 759-step validation partition becomes two blocks of 379 and 380 steps, each shorter than the required $96 + 336 = 432$ steps; its three extension runs exited before training. The extension covers the remaining seven datasets, with no replacement or imputed result.

All rules share scaling and the Solar phase exemption fitted on the full original training partition. Within that partition, the first 75% of timestamps supervise the bank and the final 25% calibrate combinations. The first and second halves of validation select the bank and combinations, respectively. Calibration and validation input windows and targets are fully contained within their own blocks. The

original chronological test partition supplies all eligible natural test windows; observed lookbacks may precede its start. All selected origin lists and validation histories are saved.

Bank fitting uses at most 8,192 evenly spaced training windows, 1,024 validation windows, 60 epochs, and patience 10. Trainable combinations use at most 1,024 calibration and 1,024 validation windows, 100 epochs, and patience 12. AdamW learning rates are 3×10^{-4} and 10^{-3}, respectively; effective batch size is 32. Eleven of 45 banks reached the epoch cap. These budgets define the reported optimization scope.

The input audit marks zero, flat, and near-constant runs at their current endpoints. It never relabels an earlier sample after observing later values. Forward repair carries the latest unflagged value; training observations determine scaling and Solar's recurring-zero exemption. Natural evaluation uses a causal cached pass through the series. Intervention evaluation restarts the same detector at each 96-step observed input, giving a window-local causal stress test. Prefix truncation, suffix perturbation, and future-target mutation checks validate the implemented information boundary.

4.3 Interventions, Aggregation, and Analysis Status

Each bank uses 96 predetermined, evenly spaced test origins. Four transformations—plateau, zeros, 24-step repetition, and level shift—affect the final 12, 24, or 48 input steps and the first $\lceil H/2 \rceil$ future steps. Shift magnitudes are 1, 2, or 4 training standard deviations. Transformations act on all variables in original observation units before training standardization; zeros therefore denotes an original-scale zero. The 12 type–severity groups are weighted equally within each dataset, horizon, seed, and combination.

The three input conditions are **no repair**, **causal forward repair**, and **source-input repair**. They retain the same observed-input audit metadata and uncertainty; only the repair tensor changes. Source-input repair supplies the retained pre-injection input as an oracle reference. Its error measures the effect of access to that retained input under the specified transformation. Raw input remains corrupted in all three arms. Each prediction receives both source and altered-target scores, producing 24,192 error rows per bank.

MAE, MSE, and SMAPE use training-standardized values; SMAPE is $100\ \mathrm{mean}[2|\hat{y}-y|/(|\hat{y}|+|y|+10^{-6})]$. Seeds are averaged within settings, then datasets equally within each horizon. Worst-group MAE is the largest of the 12 group means, computed per dataset and seed before averaging. The two horizons retain separate summaries.

The archived protocol specified four comparisons before its reporting runs: Q1 versus Q0 on natural data, and Control-Q1 versus Gate-Q1, Static, and Single on source-target MAE with causal repair. We report their dataset-cluster paired bootstrap intervals and exact two-sided sign-flip tests, with Holm correction across the four primary comparisons. The target-ranking, margin, and repair-interaction analyses were developed after inspecting these outputs and are descriptive post-hoc analyses. Pairwise rankings share predictions, seeds, and datasets; their dependent counts describe this finite panel. Appendix A.1 details uncertainty and the incomplete extension.

5 Results: Targets, Repair, and Combination

5.1 Changing the Target Changes Forecast Comparisons

Table 1 compares the same seven combination forecasts against the two targets under causal repair. At horizon 96, 73 of 168 pairwise rankings reverse, and the lowest-error combination changes in six of eight datasets. Requiring an error difference of at least 1% on *each* target retains 49 reversals; a 5% margin retains 18. Both comparison directions agree across all three seeds for 37 reversals, including 33 of the 49 with a 1% margin.

The horizon-336 extension has 52 reversals among 147 comparisons and four changes of the lowest-error combination. Twelve reversals meet the 1% margin, one meets 5%, and nine retain both directions across seeds. Three satisfy both the 1% margin and the seed-consistency condition. Figure 3 shows all pairwise differences, including their magnitudes on both targets.

Table 1: Target-dependent MAE rankings under causal repair. Each horizon averages three seeds per dataset. Margin columns require the absolute pair difference divided by the pair mean to exceed the threshold on both targets. The last column requires both target comparison signs to agree across all three seeds. Counts are descriptive; the pairwise comparisons are dependent.

H	Datasets	Winner changes	Reversals	$\geq 1\%$	$\geq 5\%$	All seeds
96	8	6/8	73/168	49	18	37
336	7	4/7	52/147	12	1	9

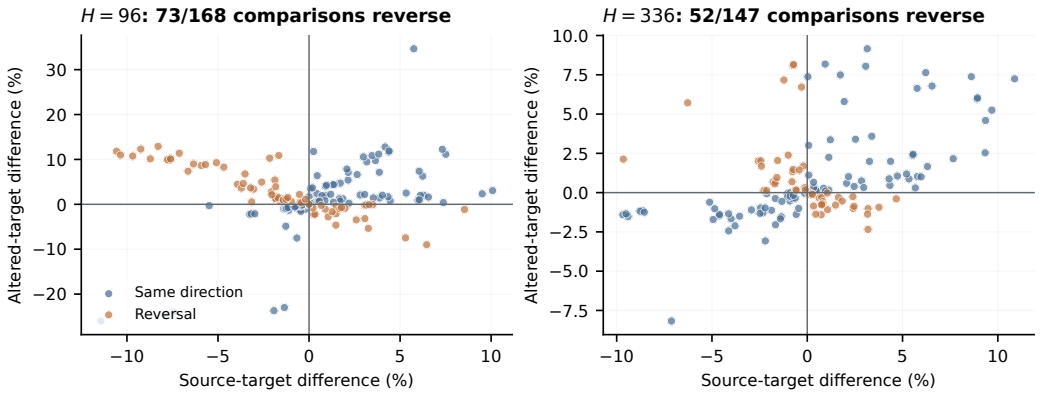


Figure 3: Target sensitivity of all pairwise combination comparisons under causal repair. Each point is one pair in one dataset after averaging three seeds and all 12 intervention groups. Axes show the signed error difference divided by the pair's mean error, in percent. Opposite signs (orange) indicate a ranking reversal when only the scoring target changes. Orientation follows a fixed combination order. Points share forecasts and are descriptive, dependent comparisons.

The pattern spans all four transformations. Analyzing them separately gives 56–77 reversals out of 168 at horizon 96 and 32–70 out of 147 at horizon 336. Removing any one type also retains ranking changes (Appendix C). The same analysis yields 63/168 and 29/147 reversals for MSE, and 95/168 and 84/147 for SMAPE. A reported ranking should therefore specify both its metric and its target.

These counts concern synthetic transformations of observed series. A transformation can represent an unwanted recording disturbance or an intended change in the observed process. Source and altered-target scores answer the corresponding questions explicitly. The results quantify their disagreement under the stated transformations; natural-series defect prevalence remains a separate measurement problem.

5.2 Repair Access Moves the Two Objectives in Opposite Directions

Table 2 reports Control-Q1 as a common reference across repair conditions, with every combination reported in Appendix B. At horizon 96, source-input repair reduces its retained-source MAE from 1.18710 to 0.28060, while altered-observation MAE increases from 0.81669 to 0.99062. The source reduction is 76.36%; the altered-target increase is 21.30%. The paired absolute source improvement has a descriptive 95% interval of [0.44830, 1.51355], and the altered-target change has the opposite direction, with improvement interval $[-0.24776, -0.11053]$.

This direction holds for all seven combinations on all eight dataset means. Across combinations, the aggregate source reduction ranges from 75.11% to 77.49%, while altered-target error increases by 12.68% to 23.23%. The seven-dataset extension retains the same direction: source reductions are 61.30–62.93%, with altered-target increases of 14.46–17.17%. These magnitudes measure the effect of access to the retained input under the prescribed transformations. The source-input arm changes predictions and is distinct from the fixed-prediction target comparison above.

Causal forward repair and no repair give close aggregate errors: source-MAE changes across the seven combinations range from -0.0114% to $+0.0246\%$ at horizon 96 and from -0.0003% to

Table 2: Two targets for the same Control-Q1 predictions within each repair condition. Horizon 96 uses eight datasets; horizon 336 uses seven. Source-input repair supplies the retained input as an oracle reference, while the raw input remains corrupted.

H	Repair condition	Source MAE	Altered MAE
96	No repair	1.18710	0.81671
96	Causal repair	1.18710	0.81669
96	Source-input repair	0.28060	0.99062
336	No repair	0.91456	0.79752
336	Causal repair	0.91451	0.79753
336	Source-input repair	0.34777	0.93326

Table 3: All seven combinations of the neutral bank. Natural MAE uses all eligible test windows. Stress and worst-group MAE use the retained-source target with causal repair. Worst-group is the maximum of 12 type–severity group means per dataset and seed, followed by averaging. Bold marks the lowest point mean in each column.

	$H = 96$, eight datasets			$H = 336$, seven datasets		
Combination	Natural	Stress	Worst	Natural	Stress	Worst
Single	0.2779	1.1911	3.9970	0.3482	**0.9104**	3.2170
Uniform	0.2837	**1.1813**	4.0395	0.3523	0.9431	3.2675
Static	**0.2731**	1.2126	4.0551	**0.3430**	0.9273	3.1766
Gate-Q0	0.2736	1.1988	3.9177	0.3436	0.9113	**3.0801**
Gate-Q1	0.2740	1.1898	3.9023	0.3439	0.9136	3.1087
Control-Q0	0.2769	1.1944	3.8954	0.3460	0.9116	3.0878
Control-Q1	0.2799	1.1871	**3.8793**	0.3469	0.9145	3.1139

$+0.0404\%$ at horizon 336. Forward filling propagates the latest unflagged value in the corrupted prefix.

5.3 Quality and Control Increments among Simple Combinations

Table 3 includes all seven prespecified combinations on natural data, source-target stress error, and worst-group source error. Static has the lowest natural MAE at both horizons. Uniform has the lowest aggregate stress MAE at horizon 96, and Single at horizon 336. Control-Q1 has the lowest average worst-group MAE at horizon 96; Gate-Q0 has the lowest at horizon 336. The same forecasts therefore favor different combinations under average and worst-group error.

The four prespecified primary effects are reported in Table 4. Quality coordinates change natural MAE by -0.00044 in the improvement convention, with interval $[-0.00152, 0.00076]$. Adding explicit control to Gate-Q1 improves source-target stress MAE by 0.00266, or 0.22%, with interval $[-0.00548, 0.01287]$. Its improvement over Static is 2.10% and over Single 0.34%. All four primary intervals cover zero and Holm-adjusted p values equal 1. These estimates quantify the incremental effects under the frozen-bank design.

Restricting both horizons to their seven common datasets gives a source-stress controller improvement of 0.28% at horizon 96 and -0.10% at horizon 336. Static remains the lowest natural-error combination in this matched coverage. Using the same seven datasets separates the horizon comparison from the change in coverage. Group-level errors and all repair/target combinations accompany the aggregate results in Appendix B.

6 Natural Forecasting Context and Audit Views

A separate end-to-end comparison covers eight datasets, four horizons $(96, 192, 336, 720)$, and three seeds (5,6,7). AIF, DLinear, PatchTST, TQNet, and iTransformer share chronological splits, training normalization, and test windows. Each fit uses at most 4,096 training windows, 1,024 validation windows, 12 epochs, patience 4, and effective batch 32. AIF receives causal audit inputs; standard baselines receive raw inputs, with matched-information controls reported separately. This comparison supplies performance context for the bank architecture under a common short budget.

Table 4: Prespecified horizon-96 MAE comparisons. Difference is comparator minus reference; positive favors the first-named rule. Intervals resample datasets and paired seeds (10,000 draws); Holm adjusts the four exact two-sided dataset sign-flip tests. Natural and stress comparisons have different error scales.

Comparison	Difference	95% interval	Positive	Holm p
Gate-Q1 vs Gate-Q0 (natural)	-0.00044	[-0.00152, 0.00076]	3/8	1.00
Control-Q1 vs Gate-Q1 (stress)	0.00266	[-0.00548, 0.01287]	4/8	1.00
Control-Q1 vs Static (stress)	0.02551	[-0.01837, 0.08743]	4/8	1.00
Control-Q1 vs Single (stress)	0.00402	[-0.05634, 0.07022]	6/8	1.00

Table 5: End-to-end context: equal-weight means over all 32 settings and three seeds under the shared short budget. These jointly trained models are separate from the neutral-bank study.

Method	MAE	MSE	SMAPE	MAE wins	SMAPE wins
TQNet	**0.34376**	**0.31876**	**57.553**	**16**	**19**
AIF	0.34862	0.32658	58.455	8	8
PatchTST	0.35260	0.32915	58.724	5	2
iTransformer	0.36246	0.33885	60.374	0	0
DLinear	0.36493	0.33672	62.593	3	3

AIF ranks second in all three aggregate metrics and attains the lowest MAE in eight settings (Table 5). TQNet has the lowest overall error. Fifty-nine of 96 AIF fits reached the epoch cap; 26 selected the final epoch. Appendix E reports seed dispersion, paired dataset-level statistics, matched-information baselines, and the four-dataset quality/control factorial. These configurations measure fixed-budget performance.

What changes when the evaluation windows change? We reuse the end-to-end models' saved per-window errors under fixed offline audit views, without changing predictions. Fifteen settings have a nonempty selected view: 12 clean-like and three phase-balanced Solar settings. The audit manifests omit the first 96 test origins of each setting because their lookbacks must also be split-contained. Separating this boundary change from view selection matters: removing those boundary origins changes two MAE winners, while applying the audit selection to covered origins changes four of 15. TQNet retains the lowest overall MAE over the selected settings (Table 6). Weather's selected view contains 47 of 10,349 covered windows; all empty views and coverage counts are retained in Appendix D.

Solar's balanced view increases the share of active-dominant target windows to one half. Across its three eligible horizons, AIF's MAE increases by 0.01540; a decomposition into phase proportions and within-phase sampling attributes 0.01541 to the former and -0.000003 to the latter. The same decomposition is reported for all five models. This is a change in the distribution of scored windows, distinct from replacing target values. The natural and injected analyses distinguish changes in evaluation coverage from changes in the scoring target.

Earlier jointly trained AIF checkpoints also provide a repair-conditioned routing comparison. With causal repair, full control improves source MAE by 10.10% relative to its pre-stabilization weights, while a validation-selected single head has lower error than full control. These runs use different bank training, sparse routing, seeds, and intervention sampling from the neutral study. Their full inverse-RevIN replay and results appear in Appendix E. The two studies jointly motivate specifying the comparator and fitting protocol with each reported controller effect.

7 Discussion and Conclusion

The controlled study identifies three practical reporting choices. First, state whether a score measures fidelity to a retained source or agreement with altered observations; identical predictions can change rank under this choice, including comparisons with at least a 1% difference on each target. Second, state the repair access, because recovered-input information moves source and altered-target errors

Table 6: Natural evaluation views on the same saved predictions, over all 15 nonempty settings. Equal-setting means average three seeds first. Selection changes the evaluated window distribution; all five methods remain in the comparison.

Method	Audit-covered MAE	Selected-view MAE
TQNet	0.34675	0.33512
AIF	0.35009	0.33796
PatchTST	0.35526	0.34064
iTransformer	0.36856	0.36183
DLinear	0.36391	0.36370

in opposite directions in these interventions. Third, place quality-conditioned gates beside single, uniform, and static combinations, so the measured increment has a clear reference.

The empirical scope is eight public datasets, a complete horizon-96 study, and a seven-dataset horizon-336 extension. The retained source is an observed trajectory, and the transformations are prescribed stress scenarios. Ranking and margin analyses are post hoc; their finite-panel counts accompany the prespecified controller comparisons. All saved natural Solar windows receive the transition label under the input-window phase rule, yielding one observed phase group in this study. Fixed training budgets and three seeds characterize optimization variability; prefix checks establish the implemented audit boundary, with streaming deployment and latency outside this evaluation.

Paired target scores, explicit repair access, and matched forecast combinations make these evaluation choices measurable. Together, they show how the scoring reference and available input information shape forecast comparisons.

AI USE STATEMENT

Generative AI tools were used only to improve the grammar and readability of author-written text. The authors reviewed all edits and take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

Code, configurations, and reproduction instructions are in the supplement.

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, 2025. OpenReview submission to ICLR 2025; arXiv:2410.10393.

Lorenzo Brigato, Rafael Morand, Knut Strømmen, Maria Panagiotou, Markus Schmidt, and Stavroula Mougiakakou. Position: There are no champions in long-term time series forecasting. *arXiv preprint arXiv:2502.14045*, 2025.

Hao Cheng, Qingsong Wen, Yang Liu, and Liang Sun. RobustTSF: Towards theory and design of robust time series forecasting with anomalies. In *International Conference on Learning Representations*, 2024.

Hansika Hewamalage, Klaus Ackermann, and Christoph Bergmeir. Forecast evaluation for data scientists: Common pitfalls and best practices. *Data Mining and Knowledge Discovery*, 37(2): 788–832, 2023.

Dongbin Kim, Youngjoo Park, Woojin Jeong, and Jaewook Lee. Local geometry attention for time series forecasting under realistic corruptions. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/476ab8f369e489c04187ba84f68cfa68-Paper-Conference.pdf.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2022.

Qianyang Li, Xingjun Zhang, Shaoxun Wang, Tao Peng, and Jia Wei. FAME: Forecastability-aware mixture of experts for heterogeneous time series forecasting. *arXiv preprint arXiv:2606.08896*, 2026.

Shengsheng Lin, Haojun Chen, Haijie Wu, Chunyun Qiu, and Weiwei Lin. Temporal query network for efficient multivariate time series forecasting. 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. 37797–37814. PMLR, 13–19 Jul 2025.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024.

Donghao Luo and Xue Wang. ModernTCN: A modern pure convolution structure for general time series analysis. In *International Conference on Learning Representations*, 2024.

Pablo Montero-Manso, George Athanasopoulos, Rob J. Hyndman, and Thiyanga S. Talagala. FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, 36(1): 86–92, 2020. doi: 10.1016/j.ijforecast.2019.02.011.

Ronghao Ni, Zinan Lin, Shuaiqi Wang, and Giulia Fanti. Mixture-of-linear-experts for long-term time series forecasting. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4672–4680, 2024.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2103.14749.

Andrew J. Patton. Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256, 2011. doi: 10.1016/j.jeconom.2010.03.034.

Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, Zhenli Sheng, and Bin Yang. TFB: Towards comprehensive and fair benchmarking of time series forecasting methods. *Proceedings of the VLDB Endowment*, 17(9): 2363–2377, 2024.

Jeremy Smith and Kenneth F. Wallis. A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, 71(3):331–355, 2009. doi: 10.1111/j.1468-0084.2008.00541.x.

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 *International Conference on Learning Representations*, 2024.

Shiyu Wang, Jiawei Li, Xiaoming Shi, Zhou Ye, Baichuan Mo, Wenze Lin, Shengtong Ju, Zhixuan Chu, and Ming Jin. TimeMixer++: A general time series pattern machine for universal predictive analysis. In *International Conference on Learning Representations*, 2025.

