GATECORRECT: RECURSIVE SELF-IMPROVEMENT FOR FROZEN TIME-SERIES FORECASTERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Even after a time-series forecaster is frozen, its errors can contain patterns predictable from the input history. The unresolved problem is sequential: accepting one correction changes the residual that supplies supervision for the next, so a basis diagnosed for the earlier target may no longer describe the useful correction directions. GateCorrect formulates this process as recursive self-improvement (RSI): each round re-diagnoses the current residual, allocates a new correction unit, and admits it only when validation supports the update. A directional path uses a low-dimensional, predictability-screened output span for uncapped amplitudes; a bounded, level-aware path fits the remaining error while retaining window-level information. Across three forecasters and seven datasets, one unit reduces test MSE by 3.74% over 252 runs. In a separate 216-run matched comparison, RSI raises mean reduction from 2.18% with one unit to 2.81%; later units improve 32 of the 37 runs that accept them. GateCorrect therefore makes residual diagnosis and depth selection part of the learning method, rather than treating successive residual fitting as a fixed stack.

1 INTRODUCTION

Historical inputs and forecast errors can train a correction for a frozen time-series forecaster without retraining its backbone (Liang et al., 2026; Xie et al., 2025). Each accepted correction, however, changes the residual that supplies supervision for the next: it can remove one history-predictable pattern while leaving another. The initial diagnosis cannot determine later updates. Each round asks what history can still predict and whether validation supports another correction.

We study that question as recursive self-improvement (RSI) in time-series forecasting. Its learning principle is that the residual is also the state of the learner: after an accepted unit, the system must re-diagnose which output directions remain predictable before assigning the next unit's capacity. This differs from ordinary boosting or repeated residual fitting with a fixed weak-learner family. Those procedures reuse a prescribed hypothesis space; RSI changes the hypothesis space itself because the target used to construct and screen the output basis has changed. The gate then decides whether the new state warrants another unit at all.

Existing adapters constrain their output updates, while corrective feature search revises the input representation (Liang et al., 2026; Xie et al., 2025; Wang et al., 2026). GateCorrect makes the state-dependent output diagnosis explicit. For a multivariate forecast, predictable error may occupy only a few joint horizon–channel directions. A first unit can remove those directions and expose another predictable component, or leave only variation that should not be fitted. RSI therefore couples three decisions that a single residual fit leaves separate: what the current residual makes learnable, where uncapped capacity may act, and whether the next task should be formed.

Residual magnitude cannot identify that span. Under squared error, only the part predictable from history is recoverable: a high-variance output mode may be poorly predictable, while a lower-variance mode may have reliably predictable amplitudes. Variance-ranked modes alone cannot distinguish the two. Free correction over the full output risks fitting unpredictable variation, whereas a uniform coordinate-wise cap can suppress a large but learnable amplitude. This calls for restricting *where* free correction acts, using historical predictability to select a low-dimensional output span.

Two measurements show the residual structure that a correction unit must address (Figure 1). On Traffic/DLinear at horizon 96, a ridge probe predicts residual-direction amplitudes from normalized history with held-out R^2 up to 0.87, supporting predictability-based direction screening. On Weather/iTransformer at horizon 192, 79% of learned correction energy lies in a component constant across the forecast horizon within each channel, yet its offset varies between windows. A fixed bias

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.

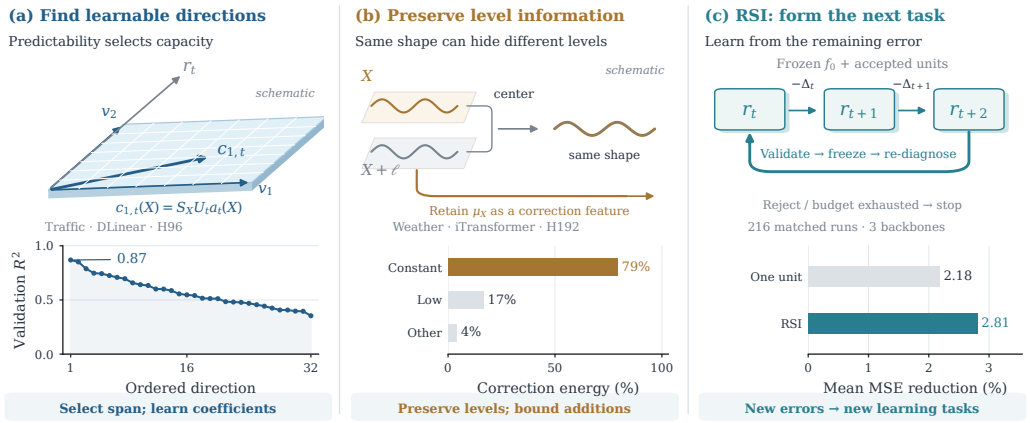

Figure 1: **RSI turns remaining errors into new learning tasks.** Top: schematic views of directional capacity, level-preserving features, and successive residual tasks. Bottom: (a) held-out direction scores; (b) correction energy by frequency band, with 79% horizon-constant; (c) mean MSE reduction relative to the frozen backbone over 216 matched runs, from 2.18% with one unit to 2.81% with RSI.

cannot follow that variation, while window-wise normalization removes the input mean and scale. The second pattern calls for a bounded path that retains raw window levels.

GateCorrect forms each task with two sequential paths. A ridge probe predicts residuals from history; SVD of those predictions proposes output directions, which the directional path screens before fitting their amplitudes without a coordinate-wise cap. The level-aware path then fits the unexplained residual using the history and prediction entering the round, with bounded output and preserved window levels. In the recursive protocol, a validation gate admits a complete unit only when a lower bound on its paired incremental MSE reduction is positive; the unit-level protocol uses a separate incremental gate for the level-aware addition. After admission, the unit is frozen into the predictor and the updated residual is diagnosed anew. Historical predictability allocates capacity; validation controls depth.

The experiments first ask whether the two-path unit can reduce a frozen predictor's error. Across three forecasters and seven datasets, one unit reduces test MSE by 3.74% on the 252-run grid. The directional path alone yields a 3.29% mean reduction, with individual cells above 20% on Electricity and Traffic. The level-aware path is admitted in 111 runs across five datasets and adds 1.02 percentage points on those runs, or 0.45 points across the full grid. On Weather, the largest dataset-level addition, mean reduction rises from 3.9% to 6.2%.

The next question is whether the updated residual merits another unit. In a separate 216-run matched comparison, RSI raises mean reduction from 2.18% with one unit to 2.81%. Validation admits further units in 37 runs; 32 improve over their matched one-unit predictor, with a mean additional reduction of 3.66 percentage points among those 37. Electricity and Traffic account for 30 of the 32 improvements. The other 179 runs retain a predictor of depth at most one. These results distinguish the benefit of a correction unit from the additional benefit of forming later residual tasks.

Contributions.

- We formulate RSI as state-dependent residual learning: each accepted unit changes the target, the output subspace to diagnose, and the next admission decision.
- We implement each RSI round with a screened directional path for uncapped low-dimensional amplitudes, a bounded level-aware path for the remainder, and a validation gate that controls both unit admission and depth.
- Unit-level, recursive, and composition experiments isolate correction, re-diagnosis, and input complementarity; the backbone stays frozen throughout.

2 Related Work

Post-hoc correction. Residual regression, forecast combination, and stacked generalization refine an existing predictor (Bates & Granger, 1969; Granger & Ramanathan, 1984; Wolpert, 1992); boosting fits successive learners to remaining errors (Friedman, 2001), while reusing a prescribed weak-learner

space. GateCorrect treats each updated residual as a new diagnostic state: it rebuilds the output space in which the next correction may act, rather than repeatedly fitting the same residual hypothesis class. Table 1 positions four post-hoc mechanisms by diagnostic target, output policy, sequential dynamics, and admission. δ-Adapter learns bounded input and output edits and can select input features (Liang et al., 2026). CRC encodes directional interactions between variables and combines a ridge correction with a constrained nonlinear addition (Xie et al., 2025). Crafter searches residual-informed input features across iterations (Wang et al., 2026). GateCorrect screens forecast-output modes for historical predictability, assigns free amplitude within their span, and fits the level-aware remainder with bounded output. Validation admits complete units and renews the residual task.

Table 1: Design choices in post-hoc forecast correction. Diagnostic target identifies the representation a method selects or structures; output policy describes its correction form and bound. Admission refers to deploying a fitted correction, beyond training-time early stopping. GateCorrect selects predictable output modes anew after each admitted unit.

Method	Diagnostic target	Output policy	Sequential dynamics	Admission test
δ-Adapter	Input mask	Bounded edits	Single pass	None
CRC	Variable links	Ridge + clipped	Single pass	Safety tests
Crafter	Input features	Tree head	Iterative search	Validation
GateCorrect	**Output modes**	**Free span + bound**	**Residual recursion**	**Paired-loss bound**

Component-wise correction, instance selection, and online adaptation address related forecasting settings (Appendix H). GateCorrect fits and admits its units before test.

Residual targets and training objectives. D3U (Li et al., 2025), FALDA (Wang et al., 2025), and TSDiff (Kollovieh et al., 2023) model residual uncertainty or diffusion-based forecast refinements. For point prediction under squared error, the relevant target is the conditional-mean residual (Section 3.1), which GateCorrect models.

3 METHOD

3.1 THE REMAINING LEARNING OPPORTUNITY

Each sliding-window example pairs an input history $X \in \mathbb{R}^{L\times C}$ of L observed steps across C channels with a target $Y \in \mathbb{R}^{F\times C}$ over the next F steps. All series are channel-wise standardized using training-set statistics. The frozen forecaster produces $\hat{Y}(X)$; we flatten its residual $r = Y - \hat{Y}(X)$ into $D = FC$ coordinates and evaluate coordinate-averaged MSE, $\|r\|^2/D$.

Write $m(X) = \mathbb{E}[r \mid X]$ for the conditional-mean residual predictable from history and $\varepsilon = r - m(X)$ for the zero-conditional-mean remainder, so that $\mathbb{E}[\varepsilon \mid X] = 0$ and

$$r = m(X) + \varepsilon, \qquad \mathbb{E}\|r\|^2 = \mathbb{E}\|m(X)\|^2 + \mathbb{E}\|\varepsilon\|^2. \tag{1}$$

For any deterministic correction $c(X)$ with finite second moment, conditional orthogonality gives $\mathbb{E}\|r - c(X)\|^2 = \mathbb{E}\|\varepsilon\|^2 + \mathbb{E}\|m(X) - c(X)\|^2$. The ε term is fixed across such corrections: reducing expected error requires approximating $m(X)$. A correction method must therefore diagnose predictability within its available output family before deciding where unrestricted capacity should act.

At round t, a candidate correction unit $\Delta_t(X)$ combines two sequential paths for the current residual r_t. The directional path screens forecast-output directions for predictability from history; the level-aware path fits the residual left by that correction. Let S_X be the diagonal map that repeats each input channel's window standard deviation over the forecast horizon, and define $\tilde{r}_t(X) = S_X^{-1} r_t(X)$. The directional path selects at most K directions in this normalized space (Section 3.3). Let $U_t \in \mathbb{R}^{D\times k_t}$ be their orthonormal basis, $k_t \le K$, and $\Pi_t = U_t U_t^\top$. Its output in the original scale is $c_{1,t}(X) = S_X U_t a_t(X)$, where $a_t(X)$ comprises network coefficients and validation-selected scaling. Thus the physical directions form the window-dependent image $S_X \operatorname{span}(U_t)$; in normalized coordinates its output lies in $\operatorname{span}(U_t)$. The predictable residual component projected onto that span is $\Pi_t \tilde{m}_t(X)$, where $\tilde{m}_t(X) = \mathbb{E}[S_X^{-1} r_t \mid X]$. The level-aware path fits $r_t - c_{1,t}(X)$ using the input window and prediction entering the round, standardized by training-set statistics to preserve window mean and scale. The complete candidate is

$$\Delta_t(X) = S_X U_t a_t(X) + b_{2,t} c_{2,t}(X), \qquad U_t^\top U_t = I_{k_t}, \quad k_t \le K, \qquad \|c_{2,t}(X)\|_\infty \le \delta, \tag{2}$$

where the second path has a coordinate-wise bound δ and $b_{2,t} \in \{0, 0.25, 0.5, 0.75, 1\}$ is selected on validation; $b_{2,t} = 0$ switches it off (Section 3.4).

After an accepted unit, the predictable residual becomes $m_{t+1}(X) = m_t(X) - \Delta_t(X)$, where $m_t(X) = \mathbb{E}[r_t \mid X]$. The residual is therefore the state that defines the next learning problem. Formally, with $\mathcal{S}_t = \mathcal{D}(X, r_t)$, the next directional span is rebuilt as

$$\mathcal{S}_{t+1} = \mathcal{D}(X, r_{t+1}), \qquad \mathcal{D}(X, r_{t+1}) \text{ need not equal } \mathcal{D}(X, r_t), \tag{3}$$

where $\mathcal{D}$ denotes probe, candidate construction, and held-out screening. Reusing $\mathcal{S}_t$ gives the fixed-basis control; it adds another fit but removes the state-dependent diagnosis that defines RSI. This distinction matters under a finite per-round budget. The bounded path can supply corrections outside the retained span, but it does not provide the same unrestricted amplitude allocation as a newly screened direction.

3.2 Forming and accepting successive learning tasks

Let $f_0 = \hat{Y}$; after t accepted units, f_t is the current predictor and $r_t = Y - f_t(X)$ its residual. The ridge probe screens candidate amplitudes in r_t using input history, the directional network fits their coefficients, and the level-aware network fits $r_t - c_{1,t}(X)$. Both correction networks use $f_t(X)$ as their current prediction feature. The fitted unit Δ_t of Eq. (2) gives the candidate predictor

$$f_t^{\text{cand}}(X) = f_t(X) + \Delta_t(X). \tag{4}$$

Both paths and their scales are fixed before testing the complete unit for acceptance. Setting $b_{2,t} = 0$ omits the level-aware term; $k_t = 0$ omits the directional term.

