AUDITING STRUCTURED FORECASTING LOSSES: ERROR CORRECTION AND EXACT-FIT OPTIMALITY

Anonymous authors
Paper under double-blind review

ABSTRACT

We audit whether structured forecasting losses preserve error correction after partial fit and favor exact agreement with fixed targets. DBLoss's trend-gradient coefficient vanishes at exact seasonal fit, leaving nonzero level biases stationary. A positive integrating factor yields a direction-equivalent scalar potential on a fixed batch, yet these biased forecasts remain global minimizers. Controlled experiments on fifteen saved models test quadratic field attenuation near this boundary and show optimizer-dependent recovery. In the published RI-Loss helper, explicitly activated in our runner, the finite-sample statistic measures residual-channel dispersion in expectation. We derive a sharp weight threshold for the full objective averaged over surrogate noise. In residual space, zero residual is the unique global minimizer for nonnegative weights at or below the threshold, and a saddle above it. Source checks, stochastic trajectories, and neural-parameter perturbations at constructed agreement test the local predictions. Matching dropout randomness largely removes the observed short-run effect of statistic replacement. Controlled extensions to Time-o1 and PS Loss probe equivalent encodings and batch-statistic scope. Five-source forecasting comparisons show heterogeneous predictive effects under the tested budgets. These results yield testable coefficient and weight conditions for error correction and exact-fit optimality.

1 INTRODUCTION

A forecast may capture the shape of a daily profile while remaining uniformly too high. A structured loss should retain a signal that corrects this bias. At exact agreement with fixed targets, we ask whether small errors along directions the model can realize increase the objective. We study these properties, *error correction* and *exact-fit stability*, in implemented forecasting objectives.[1]

These properties depend on interactions among loss terms. Adaptive balancing adjusts component-gradient strengths (Chen et al., 2018); in the rule studied here, fitting one component can eliminate another's correction signal. A finite-sample dependence statistic can also induce residual dispersion whose curvature competes with an MSE anchor. We therefore distinguish the scalar loss, the field supplied to the optimizer, and the resulting parameter update.

We identify two mechanisms that produce zero gradients at different prediction states. In DBLoss (Qiu et al., 2025), exact seasonal fit makes the trend-gradient coefficient vanish, leaving a nonzero level bias stationary. A positive integrating factor maps the fixed-batch field to the gradient of a potential whose global minima include these biased forecasts. In the published RI-Loss helper (Wang et al., 2026c), a biased finite-sample kernel statistic favors dispersion among residual channels. For the full expected objective, we derive a sharp auxiliary-weight threshold: zero residual is the unique global minimizer for nonnegative weights at or below the threshold, and a saddle in residual space above it.

[1]Code and reproduction instructions are available at `https://anonymous.4open.science/r/forecasting-loss-audit-1FAB/`.

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.

The distinction between stationarity and local optimality is well established (Dauphin et al., 2014), and work on alignment losses motivates explicit checks at prediction–target agreement (Blondel et al., 2021). We characterize the component gate and derive the sharp global optimality threshold, then test their predicted local responses. DBLoss recovery experiments distinguish attenuation of the update field from the effects of Adam's adaptive rescaling. RI experiments test predicted curvature in neural parameter space at constructed agreement and use matched random streams to isolate the effects of statistic replacement from changes in dropout randomness. Time-o1 and PS Loss extend these checks to equivalent encodings and the scope over which batch statistics are computed, respectively.

The contributions are:

- **A condition for error correction after partial fit.** We characterize DBLoss's stationary-bias boundary and nearby quadratic attenuation of the trend-correction field. On a fixed batch, a positive integrating factor yields a direction-equivalent scalar potential whose global minimizers include these biased forecasts. Controlled recovery experiments test the effect of retaining a positive trend coefficient.
- **A sharp boundary for exact-fit optimality.** For the RI helper's full expected objective, we establish a sharp threshold for the auxiliary weight. For nonnegative weights at or below this threshold, zero residual is the unique global minimizer; above it, zero residual is a saddle in residual space.
- **Controlled validation of learning mechanisms.** Optimizer measurements, matched random streams, and restoration controls connect selected diagnoses to executed updates. A ten-case audit provides broader coverage across four complementary perspectives: targets, representations, sample coupling, and field geometry.

2 Prediction properties and the learning rule

Let $r = f_\theta(X) - Y \in \mathbb{R}^{H\times D}$ denote the forecast residual, where H is the forecast horizon and D the number of channels. For a batch B, the implemented backward rule produces a field $g_B(\theta)$, which the optimizer combines with its internal state to determine a parameter update Δ. In each probe, we hold the targets fixed, specify a reference prediction state, derive the corresponding local response, and test that response in the implementation (Figure 1). Exact agreement provides a diagnostic reference for pointwise fit and, when attainable within the model class, for model-level analysis. For noisy future observations, conditional-risk optimality must instead be assessed relative to the specified statistical target (Gneiting, 2011; Gneiting & Raftery, 2007).

Keep the remaining error correctable. For a differentiable error component E, $-g_B$ is a first-order descent direction when $\langle \nabla E, g_B\rangle > 0$. For an actual optimizer step Δ, we instead evaluate the finite change $E(\theta+\Delta) - E(\theta)$. We use a constant forecast bias to isolate level error while preserving seasonal shape. More generally, consider the two-component field $g = a(S,T)\nabla S + b(S,T)\nabla T$, where S and T denote seasonal and trend errors. At a boundary where $\nabla S = 0$ and $\nabla T \neq 0$,

$$\langle \nabla T, g\rangle = b(S,T)\|\nabla T\|^2.$$

Thus, at this boundary, $-g$ decreases the trend error to first order if and only if $b(S,T) > 0$. We apply this criterion to the adaptive ratio, fixed, and softmax weighting rules.

Protect agreement against attainable errors. For a twice-differentiable scalar objective, a zero gradient at agreement establishes stationarity. Negative curvature identifies nearby errors that lower the objective, whereas a global bound can establish optimality even when the Hessian is degenerate. For the RI helper's full expected objective, these analyses identify the same sharp weight threshold. Let H_R denote the residual Hessian and J the residual Jacobian of a twice-differentiable model. At a stationary point in residual space, the parameter Hessian is $J^\top H_R J$. We identify attainable directions of negative curvature and test the predicted responses using finite perturbations and stochastic dynamics.

Derive controls with testable predictions. A positive coefficient floor tests whether a vanishing weight blocks error correction; an auxiliary weight below the curvature threshold tests the predicted

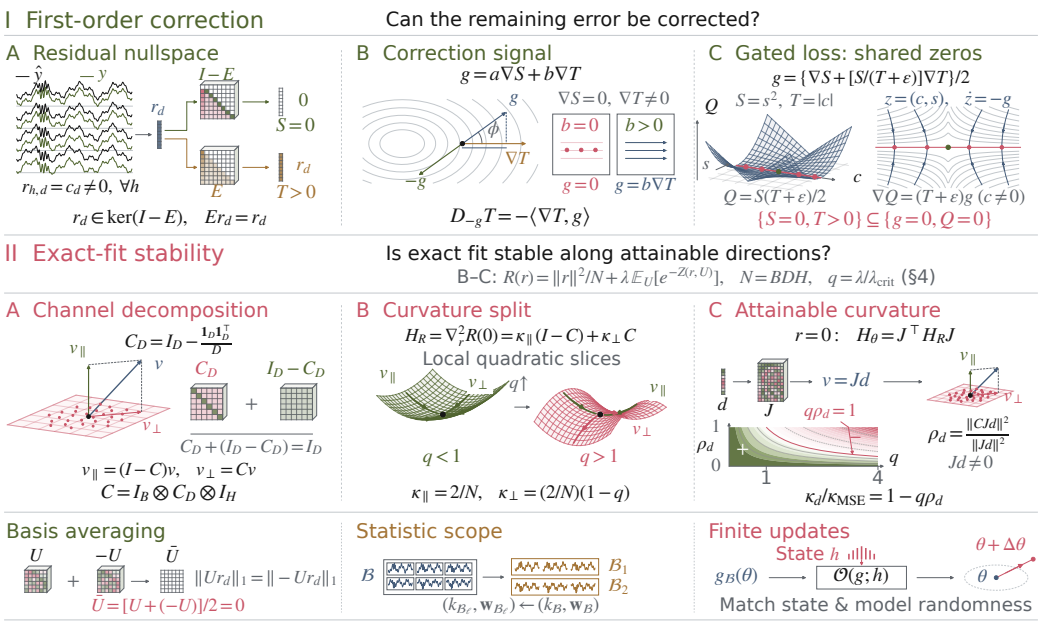


Figure 1: **From prediction states to learning responses.** Two complementary checks assess whether remaining errors receive corrective updates and whether exact fit is locally stable along attainable directions. The schematics connect residual structure and implemented updates to local curvature, with additional probes of basis averaging, statistic scope, and finite optimizer steps. All panels are analytic schematics.