Alexander Windmann, Philipp Wittenberg, Gianluca Manca, Marcel Dix, Jens U. Brandt, and Oliver Niggemann. Benchmarking sensor-fault robustness in forecasting. *arXiv preprint arXiv:2605.10822*, 2026.

Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: Modeling time series with 10k parameters. In *International Conference on Learning Representations*, 2024.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, volume 37, pp. 11121–11128, 2023.

Yuyang Zhao, Lian Xu, Hao Miao, Chenxi Liu, and Hao Xue. TS-Fault: Benchmarking time series forecasters against structural faults. *arXiv preprint arXiv:2606.18539*, 2026.

A NEUTRAL-BANK PROTOCOL AND IMPLEMENTATION

This is the completed fixed-head-v2 study. It follows the earlier end-to-end study but uses new reporting seeds and a separate training protocol. The manifest lists 48 banks: eight datasets, two horizons, and seeds 17,23,29. Forty-five completed. Three Exchange horizon-336 tasks exited before fitting because neither validation half contains a full 432-step input–target span. The missing cells remain explicit. The earlier interrupted v1 banks produced no reportable test results and are excluded.

Normalization and output identity. All losses, validation decisions, and combinations use complete inverse-RevIN head outputs. If an exported auxiliary head is $g_k = \sigma z_k + \mu$, the complete prediction is $f_k = (g_k - \mu - \sigma b)/\max(w, \epsilon) + \mu$, where w, b are the learned affine parameters and μ, σ are the current repaired-input moments. Recombining these outputs with the model's deployed weights is checked against its deployed prediction. This interface is also used for the separate replay of the earlier end-to-end checkpoints. The ordinary end-to-end benchmark scores retain their original full-model predictions.

Training and selection. Bank training uses effective batch 32, BF16, AdamW weight decay 10^{-5}, gradient clipping at 1, and learning rate 3×10^{-4}. Its objective is the mean of seven complete-head losses, each $0.7\mathrm{MAE} + 0.3\mathrm{MSE}$. The best checkpoint minimizes mean individual-head validation MAE. All bank parameters are then frozen; selected head outputs and 18 router statistics are cached for combination fitting. Static and dense gates use AdamW at 10^{-3}, batch 32, the same forecasting objective, and minimum combination-validation MAE. The single head minimizes its own MAE on this same selection block. No test-time oracle head selection is used.

Scaling statistics and the Solar phase exemption are fitted on the full original training partition and shared by all rules. The first 75% of its timestamps supervise bank fitting; the final 25% calibrate combinations. This subdivision separates supervised fitting after shared train-only preprocessing. Validation is halved for bank and combination selection. Targets lie wholly within their subpartition. Calibration and both validation lookbacks also lie wholly within their subpartition. The fixed lookback creates an observation gap at internal boundaries. Bank training has at most 8,192 windows; all calibration/validation caps are 1,024. Caps select evenly spaced origins. Natural testing keeps all eligible origins of the original test partition. Bank epoch counts range from 11 to 60 at horizon 96 (median 22) and 14 to 60 at horizon 336 (median 25). Eleven of 45 banks reach the cap. These epoch counts describe the optimization budget used for the reported predictions.

Gate inputs and controls. Q0 zeros reliability coordinates 2,10,11,12,13,14,15 (zero-based indices) of the 18-statistic vector after normalization fitted on calibration inputs. Q1 retains the full vector. The matching gates use identical initialization and batch ordering, an 18-to-32-to-7 MLP, GELU, and dense softmax. Their direct-route floor is identical before the controlled comparison. The controller applies the inherited artifact/variate caps, mass redistribution, and spectral reservation to each fitted gate; it always uses the actual input statistics. The forecast heads also retain their shared audit inputs. Q0 therefore measures the effect of these gate coordinates while preserving the other audit inputs. Shared feature modules are trained by individual-head losses; independence refers to the absence of a routing loss in bank fitting.

The ordered coordinates 0–17 are: normalized horizon ID, H/L, log-scaled support ID, input mean, input standard deviation, clipped skewness, absolute endpoint drift, half-window mean difference, spectral concentration, spectral entropy, mean absolute off-diagonal channel correlation, dominant covariance-component ratio, clean-support ratio, mean uncertainty, mask density, artifact score, $\log(1+C)/8$, and $\log(1+L)/8$. Input and spectral statistics use the repaired input after RevIN. Thus Q0 masks support ID, both channel statistics, clean support, uncertainty, mask density, and artifact score; the other eleven coordinates remain. Calibration means and population standard deviations normalize the gate inputs, with standard deviations floored at 10^{-3}.

Intervention construction. From each test partition we select 96 unique evenly spaced origins. The observation segment contains 96 past steps and H future steps. Transformations are applied in original observation units before training standardization; a zero block sets the original-scale values to zero. For severity 1/2/3, a transformation starts 12/24/48 steps before the origin and ends after $\lceil H/2 \rceil$ future steps. Plateau copies the last value before the transformed span. Zeros sets the span to

zero. Repetition tiles the preceding 24-step block. Shift adds 1/2/4 training standard deviations per variable. All channels are transformed. These are controlled observation transformations; pre-existing irregularities in the retained source persist.

The causal detector is reset at the beginning of each corrupted observed window. All three repair arms receive the same raw corrupted input, causal metadata, and uncertainty. Only the repaired-input tensor differs: corrupted input for none, forward-filled input for causal, and retained source input for the oracle arm. Targets are supplied only to error calculation. This window-local stress evaluation is distinct from the continuous causal audit cache used for natural evaluation. Causal thresholds and training-fitted Solar exemption are specified in Appendix E.

Validation and coverage. Each completed bank exports 24,192 corruption rows: 96 origins, four types, three severities, three repair arms, and seven combinations. The 45 completed banks yield 1,088,640 rows. Checks verify unique keys, the prescribed origin sets, finite nonnegative errors, normalized nonnegative weights, best-validation selection, frozen-bank hashes, effective Q0/Q1 toggles, effective control, and prediction isolation from future scoring arrays. The natural phase grouping uses the training-fitted Solar zero-phase profile over the observed input window; all saved Solar windows fall in the transition category under the prespecified thresholds. Burden-group support is reported separately in the analysis outputs. Detector flags summarize the prescribed observation patterns; their fault interpretation depends on the measurement process.

A.1 Statistics and Analysis Status

The prespecified four-comparison family is given in Table 4. For horizon 96, each paired comparison contains eight datasets and three matched seeds. The bootstrap draws eight datasets with replacement and resamples three paired seeds within each sampled dataset, using 10,000 draws. Exact two-sided sign flips enumerate the 2^8 assignments of signs to dataset-mean differences. Holm correction covers the four comparisons. The missing Exchange runs leave the prespecified eight-dataset horizon-336 family incomplete; no complete-family test is reported for that extension. Supplemental seven-dataset paired intervals are descriptive. For the common-seven analysis, horizon 96 also excludes Exchange solely to match extension coverage; the complete eight-dataset primary result remains unchanged.

The target-ranking analysis is post hoc. For each dataset, horizon, and repair condition, it first averages the three seeds and all 12 intervention groups for each rule. Each of the $\binom{7}{2} = 21$ rule pairs contributes a reversal if its source and altered-target error differences have opposite signs. For errors e_a, e_b, the relative margin is $|e_a - e_b|/[(e_a + e_b)/2]$, required on both targets for the 1% and 5% summaries. Seed consistency requires that each target's comparison sign matches its seed-averaged sign in all three seeds. Pair counts are dependent descriptions of the observed panel, with no binomial or independent-pair significance test. The lowest-error rule on each target is an ex-post descriptive ranking, never used for fitting or deployment.

Repair effects and target interactions also use paired dataset/seed bootstrap intervals, labelled descriptive and marginal. The four prespecified comparisons define the confirmatory testing family. All metrics, all repair conditions, all combinations, each corruption type, and each leave-one-type-out analysis are exported. Worst-group error takes the maximum over the 12 predefined type–severity means separately for each bank and combination, then averages datasets and seeds. The maximization occurs before averaging, so it differs from the largest group in an already aggregated table.

B Complete Neutral-Bank Results

All values below derive from the completed v2 outputs. Horizon 96 has eight datasets and horizon 336 seven. Source and altered targets are scored on the same prediction within each condition. Standard deviations in setting tables describe the three training seeds; target-ranking counts use their means.

Table 7: Natural MAE: every completed dataset–horizon setting. Each cell shows the three-seed mean above ± sample standard deviation. Q0/Q1 denote the paired gates; C0/C1 apply their controllers.

Dataset	H	Single	Uniform	Static	Q0	Q1	C0	C1
ETTh1	96	0.3938 ±0.0022	0.4086 ±0.0027	0.3928 ±0.0010	0.3948 ±0.0014	0.3973 ±0.0014	0.3948 ±0.0015	0.3972 ±0.0014
ETTh1	336	0.4452 ±0.0027	0.4579 ±0.0023	0.4433 ±0.0016	0.4444 ±0.0011	0.4440 ±0.0010	0.4444 ±0.0012	0.4442 ±0.0011
ETTh2	96	0.3349 ±0.0001	0.3297 ±0.0006	0.3338 ±0.0001	0.3326 ±0.0031	0.3342 ±0.0010	0.3308 ±0.0010	0.3325 ±0.0006
ETTh2	336	0.4267 ±0.0008	0.4123 ±0.0015	0.4217 ±0.0007	0.4197 ±0.0009	0.4240 ±0.0006	0.4193 ±0.0015	0.4214 ±0.0004
ETTm1	96	0.3705 ±0.0018	0.3575 ±0.0027	0.3551 ±0.0031	0.3552 ±0.0034	0.3550 ±0.0032	0.3556 ±0.0034	0.3554 ±0.0033
ETTm1	336	0.4105 ±0.0008	0.4028 ±0.0029	0.4015 ±0.0020	0.4013 ±0.0023	0.4012 ±0.0023	0.4013 ±0.0026	0.4012 ±0.0026
ETTm2	96	0.2591 ±0.0002	0.2521 ±0.0013	0.2520 ±0.0014	0.2516 ±0.0017	0.2517 ±0.0017	0.2516 ±0.0017	0.2517 ±0.0017
ETTm2	336	0.3384 ±0.0001	0.3328 ±0.0012	0.3327 ±0.0012	0.3326 ±0.0011	0.3327 ±0.0011	0.3326 ±0.0011	0.3327 ±0.0011
Electricity	96	0.2400 ±0.0005	0.2763 ±0.0012	0.2390 ±0.0004	0.2387 ±0.0008	0.2399 ±0.0006	0.2558 ±0.0002	0.2570 ±0.0004
Electricity	336	0.2703 ±0.0006	0.3011 ±0.0009	0.2685 ±0.0002	0.2696 ±0.0016	0.2698 ±0.0009	0.2794 ±0.0007	0.2803 ±0.0004
Exchange	96	0.2041 ±0.0064	0.2003 ±0.0019	0.2003 ±0.0020	0.2004 ±0.0023	0.2004 ±0.0023	0.2004 ±0.0023	0.2004 ±0.0024
Solar	96	0.2159 ±0.0061	0.2360 ±0.0023	0.2104 ±0.0037	0.2131 ±0.0049	0.2108 ±0.0018	0.2243 ±0.0043	0.2423 ±0.0015
Solar	336	0.2564 ±0.0045	0.2713 ±0.0020	0.2499 ±0.0025	0.2534 ±0.0010	0.2515 ±0.0028	0.2607 ±0.0004	0.2643 ±0.0064
Weather	96	0.2048 ±0.0009	0.2094 ±0.0007	0.2014 ±0.0002	0.2021 ±0.0002	0.2027 ±0.0002	0.2021 ±0.0002	0.2027 ±0.0002
Weather	336	0.2895 ±0.0006	0.2878 ±0.0003	0.2836 ±0.0004	0.2842 ±0.0006	0.2843 ±0.0006	0.2842 ±0.0006	0.2843 ±0.0006

Table 8: Retained-source stress MAE with causal repair: every completed dataset–horizon setting. Each cell shows the three-seed mean above ± sample standard deviation. Q0/Q1 denote the paired gates; C0/C1 apply their controllers.

Dataset	H	Single	Uniform	Static	Q0	Q1	C0	C1
ETTh1	96	0.8986 ±0.0050	0.8594 ±0.0110	0.8737 ±0.0055	0.8436 ±0.0354	0.8716 ±0.0154	0.8453 ±0.0338	0.8709 ±0.0162
ETTh1	336	0.8267 ±0.0142	0.8008 ±0.0126	0.8133 ±0.0015	0.8070 ±0.0263	0.8216 ±0.0083	0.8067 ±0.0249	0.8205 ±0.0057
ETTh2	96	0.8995 ±0.0100	0.9055 ±0.0054	0.9015 ±0.0093	0.8901 ±0.0268	0.8819 ±0.0012	0.8863 ±0.0195	0.8863 ±0.0047
ETTh2	336	0.8204 ±0.0192	0.8810 ±0.0209	0.8279 ±0.0202	0.8251 ±0.0211	0.7901 ±0.0214	0.8316 ±0.0277	0.8083 ±0.0251
ETTm1	96	0.7911 ±0.0101	0.8454 ±0.0164	0.8632 ±0.0144	0.8794 ±0.0294	0.8677 ±0.0294	0.8774 ±0.0270	0.8716 ±0.0260
ETTm1	336	0.8287 ±0.0413	0.7548 ±0.0159	0.7675 ±0.0236	0.7856 ±0.0555	0.7904 ±0.0495	0.7821 ±0.0550	0.7866 ±0.0487
ETTm2	96	0.7395 ±0.0111	0.7846 ±0.0169	0.7879 ±0.0189	0.7992 ±0.0204	0.7979 ±0.0216	0.7992 ±0.0204	0.7979 ±0.0216
ETTm2	336	0.7058 ±0.0067	0.7695 ±0.0161	0.7705 ±0.0165	0.7775 ±0.0054	0.7754 ±0.0088	0.7762 ±0.0075	0.7688 ±0.0078
Electricity	96	0.8905 ±0.0296	0.9073 ±0.0145	0.9190 ±0.0190	0.8799 ±0.0257	0.8841 ±0.0279	0.8524 ±0.0110	0.8541 ±0.0124
Electricity	336	0.8405 ±0.0234	0.8592 ±0.0022	0.8586 ±0.0160	0.8378 ±0.0163	0.8802 ±0.0798	0.8316 ±0.0092	0.8730 ±0.0718
Exchange	96	2.8497 ±0.2415	2.8103 ±0.0758	2.8122 ±0.0778	2.8242 ±0.1496	2.8473 ±0.1665	2.8226 ±0.1473	2.8444 ±0.1650
Solar	96	0.4315 ±0.0230	0.4227 ±0.0025	0.3962 ±0.0023	0.4009 ±0.0028	0.4152 ±0.0091	0.4091 ±0.0020	0.4305 ±0.0083
Solar	336	0.5858 ±0.0368	0.5930 ±0.0054	0.5746 ±0.0110	0.5681 ±0.0137	0.5677 ±0.0125	0.5757 ±0.0137	0.5803 ±0.0158
Weather	96	2.0286 ±0.6597	1.9151 ±0.2809	2.1473 ±0.4690	2.0728 ±0.4547	1.9524 ±0.4743	2.0625 ±0.4421	1.9412 ±0.4589
Weather	336	1.7646 ±0.3055	1.9436 ±0.0994	1.8789 ±0.2272	1.7778 ±0.2661	1.7699 ±0.5087	1.7774 ±0.2665	1.7640 ±0.4986

Table 9: All repair arms, combinations, and metrics at $H = 96$. Oracle denotes the source-input repair reference.

		Retained source			Altered observation		
Repair	Rule	MAE	MSE	SMAPE	MAE	MSE	SMAPE
none	Single	1.1911	21.1661	89.0077	0.8814	11.9808	70.0645
none	Uniform	1.1815	18.6362	90.9599	0.7980	6.6355	69.4355
none	Static	1.2129	23.8088	90.2738	0.8399	11.3853	69.2406
none	Gate-Q0	1.1986	20.8723	90.2866	0.8340	10.0449	69.3918
none	Gate-Q1	1.1898	18.6463	90.4386	0.8292	9.5419	69.1472
none	Control-Q0	1.1945	20.6978	90.3718	0.8239	9.9041	68.9351
none	Control-Q1	1.1871	18.3925	90.6642	0.8167	9.3293	68.4295
causal	Single	1.1911	21.1661	89.0089	0.8814	11.9808	70.0666
causal	Uniform	1.1813	18.6262	90.9603	0.7979	6.6321	69.4382
causal	Static	1.2126	23.7739	90.2742	0.8397	11.3593	69.2432
causal	Gate-Q0	1.1988	20.9546	90.2848	0.8343	10.1011	69.3898
causal	Gate-Q1	1.1898	18.6495	90.4377	0.8292	9.5426	69.1476
causal	Control-Q0	1.1944	20.6887	90.3695	0.8239	9.8984	68.9327
causal	Control-Q1	1.1871	18.3949	90.6627	0.8167	9.3294	68.4280
oracle	Single	0.2768	0.2238	50.4544	0.9932	14.4676	79.9439
oracle	Uniform	0.2940	0.2372	54.1893	0.9833	14.4110	80.6237
oracle	Static	0.2730	0.2140	50.3264	0.9889	14.4491	79.8121
oracle	Gate-Q0	0.2724	0.2136	50.4108	0.9894	14.4512	79.9676
oracle	Gate-Q1	0.2748	0.2175	50.6987	0.9915	14.4559	80.2065
oracle	Control-Q0	0.2758	0.2166	51.0115	0.9892	14.4480	80.1476
oracle	Control-Q1	0.2806	0.2221	51.7265	0.9906	14.4508	80.4296

Table 10: All repair arms, combinations, and metrics at $H = 336$. Oracle denotes the source-input repair reference.

		Retained source			Altered observation		
Repair	Rule	MAE	MSE	SMAPE	MAE	MSE	SMAPE
none	Single	0.9104	13.7874	89.9162	0.8177	8.8746	77.5841
none	Uniform	0.9435	15.6459	92.3381	0.8177	9.0212	79.1921
none	Static	0.9277	14.7980	91.2540	0.7994	7.6864	78.0040
none	Gate-Q0	0.9114	12.4878	91.1407	0.7949	6.7704	77.9847
none	Gate-Q1	0.9136	13.4399	91.2387	0.7976	7.6793	77.8854
none	Control-Q0	0.9118	12.4990	91.3090	0.7967	6.7699	78.1634
none	Control-Q1	0.9146	13.3343	91.4563	0.7975	7.5423	78.0021
causal	Single	0.9104	13.7874	89.9152	0.8177	8.8746	77.5861
causal	Uniform	0.9431	15.6010	92.3421	0.8178	9.0168	79.2079
causal	Static	0.9273	14.7501	91.2533	0.7994	7.6781	78.0090
causal	Gate-Q0	0.9113	12.4713	91.1447	0.7949	6.7720	77.9972
causal	Gate-Q1	0.9136	13.4307	91.2390	0.7976	7.6800	77.8896
causal	Control-Q0	0.9116	12.4833	91.3104	0.7968	6.7714	78.1717
causal	Control-Q1	0.9145	13.3252	91.4559	0.7975	7.5425	78.0055
oracle	Single	0.3523	0.3440	61.1019	0.9447	14.9225	86.4002
oracle	Uniform	0.3550	0.3377	63.0702	0.9361	14.8745	87.4777
oracle	Static	0.3437	0.3268	60.6151	0.9335	14.8844	85.9311
oracle	Gate-Q0	0.3436	0.3259	60.6717	0.9314	14.8770	85.8121
oracle	Gate-Q1	0.3448	0.3281	60.8305	0.9331	14.8779	85.9832
oracle	Control-Q0	0.3459	0.3284	61.0965	0.9320	14.8760	86.0573
oracle	Control-Q1	0.3478	0.3309	61.3663	0.9333	14.8745	86.2217

Table 11: All twelve intervention groups: source MAE with causal repair at $H = 96$. Entries average datasets and three seeds. Q0/Q1 are gates and C0/C1 their controlled versions.