To separate candidate development from complete-unit acceptance, we split validation-window origins chronologically into A for diagnosis and candidate selection and B for acceptance, with an embargo preventing target overlap across the boundary (Angelopoulos et al., 2022). Training windows fit network weights; A supplies early stopping, direction screening, coefficient calibration, and scale selection. We divide B into at most three disjoint groups of origins, each containing at least $2F$ origins, and use a fresh group for each attempted unit. A circular moving-block bootstrap with block length F resamples paired MSE reductions on that group, defined as the current predictor's loss minus the candidate predictor's loss (positive means improvement). It admits the complete unit only when $\text{LB}_\alpha(\Delta\mathcal{L}_{\text{val}}) > 0$, where LB_α is the empirical α quantile of the resampled mean reduction. Appendix C.1 gives the quantile level and group availability at each depth.

After acceptance, freeze the fitted unit and update

$$f_{t+1} = f_t + \Delta_t, \qquad r_{t+1} = r_t - \Delta_t, \qquad m_{t+1}(X) = m_t(X) - \Delta_t(X). \tag{5}$$

The updated residual is the next supervised target. GateCorrect refits the probe to this target before selecting another directional budget. An empty screen stops; otherwise, the continuation rule stops if the maximum screened score is below 0.05, or if it is below 0.10 with fewer than eight screened directions (Appendix C.1); rejection of a complete unit or exhaustion of the round budget also stops the chain. At deployment, the frozen chain uses observed history alone, without labels or refitting. Figure 2 shows the unit and its reuse on the updated residual; Sections 3.3–3.4 detail its paths.

3.3 The directional path: uncapped capacity on screened directions

Within a unit, let $\hat{Y} = f_t(X)$ be the current prediction and $r = r_t$ its target residual. The directional path screens output directions for predictability from history before fitting their amplitudes. For each input channel, let μ_X and σ_X denote the mean and standard deviation of its input window. The path centers and scales input and prediction with these statistics but scales the residual without subtracting the window mean: $X_{\text{norm}} = (X - \mu_X)/\sigma_X$, $\hat{Y}_{\text{norm}} = (\hat{Y} - \mu_X)/\sigma_X$, and $r_{\text{norm}} = r/\sigma_X$. These features describe within-window variation on a common scale (Kim et al., 2022); the level-aware path retains window levels for the remainder.

A ridge probe (Hoerl & Kennard, 1970) predicts r_{norm} from flattened X_{norm}. Singular value decomposition (SVD) of its centered training predictions yields at most $K = 32$ orthonormal candidate output directions. The candidates reflect residual variation predicted from history; the validation screen below tests whether their amplitudes remain predictable on held-out windows. We fit the ridge probe and any PCA reduction of its input features on training data; Appendix A specifies the PCA rank and ridge penalty selection.

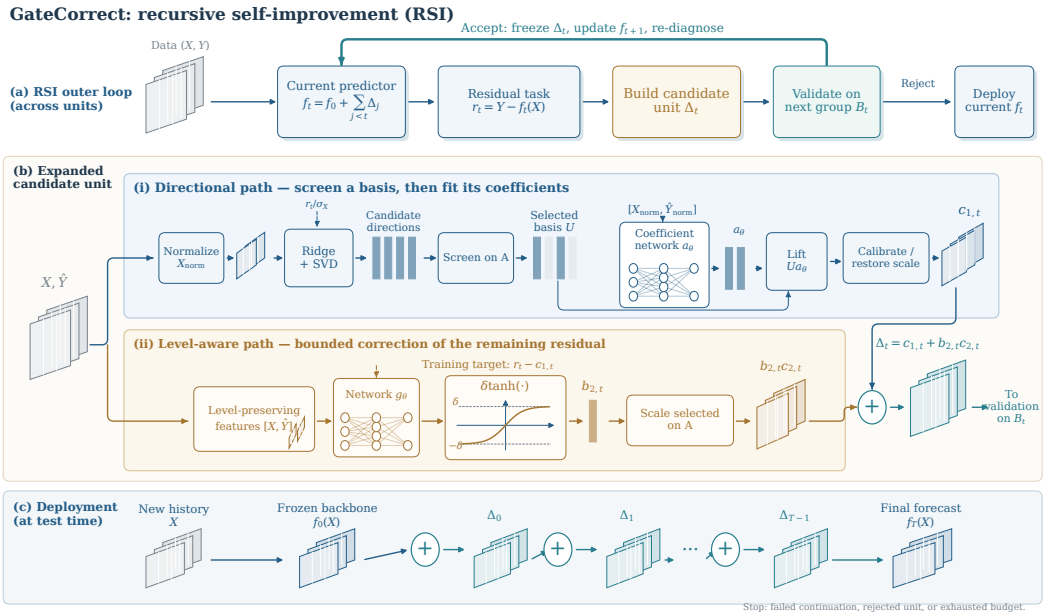

Figure 2: **RSI learns a correction unit, then renews its residual task.** (a) The accepted chain defines the current target. (b) Ridge prediction and SVD propose directions; screening on A selects the basis for coefficient regression. A level-aware network fits the remaining residual with bounded outputs. The complete unit is accepted on the next B group, denoted B_t. (c) Deployment evaluates the frozen backbone and accepted units.

For candidate u_k, define the validation residual amplitude $t_{ik} = (r_{i,\text{norm}} - \bar{r}_{\text{train}})^\top u_k$ and its ridge prediction $\hat{t}_{ik} = (\hat{r}_{i,\text{norm}} - \bar{r}_{\text{train}})^\top u_k$, where $\hat{r}_{i,\text{norm}}$ is the probe prediction and $\bar{r}_{\text{train}}$ the training residual mean. Predictive accuracy against the training-mean reference is

$$R_k^2 = 1 - \frac{\sum_i (t_{ik} - \hat{t}_{ik})^2}{\sum_i t_{ik}^2}. \tag{6}$$

A positive score improves on that reference. The screen compares R_k^2 with scores from circularly shifting the validation residual-window sequence relative to inputs; both circular distances from the original alignment must be at least F windows. Benjamini–Hochberg (BH) selection (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001) applies nominal false discovery rate (FDR) target $q = 0.1$ to the shift-based p-values. FDR is the expected fraction of null directions among those selected. Appendix D.2 gives the statistic and abstention rule.

The selected directions form the columns of U. A regression network a_θ predicts their coefficients from $[X_{\text{norm}}, \hat{Y}_{\text{norm}}]$, where brackets denote concatenation. We train it with FreDF's frequency-domain L_1 loss on the resulting output-space residual prediction (Wang et al., 2024). Validation fits one least-squares coefficient γ_k per direction and chooses an overall scale $b_1 \in \{0, 0.25, 0.5, 0.75, 1, 1.25, 1.5\}$. The resulting correction is

$$c_1(X) = b_1 S_X U \ \text{diag}(\gamma)\, a_\theta([X_{\text{norm}}, \hat{Y}_{\text{norm}}]), \tag{7}$$

where S_X repeats each channel's σ_X across forecast positions, restoring the evaluation scale. Screening selects residual patterns; a_θ predicts window-dependent coefficients, and validation calibrates them with γ and b_1. This path has no coordinate-wise size cap; $c_1 = 0$ when U is empty or $b_1 = 0$.

3.4 The level-aware path: bounded correction of the remainder

After the directional path fits amplitudes in its screened span, predictable error can remain, including within that span. The level-aware path fits this remainder $r - c_1$ by minimizing MSE across every output coordinate. Its inputs are the history and current prediction, standardized with training-set statistics rather than each window's statistics. This preserves between-window level variation needed

to predict window-dependent offsets (Appendix E.2). We bound every output coordinate of its regression network g_θ:

$$c_2(X) \;=\; \delta \cdot \tanh\big(g_\theta([X, \hat{Y}(X)])\big), \qquad \delta = 0.075 \text{ fixed}, \tag{8}$$

so each coordinate lies in $(-\delta, \delta)$ on the standardized scale. Appendix E.2 compares cap settings and their observed MSE effects. In RSI, early stopping and the scale $b_2 \in \{0, 0.25, 0.5, 0.75, 1\}$ use A; in unit-level studies V_1 selects them and disjoint V_2 tests admission.

The incremental gate asks whether the fitted level-aware addition reduces MSE after the directional correction. For window i, write $r_i^{(1)} = r_i - c_1(X_i)$ and $c_{2i} = c_2(X_i)$. The paired MSE change is

$$d_i = \frac{\|r_i^{(1)} - b_2 c_{2i}\|^2 - \|r_i^{(1)}\|^2}{D} = \frac{b_2^2 \|c_{2i}\|^2 - 2b_2 \langle r_i^{(1)}, c_{2i} \rangle}{D}. \tag{9}$$

The addition improves MSE when alignment with the remaining error outweighs its squared magnitude. The cap bounds that cost by $b_2^2 \|c_{2i}\|^2 / D \le b_2^2 \delta^2$; the paired-loss test evaluates whether alignment exceeds the cost across windows.

The unit-level analysis writes this comparison as a loss increase d_i, whereas the recursive protocol uses its negation, the loss reduction $\Delta\mathcal{L}_{\text{val}}$. In the unit-level studies, a paired block bootstrap on V_2 (Künsch, 1989; Politis & Romano, 1994) admits the level-aware path only when $p < 0.05$ and the 95% interval's upper endpoint is below zero. The latter selects the nominal 2.5% improvement tail of a two-sided comparison. Appendix D examines admission behavior under this allocation. RSI follows a distinct protocol: it selects candidates on A and tests each complete unit on its assigned B group (Section 3.2).

The level-aware path retains window-level information while fitting the remaining residual, so its output may overlap directions selected by the directional path. The unit-level experiments measure this overlap and compare the complete unit with the directional path alone.

4 Experiments

4.1 Evaluation setup

We evaluate post-hoc correction on three frozen forecasters spanning linear decomposition and Transformer architectures (Lim & Zohren, 2021): DLinear (Zeng et al., 2023), PatchTST (Nie et al., 2023), and iTransformer (Liu et al., 2024). We train them with the benchmark library's reference recipes where available (Wu et al., 2023); Appendix A reports configurations and frozen-model MSE. The seven datasets are ETTh1, ETTh2, ETTm1, and ETTm2 (electricity transformer temperature), Weather, Electricity (consumption), and Traffic (road occupancy). Each input contains 96 steps; horizons are 96, 192, 336, and 720 with the same correction recipe.

A *cell* is one model–dataset–horizon combination, with three correction seeds sharing its frozen checkpoint and test windows. The unit-level study covers the full 252-run grid. A separate matched recursive comparison includes 216 paired runs that meet its block requirement. The 36 unavailable runs are the three seeds for each $F = 720$ cell on ETTh1, ETTh2, Electricity, and Traffic; all four horizons remain represented on ETTm1, ETTm2, and Weather. The one-unit ablation shares RSI's first proposal and gate on the same 216 runs. Appendix C.1 details their coverage.

The primary metric is equal-weight mean percentage MSE change from each frozen predictor (negative is better); MAE is secondary. A change above 10^{-4} percentage points counts as harm.

The 252-run unit-level studies evaluate the two-path design, direction screening, and level-aware admission under the incremental protocol in Appendix D. This full-grid study is separate from the 216-run matched recursive comparison. The level-aware cap is fixed at $\delta = 0.075$ across runs, and the recursive continuation rule applies prespecified thresholds to validation scores (Appendix C.1). For cellwise inference, we average paired loss differences across three seeds, use 10,000 moving-block bootstrap resamples with horizon-length blocks, and adjust the 84 comparisons with BH at $q = 0.05$ (Diebold & Mariano, 1995; Giacomini & White, 2006).

Each correction unit fits both networks on a single 40 GB GPU. The optimization keeps the expensive frozen forecaster outside the correction training loop: the directional path predicts at most $K = 32$ basis coefficients, while the full-coordinate level-aware network is restricted by the coordinate-wise cap and trained as a separate remainder fit. PCA reduces the history feature dimension where needed, validation selects scale and stopping, and accepted units are frozen before the next round. Thus high output dimension changes the cost of the bounded remainder fit, not the dimension of the uncapped search space; rejected candidates add no deployment cost. Fitting, screening, admission,

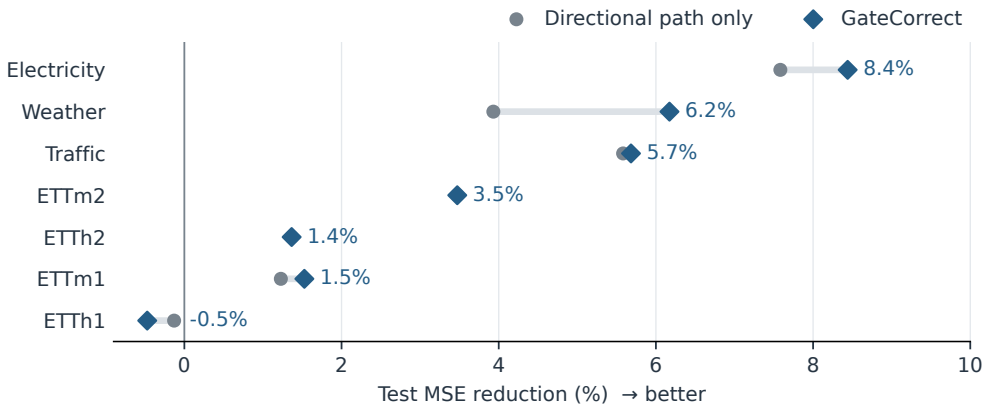

Figure 3: Sequential MSE gains. Each row connects the directional path alone (gray) to the complete unit (blue), averaged over 36 runs per dataset under the unit-level protocol. Movement to the right measures the additional reduction from the level-aware path. Negative values indicate an increase in test MSE relative to the frozen backbone.

and continuation decisions are completed before test-time deployment. Test windows provide no selection feedback; deployment evaluates the frozen predictor and accepted units sequentially.