local optimality of agreement. Matched random streams isolate the effects of statistic replacement from changes in dropout randomness. In the extension cases, controls for equivalent encodings and restoration of whole-batch statistics identify which computations alter the learning signal. Appendix F.1 presents four supporting perspectives: targets, representations, sample coupling, and field geometry. The main analysis follows the two prediction properties and the controls they motivate.

3 DBLoss: Retaining Correction After Partial Fit

DBLoss uses a linear exponential moving average (EMA) to decompose predictions and targets. Let S denote seasonal MSE and T trend MAE. With the adaptive ratio detached, the implemented update field is

$$G = \beta \nabla S + (1-\beta)\frac{S}{T+\epsilon}\nabla T.$$

The ratio is intended to align component scales adaptively, while detachment prevents backpropagation through it (Qiu et al., 2025, Section 3.3). We analyze this adaptive rule; the original paper's fixed-weight analysis concerns a different objective.

Exact seasonal fit can leave a level bias stationary. The EMA preserves sequences that are constant across the forecast horizon. For a residual $r = c\mathbf{1}$ with $c \neq 0$, this implies $S = 0$ and $\nabla S = 0$, so $G = 0$ despite positive MSE and trend error. Thus, the implemented field supplies no correction signal for this nonzero bias (Appendix A.2).

A scalar potential retains the same blind spot. For fixed $\beta \in (0,1)$ and $\epsilon > 0$, the field is nonconservative on smooth regions where the component gradients are linearly independent. On a two-dimensional DLinear parameter slice, the field Jacobian J_G satisfies $\|J_G - J_G^\top\|_F = 1.8313$ at the audited point. Finite differences and nonzero loop work corroborate nonconservativity (Appendix A.5). Nevertheless, on a fixed batch, a positive integrating factor converts G into the gradient of a scalar potential.

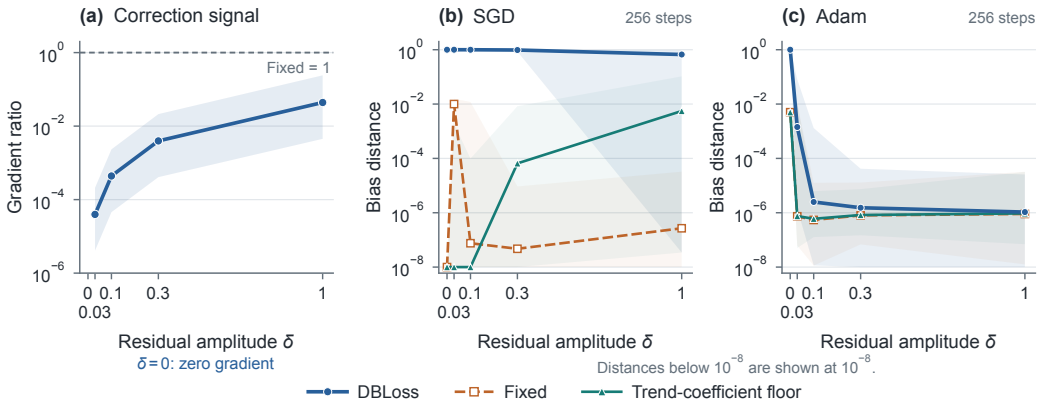


Figure 2: **Trend correction depends on weighting and optimizer choice.** Fifteen frozen DLinear models, with two initial bias offsets each (30 cases). (a) Initial bias-gradient magnitude relative to Fixed. (b–c) Distance to the trend-optimal bias after 256 steps. Lines show medians; bands show full ranges. $\delta = 1$ uses saved residuals plus a bias; $0 \le \delta < 1$ uses controlled, label-based residual interpolation. The boundary $\delta = 0$ is shown separately. Adam starts with fresh state.

Proposition 1 (Positive integrating factor)**.** *Fix a batch and constants β, ϵ, τ with $0 < \beta < 1$ and $\tau + \epsilon > 0$. Let $S, T \in C^1(U)$, where U is an open parameter domain, and suppose $T + \epsilon > 0$ on U. Set $p = (1-\beta)/\beta$ and define*

$$Q_{\beta,\tau} = \beta S \left(\frac{T+\epsilon}{\tau+\epsilon} \right)^p .$$

Then

$$\nabla Q_{\beta,\tau} = \left(\frac{T+\epsilon}{\tau+\epsilon} \right)^p G.$$

Along any solution of $\dot{\theta} = -G$ in U, $Q_{\beta,\tau}$ strictly decreases wherever $G \neq 0$.

Autograd checks across fifteen saved DLinear models, four batches per model, and three β values confirm this identity with a maximum relative error of 8.56×10^{-16}. The positive factor can vary across batches, so the averages of $\nabla Q_{\beta,\tau}$ and G need not be collinear (Appendix A.4).

At the constant-bias boundary, $Q_{\beta,\tau} = 0$ and $G = 0$. Since $S \ge 0$ implies $Q_{\beta,\tau} \ge 0$, these biased predictions are global minimizers in residual space. The direction-equivalent potential therefore retains the same correction blind spot.

The trend-correction field attenuates quadratically near the boundary. For a saved residual r_0, consider $r = \delta r_0 + c\mathbf{1}$ and optimize only c. Writing S_0 for the seasonal error of r_0, we have $S_\delta = \delta^2 S_0$ and $\partial_c S_\delta = 0$, so

$$G_c = (1-\beta) \frac{\delta^2 S_0}{T+\epsilon} \, \partial_c T.$$

For fixed $c \neq 0$ and $S_0 > 0$, the magnitude of this field decays quadratically as $\delta \to 0$. The Fixed control retains $(1-\beta)\partial_c T$. The coefficient test in Section 2 identifies the mechanism: at $S = 0$, the trend coefficient vanishes even when $\partial_c T \neq 0$.

Using validation residuals from fifteen saved DLinear models across five sources, we freeze the backbones and, for $\delta < 1$, interpolate predictions toward fixed targets before adding bias. The scalar adapter starts one unit above or below the trend-optimal bias. For each of five amplitudes, we run 256 updates at learning rate 0.05 with SGD or fresh-state Adam. A control floors the trend coefficient at 0.1 to test whether retaining the signal restores correction.

At $\delta = 0$, both optimizers leave DBLoss's initial bias distance at one. At $\delta = 0.03$, median distances to the trend-optimal bias under SGD are 0.99975 for DBLoss and 0.01001 for Fixed; Adam reduces DBLoss's median to 0.00143. At $\delta = 1$, Adam reaches medians of 1.06×10^{-6} for DBLoss and

9.04×10^{-7} for Fixed, while SGD leaves DBLoss at 0.6664 (Figure 2; Appendix A.3). In these runs, Adam's moment normalization (Kingma & Ba, 2015) largely offsets attenuation of a nonzero correction field; a zero coefficient prevents correction under either optimizer.