Type	Severity	Single	Uniform	Static	Q0	Q1	C0	C1
duplicate	1	0.3349	0.3373	0.3292	0.3299	0.3337	0.3314	0.3360
duplicate	2	0.3464	0.3493	0.3419	0.3417	0.3460	0.3428	0.3469
duplicate	3	0.3979	0.3972	0.3936	0.3923	0.3955	0.3919	0.3944
plateau	1	0.3483	0.3444	0.3413	0.3404	0.3464	0.3404	0.3459
plateau	2	0.3758	0.3781	0.3726	0.3729	0.3805	0.3749	0.3834
plateau	3	0.4390	0.4398	0.4350	0.4352	0.4423	0.4395	0.4447
shift	1	0.6042	0.5931	0.6076	0.6128	0.6203	0.5930	0.5971
shift	2	1.1999	1.1896	1.1899	1.2141	1.2074	1.1857	1.1813
shift	3	2.7226	2.7797	2.7723	2.7359	2.7666	2.7101	2.7485
zeros	1	2.3660	2.4644	2.5616	2.4959	2.3936	2.4931	2.3967
zeros	2	2.4829	2.4303	2.5348	2.4832	2.4010	2.5117	2.4358
zeros	3	2.6754	2.4727	2.6716	2.6310	2.6439	2.6179	2.6344

Table 12: All twelve intervention groups: corrupted MAE with causal repair at $H = 96$. Entries average datasets and three seeds. Q0/Q1 are gates and C0/C1 their controlled versions.

Type	Severity	Single	Uniform	Static	Q0	Q1	C0	C1
duplicate	1	0.3023	0.3045	0.2982	0.2990	0.2972	0.2972	0.2941
duplicate	2	0.3202	0.3214	0.3155	0.3159	0.3164	0.3130	0.3105
duplicate	3	0.3140	0.3189	0.3124	0.3146	0.3136	0.3117	0.3081
plateau	1	0.3351	0.3210	0.3289	0.3291	0.3294	0.3238	0.3202
plateau	2	0.3329	0.3220	0.3273	0.3273	0.3277	0.3223	0.3173
plateau	3	0.3271	0.3197	0.3209	0.3206	0.3261	0.3152	0.3148
shift	1	0.5195	0.5124	0.5146	0.5249	0.5203	0.5197	0.5145
shift	2	0.9397	0.8997	0.9141	0.9238	0.9133	0.9109	0.8978
shift	3	2.0029	1.8769	1.8810	1.8398	1.8192	1.8193	1.8063
zeros	1	1.7726	1.5066	1.6973	1.6660	1.6336	1.6459	1.6122
zeros	2	1.7290	1.4036	1.5852	1.5802	1.5799	1.5534	1.5497
zeros	3	1.6814	1.4682	1.5807	1.5703	1.5737	1.5537	1.5547

Table 13: All twelve intervention groups: source MAE with causal repair at $H = 336$. Entries average datasets and three seeds. Q0/Q1 are gates and C0/C1 their controlled versions.

Type	Severity	Single	Uniform	Static	Q0	Q1	C0	C1
duplicate	1	0.3887	0.3951	0.3863	0.3881	0.3885	0.3896	0.3903
duplicate	2	0.4094	0.4119	0.4044	0.4042	0.4041	0.4056	0.4063
duplicate	3	0.4707	0.4638	0.4629	0.4566	0.4593	0.4566	0.4589
plateau	1	0.4012	0.4007	0.3949	0.3960	0.3997	0.3965	0.3998
plateau	2	0.4324	0.4342	0.4268	0.4286	0.4320	0.4304	0.4350
plateau	3	0.5068	0.5038	0.4988	0.4997	0.5042	0.5022	0.5073
shift	1	0.5229	0.5299	0.5251	0.5336	0.5260	0.5263	0.5208
shift	2	0.9358	1.0310	0.9757	0.9984	0.9814	0.9906	0.9799
shift	3	2.3790	2.6074	2.4956	2.4647	2.5225	2.4738	2.5330
zeros	1	1.2871	1.4072	1.3717	1.2463	1.2089	1.2500	1.2165
zeros	2	1.2805	1.4076	1.3629	1.3179	1.3407	1.3408	1.3541
zeros	3	1.9098	1.7248	1.8228	1.8010	1.7960	1.7770	1.7721

Table 14: All twelve intervention groups: corrupted MAE with causal repair at $H = 336$. Entries average datasets and three seeds. Q0/Q1 are gates and C0/C1 their controlled versions.

Type	Severity	Single	Uniform	Static	Q0	Q1	C0	C1
duplicate	1	0.3589	0.3610	0.3524	0.3524	0.3516	0.3525	0.3519
duplicate	2	0.3588	0.3618	0.3534	0.3527	0.3516	0.3525	0.3518
duplicate	3	0.3586	0.3552	0.3499	0.3438	0.3464	0.3427	0.3447
plateau	1	0.4156	0.4044	0.4063	0.4068	0.4092	0.4060	0.4081
plateau	2	0.4038	0.3939	0.3956	0.3963	0.3972	0.3946	0.3954
plateau	3	0.3836	0.3748	0.3749	0.3754	0.3749	0.3739	0.3733
shift	1	0.6081	0.6216	0.6063	0.6075	0.6063	0.6163	0.6145
shift	2	1.0053	1.0318	0.9994	1.0028	0.9992	1.0147	1.0088
shift	3	1.9146	1.9451	1.8990	1.9015	1.9059	1.9159	1.9149
zeros	1	1.3739	1.3482	1.3226	1.2770	1.2941	1.2773	1.2924
zeros	2	1.3209	1.3049	1.2594	1.2526	1.2810	1.2507	1.2731
zeros	3	1.3099	1.3107	1.2735	1.2703	1.2542	1.2643	1.2416

C TARGET-RANKING SENSITIVITY

Table 15: Target-dependent rankings for every metric, repair condition, and horizon. Margin columns require the stated relative difference on both targets. All-seed consistency refers to the two comparison directions; it is separate from the margin filter.

Metric	Repair	H	Reversals	$\geq 1\%$	$\geq 5\%$	All seeds	Winners change
MAE	causal	96	73/168	49	18	37	6/8
MAE	causal	336	52/147	12	1	9	4/7
MAE	none	96	73/168	49	18	37	6/8
MAE	none	336	52/147	13	1	9	4/7
MAE	oracle	96	39/168	9	6	13	5/8
MAE	oracle	336	25/147	0	0	8	3/7
MSE	causal	96	63/168	41	25	36	6/8
MSE	causal	336	29/147	10	1	3	4/7
MSE	none	96	63/168	41	25	36	6/8
MSE	none	336	29/147	10	1	3	4/7
MSE	oracle	96	76/168	13	6	37	7/8
MSE	oracle	336	52/147	2	0	26	6/7
SMAPE	causal	96	95/168	31	7	38	7/8
SMAPE	causal	336	84/147	9	0	20	7/7
SMAPE	none	96	95/168	31	7	38	7/8
SMAPE	none	336	85/147	9	0	22	7/7
SMAPE	oracle	96	35/168	6	0	6	5/8
SMAPE	oracle	336	19/147	0	0	0	5/7

Table 16: Every completed setting: lowest-MAE rules on each target and pairwise reversals under causal repair. Winners describe retrospective test-score rankings; validation determines model selection.

Dataset	H	Source winner	Altered winner	Reversals	$\geq 1\%$	$\geq 5\%$
ETTh1	96	Gate-Q0	Static	7/21	1	0
ETTh1	336	Uniform	Single	17/21	5	0
ETTh2	96	Gate-Q1	Single	7/21	3	0
ETTh2	336	Gate-Q1	Single	6/21	0	0
ETTm1	96	Single	Gate-Q1	18/21	12	6
ETTm1	336	Uniform	Uniform	4/21	0	0
ETTm2	96	Single	Control-Q1	17/21	12	6
ETTm2	336	Single	Single	4/21	0	0
Electricity	96	Control-Q0	Control-Q1	5/21	3	0
Electricity	336	Control-Q0	Gate-Q0	6/21	3	0
Exchange	96	Uniform	Uniform	0/21	0	0
Solar	96	Static	Control-Q1	16/21	15	5
Solar	336	Gate-Q1	Gate-Q1	6/21	2	0
Weather	96	Uniform	Uniform	3/21	3	1
Weather	336	Control-Q1	Control-Q0	9/21	2	1

Table 17: All per-type and leave-one-type-out MAE analyses under causal repair. Each panel preserves every completed dataset and all seven combinations.

Panel	Type	H	Reversals	Winner changes
Type alone	duplicate	96	76/168	7/8
Type alone	duplicate	336	55/147	6/7
Type alone	plateau	96	75/168	6/8
Type alone	plateau	336	56/147	6/7
Type alone	shift	96	56/168	4/8
Type alone	shift	336	32/147	4/7
Type alone	zeros	96	77/168	6/8
Type alone	zeros	336	70/147	6/7
Omit type	duplicate	96	72/168	6/8
Omit type	duplicate	336	53/147	4/7
Omit type	plateau	96	70/168	5/8
Omit type	plateau	336	53/147	4/7
Omit type	shift	96	92/168	7/8
Omit type	shift	336	67/147	6/7
Omit type	zeros	96	67/168	6/8
Omit type	zeros	336	30/147	3/7

Table 18: Common-seven supplemental comparisons. Difference is the error of the second rule minus that of the first; positive values favor the first rule. These post-hoc paired dataset/seed bootstrap intervals are descriptive; the prespecified horizon-336 eight-dataset family remains incomplete.

H	Scope	Pair (first / second)	Difference	95% interval
96	natural	Gate-Q1 / Gate-Q0	-0.00051	[-0.00167, 0.00085]
96	stress	Control-Q1 / Gate-Q1	0.00263	[-0.00636, 0.01428]
96	stress	Control-Q1 / Gate-Q0	0.01621	[-0.01524, 0.06115]
96	stress	Control-Q1 / Static	0.03377	[-0.01181, 0.10163]
96	stress	Control-Q1 / Uniform	-0.00175	[-0.04083, 0.03497]
96	stress	Control-Q1 / Single	0.00384	[-0.04286, 0.06847]
96	worst group	Control-Q1 / Gate-Q1	0.02745	[-0.03787, 0.12156]
96	worst group	Control-Q1 / Gate-Q0	0.03473	[-0.12628, 0.20821]
96	worst group	Control-Q1 / Static	0.18712	[-0.05085, 0.51085]
96	worst group	Control-Q1 / Uniform	0.16922	[-0.04954, 0.42495]
96	worst group	Control-Q1 / Single	0.05963	[-0.22024, 0.47173]
336	natural	Gate-Q1 / Gate-Q0	-0.00034	[-0.00187, 0.00086]
336	stress	Control-Q1 / Gate-Q1	-0.00091	[-0.00900, 0.00644]
336	stress	Control-Q1 / Gate-Q0	-0.00324	[-0.03953, 0.03440]
336	stress	Control-Q1 / Static	0.01282	[-0.02238, 0.07363]
336	stress	Control-Q1 / Uniform	0.02861	[-0.02441, 0.12087]
336	stress	Control-Q1 / Single	-0.00415	[-0.11223, 0.08431]
336	worst group	Control-Q1 / Gate-Q1	-0.00515	[-0.04753, 0.03729]
336	worst group	Control-Q1 / Gate-Q0	-0.03371	[-0.25463, 0.18166]
336	worst group	Control-Q1 / Static	0.06279	[-0.18917, 0.40980]
336	worst group	Control-Q1 / Uniform	0.15362	[-0.21617, 0.54874]
336	worst group	Control-Q1 / Single	0.10317	[-0.37045, 0.80184]

D NATURAL EVALUATION VIEWS ON FIXED SAVED PREDICTIONS

This post-hoc analysis reuses the 480 complete end-to-end runs, comprising five methods, eight datasets, four horizons, and seeds 5,6,7. It operates on their saved per-origin MAE, MSE, and SMAPE. The forecasts stay fixed, and every method and seed uses the same test origins within each dataset–horizon setting. The per-window errors support selection and reweighting; the saved files do not contain the prediction tensors needed for new element-level scoring.

The fixed offline audit manifests include target-side labels used only for evaluation selection. We distinguish (i) all raw test origins, (ii) origins covered by those manifests, and (iii) the selected clean-like view, or the phase-balanced view for Solar. Every setting has exactly 96 initial raw origins absent from the manifest because the older audit required split-contained lookbacks. Reporting (i)–(ii) separately prevents attributing boundary coverage changes to audit selection.

Twelve clean-like settings and three balanced Solar settings are nonempty; all other cells remain empty. Weather horizon 96 has 47 selected windows. Across these fifteen nonempty settings, going from covered to selected origins changes four MAE winners and 22 of 150 pairwise method comparisons. Going from raw to covered origins changes two other MAE winners. TQNet retains the lowest aggregate MAE on the fifteen selected settings, followed by AIF. The historical fourteen-member subset excludes Weather; it changes three winners under covered-to-selected scoring and also retains TQNet as the aggregate leader. Historical offline-model scores use a different training protocol and are archived separately.

Phase composition. Solar uses target-window dominant-phase labels. These identify the largest phase share in each mixed window, which can contain entries from multiple phases. The three balanced horizons contain active- and night-dominant windows; horizon 720 has only the latter dominant label and therefore no balanced view. The selected balanced sample takes equal numbers from the two strata and uses an unweighted mean. Applying the old inverse-frequency weights again to this already balanced sample is a different scoring rule. We export that sensitivity without substituting it for the defined balanced view.

For phase proportions p_j and within-phase errors e_j, the change between two scoring views decomposes exactly as

$$\bar{e}_{\text{after}} - \bar{e}_{\text{before}} = \sum_j (p_{j,\text{after}} - p_{j,\text{before}}) e_{j,\text{before}} + \sum_j p_{j,\text{after}} (e_{j,\text{after}} - e_{j,\text{before}}).$$

The first term changes phase composition; the second captures within-phase subsampling. The table below gives all five methods. These decompositions account for changes in the same saved observed-target errors. The analysis retains complete coverage and method-pair outputs. Supplemental uncertainty resamples datasets as clusters; Solar alone has one cluster and receives no cluster confidence interval.

Table 19: Complete evaluation-view coverage for the end-to-end runs. A dash means an empty selected view. Counts are identical across methods and seeds.

Dataset	H	Raw origins	Audit-covered	Clean-like	Balanced
ETTh1	96	2785	2689	1982	–
ETTh1	192	2689	2593	1757	–
ETTh1	336	2545	2449	1376	–
ETTh1	720	2161	2065	635	–
ETTh2	96	2785	2689	–	–
ETTh2	192	2689	2593	–	–
ETTh2	336	2545	2449	–	–
ETTh2	720	2161	2065	–	–
ETTm1	96	11425	11329	9943	–
ETTm1	192	11329	11233	9792	–
ETTm1	336	11185	11089	9175	–
ETTm1	720	10801	10705	7933	–
ETTm2	96	11425	11329	–	–
ETTm2	192	11329	11233	–	–
ETTm2	336	11185	11089	–	–
ETTm2	720	10801	10705	–	–
Electricity	96	5166	5070	–	–
Electricity	192	5070	4974	–	–
Electricity	336	4926	4830	–	–
Electricity	720	4542	4446	–	–
Exchange	96	1423	1327	651	–
Exchange	192	1327	1231	555	–
Exchange	336	1183	1087	319	–
Exchange	720	799	703	–	–
Solar	96	10417	10321	–	9346
Solar	192	10321	10225	–	7100
Solar	336	10177	10081	–	4900
Solar	720	9793	9697	–	–
Weather	96	10445	10349	47	–
Weather	192	10349	10253	–	–
Weather	336	10205	10109	–	–
Weather	720	9821	9725	–	–

Table 20: All five end-to-end methods on the same fifteen nonempty settings. Means weight settings equally, averaging seeds first. Predictions stay fixed; selected windows change.

Evaluation view	Method	MAE	MSE	SMAPE
audit covered	AIF	0.35009	0.33100	60.985
audit covered	DLinear	0.36391	0.33669	64.702
audit covered	PatchTST	0.35526	0.32939	61.563
audit covered	TQNet	0.34675	0.32371	60.270
audit covered	iTransformer	0.36856	0.34539	63.350
selected	AIF	0.33796	0.30543	61.008
selected	DLinear	0.36370	0.32661	65.552
selected	PatchTST	0.34064	0.29909	61.265
selected	TQNet	0.33512	0.29821	60.405
selected	iTransformer	0.36183	0.32573	64.101

Table 21: Solar MAE decomposition: balanced minus audit-covered score, averaged over three eligible horizons and three seeds. Positive values increase the reported error.

Method	Total change	Phase composition	Within-phase sampling
AIF	0.015402	0.015405	-0.000003
DLinear	0.001151	0.001175	-0.000024
PatchTST	0.014776	0.014786	-0.000010
TQNet	0.015485	0.015491	-0.000006
iTransformer	0.017107	0.017128	-0.000021

E Complete Causal Comparison and Controls

This section reports the completed fast48h-v1 experiments. The end-to-end board contains 480 fits; matched-information/repair controls contain 72, causal/offline/raw controls 36, the routing factorial 48, official TimeMixer 24, and matched AEF 24, for 684 fits in total. Reporting seeds are 5,6,7. Timing and smoke runs are excluded. These results use the causal short-budget protocol in Section 6; the historical offline results in Appendix H retain their original budgets and seeds.

Causal preprocessing. Raw values are normalized with training mean and standard deviation. Zero runs are flagged at length 4 using raw absolute tolerance 10^{-8}; flat runs at length 8 use standardized adjacent differences at most 10^{-5}. A 12-step standardized range at most 0.01 flags near-constant segments. Detection assigns the current endpoint and never changes earlier labels. Repair carries the latest unflagged value forward; the initial value initializes that state. Uncertainty is 0.25 on flagged entries and zero elsewhere. For Solar, a training phase profile exempts recurring zero phases with at least 80% zero observations. Metadata summarizes the input mask and phase profile. Perturbing observations after selected origins and truncating the series at those origins preserves all prefix audit arrays; prediction checks also verify invariance to changed future observations.

Configuration and selection. All main fits use the same evenly spaced training/validation caps, all eligible test windows, BF16, gradient clipping at 1, and effective batch 32. AIF uses microbatch 8 on Solar/Electricity and 32 otherwise; baseline microbatch is 32. The saved configuration and history identify the chosen minimum-validation-MAE checkpoint. The 12-epoch cap and patience 4 apply to every main method. The official TimeMixer comparison uses commit `e24610583b36fdd8c76cc17a8df4e65759a5f460`, at $H = 96$, with the same sample and training caps. It is distinct from the archived TM++-local implementation.