Why two evaluation protocols. The full 252-run unit-level protocol asks whether one complete correction unit works across the benchmark grid and whether its two paths contribute. It uses the incremental level-aware gate because that protocol isolates the addition after the directional path. The 216-run recursive protocol asks a separate question: after an accepted unit changes the residual, can a fresh task be diagnosed and admitted without reusing the same acceptance windows? It therefore partitions validation into chronological A/B roles and assigns a fresh block of B to each attempted round. The protocols are complementary: the first estimates unit-level coverage and path composition; the second measures state-dependent continuation and RSI depth. They share the frozen checkpoints, test windows, correction architecture, and test-time composition, but their gates are not interchanged.

4.2 Establishing the effectiveness of a correction unit

Across 252 runs under the incremental level-aware protocol, one correction unit reduces test MSE by 3.74% on average. The directional path alone yields a 3.29% mean reduction (Figure 3; dataset–model results in Table 7). The level-aware gate opens in 111 runs. On these runs, the level-aware path adds 1.02 percentage points of mean reduction; averaged over all 252 runs, its contribution is 0.45 points. Test MSE does not rise in 93 of its 111 admitted runs. Weather has the largest dataset-level addition: reduction rises from 3.9% with the directional path to 6.2% with the full unit. For Weather/iTransformer, the corresponding reductions are 1.62% and 5.13%, so the level-aware path supplies most of the full unit's gain in this configuration. Across the full grid, 54 of 84 cells improve significantly and two worsen significantly after FDR correction (Appendix B).

A fixed per-channel bias, offered to the same gate, is rejected in all 36 Weather runs, whereas the input-dependent network supplies the additional gain (Appendix E.2). The two paths can overlap: across the 24 Weather runs with an open level-aware gate, 47% of that path's correction energy lies in the screened directional span. The candidate-rank control is reported in Appendix E.2; together, these tests distinguish level-aware fitting from a fixed bias or a wider directional candidate set.

The unit-level evidence shows what one correction can learn. Because that correction changes the residual, its result also creates a new question: when does the remaining error merit another unit?

4.3 Improving through successive residual tasks

We isolate the effect of further residual tasks in a separate 216-run matched comparison: RSI and the one-unit ablation share their first proposal and acceptance decision (Table 2).

Across the 216 paired runs, recursive correction raises mean test-MSE reduction relative to the frozen predictors from 2.18% with one unit to 2.81% (Table 2). The 0.63-percentage-point increment measures the value of allowing further residual tasks. The increments are 1.24 points for DLinear, 0.57 for PatchTST, and 0.07 for iTransformer.

Table 2: Test-MSE change relative to the frozen model (%, negative is better; mean ± standard deviation over the runs of each dataset–model pair, a dispersion summary across seeds and horizons). RSI and its one-unit counterpart share the first proposal and gate, so their paired difference (lower panel, mean ± SD of paired changes) isolates the gain from further residual tasks. The 216-run coverage excludes $F = 720$ on ETTh1, ETTh2, Electricity, and Traffic under the block requirement.

	DLinear		PatchTST		iTransformer	
Dataset	RSI	One unit	RSI	One unit	RSI	One unit
ETTh1	$+0.22 \pm 0.36$	$+0.22 \pm 0.36$	$+0.12 \pm 0.25$	$+0.12 \pm 0.25$	$+0.07 \pm 0.44$	$+0.07 \pm 0.44$
ETTh2	0.00 ± 0.00	0.00 ± 0.00	-0.12 ± 0.37	-0.12 ± 0.37	-0.17 ± 0.26	-0.17 ± 0.26
ETTm1	-0.12 ± 1.40	-0.22 ± 1.49	-0.07 ± 0.50	-0.07 ± 0.50	-1.24 ± 0.83	-1.19 ± 0.78
ETTm2	$+0.03 \pm 1.92$	$+0.03 \pm 1.92$	-1.53 ± 0.56	-1.53 ± 0.56	-1.69 ± 0.58	-1.69 ± 0.58
Weather	-6.15 ± 6.10	-6.15 ± 6.10	-4.76 ± 1.64	-4.76 ± 1.64	-3.91 ± 2.11	-3.75 ± 1.91
Electricity	-13.85 ± 5.39	-11.00 ± 2.78	-10.96 ± 3.24	-7.44 ± 0.44	-1.56 ± 0.85	-1.40 ± 0.78
Traffic	-11.56 ± 8.29	-4.33 ± 0.95	-3.22 ± 2.12	-2.19 ± 0.99	-0.38 ± 0.83	-0.28 ± 0.59
Additional MSE change from recursion (percentage points)						
ETTh1	0.00 ± 0.00		0.00 ± 0.00		0.00 ± 0.00	
ETTh2	0.00 ± 0.00		0.00 ± 0.00		0.00 ± 0.00	
ETTm1	$+0.10 \pm 0.17$		0.00 ± 0.00		-0.05 ± 0.17	
ETTm2	0.00 ± 0.00		0.00 ± 0.00		0.00 ± 0.00	
Weather	0.00 ± 0.00		0.00 ± 0.00		-0.17 ± 0.57	
Electricity	-2.85 ± 3.55		-3.53 ± 3.06		-0.17 ± 0.34	
Traffic	-7.23 ± 7.39		-1.04 ± 1.35		-0.10 ± 0.24	
Average	-1.24 ± 3.70		-0.57 ± 1.63		-0.07 ± 0.28	

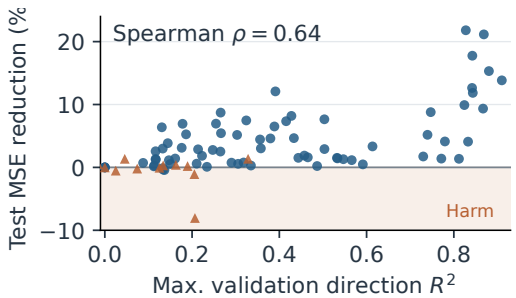

Figure 4: Directional predictability and realized gain across 84 unit-level cells. Larger pre-corrector validation R^2 is associated with greater test-MSE reduction ($\rho = 0.64$). Triangles mark ETTh1, and shading indicates higher test MSE.

Validation-gated continuation produces depths zero, one, two, and three in 55, 124, 22, and 15 of the 216 paired runs, respectively. Among the 37 validation-selected runs that continue beyond one unit, RSI adds 3.66 percentage points of mean MSE reduction over the matched one-unit predictor: 32 improve and five worsen. The other 179 runs retain identical zero- or one-unit predictors. This selectivity is part of the result: the gate leaves most predictions at the depth justified by the first task, while concentrating additional training on the smaller set of residuals that earns another task. The corpus-wide 0.63-point increment and the 3.66-point increment among continued runs describe these two distinct effects.

Electricity and Traffic account for 30 of the 32 improvements and nearly all of the net increment. A separate unit-level direction sweep finds high held-out predictability scores in these datasets (Table 8). This places most recursive improvements in datasets that also show strong first-unit directional signal; the matched continuation experiment supplies the evidence for later tasks. Elsewhere, the rule usually retains the frozen or one-unit predictor.

The validation gate targets squared error. Among the 207 paired runs with MAE recorded, RSI lowers MAE relative to one unit in 19, raises it in 15, and leaves it unchanged in 173. Appendix C details the matched protocol. Direction screening and validation admission determine which of these tasks become accepted corrections.

4.4 Diagnosing and Admitting Useful Corrections

Across 84 unit-level cells, the direction probe's maximum held-out R^2 is positively associated with realized test-MSE reduction (Spearman $\rho = 0.64$; Figure 4; Table 8). The pre-fit score thus indicates where directional correction tends to yield larger gains.

To isolate the two tests while holding the correction architecture fixed, we compare matched unit-level ablations on 42 cells at horizons 96 and 336 (126 runs per arm). With both tests, the mean test-MSE reduction is 4.78% with 13 harmed runs; removing direction screening keeps the mean at 4.78%

but raises harm to 18, while removing the aggregate gate lowers reduction to 4.43% and raises harm to 27. Removing both yields 4.61% with 23 harmed runs, and the worst cell-level increase grows from 1.07% to 7.04–8.59%. Randomized-target controls disrupt the input–residual relation: screening selects directions in only 10–12 of 108 runs and the aggregate gate deploys in two or zero. These controls separate direction selection from complete-unit admission; protocols and fitted-model controls are in Appendix D.

Re-diagnosing the residual. The 216-run, three-seed comparison establishes the system-level RSI contribution: the recursive arm shares the first unit with the ablation, then decides whether the changed residual merits another task. To isolate what makes that task recursive, we ran a separate 240-run mechanism cohort (80 matched cells at three corrector seeds) with the first unit and network budgets held fixed. The fixed-basis arm refits on each residual but retains the first screened span; the RSI arm rebuilds the probe, candidate directions, and screen at every round. Across this cohort, one unit, fixed-basis recursion, and re-diagnosed recursion yield $3.08 \pm 5.10\%$, $2.79 \pm 6.08\%$, and $2.85 \pm 6.19\%$ mean MSE reduction, respectively. Re-diagnosis adds 0.066 ± 0.512 percentage points on average over the fixed basis (68 wins, 126 ties, and 46 losses in paired cell-seed comparisons). The paired performance contrast is complemented by a direct structural measurement: the first unit and network budgets are shared, so the re-diagnosed arm changes the residual-conditioned output subspace and the capacity assigned to it. Across 256 accepted transitions, consecutive screened bases have normalized overlap 0.517 ± 0.283 (median 0.495), with 131 below 0.5; the pooled overlap with the initial basis is 0.471 ± 0.234 (Appendix C). Thus RSI changes the directions and rank available to later units, rather than merely refitting a fixed span. The full 216-run result measures RSI as a system, while this matched multi-seed control isolates renewed diagnosis as its mechanism.

4.5 Comparative effectiveness and composition

On the matched 84-cell output-side grid, GateCorrect reduces mean test MSE by 3.74% versus 1.32% for AdaY and leads on 66 cells (Table 3). Removing the adapter's input network makes both methods correct the same frozen forecast, isolating output-capacity allocation; on Traffic/DLinear/96, the reductions are 21.1% versus 5.0% (Appendix Figure 8). A second scope tests composition with input adaptation: across 13 matched cells and three seeds, X_δ+GateCorrect raises its mean reduction from 4.00% to 5.25%, improving all cell means and 38/39 runs; the full input-plus-output δ-Adapter reaches 4.35%. Thus composition adds 0.90 points and leads on 6/13 cells. The two scopes separately test the correction mechanism and its composition with a trained input intervention (Appendix F).

Table 3: Matched post-hoc comparisons. Entries are test-MSE change from the frozen backbone (%, negative is better; mean ± SD over three seeds). GateCorrect leads on 66 of 84 output-side cells; the composition control uses 13 matched cells.

	ETTh1	ETTh2	ETTm1	ETTm2	Weather	Electricity	Traffic	Avg.
AdaY	$+2.18 \pm .07$	$+2.16 \pm .10$	$-.71 \pm .08$	$-.06 \pm .19$	$-5.08 \pm .16$	$-5.04 \pm .05$	$-2.70 \pm .01$	$-1.32 \pm .003$
GateCorrect	$+.47 \pm .06$	$-1.37 \pm .10$	$-1.53 \pm .05$	$-3.47 \pm .09$	$-6.17 \pm .05$	$-8.44 \pm .15$	$-5.68 \pm .09$	$-3.74 \pm .02$

Input-adapter composition: 13 matched cells, 3 seeds			
GateCorrect	Full δ-Adapter	X_δ + GateCorrect	Additional reduction
Mean ± SD	$-4.00 \pm .05$	$-4.35 \pm .13$	$-5.25 \pm .20$	$+0.90 \pm .18$

The gain persists across horizons 96–720 (6.37%, 3.85%, 3.19%, and 1.56% mean MSE reduction; Table 9); the full grid, MAE, and abstention analysis are in Appendices B–D.2. A separate 27-run PEMS/Solar check reduces MSE by 27.5% and MAE by 21.8%, with no harmed run, illustrating directional and level-aware regimes (Appendix G).

5 Conclusion

GateCorrect turns a frozen forecaster's updated residual into the next learning task: it screens predictable output directions for uncapped amplitudes, fits a bounded level-aware remainder, and admits complete units by validation. One unit reduces mean test MSE by 3.74% over 252 runs. Across 216 matched recursive runs, further units raise the mean reduction from 2.18% to 2.81%, with 32 of 37 continuing runs improving. The unit study validates the two capacity policies; the three-seed fixed-basis control isolates RSI's additional contribution as renewed diagnosis after each accepted unit. Thus the frozen backbone remains unchanged while the residual state determines whether another learning task is worth its cost.

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

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I. Jordan. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2022.

J. M. Bates and C. W. J. Granger. The combination of forecasts. *Journal of the Operational Research Society*, 20(4):451–468, 1969.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

Chao Chen, Karl Petty, Alexander Skabardonis, Pravin Varaiya, and Zhanfeng Jia. Freeway performance measurement system: Mining loop detector data. *Transportation Research Record*, 1748(1): 96–102, 2001.

Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2019.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

Raffaella Giacomini and Halbert White. Tests of conditional predictive ability. *Econometrica*, 74(6): 1545–1578, 2006.

C. W. J. Granger and R. Ramanathan. Improved methods of combining forecasts. *Journal of Forecasting*, 3(2):197–204, 1984.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

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 (ICLR)*, 2022.

Marcel Kollovieh, Abdul Fatir Ansari, Michael Bohlke-Schneider, Jasper Zschiegner, Hao Wang, and Yuyang Wang. Predict, refine, synthesize: Self-guiding diffusion models for probabilistic time series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Hans R. Künsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989.

S. N. Lahiri. *Resampling Methods for Dependent Data*. Springer, 2003.

Qi Li, Zhenyu Zhang, Lei Yao, Zhaoxia Li, Tianyi Zhong, and Yong Zhang. Diffusion-based decoupled deterministic and uncertain framework for probabilistic multivariate time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2025.