Detachment need not produce a nonconservative field. PMLF uses detached softmax weights $w = \mathrm{softmax}(\beta\,\mathrm{sg}([S, T]))$ (Chen et al., 2025):

Proposition 2 (Detached softmax potential)**.** *For differentiable S, T and fixed $\beta > 0$, the detached softmax field satisfies*

$$w_S \nabla S + w_T \nabla T = \nabla \Phi_\beta, \qquad \Phi_\beta = \frac{1}{\beta} \log\left(\frac{e^{\beta S} + e^{\beta T}}{2}\right).$$

This normalized log-sum-exp is a potential for the executed field. Differentiating the displayed weighted average $w_S S + w_T T$ through its weights generally produces a different gradient. At $S = 0$, we have $\nabla S = 0$ and $w_T > 0$, so the negative field decreases T to first order whenever $\nabla T \neq 0$.

The practical check is whether the remaining component's gradient retains a positive coefficient after the other component is fitted. Fixed weights, the diagnostic floor, and the softmax rule all satisfy this condition. Their training behavior also depends on coefficient scale and optimizer dynamics.

4 RI-Loss: A sharp exact-fit optimality boundary

RI-Loss motivates residual–noise dependence as a way to extract predictable structure (Wang et al., 2026c). Its published DLinear helper computes a biased centered-kernel statistic on $[B, D, H]$ residuals, using channels as samples and horizon positions as features. The pinned default training entry uses MSE; our controlled runner explicitly activates this helper (Appendix B.4).

Proposition 3 (Finite-sample conditional expectation)**.** *Let $D > 1$, $C_D = I - \mathbf{1}\mathbf{1}^\top/D$, and fix a residual kernel K. For iid surrogate rows $U_i \sim \mathrm{Unif}([0,1]^H)$ independent of K, define $a_H = \mathbb{E}\exp(-\|U_i - U_j\|^2)$ for $i \neq j$. For each example, the published statistic satisfies*

$$\mathbb{E}[\widehat{\mathrm{HSIC}} \mid K] = \frac{(1 - a_H)\,\mathrm{tr}(C_D K C_D)}{(D-1)^2}.$$

For unit-diagonal K, $\mathrm{tr}(C_D K C_D) = D - D^{-1}\sum_{ij} K_{ij}$. This expectation measures dispersion among residual-channel trajectories and vanishes when they coincide, including at exact agreement. Population independence therefore does not imply a zero finite-sample expectation. At unit kernel scale, $a_H = a_1^H$ with $a_1 = 0.8615277\ldots$ and $a_{96} \simeq 6.11 \times 10^{-7}$.

From dispersion to global optimality. Let $r \in \mathbb{R}^{B \times D \times H}$, $N = BDH$, and let C center channels within each example. The *full expected objective* is $R(r) = \|r\|_F^2/N + \lambda \mathbb{E}_U[e^{-Z(r,U)}]$, where Z is the batch-summed statistic computed with a unit-scale Gaussian residual kernel.

Proposition 4 (Sharp exact-fit optimality boundary)**.** *For positive integers B, H, integer $D > 1$, and $\lambda \geq 0$, zero residual is the unique global minimizer of R exactly when $\lambda \leq \lambda_{\mathrm{crit}}$, where*

$$\lambda_{\mathrm{crit}} = \frac{(D-1)^2}{2BDH(1 - a_H)}.$$

Above this threshold, it is a saddle in residual space. For any direction v, with $c_H = (1 - a_H)/(D-1)^2$,

$$R(tv) - R(0) = t^2\left(\frac{\|v\|_F^2}{N} - 2\lambda c_H \|Cv\|_F^2\right) + O(t^4).$$

At equality, centered directions have zero quadratic curvature, yet zero residual remains strictly globally optimal.

At agreement, $Z = \nabla Z = 0$ for every draw, so $\mathbb{E}[e^{-Z}]$ and $e^{-\mathbb{E}Z}$ share the same quadratic term. Every stochastic loss is stationary there, but above the threshold the expected objective has negative centered curvature and positive channel-shared curvature. For $B = 32$, $H = 96$, λ_{crit} is 8.37×10^{-4} for $D = 7$ and 3.10×10^{-3} for $D = 21$, both below the helper weight 10. In a twice-differentiable

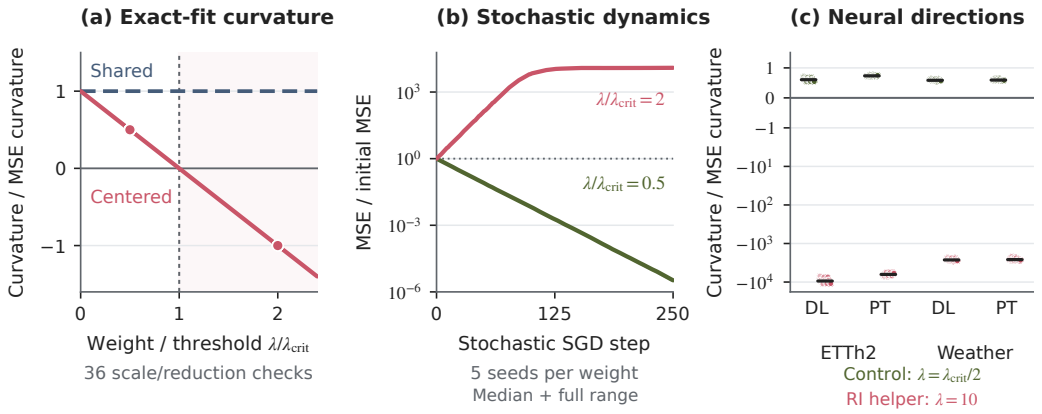


Figure 3: **Curvature predicts local responses in output and parameter spaces.** (a) Exact-fit curvature and all 36 scale/reduction checks. (b) Five stochastic bias-model runs per weight, with median lines and full-range bands. (c) Source/MSE curvature ratios along all 96 full-parameter directions at constructed agreement, with a positive-curvature control below threshold. DL/PT denote DLinear/PatchTST. Targets are zero in (b) and frozen model predictions in (c). In (c), black ticks mark medians; the symmetric log scale is linear on $[-1, 1]$.

model, the parameter Hessian at agreement is $J^\top H_R J$, so negative curvature must be attainable through its parameters.

Float64 source Hessian-vector checks match the fixed-surrogate formula with maximum absolute error 8.88×10^{-16}. Further checks cover three (B, D, H) settings, three residual-kernel scales, and sum/mean batch reductions. In a realizable two-channel bias model, five stochastic SGD trajectories per weight start from a small nonzero bias. After 250 steps, the final/initial MSE ratio has a median of about 3.3×10^{-6} at $\lambda/\lambda_{\text{crit}} = 1/2$ and exceeds 1.18×10^4 in every run at ratio 2 (Figure 3). Appendix B.2 gives the proof, the general-scale threshold, and all checks.

Test attainability in neural parameter space. For a parameter direction d with $v = Jd \neq 0$, let $\rho_d = \|Cv\|^2/\|v\|^2$. At exact agreement, its expected curvature relative to MSE is

$$\frac{d^\top J^\top H_R J d}{2\|Jd\|^2/N} = 1 - \frac{\lambda}{\lambda_{\text{crit}}}\rho_d.$$

We test twelve saved DLinear/PatchTST models on ETTh2/Weather, using four validation input batches per model and two prespecified full-parameter directions per batch. The saved models' predictions define fixed targets, giving exact agreement at the saved parameters. With a fixed surrogate draw per direction, all 96 directions have negative source curvature at $\lambda = 10$ and positive curvature at $\lambda = \lambda_{\text{crit}}/2$. Using the same draws, perturbations $\pm 10^{-3} d$ lower the full loss at $\lambda = 10$ and raise it under the control. Figure 3c and Appendix B.2.1 report all directions and source/pullback checks.

The auxiliary remains active at fitted states. Across 192 validation batches from twelve fitted RI models, the statistic's maximum observed standard deviation over 64 surrogate draws per batch is 2.10×10^{-6} at $H = 96$. On Weather, median auxiliary-to-MSE gradient-norm ratios are 0.890 for DLinear and 1.366 for PatchTST; the corresponding gradient cosines are -0.622 and -0.333. The negative combined raw gradient would increase MSE to first order on $13/48$ and $20/48$ batches, respectively. Small sampling variation thus coexists with a substantial opposing gradient (Appendix B.3).