Matched information and factorial definitions. The audit adapter is a shared pointwise 15-to-32-to-1 MLP: raw input, repaired input, uncertainty, and 12 input summaries form its inputs; its output is added to repaired input before the baseline backbone. The last layer starts at zero. Repair-only baselines use the same repaired input without the adapter. The factorial independently removes the seven designated reliability coordinates from the learned gate and toggles explicit stabilization, preserving head architectures and the direct-route floor. Statistical routing priors are off in all four factorial groups. When stabilization is active it still reads the true statistics. Offline/causal/raw comparisons use identical split-contained windows. Matched AEF uses the same temporal architecture and a fixed within-split metadata permutation for the shuffled diagnostic.

Uncertainty. Setting tables show mean and sample standard deviation across three seeds. Paired inference uses eight dataset clusters, with horizons averaged within datasets and paired seeds resampled within each selected cluster. Exact sign flips enumerate all 256 sign assignments of dataset means. Main Holm families contain four baseline comparisons per metric. The four-dataset mechanism tests use 16 sign assignments. Corruption intervals use the eight dataset clusters; the additional nine-comparison sensitivity family combines three repair modes and three routing comparators separately for each target/metric. Its scope was added during analysis and is labelled accordingly.

Fixed-head output verification. The corruption results above use complete inverse-RevIN head outputs, including learned affine parameters. The 24 original checkpoints were replayed with unchanged corruption cases and validation windows; validation single heads were selected again on that scale. All full-model corruption errors match the original outputs, and weighted head reconstructions match the deployed prediction within floating-point tolerance. The main-board model predictions and their reported scores are unchanged.

E.1 Matched Information and Mechanism Controls

Shared audit inputs. Across eight datasets at $H = 96$, AIF obtains mean MAE 0.28343, compared with 0.28161 for TQNet with the audit adapter and 0.27670 for raw-input TQNet. The paired TQNet+audit-minus-AIF effect is -0.00183 with 95% interval $[-0.00986, 0.00638]$. PatchTST+audit obtains 0.53842; its ETTm2 seed values are 3.06995, 1.41549, and 0.77249. Appendix E reports these configuration-specific variations with repair-only controls. An additional official TimeMixer

Table 22: Complete 32-setting MAE results: three-seed mean ± sample standard deviation. All settings and methods share the stated short budget.

Dataset	H	AIF	DLinear	PatchTST	TQNet	iTransformer
ETTh1	96	0.3936 ± 0.0019	0.3942 ± 0.0046	0.3959 ± 0.0031	0.3926 ± 0.0014	0.4097 ± 0.0024
ETTh1	192	0.4198 ± 0.0004	0.4280 ± 0.0024	0.4302 ± 0.0030	0.4243 ± 0.0004	0.4377 ± 0.0028
ETTh1	336	0.4557 ± 0.0219	0.4563 ± 0.0052	0.4569 ± 0.0005	0.4505 ± 0.0062	0.4631 ± 0.0058
ETTh1	720	0.4683 ± 0.0002	0.5000 ± 0.0044	0.4958 ± 0.0339	0.4739 ± 0.0103	0.4904 ± 0.0127
ETTh2	96	0.3398 ± 0.0015	0.3434 ± 0.0030	0.3430 ± 0.0030	0.3360 ± 0.0019	0.3616 ± 0.0054
ETTh2	192	0.3891 ± 0.0012	0.4097 ± 0.0080	0.3935 ± 0.0025	0.3922 ± 0.0038	0.4058 ± 0.0064
ETTh2	336	0.4338 ± 0.0124	0.4516 ± 0.0061	0.4303 ± 0.0023	0.4307 ± 0.0036	0.4428 ± 0.0057
ETTh2	720	0.4389 ± 0.0003	0.5761 ± 0.0022	0.4530 ± 0.0017	0.4410 ± 0.0010	0.4527 ± 0.0016
ETTm1	96	0.3527 ± 0.0014	0.3574 ± 0.0009	0.3566 ± 0.0049	0.3525 ± 0.0047	0.3728 ± 0.0006
ETTm1	192	0.3743 ± 0.0007	0.3802 ± 0.0003	0.3752 ± 0.0015	0.3748 ± 0.0008	0.3934 ± 0.0023
ETTm1	336	0.3984 ± 0.0013	0.4043 ± 0.0020	0.3984 ± 0.0031	0.3986 ± 0.0004	0.4140 ± 0.0025
ETTm1	720	0.4335 ± 0.0003	0.4442 ± 0.0027	0.4343 ± 0.0013	0.4355 ± 0.0013	0.4476 ± 0.0025
ETTm2	96	0.2556 ± 0.0004	0.2635 ± 0.0020	0.2603 ± 0.0019	0.2538 ± 0.0008	0.2805 ± 0.0008
ETTm2	192	0.2986 ± 0.0012	0.3098 ± 0.0041	0.3039 ± 0.0009	0.2975 ± 0.0009	0.3178 ± 0.0014
ETTm2	336	0.3371 ± 0.0019	0.3588 ± 0.0076	0.3428 ± 0.0005	0.3368 ± 0.0015	0.3517 ± 0.0017
ETTm2	720	0.3927 ± 0.0008	0.4331 ± 0.0072	0.4027 ± 0.0007	0.3947 ± 0.0007	0.4075 ± 0.0009
electricity	96	0.2572 ± 0.0012	0.2719 ± 0.0001	0.2547 ± 0.0006	0.2368 ± 0.0005	0.2459 ± 0.0005
electricity	192	0.2721 ± 0.0001	0.2751 ± 0.0002	0.2636 ± 0.0005	0.2525 ± 0.0005	0.2572 ± 0.0003
electricity	336	0.2871 ± 0.0018	0.2903 ± 0.0004	0.2803 ± 0.0006	0.2703 ± 0.0005	0.2744 ± 0.0003
electricity	720	0.3191 ± 0.0023	0.3225 ± 0.0007	0.3115 ± 0.0008	0.3040 ± 0.0011	0.3102 ± 0.0016
Exchange	96	0.2011 ± 0.0023	0.2013 ± 0.0047	0.2033 ± 0.0018	0.2084 ± 0.0076	0.2263 ± 0.0084
Exchange	192	0.2965 ± 0.0010	0.2820 ± 0.0008	0.3072 ± 0.0009	0.2974 ± 0.0024	0.3198 ± 0.0045
Exchange	336	0.4122 ± 0.0063	0.3756 ± 0.0033	0.4314 ± 0.0112	0.4271 ± 0.0137	0.4286 ± 0.0095
Exchange	720	0.7040 ± 0.0012	0.6280 ± 0.0196	0.7523 ± 0.0370	0.7090 ± 0.0050	0.7121 ± 0.0158
Solar	96	0.2572 ± 0.0035	0.3134 ± 0.0004	0.2587 ± 0.0039	0.2297 ± 0.0044	0.2840 ± 0.0181
Solar	192	0.2792 ± 0.0047	0.3328 ± 0.0015	0.2802 ± 0.0077	0.2594 ± 0.0024	0.2939 ± 0.0031
Solar	336	0.2968 ± 0.0089	0.3448 ± 0.0004	0.2942 ± 0.0092	0.2692 ± 0.0038	0.3107 ± 0.0057
Solar	720	0.3000 ± 0.0060	0.3392 ± 0.0007	0.2979 ± 0.0015	0.2720 ± 0.0013	0.3098 ± 0.0004
weather	96	0.2102 ± 0.0083	0.2376 ± 0.0012	0.2061 ± 0.0002	0.2037 ± 0.0017	0.2369 ± 0.0065
weather	192	0.2477 ± 0.0002	0.2716 ± 0.0010	0.2464 ± 0.0003	0.2471 ± 0.0007	0.2732 ± 0.0025
weather	336	0.2909 ± 0.0049	0.3121 ± 0.0022	0.2858 ± 0.0002	0.2876 ± 0.0008	0.3099 ± 0.0021
weather	720	0.3427 ± 0.0010	0.3692 ± 0.0067	0.3366 ± 0.0009	0.3408 ± 0.0007	0.3567 ± 0.0035

Table 23: Complete 32-setting MSE results: three-seed mean ± sample standard deviation. All settings and methods share the stated short budget.

Dataset	H	AIF	DLinear	PatchTST	TQNet	iTransformer
ETTh1	96	0.3851 ± 0.0024	0.3826 ± 0.0025	0.3778 ± 0.0026	0.3819 ± 0.0031	0.3957 ± 0.0044
ETTh1	192	0.4367 ± 0.0004	0.4355 ± 0.0008	0.4332 ± 0.0040	0.4362 ± 0.0031	0.4433 ± 0.0051
ETTh1	336	0.4927 ± 0.0230	0.4804 ± 0.0041	0.4888 ± 0.0111	0.4911 ± 0.0103	0.4944 ± 0.0087
ETTh1	720	0.4828 ± 0.0013	0.5042 ± 0.0045	0.5293 ± 0.0668	0.4939 ± 0.0221	0.5092 ± 0.0181
ETTh2	96	0.2921 ± 0.0008	0.2914 ± 0.0017	0.2948 ± 0.0064	0.2892 ± 0.0035	0.3165 ± 0.0075
ETTh2	192	0.3754 ± 0.0019	0.3895 ± 0.0091	0.3710 ± 0.0056	0.3746 ± 0.0074	0.3942 ± 0.0084
ETTh2	336	0.4259 ± 0.0071	0.4482 ± 0.0070	0.4172 ± 0.0048	0.4309 ± 0.0135	0.4419 ± 0.0098
ETTh2	720	0.4221 ± 0.0011	0.6614 ± 0.0011	0.4479 ± 0.0087	0.4263 ± 0.0025	0.4403 ± 0.0023
ETTm1	96	0.3266 ± 0.0025	0.3346 ± 0.0006	0.3287 ± 0.0114	0.3234 ± 0.0053	0.3428 ± 0.0008
ETTm1	192	0.3734 ± 0.0015	0.3753 ± 0.0002	0.3608 ± 0.0033	0.3659 ± 0.0024	0.3885 ± 0.0035
ETTm1	336	0.4096 ± 0.0044	0.4086 ± 0.0007	0.3907 ± 0.0029	0.4001 ± 0.0013	0.4192 ± 0.0042
ETTm1	720	0.4754 ± 0.0009	0.4709 ± 0.0018	0.4539 ± 0.0019	0.4640 ± 0.0032	0.4818 ± 0.0035
ETTm2	96	0.1778 ± 0.0004	0.1814 ± 0.0001	0.1783 ± 0.0031	0.1740 ± 0.0008	0.1991 ± 0.0012
ETTm2	192	0.2428 ± 0.0017	0.2465 ± 0.0012	0.2455 ± 0.0024	0.2412 ± 0.0010	0.2609 ± 0.0017
ETTm2	336	0.3039 ± 0.0021	0.3124 ± 0.0051	0.3071 ± 0.0020	0.3009 ± 0.0017	0.3189 ± 0.0022
ETTm2	720	0.4005 ± 0.0017	0.4276 ± 0.0058	0.4070 ± 0.0020	0.4032 ± 0.0013	0.4190 ± 0.0023
electricity	96	0.1720 ± 0.0014	0.1969 ± 0.0001	0.1679 ± 0.0003	0.1454 ± 0.0002	0.1572 ± 0.0010
electricity	192	0.1926 ± 0.0002	0.1957 ± 0.0002	0.1786 ± 0.0007	0.1634 ± 0.0006	0.1704 ± 0.0006
electricity	336	0.2060 ± 0.0044	0.2090 ± 0.0006	0.1948 ± 0.0005	0.1811 ± 0.0003	0.1872 ± 0.0005
electricity	720	0.2448 ± 0.0041	0.2437 ± 0.0005	0.2336 ± 0.0009	0.2220 ± 0.0012	0.2303 ± 0.0022
Exchange	96	0.0839 ± 0.0018	0.0843 ± 0.0044	0.0862 ± 0.0008	0.0896 ± 0.0061	0.1014 ± 0.0079
Exchange	192	0.1773 ± 0.0020	0.1524 ± 0.0015	0.1881 ± 0.0010	0.1748 ± 0.0037	0.1980 ± 0.0048
Exchange	336	0.3284 ± 0.0108	0.2540 ± 0.0053	0.3629 ± 0.0224	0.3507 ± 0.0221	0.3472 ± 0.0177
Exchange	720	0.8768 ± 0.0068	0.7038 ± 0.0526	1.0275 ± 0.1181	0.8899 ± 0.0201	0.8807 ± 0.0383
Solar	96	0.2423 ± 0.0086	0.2856 ± 0.0010	0.2257 ± 0.0036	0.2038 ± 0.0036	0.2649 ± 0.0211
Solar	192	0.2797 ± 0.0028	0.3160 ± 0.0014	0.2606 ± 0.0116	0.2478 ± 0.0021	0.2908 ± 0.0032
Solar	336	0.3020 ± 0.0097	0.3570 ± 0.0006	0.2890 ± 0.0125	0.2670 ± 0.0083	0.3225 ± 0.0110
Solar	720	0.3096 ± 0.0197	0.3662 ± 0.0007	0.3025 ± 0.0067	0.2736 ± 0.0015	0.3207 ± 0.0033
weather	96	0.1726 ± 0.0096	0.1996 ± 0.0025	0.1652 ± 0.0009	0.1648 ± 0.0014	0.1969 ± 0.0079
weather	192	0.2140 ± 0.0007	0.2358 ± 0.0015	0.2101 ± 0.0010	0.2135 ± 0.0011	0.2450 ± 0.0039
weather	336	0.2735 ± 0.0073	0.2809 ± 0.0010	0.2655 ± 0.0002	0.2683 ± 0.0008	0.2950 ± 0.0019
weather	720	0.3521 ± 0.0037	0.3437 ± 0.0038	0.3426 ± 0.0012	0.3475 ± 0.0009	0.3694 ± 0.0037

Table 24: Complete 32-setting SMAPE results: three-seed mean ± sample standard deviation. All settings and methods share the stated short budget.

Dataset	H	AIF	DLinear	PatchTST	TQNet	iTransformer
ETTh1	96	72.9015 ± 0.2706	73.0850 ± 0.8580	72.7110 ± 0.1100	72.2904 ± 0.4006	74.7787 ± 0.6899
ETTh1	192	75.5748 ± 0.1829	77.8597 ± 0.7441	76.6691 ± 0.3291	76.2803 ± 0.0494	78.2164 ± 0.7607
ETTh1	336	80.2980 ± 3.4621	81.8338 ± 0.9123	80.0342 ± 0.2101	78.9900 ± 1.2129	81.2781 ± 2.1438
ETTh1	720	84.0112 ± 0.1850	90.1527 ± 0.7037	86.8940 ± 5.9348	84.5216 ± 1.9437	87.3841 ± 3.0387
ETTh2	96	52.2956 ± 0.4571	52.2963 ± 0.3299	52.5991 ± 0.2706	51.6927 ± 0.0619	55.6175 ± 0.7349
ETTh2	192	57.3258 ± 0.3215	59.1815 ± 0.6468	58.3309 ± 0.2148	57.9541 ± 0.1441	60.4997 ± 0.6326
ETTh2	336	63.9385 ± 2.2120	64.6099 ± 0.4822	63.3247 ± 0.1935	62.7967 ± 0.1081	65.5868 ± 0.3669
ETTh2	720	66.5861 ± 0.0822	77.3161 ± 0.1859	67.3289 ± 0.1266	66.4642 ± 0.0358	68.6014 ± 0.4317
ETTm1	96	68.3833 ± 0.4607	69.4450 ± 0.1899	68.9039 ± 0.4068	68.8844 ± 0.5948	71.8601 ± 0.5512
ETTm1	192	70.6736 ± 0.3498	72.2294 ± 0.1136	71.3636 ± 0.5825	71.0216 ± 0.3216	73.4148 ± 0.1210
ETTm1	336	73.9016 ± 0.2508	75.9374 ± 0.2435	74.2887 ± 0.2708	73.8591 ± 0.1484	75.8118 ± 0.4816
ETTm1	720	78.1376 ± 0.0995	81.9047 ± 0.3889	78.4776 ± 0.2279	78.4662 ± 0.5116	79.2925 ± 0.5837
ETTm2	96	42.6322 ± 0.1167	43.8670 ± 0.2699	43.8611 ± 0.0886	42.6016 ± 0.1611	46.7756 ± 0.3116
ETTm2	192	47.1798 ± 0.1113	48.6820 ± 0.2693	48.1752 ± 0.1557	46.9986 ± 0.1195	50.4995 ± 0.1312
ETTm2	336	51.1165 ± 0.2904	53.5997 ± 0.5526	51.9391 ± 0.0408	51.3677 ± 0.1789	53.8357 ± 0.2957
ETTm2	720	57.0742 ± 0.1404	61.3417 ± 0.5122	58.3555 ± 0.2713	57.2787 ± 0.1279	59.5794 ± 0.0269
electricity	96	50.2347 ± 0.1778	52.0147 ± 0.0211	49.9007 ± 0.1384	47.1779 ± 0.1149	48.3201 ± 0.1195
electricity	192	51.9152 ± 0.0786	52.2803 ± 0.0206	50.7845 ± 0.0333	49.2128 ± 0.1315	49.6858 ± 0.0178
electricity	336	54.0392 ± 0.2013	54.5611 ± 0.1133	53.1625 ± 0.0995	51.6705 ± 0.1306	52.0509 ± 0.0350
electricity	720	58.2871 ± 0.2679	58.9100 ± 0.1131	57.2069 ± 0.0595	56.1565 ± 0.1511	56.8875 ± 0.2041
Exchange	96	29.0763 ± 0.3343	28.7671 ± 0.1568	29.2717 ± 0.1149	30.1008 ± 0.8897	31.8678 ± 0.7617
Exchange	192	39.3919 ± 0.0633	39.2900 ± 0.3310	40.1961 ± 0.1085	39.7906 ± 0.1646	41.8385 ± 0.5137
Exchange	336	49.2591 ± 0.4112	48.7683 ± 0.3771	49.7938 ± 0.6114	49.9172 ± 1.1542	50.5863 ± 0.7242
Exchange	720	65.8120 ± 0.3541	78.7080 ± 3.1280	66.0738 ± 0.7222	64.9592 ± 0.3195	66.4156 ± 0.4671
Solar	96	42.8159 ± 0.5748	54.0675 ± 0.1116	43.5738 ± 0.4582	38.3897 ± 0.4028	45.9155 ± 2.3463
Solar	192	45.1313 ± 0.8520	56.5500 ± 0.3683	46.0992 ± 1.2041	41.8198 ± 0.3528	46.9018 ± 0.3651
Solar	336	47.0780 ± 1.3852	56.2786 ± 0.1392	46.9982 ± 1.5002	42.2026 ± 0.4169	48.3037 ± 0.6180
Solar	720	46.8175 ± 0.4737	54.1461 ± 0.0266	46.5299 ± 0.6673	42.0562 ± 0.3144	47.7365 ± 0.1626
weather	96	53.0069 ± 1.2659	59.1865 ± 0.3256	52.5049 ± 0.5465	51.9625 ± 0.4794	57.5334 ± 1.4003
weather	192	59.1949 ± 0.1921	66.0119 ± 0.5320	58.9823 ± 0.2485	58.8557 ± 0.1726	63.0203 ± 0.3611
weather	336	64.9126 ± 0.5748	74.0209 ± 0.5702	63.9290 ± 0.2018	64.4052 ± 0.1600	67.8586 ± 0.3103
weather	720	71.5420 ± 0.0663	86.0687 ± 2.0896	70.8889 ± 0.1595	71.5431 ± 0.2230	74.0268 ± 0.6677

Table 25: Dataset-cluster paired effects on the complete board. Positive differences favor AIF. Intervals are 95% bootstrap intervals; Holm families contain four comparators per metric.