Daojun Liang, Qi Li, Yinglong Wang, Jing Chen, Hu Zhang, Xiaoxiao Cui, Qizheng Wang, and Shuo Li. The forecast after the forecast: A post-processing shift in time series. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2601.20280.

Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A*, 379(2194):20200209, 2021. doi: 10.1098/rsta.2020.0209.

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 (ICLR)*, 2024.

Max Planck Institute for Biogeochemistry. Weather station data. `https://www.bgc-jena.mpg.de/wetter/`, 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 (ICLR)*, 2023.

Dimitris N. Politis and Joseph P. Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313, 1994.

Joseph P. Romano and Marius A. Tirlea. Permutation testing for dependence in time series. *Journal of Time Series Analysis*, 43(5):781–807, 2022.

Fangxin Wang, Ziyi Zhang, Diyi Zhuang, Langzhou He, Shiyu Wang, Baichuan Mo, and Philip S. Yu. When do corrective features help? an agent for corrective feature discovery on black-box forecasters. *arXiv preprint arXiv:2608.05207*, 2026.

Hao Wang, Licheng Pan, Zhichao Chen, Degui Yang, Sen Zhang, Yifei Yang, Xinggao Liu, Haoxuan Li, and Dacheng Tao. Fredf: Learning to forecast in the frequency domain. *arXiv preprint arXiv:2402.02399*, 2024.

Xinyan Wang, Rui Dai, Kaikui Liu, and Xiangxiang Chu. Effective probabilistic time series forecasting with fourier adaptive noise-separated diffusion. *arXiv preprint arXiv:2505.11306*, 2025.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. doi: 10.1016/S0893-6080(05)80023-1.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations (ICLR)*, 2023.

Jianxiang Xie, Yuncheng Hua, Mingyue Cheng, Flora Salim, and Hao Xue. Causality-inspired safe residual correction for multivariate time series. *arXiv preprint arXiv:2512.22428*, 2025.

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

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021.

A *run* specifies a frozen model, dataset, horizon, and correction seed; a *cell* groups the three seeds for one model–dataset–horizon combination. Unless stated otherwise, tables report signed loss changes (negative is better), while figures report reductions (positive is better).

A Datasets, frozen models, and training setup

Datasets. Table 4 lists the seven benchmark datasets. Statistics are read from the released files and the standard data loader: the ETT datasets use the loader's fixed 12/4/4-month train/validation/test split; the remaining datasets use its 70/10/20 chronological split. Dataset sources: ETT (Zhou et al., 2021), Electricity from the UCI repository (Dua & Graff, 2019), Traffic from Caltrans PeMS (Chen et al., 2001), and Jena Weather (Max Planck Institute for Biogeochemistry, 2024); preprocessing follows the benchmark library (Wu et al., 2023) throughout. All main-suite experiments use input length 96 and forecast horizons 96, 192, 336, and 720. The separate PEMS and Solar evaluation uses horizon 96.

Frozen models. We follow the benchmark-library reference scripts (Wu et al., 2023) where available. The library provides no DLinear reference scripts for Traffic or Electricity; those pairs use the learning rate 0.05 from the original DLinear paper (Zeng et al., 2023). Table 5 reports each frozen model's test MSE before correction.

Table 4: Benchmark datasets. Channels count the forecast variables; length counts time steps.

Dataset	Channels	Sampling	Length	Split
ETTh1, ETTh2	7	hourly	17,420	12/4/4 months
ETTm1, ETTm2	7	15 min	69,680	12/4/4 months
Weather	21	10 min	52,696	70/10/20
Electricity (ECL)	321	hourly	26,304	70/10/20
Traffic	862	hourly	17,544	70/10/20

Table 5: Test MSE of the frozen models before correction. The frozen model is identical across correction seeds and its evaluation is deterministic, so each cell is a single value.

Dataset	Horizon	DLinear	PatchTST	iTransformer
ETTh1	96	0.396	0.379	0.395
ETTh1	192	0.445	0.427	0.449
ETTh1	336	0.487	0.478	0.492
ETTh1	720	0.516	0.522	0.522
ETTh2	96	0.340	0.305	0.299
ETTh2	192	0.478	0.377	0.377
ETTh2	336	0.599	0.417	0.426
ETTh2	720	0.841	0.440	0.437
ETTm1	96	0.346	0.322	0.343
ETTm1	192	0.382	0.371	0.381
ETTm1	336	0.413	0.398	0.417
ETTm1	720	0.473	0.455	0.488
ETTm2	96	0.191	0.183	0.187
ETTm2	192	0.282	0.246	0.249
ETTm2	336	0.386	0.310	0.315
ETTm2	720	0.540	0.417	0.418
Weather	96	0.196	0.175	0.174
Weather	192	0.236	0.222	0.224
Weather	336	0.283	0.279	0.285
Weather	720	0.347	0.357	0.361
ECL	96	0.195	0.180	0.148
ECL	192	0.194	0.187	0.164
ECL	336	0.207	0.204	0.179
ECL	720	0.243	0.246	0.212
Traffic	96	0.649	0.460	0.393
Traffic	192	0.599	0.467	0.412
Traffic	336	0.607	0.483	0.424
Traffic	720	0.646	0.515	0.460

Corrector implementation. The directional path first normalizes each input channel using its window mean μ_X and standard deviation σ_X: $X_{\text{norm}} = (X - \mu_X)/\sigma_X$ and $\hat{Y}_{\text{norm}} = (\hat{Y} - \mu_X)/\sigma_X$. Its residual target is $(Y - \hat{Y})/\sigma_X$, with channels flattened before horizon positions. The ridge probe uses flattened normalized history, reduced by PCA to 2,048 principal components on high-dimensional datasets. Its penalty is selected on a chronological 70/30 training split; PCA and ridge coefficients are then fit using training data only. SVD of the centered training predictions supplies 32 candidate directions.

Validation screening uses Equation 6, with the training residual mean as its reference; validation residuals are not recentered.

The directional corrector takes the concatenated normalized history and frozen prediction, with a separate training-fitted feature PCA to 4,096 components where configured. It has three linear layers and two hidden layers of width 512. Training uses the FreDF frequency-domain L_1 loss on the lifted residual prediction. Each admitted direction receives a validation least-squares coefficient $\gamma_k = \langle \hat{e}_k, e_k \rangle / \langle \hat{e}_k, \hat{e}_k \rangle$, where $e_k = r_{\text{norm}}^\top u_k$ is the uncentered residual coordinate. Validation selects b_1 from $\{0, 0.25, 0.5, 0.75, 1.0, 1.25, 1.5\}$ before rescaling by σ_X.

The level-aware path has three linear layers and two hidden layers of width 512. It maps the concatenated dataset-standardized input and frozen prediction to all D output coordinates, with output $\delta \tanh(\cdot)$, $\delta = 0.075$. It uses AdamW, learning rate 3×10^{-4}, weight decay 10^{-4}, and batch size 256 (128 when $D > 2 \times 10^5$). The first validation half selects the stopping epoch and scale $b_2 \in \{0, 0.25, 0.5, 0.75, 1.0\}$.

Optimization and seeds. The level-aware head operates in the full output space, but only for the bounded remainder; the directional path predicts at most $K = 32$ coefficients. The two paths are optimized separately with training-only feature reduction, early stopping, and validation scale selection, while the frozen forecaster receives no gradient or optimizer state. Every cell uses correction seeds 0, 1, and 2; rejected candidates are never added to the deployed chain.

B FULL RESULTS AND WHERE THE GAINS COME FROM

All 84 unit-level cells. Table 6 reports the complete single-unit grid under the incremental level-aware protocol. The 252-run unit-level study uses the full validation set for direction screening and adjacent validation halves V_1, V_2 for level-aware tuning and admission, respectively. The matched recursive evaluation uses the separate protocol in Appendix C.1. Markers carry the unit-level per-cell statistics: 54 cells are FDR-significantly better and 2 worse, and the sign-consistency marker shows where all three seeds agree.

Table 6: Test-MSE change relative to the frozen model for all 84 cells (%, mean ± std over 3 seeds). *: all three seeds share the sign; †: cell-level block bootstrap $p < 0.05$; ‡: significant after BH–FDR at $q = 0.05$ over the 84 cells.

Dataset	Horizon	DLinear	PatchTST	iTransformer
ETTh1	96	-1.37 ± 0.18*	+0.00 ± 0.00*	-0.32 ± 0.08*
ETTh1	192	-0.21 ± 0.38	+0.19 ± 0.12*	-0.40 ± 0.11*
ETTh1	336	+1.07 ± 0.13*	+0.51 ± 0.91†‡	+0.05 ± 1.16
ETTh1	720	+8.05 ± 0.19*†‡	-0.52 ± 0.26*	-1.37 ± 0.73*†‡
ETTh2	96	-0.75 ± 0.09*†‡	-0.17 ± 0.06*	-1.88 ± 0.02*†‡
ETTh2	192	-0.29 ± 0.12*	-0.71 ± 0.07*	-1.52 ± 0.06*†‡
ETTh2	336	-0.21 ± 0.29	-0.59 ± 0.14*†‡	-1.60 ± 0.15*†‡
ETTh2	720	-7.65 ± 0.54*†‡	+0.42 ± 1.33	-1.43 ± 0.19*†‡
ETTm1	96	-6.38 ± 0.09*†‡	-1.24 ± 0.53*	-3.85 ± 0.28*†‡
ETTm1	192	-0.54 ± 0.41*	-0.27 ± 1.15	-1.32 ± 0.30*
ETTm1	336	+0.34 ± 0.21*	+0.12 ± 0.47	-2.57 ± 0.10*†‡
ETTm1	720	-0.09 ± 0.08	-1.41 ± 0.35*	-1.14 ± 0.14*
ETTm2	96	-5.16 ± 0.52*†‡	-6.95 ± 0.51*†‡	-7.34 ± 0.50*†‡
ETTm2	192	-2.99 ± 0.64*†‡	-2.77 ± 0.17*†‡	-2.52 ± 0.87*†‡
ETTm2	336	-0.57 ± 1.12	-3.12 ± 0.08*†‡	-4.60 ± 0.17*†‡
ETTm2	720	-0.78 ± 0.62*†‡	-1.84 ± 0.01*†‡	-3.02 ± 0.52*†‡
Weather	96	-12.08 ± 0.24*†‡	-8.71 ± 0.28*†‡	-6.93 ± 0.13*†‡
Weather	192	-8.19 ± 0.31*†‡	-7.47 ± 0.15*†‡	-5.26 ± 0.05*†‡
Weather	336	-4.65 ± 0.06*†‡	-6.51 ± 0.53*†‡	-5.44 ± 0.36*†‡
Weather	720	-1.52 ± 0.14*	-4.44 ± 0.23*†‡	-2.87 ± 0.61*†‡
ECL	96	-21.81 ± 0.57*†‡	-17.76 ± 0.03*†‡	-2.91 ± 0.16*†‡
ECL	192	-12.63 ± 0.15*†‡	-11.85 ± 0.68*†‡	-1.30 ± 0.10*†‡
ECL	336	-8.79 ± 0.20*†‡	-9.91 ± 0.11*†‡	-1.15 ± 0.42*†‡
ECL	720	-3.34 ± 1.18*†‡	-9.35 ± 0.33*†‡	-0.48 ± 0.13*†‡
Traffic	96	-21.15 ± 1.10*†‡	-5.18 ± 0.01*†‡	-1.73 ± 0.01*†‡
Traffic	192	-15.31 ± 0.02*†‡	-4.12 ± 0.02*†‡	-1.40 ± 0.20*†‡
Traffic	336	-13.84 ± 0.03*†‡	-4.09 ± 0.01*†‡	-1.36 ± 0.00*†‡
Traffic	720	+0.00 ± 0.00*	+0.00 ± 0.00*	+0.00 ± 0.00*

Sequential path contributions. Table 7 compares the directional path alone with the full method by dataset and frozen model. The directional path provides double-digit gains for DLinear on Traffic

and Electricity. On Weather/iTransformer, the level-aware path adds 3.5 percentage points; averaged across Weather models, the reduction rises from 3.9% to 6.2%. Each contribution is measured relative to the preceding fitted path.

Table 7: Sequential path contributions by dataset and frozen model (12 runs each): directional and full-method changes, their difference in percentage points, level-aware admissions, and harmed runs.

Dataset	Model	Directional (%)	Full (%)	Level-aware (pp)	Passes	Harmed
ETTh1	DLinear	+0.97	+1.88	+0.91	3/12	7/12
	PatchTST	+0.04	+0.04	+0.00	4/12	5/12
	iTransformer	-0.62	-0.51	+0.11	6/12	1/12
ETTh2	DLinear	-2.23	-2.23	+0.00	0/12	1/12
	PatchTST	-0.26	-0.26	+0.00	0/12	2/12
	iTransformer	-1.61	-1.61	+0.00	0/12	0/12
ETTm1	DLinear	-1.93	-1.67	+0.26	11/12	3/12
	PatchTST	-0.46	-0.70	-0.24	11/12	3/12
	iTransformer	-1.29	-2.22	-0.92	9/12	0/12
ETTm2	DLinear	-2.37	-2.37	+0.00	0/12	1/12
	PatchTST	-3.67	-3.67	+0.00	0/12	0/12
	iTransformer	-4.37	-4.37	+0.00	0/12	0/12
Weather	DLinear	-6.61	-6.61	+0.00	0/12	0/12
	PatchTST	-3.56	-6.78	-3.22	12/12	0/12
	iTransformer	-1.62	-5.13	-3.50	12/12	0/12
ECL	DLinear	-10.70	-11.64	-0.94	5/12	0/12
	PatchTST	-10.72	-12.22	-1.50	8/12	0/12
	iTransformer	-1.33	-1.46	-0.13	8/12	0/12
Traffic	DLinear	-12.42	-12.58	-0.16	4/12	0/12
	PatchTST	-3.28	-3.35	-0.07	9/12	0/12
	iTransformer	-1.05	-1.12	-0.08	9/12	0/12

Predictability and realized gain. Table 10 reports, per dataset and model, how much structure the direction test finds (number of admitted directions, largest held-out R^2, fraction of error energy in the admitted span) next to the realized gain. The association between these diagnostics and gain is resolved at cell level in Table 8; Table 9 gives the horizon split.