Match randomness and measure the optimizer step. Replacing Z with its conditional expectation retains the dispersion auxiliary, but omitting surrogate draws can alter subsequent dropout randomness. We run five arms with three seeds for each ETTh2/Weather $\times$ DLinear/PatchTST combination: sixty runs of 256 Adam updates. Two controls match dropout randomness within each pair at every forward

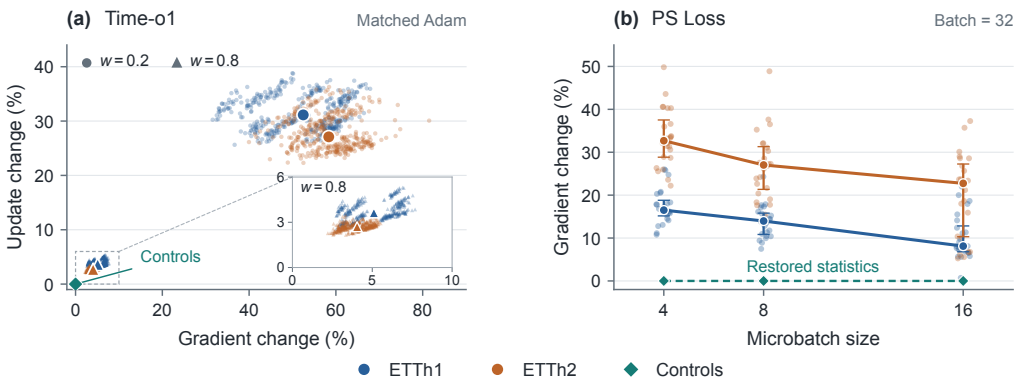


Figure 4: **Learning signals can change at fixed predictions and examples.** (a) Time-o1: relative changes in full parameter gradients and state-matched Adam updates across all sign interventions, with medians and an inset for the stronger MSE weight $w = 0.8$. Shared-sign and channelwise-projection controls give zero paired changes within their respective encodings. (b) PS Loss: relative changes in full parameter gradients across all 24 batches per dataset/partition, with medians and interquartile ranges. Restoring whole-batch statistics recovers the gradient to numerical precision.

call by retaining surrogate draws or using a separate surrogate generator. Unmatched replacement changes mean PatchTST validation MSE over three seeds by -0.436% on ETTh2 and -0.045% on Weather. In both matched contrasts, relative changes in mean validation MSE remain within $\pm 0.000671\%$ across all four combinations. These controls trace the short-run differences primarily to dropout randomness while retaining the dispersion auxiliary.

For measured Adam steps on Weather, same-mask batch MSE increases after two of 24 stream-isolated RI updates and one of 24 MSE updates, each evaluated at the method's own states. Appendix B.4.1 reports all arms and measured steps.

The ratio $\lambda/\lambda_{\text{crit}}$ measures auxiliary strength relative to the exact-fit boundary, accounting for batch size and reduction, kernel denominator, horizon, and channel count. Matching the auxiliary's value or sampling variance does not ensure matching curvature. The boundary identifies the auxiliary weights for which exact fit is the unique global minimizer of the full expected objective.

5 Preserving the Signal Through the Training Computation

The two core diagnoses connect a prediction property to the computation that changes it. The same approach tests whether equivalent encodings or different partitions of the same examples preserve supervision. These interventions trace their effects through complete model updates (Figure 4).

Equivalent bases can cancel the auxiliary signal. Time-o1's shared-basis mode averages channelwise PCA bases before computing transformed MAE (Wang et al., 2025b). Equivalent full-rank encodings (U, U) and $(U, -U)$ average to U and zero, respectively. The second encoding gives zero auxiliary gradient for every residual. In a constructed example, sign flips of the published helper's fitted bases reproduce this cancellation (Proposition 7). We evaluate the complete objective $L = w, \text{MSE} + (1 - w)L_{\text{PCA}}$ on ETTh1/ETTh2, fitting PCA only on training data. Six saved DLinear models receive sixteen updates separately at each fixed $w \in 0.2, 0.8$, yielding twelve adapted states. Four batches and 32 sign encodings per state give median relative changes of $52.48\%/58.35\%$ in full parameter gradients at $w = 0.2$ and $31.13\%/27.10\%$ in state-matched Adam updates (ETTh1/ETTh2). At $w = 0.8$, the Adam changes are $3.63\%/2.70\%$. $w = 0.2$ is transferred from a published configuration to this controlled ETT/DLinear setting. Shared-sign and channelwise-projection controls preserve every paired gradient and update within their respective encodings. The MSE anchor retains residual sensitivity; these controls isolate how basis averaging alters auxiliary supervision (Appendix E.1.1).

Restoring batch statistics restores the PS trajectory. PS Loss determines patch length and detached component weights from batch statistics (Kudrat et al., 2025). At six saved DLinear states, we partition each 32-example batch into groups of eight. Recomputing both quantities locally produces median relative changes of $13.97\%/27.03\%$ in the full parameter gradient on ETTh1/ETTh2. Restoring both whole-batch quantities recovers the gradient to numerical precision. Across eighteen 128-update continuations, all six restored trajectories also match their whole-batch counterparts to numerical precision. Local recomputation has mixed effects on validation MSE relative to whole-batch training (-1.02% to $+0.76\%$). This intervention shows how the scope of batch statistics defines the learning rule (Appendix E.2).

The checks distinguish preserved and suppressed signals. On all twelve tested inputs, AliO's detached field for overlapping forecasts (Park et al., 2025) has a nonsymmetric Jacobian, yet its negative is a strict output-MSE descent direction (Appendix F.10). For fixed-graph CvLoss, a positive MSE anchor makes negative subgradients descent directions for squared residual norm away from agreement (Ding et al., 2026). In AMRC's published ESP helper, normalization over two-example batches suppresses the relational signal for fixed nondegenerate pairs (Liang et al., 2025). The CvLoss and AMRC findings come from formula analysis and source-function checks, respectively (Appendices F.14 and F.13). Table 27 summarizes validation depth across the ten audited objectives.

6 Forecasting evidence under specified budgets

Comparison design. We evaluate five held-out sources (Electricity, Traffic, Exchange, Solar, and ILI) with chronological 70/10/20 splits and train-only channel scaling. DLinear (Zeng et al., 2023) covers all five; pinned PatchTST (Nie et al., 2023) and a structurally matched iTransformer-style backbone (Liu et al., 2024) each cover three. Sources are comparison units; seeds and architectures are repeated conditions. The five-source comparison covers MSE, MAE, Fixed, DBLoss, and Ψ_τ. Validation selects learning rates; DBLoss fixes its EMA and mixture parameters at $\alpha = 0.2, \beta = 0.5$. RI's four-anchor studies appear in Section 4. These settings support mechanism comparisons; loss-specific tuning addresses broader performance questions.

The initial budget caps training at ten epochs with patience two, and transfer backbones at 256 updates per epoch. All 125 primary, 90 transfer, six limited-diagnostic, and 399 stress fits completed. Q_τ preserves DBLoss's within-batch field direction; Ψ_τ preserves its trend coefficient but can make the seasonal coefficient negative (Table 2). Detailed Ψ_τ results appear in the appendix.

Table 1: Five-source clean-MSE effects on DLinear. Negative values favor the first method. The source-bootstrap interval refers to the mean; the median is reported separately.