Metric	Comparator	Difference	CI lower	CI upper	Holm p
MAE	DLinear	0.01632	-0.00079	0.03202	0.37500
MSE	DLinear	0.01014	-0.01653	0.03473	0.82812
SMAPE	DLinear	4.13833	2.20264	6.55956	0.03125
MAE	PatchTST	0.00398	-0.00130	0.01009	0.46875
MSE	PatchTST	0.00257	-0.00877	0.01855	0.87500
SMAPE	PatchTST	0.26900	-0.22997	0.78852	0.35938
MAE	TQNet	-0.00485	-0.01260	0.00159	0.46875
MSE	TQNet	-0.00783	-0.01849	0.00153	0.72656
SMAPE	TQNet	-0.90177	-2.10436	0.04750	0.35938
MAE	iTransformer	0.01384	0.00609	0.01881	0.06250
MSE	iTransformer	0.01227	0.00275	0.01889	0.18750
SMAPE	iTransformer	1.91988	0.66511	2.77677	0.09375

Table 26: Complete routing factorial on ETTh2, ETTm1, Exchange, and Solar, $H = 96$, three seeds. Quality is the learned gate's reliability input; control is explicit stabilization.

Gate quality / explicit control	MAE	MSE	SMAPE
Off / Off	0.28901	0.23809	48.272
On / Off	0.28668	0.23587	47.937
Off / On	0.28879	0.23826	48.349
On / On	0.28818	0.23737	48.105

Table 27: Both scoring targets for the completed causal corruption study. Source is the retained pre-injection observed target; altered is the injected observed target.

Repair	Route	Source MAE	Source MSE	Source SMAPE	Altered MAE	Altered MSE	Altered SMAPE
causal	full	1.08446	7.27592	89.61160	0.83163	8.96895	68.44942
causal	pre stabilization	1.20634	8.76744	92.13183	0.93851	10.21497	70.59624
causal	uniform	1.03155	7.95996	92.70847	0.79993	5.76041	72.76309
causal	validation single	1.01008	5.99100	88.36640	0.82975	9.83494	71.31019
none	full	1.07752	7.29792	89.27307	0.83218	8.97076	68.64037
none	pre stabilization	1.19172	8.73095	91.48541	0.93612	10.20143	70.60032
none	uniform	1.03428	7.96884	92.84484	0.80335	5.80288	72.94532
none	validation single	1.01008	5.99101	88.36575	0.82975	9.83494	71.30780
oracle	full	0.28548	0.23430	51.41011	0.96039	14.10288	81.27154
oracle	pre stabilization	0.29212	0.24086	51.97391	0.96710	14.11769	81.65111
oracle	uniform	0.34196	0.30264	61.28097	0.97985	14.11771	86.58725
oracle	validation single	0.30243	0.25731	53.02690	0.96664	14.11697	81.53355

Table 28: Matched-information and auxiliary comparisons by dataset at $H = 96$. Entries give MAE mean $\pm$ sample standard deviation over three seeds. Empty cells indicate settings outside the registered suite.

Suite	Method/input	ETTh2	ETTm1	Exchange	Solar
aef matched	AEFMatched/causal	0.3456 ± 0.0029	0.3629 ± 0.0010	0.2208 ± 0.0042	0.2441 ± 0.0029
aef matched	AEFMatched/shuffled	0.3482 ± 0.0039	0.3552 ± 0.0009	0.2150 ± 0.0072	0.2356 ± 0.0131
audit boundary	AIFCausal/causal	0.3450 ± 0.0014	0.3509 ± 0.0009	0.2007 ± 0.0031	0.2569 ± 0.0034
audit boundary	AIFOffline/offline	0.8146 ± 0.0226	0.3530 ± 0.0032	0.2053 ± 0.0037	0.2611 ± 0.0052
audit boundary	AIFRaw/raw	0.3430 ± 0.0037	0.3496 ± 0.0020	0.2013 ± 0.0030	0.2544 ± 0.0062
official baseline	TimeMixerOfficial/raw	0.3455 ± 0.0042	0.3604 ± 0.0074	0.2214 ± 0.0168	0.2561 ± 0.0050
same information	PatchTST/audit	0.3917 ± 0.0621	0.4072 ± 0.0186	0.3630 ± 0.0557	0.3118 ± 0.0245
same information	PatchTST/repair	0.3433 ± 0.0025	0.3566 ± 0.0049	0.2042 ± 0.0028	0.2563 ± 0.0029
same information	TQNet/audit	0.3584 ± 0.0054	0.3521 ± 0.0028	0.2054 ± 0.0020	0.2414 ± 0.0048
same information	TQNet/repair	0.3361 ± 0.0019	0.3525 ± 0.0047	0.2035 ± 0.0001	0.2300 ± 0.0041

Table 29: Remaining dataset cells for the eight-dataset matched-information and official TimeMixer suites, $H = 96$, three-seed MAE mean $\pm$ sample standard deviation.

Suite	Method/input	ETTh1	ETTm2	Weather	Electricity
official baseline	TimeMixerOfficial/raw	0.4058 ± 0.0047	0.2587 ± 0.0001	0.2081 ± 0.0036	0.2596 ± 0.0022
same information	PatchTST/audit	0.4379 ± 0.0172	1.7526 ± 1.1853	0.2622 ± 0.0250	0.3810 ± 0.0412
same information	TQNet/audit	0.3965 ± 0.0007	0.2595 ± 0.0008	0.2022 ± 0.0018	0.2374 ± 0.0011

Table 30: Mean retained-source MAE under corrupted inputs, eight datasets, $H = 96$, three seeds. Each repair condition includes all four corruption types. Bold denotes the lowest error within a column.

Route	No repair	Causal repair	Oracle repair
Pre-stabilization	1.19172	1.20634	0.29212
Uniform	1.03428	1.03155	0.34196
Validation-selected single	**1.01008**	**1.01008**	0.30243
Full control	1.07752	1.08446	**0.28548**

comparison covers all eight datasets at $H = 96$: its mean MAE is 0.28946 versus 0.28343 for AIF, a 2.08% difference. This comparison uses TimeMixer; the archived custom model is labelled TM++-local.

Quality features and explicit control. The factorial study uses ETTh2, ETTm1, Exchange-Rate, and Solar-Energy at $H = 96$, retaining all four combinations of gate quality features and explicit stabilization. With stabilization off, quality features reduce mean MAE from 0.28901 to 0.28668, a 0.81% improvement with the same direction on all four datasets. The paired interval is [0.00066, 0.00484]. With stabilization on, the quality-feature gain is 0.00061, with interval [-0.00348,0.00340]. With quality features enabled, mean MAE is 0.28668 without stabilization and 0.28818 with stabilization; the unstabilized-minus-stabilized difference has interval [-0.00475, 0.00189]. The statistics prior is disabled in all four groups; active explicit control still receives the actual quality statistics. Four dataset clusters give a minimum exact two-sided sign-flip p of 0.125, so these effects are descriptive mechanism evidence.

Causal and offline preprocessing. On the same four datasets with matched split-contained windows, causal, offline, and raw-input AIF obtain mean MAE 0.28835, 0.40852, and 0.28707, respectively. Causal preprocessing improves on the offline variant in all four datasets, with the largest difference on ETTh2. Its mean difference from raw-input AIF is 0.00129; the paired interval includes zero. Prefix checks separately establish the temporal information boundary. A matched-capacity AEF diagnostic obtains MAE 0.29334 with aligned input metadata and 0.28851 with shuffled metadata; its shuffled-minus-aligned effect has interval [-0.01014,0.00045]. These measurements distinguish feature alignment, preprocessing, and route control.

E.2 Corrected Replay of Jointly Trained Checkpoints

We inject plateaus, zero blocks, 24-step repetitions, and level shifts into 32 predetermined test origins for each of the 24 AIF $H = 96$ checkpoints. The altered span covers the final 24 input steps and first 48 forecast steps. Retained-source and altered-observation targets enter scoring only. The no-repair condition uses the corrupted input with zero audit metadata/uncertainty; causal repair uses endpoint detection and forward filling; oracle repair supplies the retained source input. Routes within each condition share input tensors and fully inverse-normalized head predictions. Recombining those heads with the deployed weights reproduces the model prediction; the replay also preserves the original full-model scores. The validation-selected route is fixed before test evaluation. Retained source denotes the pre-injection observed series, whose natural recording artifacts remain.

With causal repair, full control reduces source-target MAE by 10.10% relative to pre-stabilization weights, with positive differences in seven of eight datasets. The mean paired gain is 0.12188, with 95% interval [0.02886, 0.25895]. The validation-selected single route has lower mean error in this condition; with oracle repair, full control has the lowest error among the four routes. This pattern makes the repair condition and comparator central to interpreting route-control gains. An additional nine-comparison Holm sensitivity analysis, covering three repair conditions and three route comparators, gives $p = 0.10938$ for each non-oracle pre-stabilization source-MAE comparison. Appendix E reports source and altered-target metrics together.

F ORIGINAL END-TO-END AIF ARCHITECTURE

This section describes the jointly trained, sparse-gate AIF used in the end-to-end comparison. The neutral-bank study retains the seven head modules and control rules but replaces their fitting and gate-training protocol as specified in Section 4.1. Figure 4 depicts the joint training objective; Figure 2 depicts the two-stage neutral protocol.

F.1 OVERVIEW OF AIF

AIF is a support-aware routed forecaster. Given an input window X, it produces route-head predictions $f_k(X)$ and combines them with router weights $\pi_k(X)$:

$$\hat{Y} = \sum_{k=1}^{K} \pi_k(X) f_k(X), \quad \pi_k(X) \geq 0, \quad \sum_{k=1}^{K} \pi_k(X) = 1. \tag{3}$$

Figure 4 summarizes the AIF data flow. AIF contains internally implemented route heads for direct linear forecasting, trend-seasonal decomposition, local patch encoding, temporal convolution, variate mixing, periodic-frequency structure, and audit residuals. These heads are lightweight modules within one forecasting architecture. The controller receives temporal representations and reliability statistics, including clean-support ratio, mask statistics, uncertainty summaries, horizon identifiers, and cross-channel correlation indicators. It then forms the convex forecast in Eq. (3). Prediction metadata follows a fixed input-side allowlist that rejects fields containing `target`, `future`, or `test`. The end-to-end comparison uses an endpoint-only causal detector: zero and flat runs are flagged once their thresholds are reached, earlier samples keep their original labels, and repair carries forward the latest unflagged value. Normalization and the Solar phase exemption are fitted on training observations. At prediction time, summaries use the observed input prefix. Suffix-perturbation and prefix-truncation checks verify this information boundary. Appendix H separately reports the earlier offline-audit setting.

Why complementary routes? The route bank separates temporal structures that need different representations. Direct and decomposition heads provide linear and trend-seasonal predictions; patch attention and temporal convolution capture local patterns; variate mixing models channel interactions; and the spectral head summarizes periodic structure. The audit-residual head adds a correction built from the raw–masked difference and audit features. All heads predict in a shared RevIN-normalized space. The controller combines their outputs before inverse RevIN restores the forecast scale, allowing the same head bank to serve windows with different quality and temporal structure. Appendix I.4 specifies the head supervision and controller statistics.

AIF uses quality evidence to constrain route usage for each window. The artifact and variate routes are the key reliability-sensitive paths. The artifact route models residual structure, with its budget tied to clean support. The variate-route weight budget increases as cross-channel dependence strengthens. These constraints translate the filtering principle in Section G into window-dependent budgets while preserving the full route bank.

F.2 RELIABILITY-AWARE ROUTE CONSTRAINTS

AIF applies two reliability-aware constraints to the unconstrained router weights. First, because the artifact-residual route is most directly associated with the nuisance channel in the observation model, AIF computes a clean-confidence score

$$\gamma(X) = \rho_{\text{clean}}(X)\,[1 - u(X)]\,[1 - m(X)]\,[1 - a_{\text{score}}(X)],$$

where ρ_{clean} denotes clean support, u audit-derived uncertainty, m mask density, and a_{score} an artifact score. The artifact route is then capped by a reliability-dependent budget

$$\tilde{\pi}_a = \min\left\{\pi_a, \gamma(X) w_a^{\text{clean}} + [1 - \gamma(X)] w_a^{\text{default}}\right\}.$$

Thus, artifact routing is tightly limited when a window appears clean and relaxed when clean support is low or uncertainty is high.

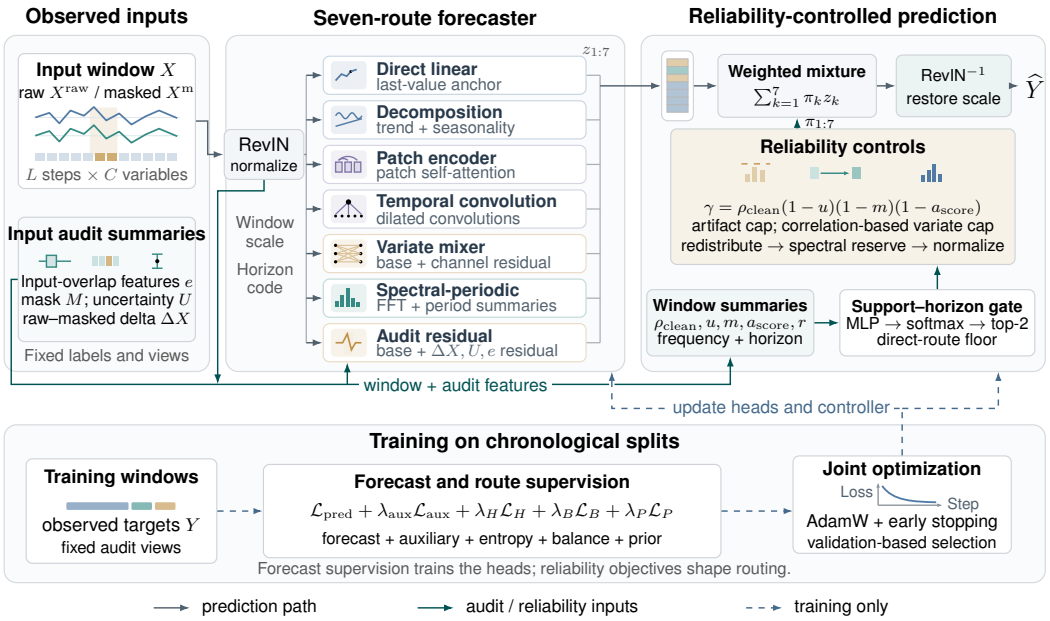


Figure 4: AIF architecture. Left: the observed window and input-side audit summaries; the end-to-end comparison computes them causally. Center: seven shared forecasting heads process the normalized window, with audit-delta inputs for the residual head; z_k denotes a normalized head prediction. Right: the support–horizon gate and reliability controls produce mixture weights; inverse RevIN restores the prediction scale. The lower band shows forecasting and routing objectives, with dashed arrows for training-only updates. Icons, waveforms, and colored tiles are schematic.

Second, AIF varies the cross-channel route budget smoothly with correlation reliability. For mean absolute input correlation $r(X) \in [0, 1]$, let

$$b_{\text{var}}(X) = w_{\text{var}}^{\text{lowcorr}} + (w_{\text{var}}^{\text{default}} - w_{\text{var}}^{\text{lowcorr}})\sigma\left(\frac{r(X) - \tau_r}{c_{\text{var}}}\right), \qquad \tilde{\pi}_{\text{var}} = \min\{\pi_{\text{var}}, b_{\text{var}}(X)\},$$

where σ is the logistic sigmoid. The budget increases from 0.08 toward 0.25 as correlation grows. Removed mass goes to the direct and decomposition heads. A separate spectral reservation uses frequency concentration, and a direct-route floor is applied before these adjustments. Together these operations define the implemented reliability controller.

The concrete thresholds, budgets, and training-time prior form are reported in Appendix I.4. We call the learned sparse weights after the direct-route floor but before reliability caps, spectral reservation, redistribution, and renormalization the *pre-stabilization router*. The reported AIF prediction uses the corresponding post-stabilization weights. This definition is used by all matched-router and controlled-corruption comparisons.

Interpreting full redistribution. For fixed head predictions, let $e = \hat{Y} - Y$ and $d = \hat{Y}' - \hat{Y}$, where $\hat{Y}'$ applies the full reliability controller. Then

$$\mathcal{R}_2(\hat{Y}') - \mathcal{R}_2(\hat{Y}) = 2\mathbb{E}\langle e, d\rangle + \mathbb{E}\|d\|_2^2. \tag{4}$$

The error alignment of the complete prediction change determines the squared-risk effect. The artifact-cap derivation is a special case under its stated assumptions. For absolute loss, $|\mathcal{R}_1(\hat{Y}') - \mathcal{R}_1(\hat{Y})| \leq \mathbb{E}\|d\|_1$; MAE and SMAPE effects of the full controller are measured in the paired experiments.

F.3 Training-time Routing Control

AIF training combines forecasting accuracy with reliability-aware route regularization. For each minibatch $\mathcal{B}$, the model receives supervised forecasting feedback together with routing-control signals

derived from the input window. The objective is

$$\mathcal{L}_{\mathrm{AIF}} = \mathcal{L}_{\mathrm{pred}} + \lambda_{\mathrm{aux}}\mathcal{L}_{\mathrm{aux}} + \lambda_H\mathcal{L}_H + \lambda_B\mathcal{L}_B + \lambda_P\mathcal{L}_P.$$

Here, $\mathcal{L}_{\mathrm{pred}}$ is the supervised prediction loss, $\mathcal{L}_{\mathrm{aux}}$ keeps route-specific heads predictive, and $\mathcal{L}_H$ and $\mathcal{L}_B$ regularize router sharpness and balanced route usage. The prior term $\mathcal{L}_P$ softly enforces the same reliability logic used at inference: it limits artifact routing under strong clean evidence, discourages variate mixing under low cross-channel correlation, and penalizes spectral-weight shortfall when periodic evidence is concentrated.

The joint objective trains predictive heads while tying their weights to window quality. The exact prior form and hyperparameters are given in Appendix I.4.

F.4 Archived Artifact Error Fitting Diagnostics

Artifact Error Fitting (AEF) evaluates observed-target predictability using artifact-oriented features and a matched feature-shuffling control. Each probe is supervised by the observed future target; the artifact residual in the theoretical model is an analytical quantity.