Regime descriptors per cell. Table 8 disaggregates the same diagnostics by horizon and adds d_c, the largest horizon-constant share of energy among the retained directions, which separates a span that carries within-window shape from one that carries window-level offsets. Across the 84 cells, the realized change Δ correlates most with the largest held-out R^2 ($\rho = -0.64$) and with d_c ($\rho = -0.54$), and less with the error energy the retained span covers ($\rho = -0.38$); negative values mean that stronger diagnostics accompany larger reductions. The two leading descriptors are not redundant: Traffic/DLinear reaches the highest predictability with a small offset share ($R^2 = 0.87$–0.91 and $d_c = 0.14$–0.31 over horizons 96–336), whereas Weather pairs a moderate R^2 with d_c above 0.72 at every horizon, which no other dataset does. The ETTh1 cells with unfavorable changes are the ones whose retained span covers little error energy ($\mathrm{ef} \leq 0.04$); validation–test reversal on this dataset appears in Figure 6.

Across the 84 benchmark cells, retained directions carry a median 16.8% of residual energy (range 0–71.2%). Four cells have an empty selected span: PatchTST/ETTh1/96 and all three Traffic/720 model cells, totaling 12 runs (Table 10).

The same runs scored by MAE. We score the same 252 unit-level outputs by MAE, with fitting, scale selection, and admission fixed by the MSE protocol. The directional path uses a Fourier-domain L_1 surrogate; the level-aware path, scale selection, and aggregate admission use MSE.

The MSE-trained correction also reduces mean MAE by 2.19% (Table 11). The two metrics agree in sign on 231 of 252 runs; 51 of 84 cells improve significantly on MAE and six worsen, compared with 54 and two on MSE. MAE increases in 40 runs and MSE in 23; 19 runs improve MSE while worsening MAE. ETTh1/DLinear/720 has the largest increase on both metrics ($+8.1\%$ MSE and $+5.0\%$ MAE).

Table 8: Per-cell diagnostics from a direction-only sweep of the same 84 cells (3 seeds each), with the realized change taken from the main grid. R^2 is the largest held-out score of the retained directions, "ef" the fraction of error energy they cover, d_c the largest horizon-constant share of their energy, and Δ the test-MSE change (%) relative to the frozen model. Cells with an empty selected span report zero throughout.

		DLinear				PatchTST				iTransformer			
Dataset	F	R^2	ef	d_c	Δ	R^2	ef	d_c	Δ	R^2	ef	d_c	Δ
ETTh1	96	0.33	0.159	0.88	-1.37	0.00	0.000	0.00	+0.00	0.13	0.157	0.94	-0.32
ETTh1	192	0.19	0.024	0.05	-0.21	0.07	0.013	0.08	+0.19	0.16	0.113	0.45	-0.40
ETTh1	336	0.21	0.018	0.06	+1.07	0.03	0.009	0.08	+0.51	0.13	0.032	0.09	+0.05
ETTh1	720	0.21	0.042	0.09	+8.05	-0.02	0.017	0.17	-0.52	0.05	0.091	0.34	-1.37
ETTh2	96	0.29	0.122	0.35	-0.75	0.11	0.097	0.36	-0.17	0.46	0.135	0.19	-1.88
ETTh2	192	0.33	0.119	0.62	-0.29	0.09	0.220	0.42	-0.71	0.53	0.115	0.37	-1.52
ETTh2	336	0.49	0.052	0.11	-0.21	0.21	0.111	0.12	-0.59	0.47	0.095	0.23	-1.60
ETTh2	720	0.50	0.089	0.76	-7.65	0.14	0.163	0.28	+0.42	0.53	0.100	0.17	-1.43
ETTm1	96	0.13	0.699	0.41	-6.38	0.12	0.598	0.51	-1.24	0.14	0.470	0.56	-3.85
ETTm1	192	0.15	0.587	0.55	-0.54	0.11	0.525	0.58	-0.27	0.12	0.487	0.48	-1.32
ETTm1	336	0.13	0.458	0.39	+0.34	0.13	0.417	0.61	+0.12	0.12	0.416	0.44	-2.57
ETTm1	720	0.23	0.338	0.45	-0.09	0.16	0.291	0.60	-1.41	0.15	0.281	0.45	-1.14
ETTm2	96	0.30	0.644	0.66	-5.16	0.25	0.712	0.73	-6.95	0.42	0.707	0.84	-7.34
ETTm2	192	0.13	0.449	0.57	-2.99	0.25	0.608	0.77	-2.77	0.26	0.554	0.78	-2.52
ETTm2	336	0.31	0.319	0.62	-0.57	0.18	0.518	0.50	-3.12	0.38	0.524	0.57	-4.60
ETTm2	720	0.32	0.220	0.25	-0.78	0.22	0.424	0.42	-1.84	0.36	0.416	0.48	-3.02
Weather	96	0.39	0.170	0.88	-12.08	0.27	0.219	0.82	-8.71	0.18	0.187	0.80	-6.93
Weather	192	0.43	0.181	0.89	-8.19	0.32	0.178	0.87	-7.47	0.19	0.192	0.85	-5.26
Weather	336	0.43	0.193	0.86	-4.65	0.39	0.223	0.89	-6.51	0.27	0.197	0.87	-5.44
Weather	720	0.44	0.157	0.72	-1.52	0.36	0.270	0.84	-4.44	0.21	0.234	0.83	-2.87
ECL	96	0.83	0.261	0.79	-21.81	0.84	0.245	0.74	-17.76	0.50	0.194	0.67	-2.91
ECL	192	0.84	0.145	0.55	-12.63	0.84	0.148	0.75	-11.85	0.55	0.106	0.78	-1.30
ECL	336	0.75	0.083	0.49	-8.79	0.82	0.091	0.64	-9.91	0.57	0.053	0.77	-1.15
ECL	720	0.61	0.073	0.77	-3.34	0.87	0.118	0.61	-9.35	0.59	0.083	0.66	-0.48
Traffic	96	0.87	0.322	0.31	-21.15	0.74	0.165	0.72	-5.18	0.73	0.095	0.59	-1.73
Traffic	192	0.88	0.249	0.18	-15.31	0.78	0.133	0.44	-4.12	0.77	0.070	0.26	-1.40
Traffic	336	0.91	0.221	0.14	-13.84	0.83	0.120	0.30	-4.09	0.81	0.062	0.23	-1.36
Traffic	720	0.00	0.000	0.00	+0.00	0.00	0.000	0.00	+0.00	0.00	0.000	0.00	+0.00

Table 9: Behavior by horizon (63 runs each). "Harmed" counts runs ending worse than the frozen model; "sig. better" counts FDR-significant cells of 21; "level-aware effect" averages the level-aware contribution over all 63 runs (declined runs contribute zero).

Horizon	Mean change	Harmed	Sig. better	Test passes	Level-aware effect
96	-6.37%	0	16	32/63	-0.61 pp
192	-3.85%	5	13	24/63	-0.33 pp
336	-3.19%	13	14	24/63	-0.29 pp
720	-1.56%	5	11	31/63	-0.57 pp

The path ablation explains the difference in objectives. The directional path alone reduces MAE by 2.22%, while the full unit reduces it by 2.19%; the level-aware addition increases MAE relative to the directional path in 48 runs even as it contributes 0.45 points to mean MSE reduction. Equation 1 concerns the conditional mean, the squared-error target; absolute error instead targets the conditional median. The fitted paths and admission test optimize MSE, so MAE is a secondary outcome rather than the criterion for admitting a unit.

C Recursive correction protocols and mechanism studies

Each accepted correction updates the residual target while the frozen forecaster and earlier units remain fixed. The matched recursive protocol produces Table 2; a separate fixed-basis comparison isolates the value of diagnosing that updated target again. The two protocols serve different evidence roles: the 252-run incremental protocol estimates the coverage and composition of one unit, whereas the 216-run chronological protocol estimates whether a changed residual supports another task. The latter uses disjoint acceptance blocks because each accepted unit consumes validation-side evidence and must be tested against a fresh block. The single-unit tables use the incremental level-aware protocol of Appendix D.

Table 10: Structure found by the direction test versus realized gain, per dataset × frozen model (12 runs each): mean admitted nonzero directions m, mean largest held-out R^2, mean fraction of error energy in the admitted span, and full-method change. Empty selected spans have zero admitted directions.

Dataset	Model	m	$R^2_{\max}$	Energy frac.	Full (%)
ETTh1	DLinear	12.6	0.23	0.061	+1.88
	PatchTST	2.1	0.02	0.010	+0.04
	iTransformer	14.7	0.12	0.098	-0.51
ETTh2	DLinear	16.6	0.40	0.096	-2.23
	PatchTST	13.8	0.14	0.148	-0.26
	iTransformer	18.5	0.50	0.111	-1.61
ETTm1	DLinear	31.3	0.16	0.521	-1.67
	PatchTST	28.7	0.13	0.458	-0.70
	iTransformer	30.5	0.13	0.414	-2.22
ETTm2	DLinear	19.8	0.26	0.408	-2.37
	PatchTST	27.8	0.23	0.565	-3.67
	iTransformer	30.2	0.35	0.550	-4.37
Weather	DLinear	13.4	0.42	0.175	-6.61
	PatchTST	14.9	0.33	0.222	-6.78
	iTransformer	13.2	0.21	0.202	-5.13
ECL	DLinear	28.0	0.76	0.140	-11.64
	PatchTST	29.4	0.84	0.150	-12.22
	iTransformer	27.4	0.55	0.109	-1.46
Traffic	DLinear	24.0	0.66	0.198	-12.58
	PatchTST	24.0	0.59	0.104	-3.35
	iTransformer	24.0	0.58	0.057	-1.12

Table 11: The same 252 runs scored by both metrics: test-MSE and test-MAE change against the frozen model (%, negative is better; mean over the 12 runs of each dataset–model pair). The MSE columns reproduce the unit-level results in Table 6.

	DLinear		PatchTST		iTransformer	
Dataset	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	$+1.88$	$+1.19$	$+0.04$	$+0.06$	-0.51	0.00
ETTh2	-2.23	-0.87	-0.26	-0.42	-1.61	-1.22
ETTm1	-1.67	-0.01	-0.70	-0.39	-2.22	-1.21
ETTm2	-2.37	-1.97	-3.67	-2.85	-4.37	-3.91
Weather	-6.61	-5.72	-6.78	-0.92	-5.13	-0.48
Electricity	-11.64	-5.48	-12.22	-7.96	-1.46	-0.63
Traffic	-12.58	-8.82	-3.35	-3.08	-1.12	-1.21
Average	-5.03	-3.10	-3.85	-2.22	-2.35	-1.24

C.1 Matched recursive evaluation

Training windows fit the probe and correction networks. Split the validation window origins at their midpoint: A retains earlier origins whose entire target interval ends before the boundary, and B retains origins at or after it. Require at least 32 origins in each portion. Direction screening, early stopping, coefficient calibration, and both scale choices use A. A second-path scale of zero omits that path. The complete candidate is fixed before unit acceptance is evaluated.

For n_B origins, set $J = \min(3, \lfloor n_B/(2F) \rfloor)$ and split B into J consecutive groups with approximately equal numbers of origins. Each attempted unit uses the next group, including the first unit. When $J = 0$, the run is structurally unavailable under this protocol. Distinct groups have disjoint origins; adjacent forecast target intervals can still overlap. For the paired per-window loss reduction d_i, draw 2,000 circular moving-block resamples of length F and accept the complete unit when the empirical α quantile of the resampled mean is positive. All earlier accepted units remain fixed. The test set supplies no fitting, selection, or stopping feedback.

The first unit is shared with the one-unit ablation. It keeps up to eight screened directions with score at least half the maximum and uses $\alpha = 0.01$ for acceptance. Subsequent units re-probe the current residual on A. Let s be the maximum screened direction score and k the number of screened directions. The deterministic continuation rule chooses:

- If $s \geq 0.20$, keep at most four directions with score at least $0.75s$, and use $\alpha = 0.01$.
- If $0.10 \leq s < 0.20$, keep at most eight directions with score at least $0.50s$, with acceptance tested at the $\alpha = 0.01$ level.
- If $0.05 \leq s < 0.10$ and $k \geq 8$, keep at most sixteen screened directions and use $\alpha = 0.05$.
- Otherwise stop and retain the current predictor.

Each continuing candidate still contains both correction paths. The policy allows at most two additional units, subject to available B groups, and stops at the first rejected complete unit. These thresholds were prespecified before the sweep; the reported depths are produced by this fixed policy.

The full grid has 252 runs. It contains 216 eligible paired runs and 36 structurally unavailable runs: all three seeds and backbones at $F = 720$ for ETTh1, ETTh2, Electricity, and Traffic. ETTm1, ETTm2, and Weather retain all four horizons. The accepted-depth counts for RSI are 55, 124, 22, and 15 at depths zero through three; the one-unit ablation has 55 at depth zero and 161 at depth one. Table 12 reports absolute errors by dataset and model on these matched runs. Means of relative improvements use each run's own frozen MSE as denominator and equal weight per run. The complementary MAE comparison covers the 207 matched runs for which both accepted predictions have MAE values.

Table 12: Absolute test MSE for the matched recursive evaluation, averaged over eligible horizons and three seeds within each dataset–model pair, with coverage specified in Appendix C.1.