Contrast	Mean (%)	Mean 95% interval	Median (%)	Better
DBLoss vs Fixed	+1.413	[-0.483, +3.980]	-0.140	3/5
DBLoss vs MSE	+3.223	[-0.035, +8.933]	+0.907	1/5
Fixed vs MSE	+1.726	[-0.774, +5.020]	+1.048	2/5

Relative to Fixed, DBLoss's test MSE change ranges from -0.900% on Solar to $+6.228\%$ on ILI, favoring DBLoss on three of five sources. The mean and median are $+1.413\%$ and -0.140% (Figure 7; Appendix D). The sign also varies with horizon, lookback, and batch size; the full batch-size sweep spans -4.008% to $+18.380\%$. In a separate post-hoc experiment under a frozen protocol, Q_τ yields a median MSE change of -0.050% relative to DBLoss, improving on three of five sources (Appendix A.7).

Extending the budget where it binds. On Electricity–PatchTST, validation selects epoch ten for all three DBLoss seeds and two each for MSE and Fixed. We rerun all three objectives with three seeds each, retaining selected learning rates and raising the cap to thirty epochs with patience five. Each new trajectory supplies the best validation checkpoint within ten epochs and within the extended budget; all fits finish before test evaluation. DBLoss versus Fixed changes from $+0.331\%$ at the new first-ten checkpoints to $+1.867\%$ at the extended checkpoints. The new cap is reached in 3/9 runs and selected as the best epoch in 1/9. Fixed's mean test MSE falls by 1.924%, versus 0.422%

for DBLoss and 0.382% for MSE (Table 21). Comparisons between these unchanged update rules therefore depend on budget. Appendix D.2 reports all trajectories and checkpoint selections.

7 Related work

Forecasting objectives encode temporal alignment, geometry, constraints, and relative error (Cuturi & Blondel, 2017; Le Guen & Thome, 2019; Yu et al., 2025; Hounie et al., 2024; Vanegas Arias et al., 2025). Elicitation and proper scoring characterize optimal predictions under expected loss (Gneiting, 2011; Gneiting & Raftery, 2007). Generative objectives model distributions over future trajectories (Liu et al., 2025; Zhang et al., 2026), while analyses of conditional uncertainty distinguish point accuracy from marginal realism (Green et al., 2026). With targets fixed, we connect specified prediction properties to objective geometry and the learning response of implemented point-forecast losses.

Soft-DTW divergence corrects self-alignment bias (Blondel et al., 2021), and saddle-point analysis distinguishes stationarity from local optimality (Dauphin et al., 2014). Adaptive gradient balancing (Chen et al., 2018) and finite-sample kernel statistics (Gretton et al., 2007) inform our analysis of weighted updates and sampled auxiliaries. In heteroscedastic regression, inverse-variance weighting can suppress mean-prediction gradients for poorly fitted examples (Seitzer et al., 2022). Our DBLoss analysis connects a vanishing correction coefficient to biased global minima of a fixed-batch potential obtained through a positive integrating factor. For the RI helper, finite-sample residual dispersion yields a sharp global-optimality boundary for the full expected objective. Neural perturbations test whether the predicted negative curvature is attainable at constructed agreement in the tested models.

Behavioral testing probes specific model properties (Ribeiro et al., 2020), while studies of federated optimization and drifting models examine when learning fields admit scalar potentials (Charles & Rush, 2022; Franz et al., 2026). Forecasting analyses study expressivity, component necessity, and implicit bias (Toner & Darlow, 2024; Tan et al., 2024; Yu et al., 2026). We derive learning responses from specified prediction properties, then test gradients, optimizer steps, and trajectories under controlled interventions. Equivalent encodings and alternative partitions of identical examples test whether the computation preserves supervision. Following comparative evaluation and reproducibility principles (Demšar, 2006; Benavoli et al., 2017; Pineau et al., 2021), we report complete forecasting results and distinguish seed variation from uncertainty across sources.

8 Scope and conclusion

We derive conditions for error correction and exact-fit optimality. At DBLoss's fitted seasonal boundary, the trend coefficient vanishes, leaving nonzero level bias stationary. These biased forecasts globally minimize a direction-equivalent fixed-batch potential. For the RI helper's full expected objective, zero residual is the unique global minimizer for $0 \leq \lambda \leq \lambda_{\mathrm{crit}}$ and a saddle in residual space above this threshold. Strict global optimality persists at the critical weight despite degenerate centered curvature. RI's global result follows from an analytic bound; controlled tests assess local geometry and training responses.

A positive trend-coefficient floor restores a first-order level-correction signal at the fitted seasonal boundary. Adam can compensate for small nonzero fields in controlled recovery tests. For RI, replacing the batch sum with a mean inside the exponential multiplies λ_{crit} by B; the threshold also scales proportionally with the residual Gaussian kernel's denominator, with other quantities fixed.

For new composite objectives, specify a prediction state and derive a control with a predicted effect on its learning signal. Time-o1 and PS extend these checks to encodings and batch statistics while preserving predictions or effective examples. These mechanism tests complement forecasting comparisons under explicit training budgets.

Neural agreement targets are constructed, and DBLoss's small-residual probes use label-based interpolation. The frequency of these states in natural training and their long-run forecasting effects remain open. Evidence depth varies across the ten audit cases; the additional AMRC check is limited to the ESP embedding input. With five held-out sources, the minimum attainable two-sided exact signed-rank p-value is 0.0625, and source-bootstrap coverage remains approximate. The proofs and controls make prediction properties directly testable in the implemented training computation.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing and improve the clarity and readability of the manuscript. The authors take full responsibility for the final content and accuracy of the submission.

ETHICS AND REPRODUCIBILITY STATEMENT

The forecasting experiments use publicly available time-series benchmarks. The supplementary material provides proofs, source-code snapshots, complete experimental records, and instructions for accessing the data and reproducing the analyses. An anonymous repository[2] also provides browsable code and reproduction instructions. The accompanying evidence index links reported findings to the corresponding scripts, protocols, and results.

REFERENCES

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6(58):1705–1749, 2005. URL https://www.jmlr.org/papers/v6/banerjee05b.html.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: A tutorial for comparing multiple classifiers through Bayesian analysis. *Journal of Machine Learning Research*, 18(77):1–36, 2017. URL https://www.jmlr.org/papers/v18/16-305.html.

Mathieu Blondel, Arthur Mensch, and Jean-Philippe Vert. Differentiable divergences between time series. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3853–3861, 2021. URL https://proceedings.mlr.press/v130/blondel21a.html.

Zachary Charles and Keith Rush. Iterated vector fields and conservatism, with applications to federated learning. In *Proceedings of the 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pp. 130–147, 2022. URL https://proceedings.mlr.press/v167/charles22a.html.

Xinghong Chen, Weilin Wu, Kunping Yang, and Guannan Chen. PMLF: A physics-guided multiscale loss framework for structurally heterogeneous time series. In *Advances in Neural Information Processing Systems*, volume 38, pp. 58505–58530, 2025. doi: 10.52202/085713-1960. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/54d2d38a56a74387d5916ee40e462295-Abstract-Conference.html.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Exploring_Simple_Siamese_Representation_Learning_CVPR_2021_paper.html.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 794–803, 2018. URL https://proceedings.mlr.press/v80/chen18a.html.

Marco Cuturi and Mathieu Blondel. Soft-DTW: A differentiable loss function for time-series. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings*

[2] https://anonymous.4open.science/r/forecasting-loss-audit-1FAB/

of Machine Learning Research, pp. 894–903, 2017. URL https://proceedings.mlr.press/v70/cuturi17a.html.

Yann N. Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems*, volume 27, 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/hash/04192426585542c54b96ba14445be996-Abstract.html.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1):1–30, 2006. URL https://www.jmlr.org/papers/v7/demsar06a.html.

Kuiye Ding, Yifan Hu, Hanchen Wang, and Hao Xue. Multivariate time series forecasting needs cross variable loss. *arXiv preprint arXiv:2608.05742*, 2026. URL https://arxiv.org/abs/2608.05742.

Jiayu Fang, Xuande Liu, Sangsha Fang, Zhen Tian, Hongwei Ma, Zhiqi Shao, and Junbin Gao. Information geometry loss for time series forecasting. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://icml.cc/virtual/2026/poster/66595.