The AEF variants separate feature alignment from model capacity. AEF-Weak is an MLP over window statistics and tabular artifact, phase, severity, variable-count, and horizon features. AEF-Control uses the same architecture and dimensionality but independently shuffles the artifact- and phase-oriented columns, breaking their alignment with each window. AEF-Plus is a temporal probe that combines a temporal trunk, boundary encoder, metadata-token encoder, cross-gated fusion, and sparse expert head. AEF uses target-side audit labels in its offline diagnostic setting and is evaluated separately from the AIF forecasting comparison.

Table 31: AEF observed-target prediction diagnostics across evaluation views. Lower MAE is better.

Evaluation view	Tasks	AEF-Plus	AEF-Weak	AEF-Control
raw	32	**0.361**	0.808	0.808
clean-like	12	**0.372**	0.849	0.805
intervened	22	**0.509**	0.721	0.719
balanced	3	**0.252**	0.339	0.366

AEF-Plus has lower error than the two smaller probes, including raw MAE 0.361 versus 0.808. This comparison measures the predictability available to an expressive diagnostic with additional inputs. The architecture-matched Weak–Control comparison is equal on raw rows, favors Control on clean-like and intervened rows, and favors Weak on balanced rows. Evidence for the incremental value of aligned artifact features is therefore view-dependent; the Plus gap also includes capacity and input differences.

G Additional Derivations

These supplementary calculations motivated the original AIF design. They use a stylized latent-target model and conditional orthogonality; the current empirical targets are the operational retained-source and altered-observation quantities in Section 3. Let $Y^\star = f^\star(X)$ denote a latent clean prediction, with observed $Y = f^\star(X) + Sa(X) + \epsilon$, $S \in \{0, 1\}$, and predictable residual $a(X)$. Define $\mathcal{R}_{\mathrm{clean}}$ as squared error relative to the latent clean outcome conditional on $S = 0$. The scalar calculation below uses the clean predictor as the clean outcome; the later cap model includes an orthogonal residual. The full redistribution identity treats the actual change of forecast separately.

G.1 Raw-risk Bias

Consider the one-parameter predictor $f_\lambda(X) = f^\star(X) + \lambda a(X)$ under the observation model $Y = f^\star(X) + Sa(X) + \epsilon$. Assume $\mathbb{E}[\epsilon \mid X, S] = 0$ and $\mathbb{E}[\|a(X)\|_2^2] > 0$. For squared loss,

$$\begin{aligned}\mathcal{R}_{\mathrm{raw}}(\lambda) &= \mathbb{E}\left[\|Y - f_\lambda(X)\|_2^2\right] \\ &= \mathbb{E}\left[\|(S-\lambda)a(X) + \epsilon\|_2^2\right] \\ &= \mathbb{E}\left[(S-\lambda)^2\|a(X)\|_2^2\right] + \mathbb{E}[\|\epsilon\|_2^2],\end{aligned} \tag{5}$$

where the cross term vanishes under the conditional zero-mean assumption. Differentiating with respect to λ gives

$$2\lambda\mathbb{E}[\|a(X)\|_2^2] - 2\mathbb{E}[S\|a(X)\|_2^2] = 0, \tag{6}$$

which yields

$$\lambda^\star_{\mathrm{raw}} = \frac{\mathbb{E}[S\|a(X)\|_2^2]}{\mathbb{E}[\|a(X)\|_2^2]}. \tag{7}$$

Thus, whenever the artifact residual has nonzero energy on artifact-affected examples, raw-risk minimization assigns positive weight to $a(X)$. On clean windows $S = 0$, the corresponding excess risk relative to $f^\star$ is

$$(\lambda^\star_{\mathrm{raw}})^2\mathbb{E}[\|a(X)\|_2^2 \mid S = 0]. \tag{8}$$

G.2 Filtering Interpretation

The raw-risk derivation shows that predictable artifact residuals can receive positive weight under observed-target training. The route-cap analysis below isolates the artifact-induced component on clean windows and explains why reducing artifact-route mass can reduce that component when clean support is strong.

A residual head may combine nuisance variation with clean-relevant structure. AIF preserves this route and restricts its mass under strong clean evidence. The route remains available when clean support is low.

G.3 Artifact-route Cap

On clean windows, let $Y^\star = h_c(X)+\epsilon$. For any nonnegative artifact-route weight q, assume the clean residual is conditionally orthogonal to the routed artifact residual, i.e., $\mathbb{E}_{S=0}[\epsilon^\top q(X)h_a(X)] = 0$ for the route weights being compared. Then

$$\mathcal{R}_{\mathrm{clean}}(q) = \mathbb{E}_{S=0}\|\epsilon - q(X)h_a(X)\|^2. \tag{9}$$

The conditional orthogonality assumption removes the cross term:

$$\mathcal{R}_{\mathrm{clean}}(q) = \mathbb{E}_{S=0}\|\epsilon\|^2 + \mathbb{E}_{S=0}\left[q(X)^2\|h_a(X)\|^2\right]. \tag{10}$$

Subtracting the capped-route risk $q = \tilde{\pi}_a$ from the uncapped-route risk $q = \pi_a$ gives

$$\mathcal{R}_{\mathrm{clean}}(\hat{Y}_\pi) - \mathcal{R}_{\mathrm{clean}}(\hat{Y}_{\tilde{\pi}}) = \mathbb{E}_{S=0}\left[(\pi_a^2 - \tilde{\pi}_a^2)\|h_a(X)\|^2\right]. \tag{11}$$

The monotonic reduction applies to the artifact-induced component isolated by $Y^\star = h_c(X) + \epsilon$. A route that also carries clean-relevant information can remain useful. AIF therefore uses a reliability-dependent cap and retains route capacity under low support or high audit uncertainty.

G.4 Full Redistribution and Metric Scope

For fixed head predictions, let $\hat{Y}$ and $\hat{Y}'$ be the mixtures before and after reliability control, and define $e = \hat{Y} - Y$ and $d = \hat{Y}' - \hat{Y}$. Expanding the square gives the exact identity

$$\mathcal{R}_2(\hat{Y}') - \mathcal{R}_2(\hat{Y}) = 2\mathbb{E}\langle e, d\rangle + \mathbb{E}\|d\|_2^2. \tag{12}$$

Thus full redistribution improves squared risk when $2\mathbb{E}\langle e, d\rangle \leq -\mathbb{E}\|d\|_2^2$. The artifact-cap result is the special case with a fixed clean component and the stated orthogonality assumption. For absolute loss, the triangle inequality gives $|\mathcal{R}_1(\hat{Y}') - \mathcal{R}_1(\hat{Y})| \leq \mathbb{E}\|d\|_1$. These identities specify the role of prediction changes and error alignment; the MAE and SMAPE comparisons measure the realized effects of the full controller.

H Archived Offline-Audit Experiments

H.1 Setup

Datasets. We evaluate AIF on eight real-world time-series datasets from diverse domains, including the ETT family, Exchange-Rate, Solar-Energy, Weather, and Electricity.

Baselines. The archived comparison includes AIF, TQNet, TimeMixer, ModernTCN, iTransformer, PatchTST, DLinear, and the local comparator TM++-local. These baselines span linear, Transformer-based, convolutional, multiscale-mixing, and efficient multivariate forecasting architectures. Unless otherwise specified, the lookback length is set to $L = 96$, and the forecasting horizons are $H \in \{96, 192, 336, 720\}$. All methods use the same chronological splits, evaluation windows, and metric implementation. The main board compares complete systems: AIF receives audit metadata, while standard baselines use their forecasting inputs. The matched same-head controls evaluate inference-time routing with head predictions held fixed.

Metrics. We report both MAE and SMAPE. Both metrics use values standardized by the training-split mean and standard deviation. MAE measures absolute deviation on this scale. SMAPE is computed as

$$\mathrm{SMAPE} = 100 \cdot \frac{1}{N} \sum_{i=1}^{N} \frac{2|\hat{y}_i - y_i|}{|\hat{y}_i| + |y_i| + \varepsilon},$$

where N is the number of evaluated target entries and $\varepsilon = 10^{-6}$ is the numerical stabilizer used in implementation. Lower values are better for both metrics. Reported main-board method metrics are means over three independent repetitions before dataset–horizon aggregation. Main-board means are unweighted averages over settings, and wins are counted separately for MAE and SMAPE. The archived primary board contains three-repetition means; seed-level dispersion is unavailable in that archive. We summarize these means using descriptive ranks and wins. Matched-routing and controlled-corruption comparisons use the prespecified seeds $\{0, 7, 13, 42, 2026\}$. Unless stated otherwise, all calculations use unrounded values.

Evaluation Views. We evaluate forecasting performance under multiple views, each targeting a distinct aspect of data quality and benchmark reliability:

- **Raw**: preserves the original test windows and corresponds to the standard benchmark protocol.
- **Clean-like**: retains windows satisfying conservative audit-support criteria and measures error against their recorded targets.
- **Intervened**: evaluates repaired or counterfactual windows when strict clean support is limited, measuring sensitivity to artifact-related changes.
- **Balanced**: stratifies semantically meaningful regimes, such as night-zero and phase-transition periods in Solar-Energy, through phase-stratified sampling.
- **Natural-clean**: is the archived name for an audit-selected view scored against recorded targets. For Electricity, it coincides with the raw rows and is reported in the raw-equivalent auxiliary setting.

Quality Audit and Main-Board Construction. The quality audit defines the dataset–horizon settings for the main comparison using fixed zero-block, flat-run, near-constant, and repeated-window rules. Detector thresholds, view rules, and semantic decisions are fixed without using any model's test error. Appendix I.6 reports the dataset-specific thresholds and the minimum training-support rules. The main board follows the fixed dataset–horizon eligibility decisions; Table 32 summarizes the observed support and decisions. ETTh1 and ETTm1 enter under clean-like evaluation. Exchange-Rate contributes three clean-like horizons despite low training support under the default rule. Solar-Energy enters under balanced evaluation because night-zero and phase-transition regimes are meaningful operating states. This low-training-support exception is part of the frozen scope. The default horizon set is $H \in \{96, 192, 336, 720\}$. Exchange-Rate and Solar-Energy contribute main-board rows at $H \in \{96, 192, 336\}$; their $H = 720$ rows remain outside the retained scope under the audit eligibility decisions. ETTh2 and ETTm2 fall below the strict support criteria; Weather has anchor-only clean-like support; Electricity uses the raw-equivalent view. The archived comparison uses these 14 settings; the remaining settings characterize boundary and stress conditions.

From events to evaluation windows. The audit merges adjacent same-type detections and overlapping variable records into events, then projects event labels onto input and target windows. Input overlap supplies the forecaster's quality features; target overlap determines evaluation eligibility. Repairable input spans receive local trend interpolation, and mask-only spans receive masking. Solar's balanced view samples equal numbers of eligible windows from each target-phase stratum and averages their errors without weighting. This preserves the eligible operating regimes in the comparison. Together, the input features and evaluation views connect observation quality to prediction and scoring.

Table 32: Quality audit summary and main-board eligibility. Clean-like and alternative shares are the horizon-averaged fractions of test windows in the conservative clean-like and intervened views, respectively; values are rounded to two decimals. Solar eligibility additionally requires phase-balanced support.

Dataset	Events	Clean-like share	Alt. share	Main-board decision
ETTh1	43	0.57	0.67	Included: clean-like
ETTh2	425	0.00	0.00	Boundary: below support rule
ETTm1	50	0.83	0.88	Included: clean-like
ETTm2	575	0.00	0.01	Boundary: below support rule
Exchange-Rate	51	0.31	0.56	Included: frozen low-support exception
Solar-Energy	1461	0.00	1.00	Included: balanced stratification
Weather	38	0.00	1.00	Stress: anchor-only support
Electricity	1	0.00	0.00	Excluded: raw-equivalent natural-clean

H.2 Main Results

Table 33 reports the archived comparison on the 14-setting offline-audit board. AIF achieves the lowest mean MAE and SMAPE, the best average ranks, and 12 MAE and 13 SMAPE wins. The complete per-setting table preserves the variation across datasets and horizons, including the two settings where PatchTST has lower MAE.

Table 33. Archived offline-audit comparison across 14 settings retained by the original protocol: clean-like rows plus phase-balanced Solar-Energy rows. MAE and SMAPE are averaged uniformly over dataset–horizon settings; SMAPE is reported in percentage points. Lower error and rank are better; higher point-estimate wins are better. Ranks and wins are computed from unrounded metric values.

Metric	AIF (ours)	TM++-local	TQNet	ModernTCN	iTransformer	TimeMixer	DLinear	PatchTST
Mean MAE ↓	**0.340**	0.537	0.350	0.372	0.360	0.366	0.405	0.350
Avg. MAE rank ↓	**1.36**	7.57	2.86	5.07	5.43	4.14	6.71	2.86
MAE wins ↑	**12**	0	0	0	0	0	0	2
Mean SMAPE (%) ↓	**60.392**	94.487	62.460	65.293	63.360	63.547	70.709	62.758
Avg. SMAPE rank ↓	**1.14**	7.71	2.93	4.57	4.86	4.50	6.79	3.50
SMAPE wins ↑	**13**	0	0	1	0	0	0	0

The per-setting results in Appendix I.8 show that AIF wins all Exchange-Rate and Solar-Energy settings, three of four ETTh1 settings, and three of four ETTm1 settings under MAE. Under SMAPE, AIF wins 13 of 14 settings. The largest relative gains occur on Solar-Energy and Exchange-Rate. The route summaries in Figure 6 show temporal-convolution emphasis on Solar-Energy and spectral emphasis on Exchange-Rate.

Matched Same-head Routing. We test the routing mechanism by freezing the trained seven-head predictions for all 14 settings and five seeds, then recombining those predictions with four controls. Table 34 reports the per-setting mean of control-minus-AIF error. Positive values favor AIF. The intervention holds heads, checkpoints, windows, and seeds fixed, so the measured differences reflect how those same predictions are combined.

Table 34: Matched routing controls over 14 settings and five seeds. Setting-level summaries use a working independence assumption across the fixed 14 settings. CIs are two-sided 95% Student-t intervals; Holm p values adjust one-sided signed-rank tests within the four-control MAE family.

Control	Mean ΔMAE	95% CI	Positive settings	Holm p
Pre-stabilization router	0.09009	[0.03703, 0.14314]	14/14	0.000244
Quality zeroed; stabilization off	0.08476	[0.03646, 0.13306]	14/14	0.000244
Uniform route mixture	0.07398	[0.03436, 0.11361]	13/14	0.000305
Quality features shuffled	0.00761	[-0.00126, 0.01648]	14/14	0.000244

The pre-stabilization comparison measures the joint effect of caps, spectral reservation, and redistribution. The zeroed-quality control also disables stabilization, so its effect combines feature removal with that change. Shuffling preserves stabilization but changes the alignment of quality features. Its effect is positive in all settings, with an interval that includes zero. These interventions support the combined inference-time controller with the training prior held fixed. Intervals and p values use setting-level working independence. The fixed board has four datasets with shared-horizon observations; we assess consistency through positive-setting counts and the dataset sensitivity summaries below.

Identifiable Controlled Corruptions. We test fidelity to a retained source target under four injected transformations: plateaus, zero blocks, duplicated segments, and level shifts. The changed span crosses the forecast boundary, linking an observed input pattern to the future target. All routing variants receive the same corrupted raw input, retained source input as an oracle repair, and fixed head predictions. This design compares how the controller combines the available predictions with the retained source target available for scoring. Appendix I.7 gives the generator settings and examples.

Table 35: Five-seed controlled-corruption results over 14 dataset–horizon settings. Each setting averages plateau, zero-block, duplicated-segment, and level-shift effects. Differences are control minus reliability-aware AIF; positive values favor AIF. CIs resample the 14 setting means under working independence; horizons share datasets. Holm adjustment applies to one-sided signed-rank tests within each two-control outcome family.

Control	Outcome	Mean difference	95% CI	Positive settings	Holm p
Pre-stabilization router	Retained-source MAE	0.17505	[0.05975, 0.31526]	14/14	0.000122
Validation-selected route	Retained-source MAE	0.01528	[0.00187, 0.03195]	11/14	0.067627
Pre-stabilization router	Gain difference (MAE)	0.00511	[0.00221, 0.00871]	10/14	0.005249
Validation-selected route	Gain difference (MAE)	0.00110	[-0.00043, 0.00301]	6/14	0.312866

Table 35 reports the retained-source MAE gain and its difference from the altered-target MAE gain. Relative to the pre-stabilization router, AIF improves source-target MAE by 0.17505 on average, with positive differences in all 14 settings. The gain difference is 0.00511, so the gain is larger against the retained source than against the corrupted observation. Against the validation-selected route, AIF improves source-target MAE by 0.01528 on average, with positive differences in 11/14 settings and Holm $p = 0.067627$; its tradeoff interval includes zero. These two comparisons assess reliability controls against learned sparse weights and a validation-selected single route, respectively. Leave-one-dataset-out means for the pre-stabilization comparison remain positive in $[0.0542, 0.2221]$, with the largest magnitude on Solar-Energy. Figure 6 separately summarizes route usage by dataset.

Table 36: Dataset summary. The source column indicates the public benchmark family used by the local raw-data bundle; audit annotations are added by our pipeline.

Dataset	Domain and frequency	Source note
ETTh1, ETTh2	Electricity transformer temperature, hourly	ETT benchmark files distributed in common long-term forecasting repositories and benchmark suites.
ETTm1, ETTm2	Electricity transformer temperature, 15-minute	Minute-level ETT counterparts from the same public benchmark family.
Exchange-Rate	Daily foreign-exchange rates	Public exchange-rate benchmark used in multivariate long-horizon forecasting comparisons.
Solar-Energy	Solar power generation, 10-minute	Public Solar-Energy benchmark; our audit treats night-zero and transition regimes as semantic phases.
Weather	Meteorological measurements, 10-minute	Public weather benchmark used as an auxiliary stress setting in this paper.
Electricity	Electricity consumption, hourly	Public electricity-load benchmark reported as a raw-equivalent boundary setting.

Human Audit Reliability. Two independent raters agreed on 44/48 panels (91.67%; Cohen's $\kappa = 0.830688$); this supports rubric reproducibility on the 48-panel sample.

H.3 Boundary and Stress Settings

The auxiliary comparison reports ETTh2 and ETTm2 below the target-clean support criteria, Weather with anchor-only clean-like support, and Electricity under raw-equivalent evaluation.

I Archived Results and Shared Architecture Details

The numerical results and optimization overrides in this section describe the archived offline-audit experiments. The route architecture and mathematical definitions are shared; the causal fixed-budget results and their configurations appear in Appendix E.

This section complements the main comparison with view-specific diagnostics, boundary and stress results, implementation details, and resource measurements. Each view specifies its scoring target.

I.1 Dataset and Baseline Sources

The experiments use public long-horizon forecasting datasets that are standard in recent benchmark suites such as TFB (Qiu et al., 2024). The local bundle preserves the benchmark naming convention and adds the audit metadata and evaluation views used by our protocol. Table 36 gives a compact source summary.

The archived implementations include DLinear, PatchTST, iTransformer, ModernTCN, TimeMixer, and TQNet (Zeng et al., 2023; Nie et al., 2023; Liu et al., 2024; Luo & Wang, 2024; Wang et al., 2024; Lin et al., 2025). TM++-local denotes the archived custom multiscale/expert model previously labelled TimeMixer++. We retain its numerical results under this implementation-specific name. The published TimeMixer++ architecture (Wang et al., 2025) uses multiresolution time imaging, time image decomposition, and multiscale mixing components. The implementation-specific label keeps this distinction explicit. The historical runs use the project configurations, including model-specific training budgets. The archived ranks describe these complete implementations and training protocols.