Dataset	Model	Frozen	One unit	RSI
ETTh1	DLinear	0.442631	0.443497	0.443497
ETTh1	PatchTST	0.428315	0.428784	0.428784
ETTh1	iTransformer	0.445231	0.445444	0.445444
ETTh2	DLinear	0.472217	0.472217	0.472217
ETTh2	PatchTST	0.366773	0.366309	0.366309
ETTh2	iTransformer	0.367299	0.366652	0.366652
ETTm1	DLinear	0.403349	0.402963	0.403324
ETTm1	PatchTST	0.386662	0.386237	0.386237
ETTm1	iTransformer	0.407158	0.402641	0.402468
ETTm2	DLinear	0.349681	0.350518	0.350518
ETTm2	PatchTST	0.289072	0.285052	0.285052
ETTm2	iTransformer	0.291967	0.287411	0.287411
Weather	DLinear	0.265387	0.252014	0.252014
Weather	PatchTST	0.258295	0.246948	0.246948
Weather	iTransformer	0.260944	0.252257	0.251970
Electricity	DLinear	0.198784	0.177013	0.171461
Electricity	PatchTST	0.190456	0.176300	0.169870
Electricity	iTransformer	0.163883	0.161662	0.161399
Traffic	DLinear	0.618341	0.591416	0.545381
Traffic	PatchTST	0.470069	0.459817	0.455028
Traffic	iTransformer	0.409498	0.408383	0.407969

C.2 Admission and renewed diagnosis

The following 240-run comparison (80 matched cells at three corrector seeds) isolates renewed diagnosis on matched candidates. Its permissive arm accepts a complete unit when $g_t > 0$ on V_{dev}; a paired circular-block replay applies a lower-bound gate at $\alpha = 0.01$ to the same candidates. The primary 216-run evaluation instead uses the separate A/B protocol and score-dependent α levels above. Means and standard deviations below are computed across cell-seed runs within this matched mechanism cohort.

Why the target changes. Condition on the fitted predictor and correction functions. Since $\Delta_t(X)$ is determined by X, conditional expectation yields $\mathbb{E}[Y - f_t(X) - \Delta_t(X) \mid X] = m_t(X) - \Delta_t(X)$. The unpredictable component $Y - \mathbb{E}[Y \mid X]$ is unchanged. Thus recursion changes the approximation

problem, rather than creating new information about the future. A correction that removes some input-dependent patterns can change their relative prominence in the remaining conditional mean. Re-fitting the probe measures that new target before allocating the next unit's directional capacity.

Data roles. Training windows fit each probe and both regression networks. Partition validation chronologically into a fitting-validation portion V_{fit} and a development portion V_{dev}, removing origins whose target intervals cross their boundary. Direction screening, early stopping, directional calibration, and scale selection use V_{fit}. The level-aware incremental comparison and acceptance of a complete unit use V_{dev}. All test outcomes are reserved for evaluation after the correction functions, gates, and stopping depth are fixed. Inherited checkpoint selection uses the original validation set.

One attempted unit. At state f_t, recompute r_t on the training and validation windows. Fit the ridge probe and screen its newly derived directions, then train and calibrate $c_{1,t}$ using the directional-path representation and FreDF loss. Use f_t in place of $\hat{Y}$ in all prediction features. Next fit the full-coordinate, level-preserving level-aware network to $r_t - c_{1,t}$. Its input remains $[X, f_t(X)]$, so both paths use the prediction at the start of the unit. Retain the scale grid and $\delta = 0.075$ of the two-path design. In this recursive evaluation, the incremental level-aware gate retains the addition when its mean development MSE is below that of $f_t + c_{1,t}$ and forms $\Delta_t = c_{1,t} + G_{2,t} b_{2,t} c_{2,t}$. An empty screened span sets $c_{1,t} = 0$ but does not skip the level-aware path.

Unit acceptance and stopping. For n_{dev} development windows, define the average loss reduction

$$g_t = \frac{1}{n_{\mathrm{dev}} D} \sum_{i \in V_{\mathrm{dev}}} \left(\|r_{t,i}\|^2 - \|r_{t,i} - \Delta_t(X_i)\|^2 \right). \tag{10}$$

Accept the complete unit when $g_t > 0$, update f_t using Equation (5), and diagnose the resulting residual. Otherwise retain f_t and stop. A predeclared maximum number of attempts T bounds the training cost. Rejection of the level-aware path leaves the directional path available for this unit-level decision; rejection of the complete unit preserves every previously accepted unit. The permissive criterion makes empirical development MSE strictly decrease along accepted steps; test outcomes are evaluated only after acceptance and depth are fixed. The level-aware cap applies to each addition, while cumulative magnitudes can grow as subsequent units contribute to the final prediction.

Test-time composition. Let $t_\star$ be the accepted depth. On a new history X, compute the frozen forecast once and apply the accepted units in order, each using the preceding corrected forecast as its input prediction. This produces $f_{t_\star}(X)$ without future labels or new fitting. Evaluate real test MSE as

$$\mathrm{MSE}_{\mathrm{test}} = \frac{1}{n_{\mathrm{test}} D} \sum_{i=1}^{n_{\mathrm{test}}} \|Y_i - f_{t_\star}(X_i)\|^2. \tag{11}$$

Here n_{test} is the number of test windows. The frozen checkpoint is reused, while each new unit fits to the residual created by the preceding accepted units.

Matched recursive comparison. We compare three procedures with identical frozen checkpoints, training windows, validation partitions, first-unit weights, and per-round network budgets. One unit stops after the shared proposal. Fixed initial basis fits subsequent units to the current residual while retaining the first screened basis. Re-diagnosed basis recomputes the ridge probe, candidate directions, and direction screen at each attempt. Both recursive arms fit fresh directional and level-aware networks, use the same round-specific random seeds, and stop at the first rejected complete unit or five attempts. Thus the fixed-basis arm changes the residual target but not the space in which uncapped capacity can act, whereas the RSI arm updates both: its U_{t+1} is built from the probe prediction of r_{t+1} rather than reusing U_t. Their difference is the value of residual-conditioned capacity allocation, not a difference in network size or round budget.

Table 13 reports all 240 cell-seed runs. The cohort contains all seven datasets and three frozen models, excluding PatchTST/Traffic/336 and the three Traffic/720 cells. This mechanism cohort uses the permissive validation gate and a five-attempt budget; the primary A/B protocol uses its chronological block gate and has the separate depth distribution reported in Appendix C.1. All reported test windows are evaluated after each chain's acceptance decisions and stopping depth are fixed.

Table 13: Matched full-unit recursion on 240 cell-seed runs. MSE reduction is relative to each frozen model; larger is better. Values are mean±standard deviation across runs. Harm counts use the 10^{-4} percentage-point threshold. Depth records the number of complete units accepted by the procedure.

Procedure	Mean MSE reduction (%)	Harmed / 240	Mean depth
One unit	3.08 ± 5.10	41	0.80 ± 0.40
Fixed initial basis	2.79 ± 6.08	59	2.10 ± 1.63
Re-diagnosed basis	2.85 ± 6.19	59	2.32 ± 1.71

Depth under permissive admission. The recursive arms accept more depth than the one-unit procedure, with mean depths of 2.10 ± 1.63 and 2.32 ± 1.71 against 0.80 ± 0.40. Harmed runs rise from 41 for one unit to 59 for each recursive arm, while the shared one-unit procedure retains the largest mean reduction (3.08% against 2.79% and 2.85%). Re-diagnosis improves the mean over a fixed initial basis by 0.066 percentage points, with 68 better cell-seed comparisons, 126 ties, and 46 worse. The primary 216-run, chronological evaluation supplies the broader RSI result; this 240-run matched cohort makes the basis policy the mechanism being contrasted.

Measured subspace reconfiguration. For each accepted transition from round t to $t+1$, let U_t and U_{t+1} denote the orthonormal screened output bases diagnosed before and after the residual update. We measure their normalized squared overlap as

$$\mathrm{ov}(U_t, U_{t+1}) = \frac{\|U_t^\top U_{t+1}\|_F^2}{\min\{\mathrm{rank}(U_t), \mathrm{rank}(U_{t+1})\}}. \tag{12}$$

Identical spans have overlap one; orthogonal spans have overlap zero. The fixed-basis control retains U_t by construction and therefore lies at one for every transition. RSI recomputes U_{t+1} after the accepted unit changes the residual. Figure 5 displays the resulting overlap distribution and the rank change at each transition. Across 256 accepted cell-seed-round transitions, consecutive-basis overlap is 0.517 ± 0.283 (median 0.495), with 131 below 0.5; the screened rank changes from 19.5 ± 9.7 to 15.6 ± 9.9 (Table 14). The first-to-later basis comparison gives pooled overlap 0.471 ± 0.234 across the same 256 pairs, measuring cumulative drift from the initial span. These are geometry measurements, not test-error gains: they show directly that a renewed residual diagnosis changes both the directions and the rank available to the next unit. The bases are computed from training residuals and screened on validation data; test losses do not define the measurements. Later rounds have fewer transitions because only accepted units create the next residual task.

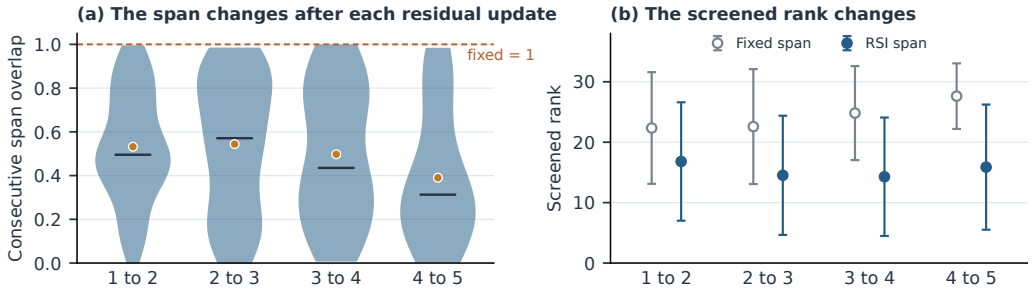

Figure 5: Subspace reconfiguration after each accepted residual update. (a) Distribution of normalized squared overlap between consecutive screened bases in RSI; the dashed line is the fixed-basis control, which reuses the preceding span (overlap one). Gold dots mark means and dark bars mark medians. (b) Mean±standard deviation of screened rank for the fixed and re-diagnosed spans on the same transition cohorts. Counts for rounds $1\to2$ through $4\to5$ are 114, 79, 39, and 24 cell-seed transitions.

D Admission under temporal dependence

D.1 Validation transfer and signed admission

Validation partitioning and signed admission matter when residual windows overlap in time. Figure 6 compares validation and test outcomes; randomized-target controls examine which corrections pass the two tests after the input–residual relationship is disrupted.

Table 14: Consecutive residual-basis transitions in the three-seed matched mechanism cohort. Overlap is defined in Eq. (12); < 0.5 counts transitions below one half. The first rank is the basis available before the residual update; the next rank is re-diagnosed after that update. Values are mean±standard deviation.

Transition	n	Overlap	< 0.5	First rank	Next rank
$1 \to 2$	114	0.532 ± 0.245	58	22.4 ± 9.2	16.8 ± 9.8
$2 \to 3$	79	0.544 ± 0.308	34	17.5 ± 9.9	14.5 ± 9.9
$3 \to 4$	39	0.498 ± 0.309	22	16.4 ± 8.9	14.3 ± 9.8
$4 \to 5$	24	0.390 ± 0.302	17	17.3 ± 9.3	15.9 ± 10.4
Pooled	256	0.517 ± 0.283	131	19.5 ± 9.7	15.6 ± 9.9

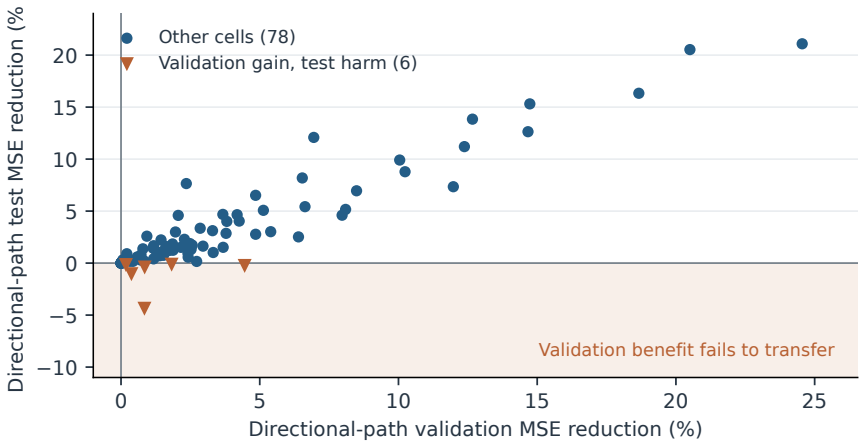

Figure 6: Validation versus test MSE reduction of the directional correction (84 cells, seed-averaged; the same correction measured on both splits; positive is better). Six cells move from a validation gain to a test loss (orange triangles in the shaded region), concentrated on ETTh1.

Cell-level significance. A cell is one model–dataset–horizon combination; its three seeds share the same test windows because the frozen model and the window grid are fixed. For each cell, the per-window squared-error differences (corrected minus frozen) of the three seeds are averaged window-wise, giving one difference sequence. That sequence undergoes a moving-block bootstrap (Lahiri, 2003): blocks of length equal to the forecast horizon are resampled with replacement to the original length, 10,000 times. Adjacent windows share $F - 1$ target steps; Table 15 also tests longer dependence scales. The statistic is the resampled mean difference divided by the frozen model's mean loss on the same resampled indices (a percentage change); the p-value is twice the smaller tail probability of zero, and the interval is the 2.5–97.5 percentile range. Benjamini–Hochberg at $q = 0.05$ is applied to the 84 cell p-values. The aggregate test of Section 3.4 uses the same bootstrap on the second validation half, comparing per-window losses with and without the level-aware corrector, and deploys only when $p < 0.05$ and the interval's upper endpoint is negative. This reproduces the signed incremental level-aware decision in the unit-level studies.