Tobias Fissler and Johanna F. Ziegel. Higher order elicitability and Osband's principle. *The Annals of Statistics*, 44(4):1680–1707, 2016. doi: 10.1214/16-AOS1439. URL https://doi.org/10.1214/16-AOS1439.

Tobias Fissler and Johanna F. Ziegel. Correction note: Higher order elicitability and Osband's principle. *The Annals of Statistics*, 49(1):614, 2021. doi: 10.1214/20-AOS2014. URL https://doi.org/10.1214/20-AOS2014.

Leonard T. Franz, Sebastian Hoffmann, Tim Weiland, Bernhard Schölkopf, and Georg Martius. Drifting fields are not conservative. *arXiv preprint arXiv:2604.06333*, 2026. URL https://arxiv.org/abs/2604.06333.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011. doi: 10.1198/jasa.2011.r10138. URL https://doi.org/10.1198/jasa.2011.r10138.

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

Tilmann Gneiting and Johannes Resin. Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams, and coefficient of determination. *Electronic Journal of Statistics*, 17(2):3226–3286, 2023. doi: 10.1214/23-EJS2180. URL https://doi.org/10.1214/23-EJS2180.

Riku Green, Zahraa S. Abdallah, and Telmo M. Silva Filho. Expectations vs. realities: The cost of MSE-optimal forecasting under conditional uncertainty. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 1298–1309, 2026. doi: 10.1145/3770855.3818087. URL https://arxiv.org/abs/2606.04342.

Arthur Gretton, Kenji Fukumizu, Choon H. Teo, Le Song, Bernhard Schölkopf, and Alex J. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, volume 20, 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/hash/d5cfead94f5350c12c322b5b664544c1-Abstract.html.

Theivaprakasham Hari, Yanan Xin, Winnie Daamen, Serge Paul Hoogendoorn, and Sascha Hoogendoorn-Lanser. Asymmetric peak-aware loss for peak-critical time series forecasting. *arXiv preprint arXiv:2607.14871*, 2026. URL https://arxiv.org/abs/2607.14871.

Ignacio Hounie, Javier Porras-Valenzuela, and Alejandro Ribeiro. Loss shaping constraints for long-term time series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19062–19084, 2024. URL https://proceedings.mlr.press/v235/hounie24a.html.

Sheo yon Jhin and Noseong Park. SALT: Structure-aligned learning for time-series forecasting, 2026. URL https://openreview.net/forum?id=d3zrIHukon. Public ICLR 2026 submission.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Agustinus Kristiadi, Felix Dangel, and Philipp Hennig. The geometry of neural nets' parameter spaces under reparametrization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 17669–17688, 2023. doi: 10.52202/075280-0775. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/395371f778ebd4854b88521100af30ad-Abstract-Conference.html.

Dilfira Kudrat, Zongxia Xie, Yanru Sun, Tianyu Jia, and Qinghua Hu. Patch-wise structural loss for time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 31841–31859, 2025. URL https://proceedings.mlr.press/v267/kudrat25a.html.

Vincent Le Guen and Nicolas Thome. Shape and time distortion loss for training deep time series forecasting models. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/466accbac9a66b805ba50e42ad715740-Abstract.html.

Hyunwook Lee, Chunggi Lee, Hongkyu Lim, and Sungahn Ko. TILDE-Q: A transformation invariant loss function for time-series forecasting. *arXiv preprint arXiv:2210.15050*, 2022. URL https://arxiv.org/abs/2210.15050.

Seunghan Lee, Jaehoon Lee, Jun Seo, Junhyeok Kang, Sangjun Han, Sungdong Yoo, Minjae Kim, Tae Yoon Lim, Dongwan Kang, Hwanil Choi, Soonyoung Lee, and Wonbin Ahn. Beyond magnitude and shape: A direction-aware loss for time series forecasting. *arXiv preprint arXiv:2608.01857*, 2026. URL https://arxiv.org/abs/2608.01857.

Renzhao Liang, Sizhe Xu, Chenggang Xie, Jingru Chen, Feiyang Ren, Shu Yang, and Takahiro Yabe. Abstain mask retain core: Time series prediction by adaptive masking loss with representation consistency. In *Advances in Neural Information Processing Systems*, volume 38, pp. 99445–99471, 2025. doi: 10.52202/085713-3324. URL https://papers.nips.cc/paper_files/paper/2025/hash/8f7ce7c30c62e92cad90cfaefdb5962c-Abstract-Conference.html.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=JePfAI8fah.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39295–39317, 2025. URL https://proceedings.mlr.press/v267/liu25be.html.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2211.14730.

Jinwoo Park, Hyeongwon Kang, and Pilsung Kang. Not all variables agree: Reliability-aware variable-wise gradient surgery for multivariate time-series forecasting. *arXiv preprint arXiv:2609.08554*, 2026. URL https://arxiv.org/abs/2609.08554.

Kwangryeol Park, Jaeho Kim, and Seulki Lee. AliO: Output alignment matters in long-term time series forecasting. In *Advances in Neural Information Processing Systems*, volume 38, pp. 119563–119608, 2025. doi: 10.52202/085713-3992. URL https://openreview.net/forum?id=AuOZDp4gy7.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164):1–20, 2021. URL https://www.jmlr.org/papers/v22/20-303.html.

Jean Ponce, Basile Terver, Martial Hebert, and Michael Arbel. Dual perspectives on non-contrastive self-supervised learning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=f5MC1G6XhB.

Xiangfei Qiu, Xingjian Wu, Hanyin Cheng, Xvyuan Liu, Chenjuan Guo, Jilin Hu, and Bin Yang. DBLoss: Decomposition-based loss function for time series forecasting. In *Advances in Neural Information Processing Systems*, volume 38, pp. 27741–27768, 2025. doi: 10.52202/085713-0936. URL https://papers.nips.cc/paper_files/paper/2025/hash/2823c4f838727564c108c85d7e6e8507-Abstract-Conference.html.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aPOpXlnV1T.

Tianyi Shi, Zhu Meng, Yue Chen, Siyang Zheng, Fei Su, Jin Huang, Changrui Ren, and Zhicheng Zhao. OLMA: One loss for more accurate time series forecasting. *arXiv preprint arXiv:2505.11567*, 2025. URL https://arxiv.org/abs/2505.11567.

Mingtian Tan, Mike A. Merrill, Vinayak Gupta, Tim Althoff, and Thomas Hartvigsen. Are language models actually useful for time series forecasting? In *Advances in Neural Information Processing Systems*, volume 37, pp. 60162–60191, 2024. doi: 10.52202/079017-1922. URL https://doi.org/10.52202/079017-1922.

William Toner and Luke Nicholas Darlow. An analysis of linear time series forecasting models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48404–48427, 2024. URL https://proceedings.mlr.press/v235/toner24a.html.

Sergio M. Vanegas Arias, Lasse Lensu, and Fredy Ruiz Palacios. A symmetric relative-error loss function for intermittent multiscale signal modelling. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 6281–6288, 2025. doi: 10.24963/ijcai.2025/699. URL https://doi.org/10.24963/ijcai.2025/699.

Hao Wang, Lichen Pan, Yuan Shen, Zhichao Chen, Degui Yang, Yifei Yang, Sen Zhang, Xinggao Liu, Haoxuan Li, and Dacheng Tao. FreDF: Learning to forecast in the frequency domain. In *International Conference on Learning Representations*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1457fb1e5d72cdc4ecd88bc10f916095-Abstract-Conference.html.

Hao Wang, Licheng Pan, Zhichao Chen, Xu Chen, Qingyang Dai, Lei Wang, Haoxuan Li, and Zhouchen Lin. Time-o1: Time-series forecasting needs transformed label alignment. In *Advances in Neural Information Processing Systems*, volume 38, pp. 8660–8690, 2025b. doi: 10.52202/085713-0294. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/0cd62dea69635f4c5b569848267fe5a8-Abstract-Conference.html.