I.2 Resource Footprint

We profile ETTh1 $96 \rightarrow 96$ on the same RTX 3090 used for the ablation runs. AIF uses 3.29M parameters, 0.17 GiB peak training memory, and 1.48 ms/window inference latency. AIF uses more resources than the lightweight baselines and fewer than TM++-local in parameters, memory, and latency; the profiling run gives 28.85 s/epoch for AIF. The extra cost mainly comes from keeping several lightweight route heads and reliability features active in one model, while the controller itself is small relative to the forecasting heads. These measurements characterize the cost of the seven-route architecture.

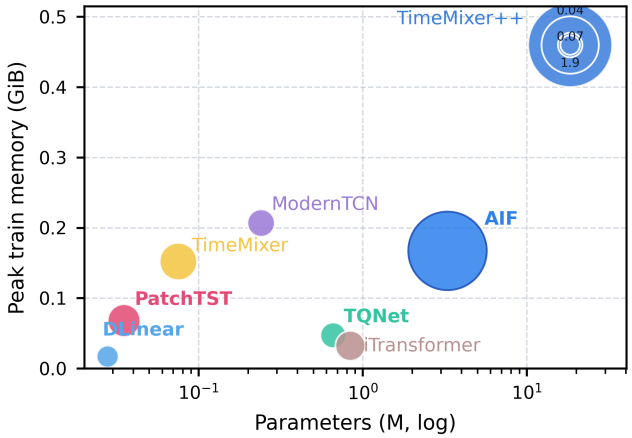


Figure 5: ETTh1 $96 \rightarrow 96$ resource footprint. The x-axis is parameter count, the y-axis is peak training memory, and bubble area encodes per-window inference latency. The archived plot label TimeMixer++ refers to TM++-local.

Table 37: Internal AIF route modules and routing controller. The seven route outputs are produced in normalized space and combined by the support-horizon controller.

Component	Role	Implementation details
Direct linear	Stable direct route	Channel-independent shared linear map from the input window to the forecast horizon with last-value anchoring; the same map is reused for all variables.
Trend-seasonal linear	Multi-scale decomposition route	A set of moving-average decompositions; each decomposition has separate seasonal and trend linear maps, and a learned kernel gate combines the decomposition heads.
Patch encoder	Local subseries route	Multiple patch-encoding branches. Each branch uses a Conv1d patch projection, positional and type embeddings, self-attention encoding, mean/max pooling, a branch gate, and an MLP prediction head.
Temporal convolution	Local temporal route	Pointwise input projection followed by depthwise-separable temporal convolution blocks; the temporal state is mean/max pooled before the MLP head.
Variate mixer	Cross-channel residual route	Per-channel states are projected and then mixed with self-attention over variables plus an FFN. The resulting cross-channel residual is gated and added to the direct linear base prediction.
Frequency-periodic	Periodic structure route	FFT-magnitude tokens encode dominant frequency content, while period-summary features encode mean, scale, and drift over configured periods. Frequency, period, and horizon states are concatenated for the prediction MLP.
Audit residual	Audit-residual route	Uses normalized audit delta and uncertainty series as temporal inputs, together with audit metadata summaries and the horizon embedding. A gated residual head predicts an additive correction to the direct linear base prediction.
Routing controller	Support-horizon controller	A feed-forward controller consumes forecast-time horizon, support, moment, drift, frequency, cross-channel, uncertainty, mask, artifact, and size statistics. Its sparse seven-route weighting is followed by the reliability controls in Appendix I.4.

I.3 AIF Internal Route-Module Design

The reported AIF model uses a single shared set of internal route modules across datasets and horizons. Dataset-specific overrides retain these module definitions. RevIN normalizes the masked input window before route prediction and denormalizes the final routed output. Each route produces a normalized-space forecast, and the routing controller forms the seven-route weighted output using the current input and its audit-derived support and reliability summaries.

I.4 Routing Constraints and Training Prior

The default AIF configuration uses the following routing budgets and thresholds:

$$w_a^{\text{clean}} = 0.04, \quad w_a^{\text{default}} = 0.25, \quad w_{\text{var}}^{\text{lowcorr}} = 0.08, \quad w_{\text{var}}^{\text{default}} = 0.25, \quad c_{\text{var}} = 0.08,$$
$$\tau_r = 0.22, \quad \lambda_P = 0.006.$$

Reliability Statistics. The router uses current-window statistics and input-side metadata. The primary comparison uses prefix-causal endpoint detection and forward repair. In the archived offline setting, event detection and labeling can depend on observations outside the input window; that preprocessing is reported separately. Events contribute prediction metadata only through overlap with the observed input interval; target-window fields are reserved for view construction and support checks. A fixed 12-field allowlist rejects names containing `target`, `future`, or `test`. Mutation tests verify prediction invariance to target-side field changes with the audit held fixed. Router statistics are clipped or normalized, and their gradients are stopped in the reported configuration. Let $M(X) \in [0,1]^{L\times C}$ denote the soft intervention mask, where larger values indicate entries replaced or softened by the audit intervention recipe. The mask-density statistic is

$$m(X) = \frac{1}{LC}\sum_{t=1}^{L}\sum_{c=1}^{C} M_{t,c}.$$

The causal studies use a binary repair mask and $U(X) = 0.25M(X)$. In the archived offline recipe, M may be soft and the uncertainty tensor $U(X) \in [0,1]^{L\times C}$ instead uses event confidence and severity. For an offline repaired span, its entries are assigned

$$U_{t,c} = \mathrm{clip}_{[0,1]}(1 - \mathrm{conf} + 0.25\,\mathrm{sev}),$$

using the audit event confidence conf and severity sev; unrepaired entries have zero uncertainty. The scalar uncertainty statistic is

$$u(X) = \frac{1}{LC}\sum_{t=1}^{L}\sum_{c=1}^{C} U_{t,c}.$$

Let $q_{\mathrm{in}}(X)$ be the input-window contamination score from the fixed audit metadata. AIF uses

$$\rho_{\mathrm{clean}}(X) = 1 - q_{\mathrm{in}}(X)$$

as its input-side clean-support ratio, clipped to $[0,1]$. The artifact score is computed from five input-side audit features,

$$a_{\mathrm{score}}(X) = \frac{1}{4}\,\mathrm{clip}_{[0,4]}\left(\frac{1}{5}\sum_{j=1}^{5}|e_j(X)|\right),$$

In the archived offline schema, e_j denotes input contamination, input event count, maximum event weight, and repairable/unrecoverable overlap indicators. The causal studies use mask-derived proxies: $(e_1,\ldots,e_5) = (b, k, \mathbb{1}[b>0], b, 0)$, where $b = \mathrm{mean}(M)$ and k is the number of mask transitions from 0 to 1 divided by the channel count, treating the preceding mask at the input boundary as zero.

The cross-channel score $r(X)$ is the mean absolute off-diagonal correlation of the RevIN-normalized masked input. If $C > 64$, AIF computes this statistic on 64 evenly spaced channels:

$$r(X) = \frac{1}{C'(C'-1)}\sum_{i\neq j}\left|\frac{1}{L}\sum_{t=1}^{L} z_{t,i}z_{t,j}\right|,$$

where z is the per-channel standardized input and $C' = \min(C, 64)$. If $C < 2$, $r(X) = 0$. The clean-confidence score used by the route cap is therefore

$$\gamma(X) = \rho_{\mathrm{clean}}(X)[1-u(X)][1-m(X)][1-a_{\mathrm{score}}(X)].$$

Route Redistribution. The reported implementation redistributes capped mass before normalizing the final router weights. Removed artifact mass is assigned 55% to the direct linear route and 45% to the decomposition route. Removed variate mass is split equally between these two stable routes. The direct-route floor of 0.1 is applied before the two caps. After redistribution, the spectral activation is

$$v_s = \sigma((c_s - 0.32)/0.08)(1-H_s)(1-a_{\mathrm{score}}), \qquad b_s = 0.07v_s,$$

where c_s is the top-frequency energy share and H_s normalized spectral entropy. A positive shortage $(b_s - \pi_s)_+$ is transferred proportionally from the other routes except the direct route. The implementation adds the full shortage even if the available donor mass is smaller, then clips and renormalizes all weights. Thus the reservation is a pre-normalization adjustment. All coefficients are fixed by the supplied configuration.

Prediction and Auxiliary Losses. For a prediction $\hat{Y}$ and target Y, the supervised loss is

$$\mathcal{L}_{\text{pred}} = \alpha_{\text{MAE}}\|\hat{Y}-Y\|_1 + \alpha_{\text{MSE}}\|\hat{Y}-Y\|_2^2 + \alpha_{\text{H}}\,\text{Huber}_\beta(\hat{Y},Y) + \alpha_{\text{tail}}\mathcal{L}_{\text{tail}} + \alpha_{\text{wtail}}\mathcal{L}_{\text{window-tail}}.$$

Here, $\mathcal{L}_{\text{tail}}$ covers entries above the configured absolute-error quantile. The window-tail term covers windows above the configured per-window MSE quantile. The default uses $\alpha_{\text{MAE}} = 0.7$, $\alpha_{\text{MSE}} = 0.3$, and zero weights for optional Huber and tail terms.

Let z_k denote a normalized head output, with input mean μ, scale s, and RevIN affine parameters w, b. The auxiliary tensor uses $g_k = sz_k + \mu$, while a head in the final prediction scale is $f_k = s(z_k - b)/\max(w,\varepsilon) + \mu$. The archived training objective uses g_k for auxiliary supervision and the full affine inverse for the final mixture. Let $\mathcal{D} = \{\text{direct}, \text{decomp}\}$, and let $\mathcal{T}_2(X)$ be the two route heads with largest detached router weights. The auxiliary route-head loss is

$$\mathcal{L}_{\text{aux}} = \frac{1}{|\mathcal{D}|+1}\left[\sum_{k\in\mathcal{D}}\mathcal{L}_{\text{pred}}(g_k(X),Y) + \frac{1}{|\mathcal{T}_2(X)|}\sum_{k\in\mathcal{T}_2(X)}\|g_k(X)-Y\|_1\right],$$

with default weight $\lambda_{\text{aux}} = 0.05$. For deployed-scale head comparisons, the full inverse is $f_k = (g_k - \mu - sb)/\max(w,\varepsilon) + \mu$, where w, b are the RevIN affine parameters. A weighted reconstruction of these f_k is checked against the deployed mixture prediction. Let π^{dense} be the dense softmax router distribution before top-k sparsification. The router entropy and balance terms are

$$\mathcal{L}_H = \mathbb{E}_{X\in\mathcal{B}}\left[-\sum_{k=1}^{K}\pi_k^{\text{dense}}(X)\log\pi_k^{\text{dense}}(X)\right], \quad \lambda_H = 0.0015,$$

$$\mathcal{L}_B = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{|\mathcal{B}|}\sum_{X\in\mathcal{B}}\pi_k^{\text{dense}}(X) - \frac{1}{K}\right)^2, \quad \lambda_B = 0.001.$$

The entropy term is a small sharpness penalty, while the balance term discourages collapse to a single dense route across a minibatch.

The training-time routing prior used in the reported configuration is

$$\mathcal{L}_P = \mathbb{E}_{X\in\mathcal{B}}\Big[\gamma(X)^2\pi_a(X)^2 + 0.35[1-r(X)]^2\pi_{\text{var}}(X)^2 + 0.50\{v_s(X)[b_s(X)-\pi_s(X)]_+\}^2\Big],$$

Here the weights are the post-stabilization weights used in prediction, and $[z]_+ = \max(z, 0)$. The spectral term penalizes any remaining shortfall after normalization. The prior is soft: it encourages reliability-consistent route usage during training, while the inference-time constraints implement the final route caps described in Section F.2.

Archived Optimization and Overrides. The historical offline runs use AdamW with mixed precision, gradient clipping at 1.0, a 2% linear warmup followed by cosine decay, and validation-based early stopping. Their default uses learning rate 1.2×10^{-4}, weight decay 0.01, batch size 48, at most 45 epochs, and patience 10. ETTh1 uses 10^{-4}, batch size 24, at most 55 epochs, and patience 12. Exchange-Rate uses 8×10^{-5}, batch size 8, gradient clipping at 0.5, at most 70 epochs, and patience 14. Solar-Energy uses 1.8×10^{-4}, batch size 24, and at most 45 epochs; ETTm1 uses the default. Solar-Energy uses $\lambda_P = 0.008$ and Weather uses 0.012; the other reported main-board datasets use the default 0.006. The archived executable configuration provides these overrides. Section 6 and Appendix E specify the causal short-budget end-to-end configuration.

I.5 Aggregate by Evaluation View

Table 38 aggregates the reference seed-0 AIF rows by evaluation view, including raw and intervened settings. It is a view diagnostic distinct from the original three-repetition main board and the five-seed

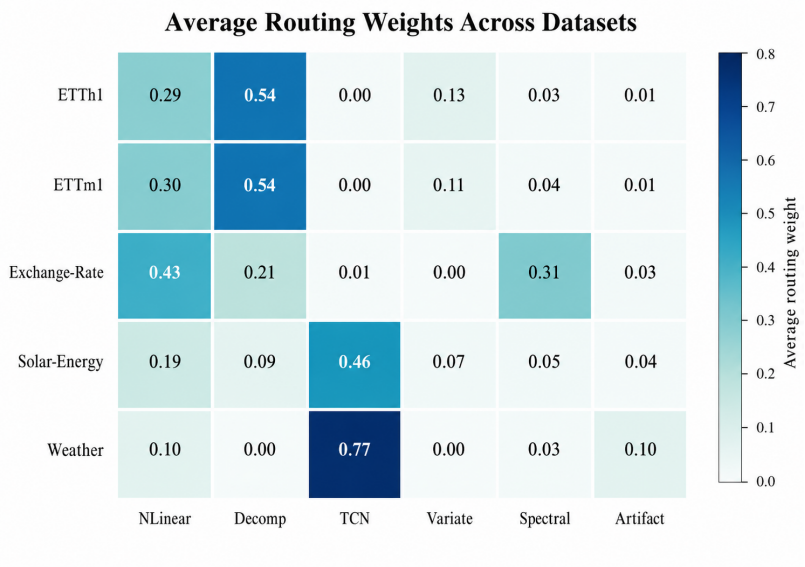


Figure 6: Archived AIF route-weight summary by dataset. The six displayed columns omit the patch route, so row sums may be below one; darker cells indicate larger weights. ETTh1 and ETTm1 emphasize linear and decomposition routes, Exchange-Rate emphasizes the spectral route, and Solar-Energy emphasizes temporal convolution. Weather is a stress setting; the artifact route remains small on main-board datasets.

Table 38: Aggregate AIF results by evaluation view.

View	Rows	Eval windows	Mean MAE	Mean SMAPE (%)	Router entropy
clean-like	12	44,165	0.356	65.598	0.666
balanced	3	21,346	0.268	43.159	0.775
intervened	18	119,764	0.413	86.550	0.618
raw	18	128,860	0.368	70.952	0.610

routing controls. The archived comparison follows the frozen 14-setting scope; Table 32 records the Exchange-Rate low-training-support exception.

Each view describes a different evaluation target. Clean-like rows measure performance on audit-selected windows, and balanced rows preserve semantic phases. Raw rows measure observed-target fidelity, while intervened rows test repair or counterfactual changes. These diagnostics estimate different risks.

I.6 Audit Protocol and Limitations

Protocol. The semi-automatic audit follows a fixed pipeline. First, candidate intervals are generated by deterministic rules in `artifact_rules.yaml`. The implemented detectors scan each numeric variable for zero blocks, exact flat runs, rolling near-constant segments, and suspicious repeated windows, using dataset-specific thresholds. The review stage then writes per-dataset flagged files, an artifact taxonomy, a manual-review template, and, for Solar-Energy, phase annotations separating night-zero, transition, and active regimes.

Second, candidate rows are merged into event-level metadata. Adjacent same-type intervals are merged, and overlapping variable records are clustered into multivariate events. Each event records validity, exploitability, confidence, severity, recoverability, policy, and evaluation view. In ETT datasets, suspicious repeated windows are corrupted and unrecoverable; plateau, zero-block, and near-constant events are usually repairable or mask-only. Weather uses lower-confidence labels for its dense candidate set. For Solar-Energy, night-zero and transition bands are valid but exploitable regimes, whereas active-period zero or constant events are suspicious.

Third, the event table is projected onto forecasting windows. Each window receives input and target contamination scores, event-overlap flags, dominant solar phase labels, and an intervention recipe. Clean-like views retain windows satisfying conservative contamination and target-validity thresholds. Intervened views keep windows whose target does not overlap unrecoverable events and apply local trend interpolation to repairable input spans or masking to mask-only spans. Solar night-zero values are preserved. The balanced view selects equal numbers of windows from each eligible target-phase stratum and reports unweighted errors on that subset; clipped phase weights are stored as metadata. Finally, a support check marks settings as supported, anchor-only, fallback-to-intervention, or intervention-primary using minimum train-window counts and ratios. The frozen main board contains clean-like and semantically valid balanced rows; its documented Exchange-Rate exception falls below the default training-support rule. Other unsupported settings are reported as boundary, stress, or raw-equivalent cases.

Fixed thresholds and support. Table 39 lists the detector lengths from the supplied configuration. Zero and exact-flat tolerances are 10^{-12}. Near-constant detection uses standard-deviation ratio 0.01 with a 10^{-6} floor; repeated-window tolerance uses ratio 0.001 with a 10^{-6} absolute floor. Anchor support requires at least 500 training windows and a training-window ratio of 0.05; conservative support uses 1000 and 0.10, respectively.

Table 39: Fixed audit detector lengths in time steps. NC denotes near-constant detection; repeat denotes repeated-window detection.

Dataset	Zero min.	Flat min.	NC window	NC min.	Repeat window	Repeat min.
ETTh1	3	6	12	12	12	24
ETTh2	3	6	12	12	12	24
ETTm1	12	8	24	24	24	48
ETTm2	12	8	24	24	24	48
Exchange-Rate	3	5	14	14	14	28
Solar-Energy	12	12	36	36	72	144
Weather	12	12	36	36	72	144
Electricity	6	8	24	24	24	48

For the retained clean-like views, the maximum input/target contamination scores are $0.10/0.02$ for ETTh1 and ETTm1 and $0.20/0.10$ for Exchange-Rate. The ETT views also exclude target intervals labeled corrupted. Exchange-Rate retains the original low-training-support exception. Solar

phase balancing requires at least two target-phase strata with at least 100 windows per stratum; the view uses equal-count deterministic sampling. The full dataset-specific predicates are supplied in `view_specs.yaml`.

Scope and Limitations. The protocol audits the specified flat, zero, near-constant, and repeated-pattern rule families using dataset-specific semantics. Its code and configuration regenerate candidate, event, view, manifest, and support summaries from user-provided benchmark files. The companion package supplies source code and configurations; datasets and intermediate outputs are obtained or generated separately. Audit coverage depends on the rule family and dataset semantics, including Solar night zeros and the support conditions in ETT. New domains require metadata or detector validation for their observation processes.