For n_B windows in V_2, the aggregate test requires at least 20 distinct length-F block starts, or $n_B - F + 1 \geq 20$. These starts can overlap; the requirement ensures resampling resolution, while the dependence checks below assess the resulting admission behavior. It also excludes the level-aware path on Traffic-720.

Sensitivity to the block length and to the multiplicity procedure. Two adjacent windows share their lookback as well as their horizon, so their loss difference can be correlated over more than F steps. We therefore recomputed every cell p-value at block lengths F, $F+96$ (lookback plus horizon), $2F$, and $4F$, and applied to each resulting family both Benjamini–Hochberg and Benjamini–Yekutieli (Benjamini & Yekutieli, 2001), the latter permitting arbitrary dependence between the cell statistics when the individual p-values are valid (Table 15). Between F and $F + 96$ the counts are unchanged (54 better, 2 worse); longer blocks raise both counts slightly. At block length F, BY identifies 51 significantly improved and one significantly worsened cell, versus 54 and two under BH. Both procedures retain the same broad pattern of significant corrections across the grid.

Table 15: Significance counts (better / worse out of 84 cells) under four bootstrap block lengths and two multiplicity procedures, both at $q = 0.05$. The main setting uses BH with horizon-length blocks.

Block length	$p < 0.05$	Benjamini–Hochberg	Benjamini–Yekutieli
F (main table)	54 / 2	54 / 2	51 / 1
$F + 96$ (lookback + horizon)	54 / 2	54 / 2	50 / 2
$2F$	55 / 5	55 / 3	52 / 3
$4F$	57 / 7	57 / 7	55 / 6

Randomized-target controls. The controls replace residual targets within training and validation while keeping inputs, the frozen forecaster, and real test outcomes unchanged. A sign-flipped null removes the input–residual alignment; an envelope-matched phase-randomized surrogate preserves a comparable residual-energy envelope while altering its temporal organization. These controls separate direction selection from the signed decision to deploy a complete correction.

On the matched 108-run grid, the direction test selected nonzero directions in 10 sign-flipped runs and 12 envelope-matched runs. The aggregate gate deployed in two and zero runs, respectively; the corresponding largest absolute test-MSE changes were 0.208% and 1.074%. The signed gate therefore rejects the randomized corrections much more often than it selects them, while the real test outcomes remain held out from fitting.

D.2 Construction of the direction test's null

Let $R^2_{k,s}$ denote the score for candidate direction k under circular shift s, and let S be the number of evaluated shifts. The empirical comparison assigns

$$p_k = \frac{1 + \#\{s : R^2_{k,s} \geq R^2_k\}}{S + 1}. \tag{13}$$

BH orders these K values and retains those through the largest rank j satisfying $p_{(j)} \leq qj/K$, with $q = 0.1$. The false discovery rate is the expected fraction of null directions among the selected ones, defined as zero for an empty selection.

The direction test circularly shifts the validation error-window sequence relative to the input sequence, retaining circular temporal structure. Under dependence, nominal q is interpreted alongside the empirical selection controls below (Romano & Tirlea, 2022).

For n validation windows, offsets from F to $n - F$ keep both circular distances at least F from the original alignment. The nominal number of admissible offsets is $A = \max(n - 2F + 1, 0)$. If $A < 19$, the correction abstains. Otherwise, it enumerates every offset when $A < 1{,}000$ and evaluates 1,000 random shifts when $A \geq 1{,}000$. Equation 13 uses the number of evaluated shifts, not the entire admissible set. The rule sets the resolution of the empirical p-values; the rank-specific BH threshold may still fall below the smallest attainable p-value.

Traffic-720 has no admissible shift, so the directional path abstains.

E Design ablations

E.1 Value of the two tests

To isolate the decisions from the correction architecture, three arms run the whole pipeline over 7 datasets $\times$ 3 models $\times$ horizons $\{96, 336\}$ $\times$ 3 seeds (42 cells, 126 runs per arm), compared against the unit-level protocol restricted to those same cells. *Direction test off* keeps all 32 candidate directions without testing any of them; the directional path's per-direction coefficients and its overall scale are still chosen on validation, so the path can still switch itself off. *Aggregate test off* deploys the level-aware path at the scale chosen on the first validation half whatever the test says. *Both off* is the plain two-path corrector, with no test anywhere. Each arm applies BH correction to its 42 cell-level comparisons.

The matched ablations separate the tests' roles (Table 17). Direction screening limits severe harmful corrections: disabling it leaves the mean change at -4.78% but raises the harmed-run count from 13 to 18 and the worst cell from $+1.07\%$ to $+7.04\%$. Aggregate admission controls level-aware deployment: the path is added in 56 of 126 runs with the test and 123 of 126 without it; disabling this test raises harmed runs to 27 while reducing mean benefit by 0.35 percentage points. Removing both tests yields 23 harmed runs and a -4.61% mean change. The full cell-level outcomes remain in Table 17.

Table 16: Randomized-target controls on one identical 108-run grid (36 cells × 3 seeds): sign-flipped, with zero randomization mean and destroyed temporal dependence, and an envelope-matched phase-randomized surrogate approximating the normalized residual-energy sequence. Per cell: runs in which the direction test admitted any direction (Dir.), runs in which the aggregate test deployed the dense addition (Agg.), and the largest absolute test-MSE change (%).

			Sign-flipped null			Semi-synthetic null		
Dataset	Horizon	Model	Dir.	Agg.	max $\|\Delta\|$	Dir.	Agg.	max $\|\Delta\|$
ETTh1	96	DLinear	0/3	0/3	0.000	0/3	0/3	0.000
ETTh1	96	PatchTST	0/3	0/3	0.000	0/3	0/3	0.000
ETTh1	96	iTransformer	0/3	0/3	0.000	0/3	0/3	0.000
ETTh1	336	DLinear	0/3	0/3	0.000	0/3	0/3	0.000
ETTh1	336	PatchTST	0/3	0/3	0.000	0/3	0/3	0.000
ETTh1	336	iTransformer	0/3	0/3	0.000	1/3	0/3	0.051
ETTh2	96	DLinear	0/3	0/3	0.000	0/3	0/3	0.000
ETTh2	96	PatchTST	0/3	0/3	0.000	1/3	0/3	0.039
ETTh2	96	iTransformer	1/3	0/3	0.000	1/3	0/3	0.532
ETTh2	336	DLinear	0/3	0/3	0.000	1/3	0/3	0.109
ETTh2	336	PatchTST	1/3	0/3	0.208	1/3	0/3	0.027
ETTh2	336	iTransformer	1/3	0/3	0.010	2/3	0/3	0.027
ETTm2	96	DLinear	1/3	0/3	0.001	1/3	0/3	0.232
ETTm2	96	PatchTST	2/3	0/3	0.000	0/3	0/3	0.000
ETTm2	96	iTransformer	0/3	0/3	0.000	1/3	0/3	0.089
ETTm2	336	DLinear	1/3	0/3	0.032	0/3	0/3	0.000
ETTm2	336	PatchTST	0/3	0/3	0.000	0/3	0/3	0.000
ETTm2	336	iTransformer	0/3	0/3	0.000	0/3	0/3	0.000
Weather	96	DLinear	0/3	0/3	0.000	0/3	0/3	0.000
Weather	96	PatchTST	0/3	0/3	0.000	0/3	0/3	0.000
Weather	96	iTransformer	1/3	0/3	0.017	0/3	0/3	0.000
Weather	336	DLinear	0/3	0/3	0.000	1/3	0/3	0.005
Weather	336	PatchTST	0/3	0/3	0.000	1/3	0/3	0.394
Weather	336	iTransformer	0/3	0/3	0.000	1/3	0/3	1.074
ECL	96	DLinear	0/3	0/3	0.000	0/3	0/3	0.000
ECL	96	PatchTST	0/3	1/3	0.001	0/3	0/3	0.000
ECL	96	iTransformer	1/3	0/3	0.000	0/3	0/3	0.000
ECL	336	DLinear	0/3	0/3	0.000	0/3	0/3	0.000
ECL	336	PatchTST	0/3	0/3	0.000	0/3	0/3	0.000
ECL	336	iTransformer	0/3	0/3	0.000	0/3	0/3	0.000
Traffic	96	DLinear	0/3	0/3	0.000	0/3	0/3	0.000
Traffic	96	PatchTST	1/3	0/3	0.000	0/3	0/3	0.000
Traffic	96	iTransformer	0/3	0/3	0.000	0/3	0/3	0.000
Traffic	336	DLinear	0/3	1/3	0.000	0/3	0/3	0.000
Traffic	336	PatchTST	0/3	0/3	0.000	0/3	0/3	0.000
Traffic	336	iTransformer	0/3	0/3	0.000	0/3	0/3	0.000

Table 17: Value of direction screening and aggregate admission on the same 42 cells (126 runs per arm). The full-method row uses the unit-level protocol restricted to these cells. "Harmed" counts runs ending worse than the frozen model; each arm adjusts its 42 cellwise significance tests for FDR.

Arm	Mean (%)	Harmed	Sig. better	Sig. worse	ETTh1 mean (%)
Both tests (full method)	−4.78	13	30	1	−0.01
Direction test off	−4.78	18	31	3	+2.69
Aggregate test off	−4.43	27	27	4	+1.04
Both tests off	−4.61	23	28	3	+3.80

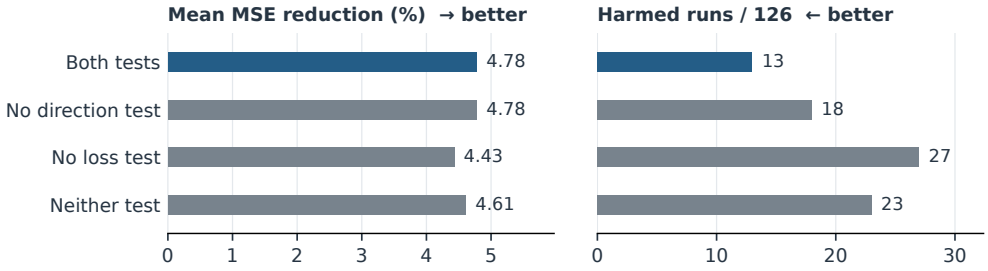

Figure 7: Direction screening and admission shape gain and harm. Across 42 matched cells and three seeds (126 runs per row), the left panel shows mean MSE reduction and the right panel counts harmed runs. Table 17 gives the values, and harm denotes an increase above 10^{-4} percentage points in test MSE relative to the frozen model.

E.2 Level-aware representation ablation

The level-aware path has to recover the part of the remaining residual that the directional path leaves behind, which sets two requirements on its representation: the fitting problem has to sit on the evaluation scale, and the input window's level has to survive normalization. Table 18 separates target scale, input features, and loss on Weather/iTransformer, with four horizons and three seeds each. Using the dataset-standardized $[X, \hat{Y}]$ features retains window-level information lost under per-window normalization and raises the horizon-720 mean from -1.07% to -5.52%. Moving the target to the dataset-standardized scale increases the horizon-96 reduction from 6.46% to 7.67%. Their combination changes the four-horizon profile; the additional time-domain MSE objective favors horizons 192 and 720.

A fixed per-channel bias does not reproduce that contribution. Offered to the same aggregate test it is declined in all 36 Weather runs, while the input-dependent network contributes -2.8 to -4.3 points on the two transformer models in those same runs. Together these results support an input-window-dependent contribution beyond a fixed bias.

Table 18: Level-aware representation ablation on Weather/iTransformer (MSE change vs. frozen model, %, averaged over three seeds per horizon). "Window-normalized" features subtract the input-window mean and divide by its standard deviation; window-normalized targets divide the residual by that deviation. "Dataset-standardized" retains training-set normalization. The first four rows use the frequency-domain loss; the last row switches to time-domain MSE.

Targets	Features	Loss	F96 / F192 / F336 / F720
window-normalized	window-normalized	FreDF	-6.46 / -3.52 / -3.80 / -1.07
dataset-standardized	window-normalized	FreDF	-7.67 / -2.73 / -1.28 / -0.56
window-normalized	dataset-standardized	FreDF	-4.80 / -2.30 / -3.64 / -5.52
dataset-standardized	dataset-standardized	FreDF	-8.68 / -7.67 / -10.25 / -4.78
dataset-standardized	dataset-standardized	MSE	-7.68 / -8.57 / -9.50 / -5.57

Target scaling and input features act on different parts of the residual: the former places fitting on the evaluation scale, while the latter preserves each window's level. The combined rows differ by loss, with small differences in their four-horizon means. The reported method uses dataset-standardized inputs and targets with time-domain MSE, the cap, and split-gated selection of Section 3.4.

The overlap of the two paths is measured directly: on Weather, projecting the trained level-aware correction onto the admitted screened span leaves 47% of its energy inside the span on average over the 24 runs whose gate opens (46.6% over all 36 runs), so the paths are not orthogonal and their contributions are reported jointly and by ablation (Section 3.4). Increasing the directional candidate rank from 32 to 256 leaves the measured outcome indistinguishable from the directional path on this configuration.

Level-aware cap across two regimes. A 24-run iTransformer control spans Weather and ETTh1 at four horizons and three seeds. As δ increases from 0.02 to 0.15, Weather's four-horizon mean MSE change moves from -2.94% to -6.22%, while the worst ETTh1 cell moves from 0.00% to $+4.66\%$. At $\delta = 0.075$, those values are -5.29% and $+0.62\%$, respectively. The cap changes both Weather's available correction and the largest ETTh1 increase within this 24-run control.