Hao Wang, Licheng Pan, Yuan Lu, Zi Chan, Tianqiao Liu, Shuting He, Zhixuan Chu, Qingsong Wen, Haoxuan Li, and Zhouchen Lin. Quadratic direct forecast for training multi-step time-series forecast models. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/d56a1a2b24b1906345fc2cd8276b11ed-Abstract-Conference.html.

Hao Wang, Licheng Pan, Yuan Lu, Zhixuan Chu, Xiaoxi Li, Shuting He, Zi Chan, Qingsong Wen, Haoxuan Li, and Zhouchen Lin. DistDF: Time-series forecasting needs joint-distribution Wasserstein alignment. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/d591ef4c631b33ac4e08dd88f1ee1dff-Abstract-Conference.html.

Jieting Wang, Xiaolei Shang, Feijiang Li, and Furong Peng. RI-Loss: A learnable residual-informed loss for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26277–26284, 2026c. doi: 10.1609/aaai.v40i31.39832. URL https://ojs.aaai.org/index.php/AAAI/article/view/39832/43793.

Annan Yu, Danielle Maddix, Boran Han, Xiyuan Zhang, Abdul Fatir Ansari, Oleksandr Shchur, Christos Faloutsos, Andrew Gordon Wilson, Michael W. Mahoney, and Bernie Wang. Understanding the implicit biases of design choices for time series foundation models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/753d9584b57ba01a10482f1ea7734a89-Abstract-Conference.html.

Mingyang Yu, Xiahui Guo, Peng Chen, Zhenkai Li, and Yang Shu. Towards measuring and modeling geometric structures in time series forecasting via image modality. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2390–2398, 2025. doi: 10.1145/3746027.3754794. URL https://doi.org/10.1145/3746027.3754794.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow Twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320, 2021. URL https://proceedings.mlr.press/v139/zbontar21a.html.

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. doi: 10.1609/aaai.v37i9.26317. URL https://ojs.aaai.org/index.php/AAAI/article/view/26317.

Yunhao Zhang, Wenyao Hu, Jiale Zheng, Lujia Pan, and Junchi Yan. MMPD: Diverse time series forecasting via multi-mode patch diffusion loss. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=NEUgHT8dvH.

A DBLoss: theory, recovery, and scalar controls

This appendix connects the field identities and biased stationary boundary to the recovery experiments and scalar controls. Direction statements concern fields before optimizer transformations and retain the specified batch and smoothness conditions.

A.1 Field geometry and the direction-equivalent potential

Proposition 5 (Jacobian antisymmetry and the rank condition)**.** *For fixed β, ϵ, $S, T \in C^2(U)$, and $T + \epsilon > 0$,*

$$J_G - J_G^\top = \frac{1-\beta}{T+\epsilon}(\nabla T \nabla S^\top - \nabla S \nabla T^\top).$$

For $\beta \neq 1$, this difference is nonzero exactly where $\operatorname{rank}[\nabla S, \nabla T] = 2$.

Table 2: Properties preserved by the decomposition controls. Statements concern the pre-optimizer field on a smooth stratum with $S \geq 0$, $T + \epsilon > 0$, and $0 < \beta < 1$.

Control	Preserved property	Changed property or boundary
Fixed	Component losses	Removes the adaptive ratio and retains a positive trend coefficient at $S = 0$.
No detach	Displayed expression	Differentiates through the ratio, scaling the trend coefficient by $\epsilon/(T + \epsilon)$.
Ψ_τ	Trend coefficient	Can make the seasonal coefficient negative.
$Q_{\beta,\tau}$	Field direction	Rescales the field by a positive, state-dependent factor and retains the zero-seasonal-error blind spot.

Proof of Proposition 5. Write $s = \nabla S$, $t = \nabla T$, and $a = 1 - \beta$. Differentiating $G = \beta s + aS(T + \epsilon)^{-1}t$ gives symmetric Hessian terms and

$$\frac{a}{T+\epsilon} ts^\top - \frac{aS}{(T+\epsilon)^2} tt^\top.$$

Subtracting the transpose cancels the Hessian terms and $tt^\top$, leaving the stated outer-product difference. Moreover,

$$\|J_G - J_G^\top\|_F^2 = \frac{2a^2}{(T+\epsilon)^2} \left(\|s\|_2^2 \|t\|_2^2 - (s^\top t)^2 \right).$$

The Gram determinant in parentheses is positive exactly when s and t are linearly independent, proving the rank condition.

Proof of Proposition 1. Set $a = ((T + \epsilon)/(\tau + \epsilon))^p > 0$. The chain rule gives $\nabla(\beta S a) = \beta a \nabla S + \beta S p a \nabla T/(T + \epsilon) = aG$. Along a differentiable fixed-batch solution of $\dot{\theta} = -G$, $dQ/dt = -a\|G\|^2 \leq 0$, with strict decrease wherever $G \neq 0$. If G is locally Lipschitz, the flows of $-G$ and $-\nabla Q$ trace the same orbits under a positive time change on their common domains of existence. This equivalence does not imply global existence or convergence. For the released MAE component, classical smooth-flow statements apply within sign strata. The algebraic identity also holds with the common autograd subgradient choice at a kink. For stochastic batches, a_B depends on the batch and parameters, so $\mathbb{E}[a_B G_B]$ need not be collinear with $\mathbb{E}[G_B]$. Finite step sizes and adaptive optimizer states further separate agreement of field directions from agreement of training trajectories.

A.2 Direction preservation does not remove biased stationary points

Let E be the linear EMA operator initialized with its first input. It preserves horizon-constant sequences, so $E\mathbf{1} = \mathbf{1}$. For any fixed target y, consider $\hat{y} = y + c\mathbf{1}$, where c may vary across examples and channels but is constant over the horizon. The seasonal residual $(I - E)(\hat{y} - y)$ vanishes, giving $S = 0$ and $\nabla_{\hat{y}} S = 0$. Hence $G = 0$ and $Q_{\beta,\tau} = 0$, even when $T = \text{mean}\,|c| > 0$ and MSE is positive. The field also vanishes in parameter space at any model state realizing these predictions. Under Proposition 1, with $\epsilon > 0$ and $S, T \geq 0$, Q is nonnegative, so these biased predictions are global minimizers of the fixed-batch scalar objective.

This construction extends the constant-residual certificate obtained from the unchanged DBLoss source. At a constant residual of 2, the reported loss and gradient norm are zero despite a trend MAE of 2. The supplementary check applies the unchanged EMA implementation from the confirmation experiments to randomly generated targets in float64. It covers three horizons (8, 24, 96), three nonzero biases (0.25, 1, 2), and three β values (0.25, 0.5, 0.75). All 27 cases have positive MSE and trend error, as predicted. DBLoss and Q gradient norms are below 5.01×10^{-16}, whereas Fixed and MSE gradients are nonzero. Appendix A.3 tests recovery near this boundary.

Proposition 6 (Coefficient-matched anchored completion)**.** *On an open rectangle $I \times J$ in (S, T) with $\tau \in J$ and $T + \epsilon > 0$, a C^2 potential satisfying $\partial_T \Psi = (1 - \beta)S/(T + \epsilon)$ and $\partial_S \Psi(S, \tau) = \beta$ is unique up to an additive constant. It is given by $\Psi_\tau = \beta S + (1 - \beta)S \log\{(T + \epsilon)/(\tau + \epsilon)\}$.*

Proof. Integrating the prescribed T-derivative gives $\Psi(S,T) = (1-\beta)S\log(T+\epsilon) + C(S)$. The anchor condition at $T = \tau$ requires $C'(S) = \beta - (1-\beta)\log(\tau+\epsilon)$. Thus $C(S) = \{\beta - (1-\beta)\log(\tau+\epsilon)\}S + c$, which yields Ψ_τ plus an arbitrary constant c.