Independent Review. The original audit study used 48 panels labeled independently by two raters under a four-class codebook. The raters agree on 44 panels, for raw agreement of 91.67% and Cohen's $\kappa = 0.830688$ for both the binary defect decision and the four-class decision. The agreement quantifies rubric consistency on this sample. The source-only supplement omits the panel data and rater outputs, and the paper retains the statistic reported in the original study.

I.7 Identifiable Controlled Corruptions

To separate fidelity to an observed artifact from fidelity to a retained source signal, we construct four semi-synthetic corruptions on source-view test windows: plateaus, zero blocks, duplicated segments, and level shifts. Transformations are applied on the training-standardized scale; a zero block sets entries to the training mean in original units. Every fourth channel is affected. The final 24 input steps and first 24 target steps are changed, except that the level shift persists across the full target. The uncorrupted source target remains available for source-target scoring, while the changed target defines corrupted-observation scoring. The source is clean relative to the injected transformation; naturally occurring defects may remain.

The corrupted input is supplied as the raw input and the retained source input as the counterfactual masked input; the changed input span also receives input-side uncertainty and contamination metadata. This construction therefore supplies both the affected span and its uncorrupted source input. It evaluates routing under oracle repair. All compared routing variants receive the same corrupted raw input, counterfactual masked input, route heads, and validation split. Figure 7 illustrates the four prespecified transformations on one held-out channel.

We evaluate all 14 main-board dataset–horizon settings with five frozen seeds. Within each setting, effects are first averaged across the four corruption types; the 14 setting means enter the working-independence summaries for this fixed board. Table 35 reports control-minus-AIF differences, so positive values favor reliability-aware routing. Relative to the matched pre-stabilization router, AIF improves source-target MAE on all 14 settings by 0.17505 on average. Its source-target gain exceeds its altered-target gain by 0.0051 in this oracle-repair setting. Against the validation-selected single route, the mean source-target gain is 0.0153, with positive differences in 11/14 settings. Its source-minus-altered gain difference is 0.00110 (95% CI $[-0.00043, 0.00301]$). Dataset-level pre-stabilization differences are 0.0858, 0.0615, 0.0024, and 0.6181 for ETTh1, ETTm1, Exchange-Rate, and Solar-Energy, respectively. Removing one dataset at a time leaves a positive overall mean between 0.0542 and 0.2221, with the pooled magnitude reflecting the larger Solar-Energy effect.

I.8 Dataset-by-View Summary

Table 40 reports AIF and baseline performance by dataset and evaluation view.

Weather is reported separately as an auxiliary stress case. The archived Solar-Energy view means are 0.268 for balanced and 0.527 for intervened evaluation. These views retain different phase compositions and intervention semantics.

The solar profile motivates phase-aware selection. Night-zero and transition intervals are frequent, structured, and semantically meaningful for solar generation. The balanced view preserves these operating regimes through equal-count stratified selection.

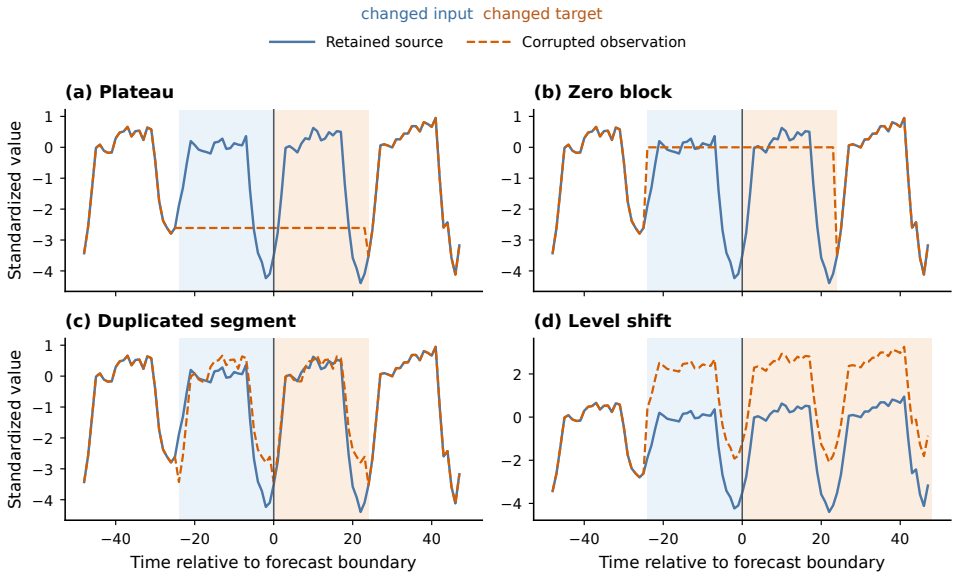


Figure 7: Controlled-corruption generator inspection. Blue curves are the retained source and dashed orange curves are the corrupted observation. Shaded regions mark the changed input and target spans; the vertical line is the forecast boundary. The example illustrates the four specified transformations.

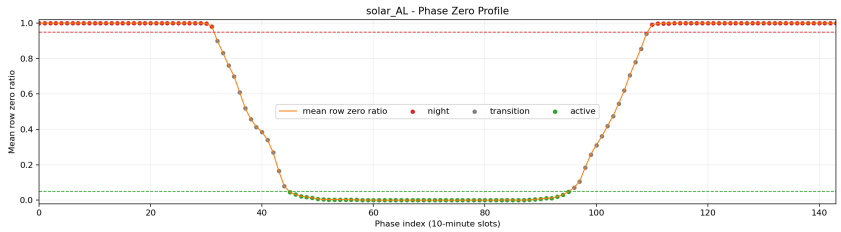


Figure 8: Solar-Energy phase profile used by the audit. The x-axis is the within-day phase index in 10-minute slots and the y-axis is the mean fraction of zero-valued variables. Red, gray, and green markers denote night, transition, and active regimes; dashed lines show the phase thresholds.

Solar-Energy and the ETT boundary settings have different semantics. The ETT audit yields few strict clean windows, while recurring solar night zeros are part of the generation process. Dataset semantics guide the distinction between candidate defects and meaningful operating regimes.

I.9 Boundary Board: ETTh2

ETTh2 is reported as a boundary setting under the strict target-clean support criteria. These rows characterize raw-target forecasting in the boundary setting.

ETTh2 documents forecasting below the strict clean-support criteria. Its raw MAE ranges from 0.856 to 0.888 across the three reported horizons, with spectral or linear routing most frequent. These rows characterize the observed-target boundary setting alongside the main board.

This audited segment illustrates the boundary-setting issue. Local flat or repeated behavior can be predictable from nearby context, allowing a raw-risk model to reproduce it. The raw evaluation measures fidelity to those observed patterns.

The ETT-family gallery shows recurring flat, zero, and repeated patterns that motivate separate reporting of audit-selected and raw windows.

Table 40: Detailed results for AIF and all baselines across the 14 clean-like/phase-balanced settings. Lower is better. Superscripts denote the view: ^{c} clean-like and ^{b} balanced.

(a) MAE

Dataset	Horizon	AIF	DLinear	PatchTST	ModernTCN	iTransformer	TimeMixer	TM++-local	TQNet
ETTh1^{c}	96	0.388	0.402	0.389	0.393	0.399	0.395	0.562	0.389
	192	0.417	0.438	0.416	0.423	0.426	0.435	0.568	0.417
	336	0.418	0.456	0.422	0.431	0.435	0.434	0.571	0.423
	720	0.428	0.475	0.451	0.454	0.446	0.434	0.595	0.434
ETTm1^{c}	96	0.343	0.360	0.360	0.355	0.369	0.353	0.547	0.344
	192	0.366	0.379	0.372	0.383	0.382	0.374	0.551	0.371
	336	0.383	0.401	0.391	0.417	0.405	0.401	0.561	0.394
	720	0.433	0.439	0.427	0.438	0.444	0.433	0.574	0.432
Exchange^{c}	96	0.176	0.240	0.185	0.196	0.207	0.200	0.223	0.191
	192	0.256	0.427	0.268	0.275	0.289	0.305	0.287	0.268
	336	0.347	0.468	0.352	0.357	0.361	0.515	0.362	0.351
Solar^{b}	96	0.246	0.383	0.268	0.339	0.262	0.256	0.700	0.272
	192	0.267	0.401	0.296	0.369	0.306	0.288	0.713	0.308
	336	0.291	0.404	0.307	0.381	0.310	0.301	0.696	0.307

(b) SMAPE (%)

Dataset	Horizon	AIF	DLinear	PatchTST	ModernTCN	iTransformer	TimeMixer	TM++-local	TQNet
ETTh1^{c}	96	70.828	73.976	72.014	71.326	73.044	72.579	104.043	71.862
	192	75.304	79.439	75.572	74.881	76.171	76.855	105.078	75.244
	336	74.743	81.834	76.224	76.188	76.819	78.030	106.352	75.535
	720	80.769	91.729	87.578	86.044	84.251	82.413	112.998	83.081
ETTm1^{c}	96	67.823	71.496	70.489	69.837	71.931	71.595	99.646	69.087
	192	70.221	73.277	71.673	73.605	73.052	72.053	100.542	71.363
	336	71.777	76.562	74.249	76.563	75.671	74.346	102.169	73.818
	720	77.887	81.992	77.992	79.569	79.576	79.115	104.272	78.449
Exchange^{c}	96	31.873	37.389	33.269	34.766	35.708	34.979	37.987	34.388
	192	42.691	56.471	44.016	45.406	46.626	48.008	47.000	44.387
	336	52.092	59.262	52.168	53.173	52.888	64.310	53.817	52.146
Solar^{b}	96	40.955	69.116	45.292	54.930	43.771	43.273	119.983	45.511
	192	43.175	69.935	48.541	58.673	49.307	45.282	117.743	50.940
	336	45.346	67.456	49.533	59.145	48.219	46.818	111.193	48.635

Table 41: Boundary-board results for ETTh2.

Dataset	Horizon	View	Eval windows	MAE	SMAPE (%)	Top-1 route
ETTh2	96	raw	2,689	0.856	115.240	spectral
ETTh2	192	raw	2,593	0.872	116.849	linear
ETTh2	336	raw	2,449	0.888	120.697	spectral

I.10 Excluded and Stress Settings

The auxiliary settings report stress tests, settings below the strict support criteria, and raw-equivalent evaluation.

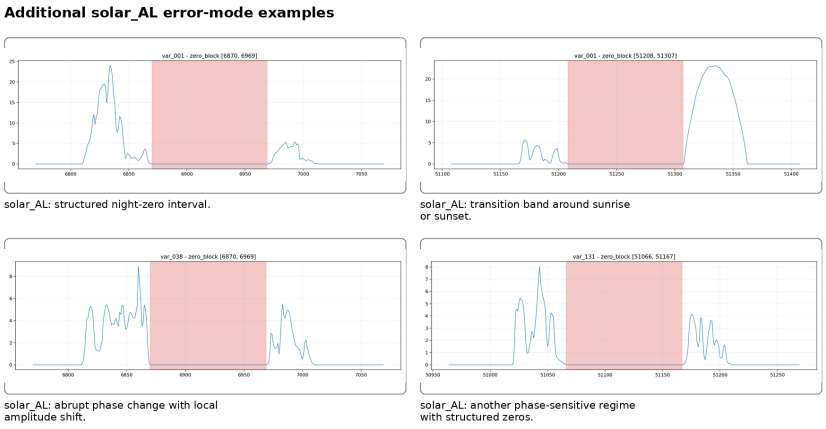


Figure 9: Solar-Energy audit examples. Each panel plots one observed variable against row index; red shading marks the rule-selected interval. These recurring zero-output and transition intervals are legitimate phase regimes, motivating phase-balanced evaluation.

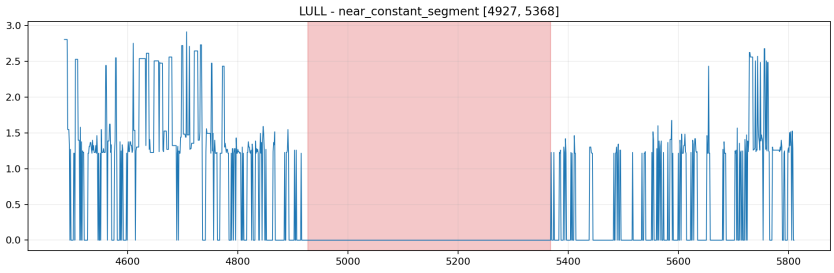


Figure 10: Representative ETTh2 LULL channel against row index. Red shading marks the rule-selected near-constant interval (rows 4927–5368), which is much flatter than the surrounding observations.

Table 42. Excluded and stress settings.

Dataset	Horizon	View	Eval windows	MAE	SMAPE (%)	Reason
Weather	96	clean-like	47	0.321	71.170	anchor-only clean support
Weather	96	intervened	10,349	0.403	135.027	anchor-only stress
Weather	96	raw	10,349	0.340	94.639	anchor-only stress
Weather	192	intervened	10,253	0.441	140.219	anchor-only stress
Weather	192	raw	10,253	0.386	104.357	anchor-only stress
Weather	336	intervened	10,109	0.467	143.529	anchor-only stress
Weather	336	raw	10,109	0.419	109.526	anchor-only stress
Weather	720	intervened	9,725	0.513	148.888	anchor-only stress
Weather	720	raw	9,725	0.465	114.245	anchor-only stress
ETTh2	96	raw	2,689	0.856	115.240	below target-clean support rule
ETTh2	192	raw	2,593	0.872	116.849	below target-clean support rule
ETTh2	336	raw	2,449	0.888	120.697	below target-clean support rule
ETTm2	96	intervened	379	0.343	59.744	fallback auxiliary
ETTm2	96	raw	11,329	1.081	141.569	fallback auxiliary
Electricity	96	natural-clean	5,070	0.417	83.613	natural-clean equals raw
Electricity	96	raw	5,070	0.417	83.613	natural-clean equals raw
Electricity	192	natural-clean	4,974	0.405	79.950	natural-clean equals raw
Electricity	192	raw	4,974	0.405	79.950	natural-clean equals raw
Electricity	336	natural-clean	4,830	0.448	88.850	natural-clean equals raw
Electricity	336	raw	4,830	0.448	88.850	natural-clean equals raw
Electricity	720	natural-clean	4,446	0.433	83.751	natural-clean equals raw
Electricity	720	raw	4,446	0.433	83.751	natural-clean equals raw

The auxiliary settings capture distinct evaluation conditions. Weather provides extensive raw and intervened coverage alongside the clean-like anchor set. ETTh2 and ETTm2 supply raw-target results



Figure 11: ETT-family audit examples. Each panel plots an observed variable against row index; red shading marks the detected near-constant, zero-block, or repeated interval. The examples illustrate both abrupt flat segments and long locally regular regions.

below the strict target-clean support criteria. Electricity has coincident natural-clean and raw rows, which measure observed-target fidelity.

I.11 Worst-Group Risk Examples

Worst-group risk (WGR) reports the highest group MAE and its gap from the mean MAE. This diagnostic tests whether quality effects are concentrated in particular event groups rather than uniformly distributed.

Table 43. Worst-group risk examples.

Dataset	Horizon	View	Worst group	Mean error	WGR	Gap	Total eval windows
ETTh1	96	clean-like	multi_var_plateau_or_zero_block	0.388	0.471	0.083	1,982
ETTh1	192	clean-like	multi_var_plateau_or_zero_block	0.417	0.490	0.072	1,757
ETTh1	336	clean-like	multi_var_plateau_or_zero_block	0.418	0.526	0.108	1,376
ETTh1	720	clean-like	mixed_suspicious_event	0.428	0.475	0.047	635
ETTm1	96	clean-like	local_near_constant_event	0.343	0.357	0.015	9,943

WGR locates error concentration across audited groups. On ETTh1 H336, the worst-group gap is 0.108, much larger than the ETTm1 H96 gap of 0.015. The last column gives the full clean-like evaluation-set size for each dataset–horizon pair. The group gaps locate the concentration of residual errors within each evaluation view.

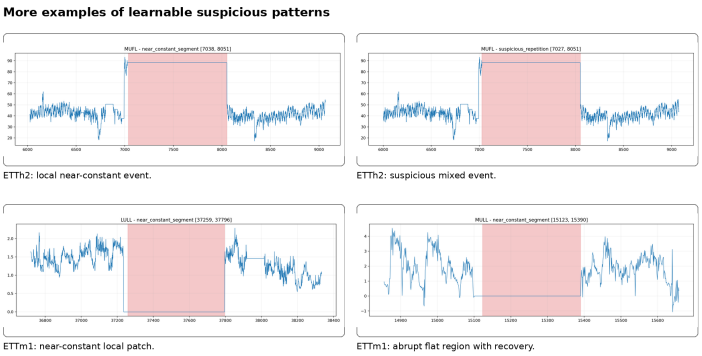


Figure 12: Additional ETT audit examples. Each panel plots one variable against row index; red shading marks the detector-selected interval. The long plateaus and abrupt zero regions recur across neighboring forecast windows and can therefore become predictable structure.

These examples illustrate event types used in group diagnostics: recurring plateaus, near-constant segments, and recovery-like regimes span multiple windows. Group reporting complements the aggregate score by locating residual error concentration.

WGR summarizes regime-specific errors after quality filtering.

I.12 Archived Single-method Ranking Diagnostic

Table 44 preserves the archived ranking output for AIF across evaluation views. The utility ranks backbone identifiers, and these rows have one backbone; both statistics are therefore zero by construction. Cross-model ranking analysis requires a shared multi-method result table.

Table 44. Archived single-method ranking output (structurally zero).

Dataset	Horizon	Compare	Top-1 flip	Mean rank shift
ETTh1	96	raw → clean-like	0	0.0
ETTh1	96	raw → intervened	0	0.0
ETTh1	192	raw → clean-like	0	0.0
ETTh1	192	raw → intervened	0	0.0
ETTh1	336	raw → clean-like	0	0.0
ETTh1	336	raw → intervened	0	0.0
ETTh1	720	raw → clean-like	0	0.0
ETTh1	720	raw → intervened	0	0.0
ETTm1	96	raw → clean-like	0	0.0
ETTm1	96	raw → intervened	0	0.0
ETTm1	192	raw → clean-like	0	0.0
ETTm1	192	raw → intervened	0	0.0
ETTm1	336	raw → clean-like	0	0.0
ETTm1	336	raw → intervened	0	0.0
ETTm1	720	raw → clean-like	0	0.0
ETTm1	720	raw → intervened	0	0.0

This table reports the single-member ranking produced by the archived utility.

J EVIDENCE MAP

Table 45: Evidence for the reconstructed study. Each result keeps its original training and evaluation scope.

Question	Evidence and scope
Scoring-target sensitivity	Section 5.1, Table 1, Figure 3; post-hoc fixed-prediction analyses in Appendix C.
Repair access	Section 5.2, Table 2; all rules and targets in Appendix B.
Quality and control increments	Table 4; prespecified complete horizon-96 comparisons and separate common-seven extension.
Natural scoring views	Section 6, Appendix D; same saved end-to-end prediction errors, explicit boundary and view coverage.
End-to-end forecasting context	Table 5, Appendix E; five methods, 32 settings, three seeds, shared short budget.
Architectural and historical context	Appendix F, Appendix H; separately scoped original joint/offline results.