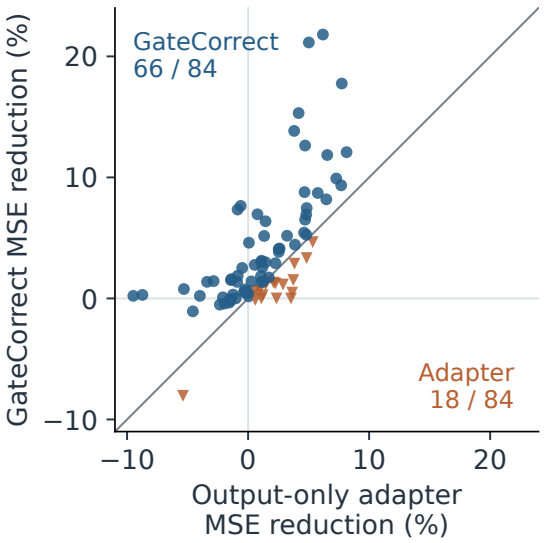

Figure 8: Output-side comparison on 84 cells (three-seed means). Both axes show signed MSE reductions at the same scale. GateCorrect leads above the diagonal (66 cells); the output-side δ-Adapter variant leads below it (18). Table 3 summarizes the same paired records by dataset.

F δ-ADAPTER COMPARISONS

Matched output-side comparison. The published δ-Adapter (Liang et al., 2026) jointly trains an input-side X network and an output-side Y network. Our full-grid comparison removes X and retains the output-side correction, so the 84 cells use identical frozen models, test windows, and seeds. This isolates how the two methods allocate correction after the same frozen forward pass.

Table 20 is the complete cell-level record: the output-side adapter reduces MSE by 1.3% on average against GateCorrect's 3.7%, and GateCorrect leads on 66 of the 84 cells. Its output-side scope keeps the frozen forward pass identical and isolates output-capacity allocation.

Full input-plus-output comparison. Table 19 reports 13 matched cells with three seeds. Both adapter networks use the released architecture and offline recipe with $\delta = 0.1$. All changes use the original frozen-model MSE. The full adapter reduces mean MSE by 4.35%, compared with 4.00% for GateCorrect, and leads on 11 of 13 cells.

Combining input adaptation with residual correction. We freeze the official X network chosen by the adapter's validation selection for each cell and fit GateCorrect on its resulting predictions. This composition reduces mean MSE by 5.25%, compared with 4.00% without the input adapter. It improves all 13 cell means and 38 of 39 individual runs. Against the full input-plus-output adapter, the composition leads on six cells, including Traffic/DLinear/96 (22.85% versus 8.65%) and three Weather horizons. Thus input adaptation can complement the two-path residual corrector.

Table 19: Input and output correction on the same 13 cells. Entries are test-MSE changes (%) relative to the original frozen model, averaged over three seeds; negative is better. Full δ trains its input-side X network and output-side Y network jointly. GC denotes GateCorrect; GC+X freezes the official input network and then runs GateCorrect on its forecast. All procedures are fixed before test.

Model	Dataset	Horizon	GC	Full δ	GC+X
DLinear	Traffic	96	-21.15	-8.65	-22.85
iTransformer	ECL	96	-2.91	-4.50	-5.10
	ECL	192	-1.30	-2.82	-2.84
	ECL	336	-1.15	-2.93	-1.79
	ECL	720	-0.48	-5.45	-3.07
	Traffic	96	-1.73	-2.55	-2.40
	Traffic	192	-1.40	-1.92	-1.82
	Traffic	336	-1.36	-1.66	-1.62
	Traffic	720	-0.00	-1.67	-0.54
	Weather	96	-6.93	-6.73	-7.88
	Weather	192	-5.26	-6.12	-6.77
	Weather	336	-5.44	-6.48	-7.80
	Weather	720	-2.87	-5.08	-3.72
Mean			-4.00	-4.35	-5.25

Table 20: Output-side δ-Adapter variant (Y network retained, input network removed), all comparison cells (3 seeds each): test-MSE change, mean output saturation at its $\tanh$ bound, and Y-network parameter count.

Dataset	Horizon	Model	Change (%)	Saturation	Parameters
ETTh1	96	DLinear	+3.39 $\pm$ 0.09	2%	951,968
ETTh1	96	PatchTST	+1.02 $\pm$ 0.10	2%	951,968
ETTh1	96	iTransformer	+1.24 $\pm$ 0.28	15%	951,968
ETTh1	192	DLinear	+3.98 $\pm$ 0.29	12%	1,640,768
ETTh1	192	PatchTST	+1.47 $\pm$ 0.11	11%	1,640,768
ETTh1	192	iTransformer	+0.00 $\pm$ 0.47	28%	1,640,768
ETTh1	336	DLinear	+4.57 $\pm$ 0.33	15%	2,673,968
ETTh1	336	PatchTST	+2.33 $\pm$ 0.15	27%	2,673,968
ETTh1	336	iTransformer	+1.49 $\pm$ 0.71	30%	2,673,968
ETTh1	720	DLinear	+5.38 $\pm$ 0.14	17%	5,429,168
ETTh1	720	PatchTST	+0.35 $\pm$ 0.07	30%	5,429,168
ETTh1	720	iTransformer	+0.90 $\pm$ 0.18	29%	5,429,168
ETTh2	96	DLinear	+0.26 $\pm$ 0.61	12%	951,968
ETTh2	96	PatchTST	-0.04 $\pm$ 0.47	7%	951,968
ETTh2	96	iTransformer	+0.86 $\pm$ 0.03	18%	951,968
ETTh2	192	DLinear	+8.73 $\pm$ 0.38	39%	1,640,768
ETTh2	192	PatchTST	-0.95 $\pm$ 0.14	9%	1,640,768
ETTh2	192	iTransformer	+1.42 $\pm$ 0.09	37%	1,640,768
ETTh2	336	DLinear	+9.48 $\pm$ 0.30	34%	2,673,968
ETTh2	336	PatchTST	-0.60 $\pm$ 0.05	30%	2,673,968
ETTh2	336	iTransformer	+1.38 $\pm$ 0.05	41%	2,673,968
ETTh2	720	DLinear	+0.61 $\pm$ 0.08	32%	5,429,168
ETTh2	720	PatchTST	+1.94 $\pm$ 0.39	15%	5,429,168
ETTh2	720	iTransformer	+2.85 $\pm$ 0.12	28%	5,429,168
ETTm1	96	DLinear	-1.44 $\pm$ 0.16	19%	951,968
ETTm1	96	PatchTST	-2.26 $\pm$ 0.16	17%	951,968
ETTm1	96	iTransformer	-2.54 $\pm$ 0.29	17%	951,968
ETTm1	192	DLinear	+0.20 $\pm$ 0.23	19%	1,640,768
ETTm1	192	PatchTST	-1.20 $\pm$ 0.17	19%	1,640,768
ETTm1	192	iTransformer	-1.48 $\pm$ 0.16	22%	1,640,768
ETTm1	336	DLinear	+1.56 $\pm$ 0.14	22%	2,673,968
ETTm1	336	PatchTST	-0.59 $\pm$ 0.30	28%	2,673,968
ETTm1	336	iTransformer	-1.18 $\pm$ 0.13	27%	2,673,968
ETTm1	720	DLinear	+2.08 $\pm$ 0.43	31%	5,429,168
ETTm1	720	PatchTST	-0.27 $\pm$ 0.20	43%	5,429,168

Continued on next page

Table 20 (continued)

Dataset	Horizon	Model	Change (%)	Saturation	Parameters
ETTm1	720	iTransformer	-1.44 ± 0.11	39%	5,429,168
ETTm2	96	DLinear	-1.33 ± 1.16	6%	951,968
ETTm2	96	PatchTST	-0.77 ± 0.75	9%	951,968
ETTm2	96	iTransformer	+0.87 ± 0.18	11%	951,968
ETTm2	192	DLinear	-1.43 ± 0.41	16%	1,640,768
ETTm2	192	PatchTST	-0.55 ± 0.43	25%	1,640,768
ETTm2	192	iTransformer	+0.49 ± 0.48	24%	1,640,768
ETTm2	336	DLinear	+0.01 ± 0.50	35%	2,673,968
ETTm2	336	PatchTST	-1.11 ± 0.22	25%	2,673,968
ETTm2	336	iTransformer	-0.07 ± 0.14	32%	2,673,968
ETTm2	720	DLinear	+5.30 ± 1.87	39%	5,429,168
ETTm2	720	PatchTST	-1.06 ± 0.23	39%	5,429,168
ETTm2	720	iTransformer	-1.08 ± 0.29	39%	5,429,168
Weather	96	DLinear	-8.14 ± 0.05	20%	2,329,568
Weather	96	PatchTST	-5.76 ± 0.37	16%	2,329,568
Weather	96	iTransformer	-4.80 ± 0.52	16%	2,329,568
Weather	192	DLinear	-6.46 ± 0.42	19%	4,395,968
Weather	192	PatchTST	-4.84 ± 0.16	18%	4,395,968
Weather	192	iTransformer	-4.81 ± 0.20	18%	4,395,968
Weather	336	DLinear	-5.34 ± 0.51	21%	7,495,568
Weather	336	PatchTST	-4.71 ± 0.30	21%	7,495,568
Weather	336	iTransformer	-4.62 ± 0.27	21%	7,495,568
Weather	720	DLinear	-3.74 ± 0.86	26%	15,761,168
Weather	720	PatchTST	-3.87 ± 0.19	25%	15,761,168
Weather	720	iTransformer	-3.83 ± 0.30	26%	15,761,168
ECL	96	DLinear	-6.18 ± 0.09	9%	31,849,568
ECL	96	PatchTST	-7.74 ± 0.06	17%	31,849,568
ECL	96	iTransformer	-2.26 ± 0.12	1%	31,849,568
ECL	192	DLinear	-4.71 ± 0.11	8%	63,435,968
ECL	192	PatchTST	-6.54 ± 0.17	15%	63,435,968
ECL	192	iTransformer	-2.11 ± 0.18	0%	63,435,968
ECL	336	DLinear	-4.67 ± 0.63	9%	110,815,568
ECL	336	PatchTST	-7.28 ± 0.21	16%	110,815,568
ECL	336	iTransformer	-2.88 ± 0.03	1%	110,815,568
ECL	720	DLinear	-4.84 ± 0.31	8%	237,161,168
ECL	720	PatchTST	-7.69 ± 0.26	15%	237,161,168
ECL	720	iTransformer	-3.64 ± 0.02	5%	237,161,168
Traffic	96	DLinear	-5.02 ± 0.01	13%	85,083,968
Traffic	96	PatchTST	-3.22 ± 0.01	3%	85,083,968
Traffic	96	iTransformer	-1.71 ± 0.00	1%	85,083,968
Traffic	192	DLinear	-4.18 ± 0.02	8%	169,904,768
Traffic	192	PatchTST	-2.60 ± 0.01	2%	169,904,768
Traffic	192	iTransformer	-1.24 ± 0.02	0%	169,904,768
Traffic	336	DLinear	-3.81 ± 0.01	6%	297,135,968
Traffic	336	PatchTST	-2.53 ± 0.01	2%	297,135,968
Traffic	336	iTransformer	-1.10 ± 0.05	0%	297,135,968
Traffic	720	DLinear	-3.54 ± 0.01	6%	636,419,168
Traffic	720	PatchTST	-2.35 ± 0.01	1%	636,419,168
Traffic	720	iTransformer	-1.10 ± 0.01	0%	636,419,168

G DATASETS OUTSIDE THE MAIN SUITE

Predictable modes and level correction outside the main suite. PEMS and Solar extend the horizon-96 evaluation to traffic-sensor and solar-energy data (27 runs; Table 21). The mean test-MSE change is -27.5% (-21.8% on MAE); no run worsens on either metric, and 8 of 9 cells are FDR-significant. PEMS is the high-score case: largest held-out directional scores range from 0.73 to 0.91, with 31–32 directions admitted per cell. Solar admits 31 directions despite a lower maximum score ($R^2_{\max} = 0.11$); its level-aware gate opens in all three runs, and the full unit changes test MSE by -3.8% (not FDR-significant). The contrast separates directional score strength from the number of admitted directions and from the level-aware decision.

Table 21: PEMS and Solar at horizon 96, 3 seeds per cell. "Host MSE" is the test MSE of the frozen model; "change" is the test-MSE change (%, negative is better); "dirs" and the next column give admitted directions and their largest held-out score; "gate" counts level-aware deployments; "val half-2" re-measures the correction on the held-out validation half.

Dataset	F	Model	Host MSE	Change (%)	dirs	$R^2_{\max}$	gate	val half-2
PEMS03	96	DLinear	0.4525	$-51.59 \pm 0.28^*$	32	0.85	3/3	-64.14
PEMS03	96	iTransformer	0.1729	$-14.30 \pm 0.11^*$	32	0.82	3/3	-18.89
PEMS04	96	DLinear	0.4236	$-58.60 \pm 0.38^*$	31	0.90	3/3	-56.52
PEMS04	96	iTransformer	0.1733	$-20.74 \pm 0.11^*$	32	0.91	3/3	-22.72
PEMS07	96	DLinear	0.5722	$-24.21 \pm 1.69^*$	32	0.83	3/3	-53.87
PEMS07	96	iTransformer	0.1381	$-19.43 \pm 0.14^*$	31	0.84	3/3	-19.05
PEMS08	96	DLinear	0.6606	$-43.26 \pm 0.83^*$	31	0.76	3/3	-13.23
PEMS08	96	iTransformer	0.2947	$-11.59 \pm 0.33^*$	31	0.73	3/3	-2.88
Solar	96	iTransformer	0.2023	-3.78 ± 0.36	31	0.11	3/3	-3.55

H Further related work

Other post-hoc methods structure components, instances, decision losses, or predefined correction choices, whereas online adaptation updates a corrector when future observations arrive. GateCorrect fixes its regressors, scales, and admission decisions before test, and diagnoses history-predictable output modes from the residual at each accepted round. Representative methods are positioned in the main text and cited there.

I Statements

Ethics statement

We evaluate statistical methods on public benchmarks under their original licenses. The study involves no human subjects or personal data.

Reproducibility statement

The appendix specifies the datasets, training configurations, correction seeds, validation allocations, and evaluation procedures. Each paired comparison uses the same frozen checkpoint and test windows.