A.3 DBLoss recovery under a controlled output bias

We use all fifteen original DLinear DBLoss checkpoints, covering five sources and three seeds, with their original dimensions and train-only scaling. For each checkpoint, we use the first complete validation batch of size 32. Let r_0 be the saved prediction residual and set $r(c) = \delta r_0 + c\mathbf{1}$. At $\delta = 1$, this adds a scalar output adapter to the saved model; at $0 \le \delta < 1$, it uses controlled, label-based residual interpolation. The backbone remains fixed. The linear EMA E gives

$$S(c) = \delta^2 \operatorname{mean} |(I-E)r_0|^2, \qquad T(c) = \operatorname{mean} |\delta E r_0 + c|.$$

One trend-optimal bias is $c_* = -\operatorname{median}(\delta E r_0)$, and each condition starts at $c_* \pm 1$. The scalar field is the trend coefficient times $\operatorname{mean}\operatorname{sign}(\delta E r_0 + c)$, with zero selected at an exact MAE kink. We compute it using sorted trend residuals and prefix sums; 150 initial checks compare it with autograd through the original decomposition.

We fix $\delta \in \{0, 0.03, 0.1, 0.3, 1\}$ and compare DBLoss, Fixed, and a diagnostic floor $\max\{S/[2(T+10^{-8})], 0.1\}$ on DBLoss's weighted trend coefficient. Each run takes 256 steps at learning rate 0.05 using SGD or fresh-state Adam ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$), yielding 900 trajectories. At every step, we record $|c - c_*|$, MSE, trend error, the trend coefficient, and the bias gradient. The learning rate and floor are fixed diagnostic choices.

Table 3: Median remaining bias distance after 256 updates (thirty cases per cell). Initial distance is one. Fixed, DBLoss and the coefficient-floor control use the same scalar adapter and step budget.

δ	SGD DB	SGD Fixed	SGD floor	Adam DB	Adam Fixed	Adam floor
0	1	6.04e-16	8.64e-16	1	0.00501	0.00501
0.03	1	0.01	1.89e-09	0.00143	7.41e-07	7.55e-07
0.1	0.997	7.55e-08	8.85e-09	2.49e-06	5.43e-07	6e-07
0.3	0.974	4.72e-08	6.56e-05	1.53e-06	8e-07	8.26e-07
1	0.666	2.68e-07	0.00555	1.06e-06	9.04e-07	9.46e-07

At the exact zero-seasonal-error boundary, DBLoss produces no correction under either SGD or fresh-state Adam. Near the boundary, $S/(T+\epsilon)$ attenuates the raw correction field, whereas Adam's adaptive normalization substantially changes recovery. At $\delta = 1$, DBLoss's median remaining bias distance is 0.6664 with SGD and 1.06×10^{-6} with Adam. The coefficient floor restores a nonzero correction signal at the boundary. These results characterize the interaction between weighting and optimization in the specified bias-adapter experiment.

A.4 From fixed-batch directions to finite batch averages

At a common parameter state, let $a_b = (T_b + \epsilon)/(\tau + \epsilon)$ for $\beta = 1/2$. For any finite collection of M batches,

$$\overline{\nabla Q} = \bar{a}\,\bar{G} + \frac{1}{M}\sum_{b=1}^{M}(a_b - \bar{a})(G_b - \bar{G}).$$

The second term is vector-valued and can change direction, so per-batch collinearity need not survive averaging. We use the first three original DBLoss seeds for each of the five primary sources, retaining the train-only scaler, horizon, lookback, and persistence anchor. For each fixed DLinear checkpoint, we sample 64 batches of 32 distinct training windows uniformly. Batches may overlap; within a source, the three checkpoints use the same sampled indices. This diagnostic involves no fitting or test evaluation. The two fields are computed directly with float64 autograd at $\beta = 0.5, \epsilon = 10^{-8}$. Across 960 batches, the per-batch identity error is at most 1.51×10^{-15}.

Table 4 reports fixed prefixes of 16, 32, and 64 batches for every source. Median angles decrease across these observed prefixes; the largest 64-batch angle is 3.97° on ILI. At 64 batches, norm ratios of the finite averages range from 0.194 to 1.130. Thus small angular differences coexist with substantial differences in magnitude.

Table 4: Angles (degrees) between the mean DBLoss gradient and mean ∇Q_τ at fixed DLinear parameters. Each source uses three saved DBLoss checkpoints and 64 random training batches of 32 windows. Entries are checkpoint medians; brackets give the three-checkpoint range at 64 batches. Prefixes 16/32/64 display finite-sample sensitivity. The final column is the median ratio $\|\overline{\nabla Q}\|/\|\bar{G}\|$ at 64 batches.

Source	16 batches	32 batches	64 batches [range]	Norm ratio
Electricity	0.796	0.458	0.357 [0.287, 0.369]	0.217
Traffic	1.328	0.795	0.547 [0.503, 0.560]	0.195
Exchange	0.541	0.358	0.275 [0.233, 0.537]	1.034
ILI	6.051	4.888	3.956 [3.910, 3.974]	0.601
Solar	1.454	0.947	0.808 [0.731, 0.986]	0.202

A.5 Self-contained parameter-space DBLoss certificate

The numerical certificate uses only ETTh1 training rows $[0, 8640)$, standardized with statistics from that prefix. The float64 channel-shared DLinear model has a lookback of 336, a horizon of 96, and 64,704 parameters. We use EMA $\alpha = 0.2$, $\beta = 0.5$, $\epsilon = 10^{-8}$, uniform $1/336$ weight initialization, and zero biases. The batch contains windows starting at indices 1024–1031.

Let u_1 be the normalized full-parameter gradient of S at initialization, and u_2 the normalized Gram–Schmidt residual of the gradient of T. We evaluate the affine slice $\theta(z) = \theta_0 + u_1 z_1 + u_2 z_2$. At the origin, $\|\nabla S\| = 0.6594$, $\|\nabla T\| = 2.2566$, and their cosine is 0.3145. At $z = (0.013, -0.009)$, the 2×2 component-gradient matrix has singular values $(2.2681, 0.7422)$ and determinant 1.6833, certifying rank two.

At this audit point, $S = 0.124724$ and $T = 0.649964$. Proposition 5 gives

$$J_G - J_G^\top = \begin{pmatrix} 0 & -1.294927 \\ 1.294927 & 0 \end{pmatrix}, \qquad \|J_G - J_G^\top\|_F = 1.831303.$$

A centered finite difference with step 3×10^{-6} gives a norm of 1.831303 for DBLoss. The corresponding values are 3.93×10^{-11} for the uncentered conservative control $\Psi_{\mathrm{u}} = \beta S + (1-\beta) S \log(T + \epsilon)$ and 7.85×10^{-11} for the fixed scalar decomposition. Counterclockwise loop work on a square of side 0.005, using 256 midpoint panels per edge, is 3.2354×10^{-5} for DBLoss. Dividing by the area gives 1.29415, close to the analytic curl magnitude 1.29493. Absolute loop integrals for Ψ_{u} and the fixed decomposition are 1.05×10^{-9} and 5.57×10^{-9}, respectively. The machine-readable certificate retains step-size and quadrature-convergence records. Its reverse-orientation value is the negative of the forward integral, an algebraic check rather than a separately traversed loop. The endpoint experiments use anchored Ψ_τ rather than this uncentered control. Together, these checks identify a nonconservative parameter point in the specified two-coordinate slice.

A.6 Decomposition controls and their interpretation

Without detachment, DBLoss's aligned term is $TS/(T+\epsilon) = ST/(T+\epsilon)$. For fixed $S, T > 0$, as $\epsilon \to 0$, its value and S-derivative approach those of S, while its T-derivative approaches zero. Removing detachment therefore restores a scalar potential but can reduce the objective to an almost purely seasonal loss. This motivates the separately frozen no-detach, season-only, fixed-component, and squared-component endpoint controls, none of which is proposed as a new loss.

Removing stop-gradient increases clean test MSE, averaged over five seeds, by 26.52–65.91% relative to DBLoss across the four discovery anchors. The corresponding season-only effects differ by at most 0.016 percentage points. Differentiating $ST/(T+\epsilon)$ scales its trend coefficient by $\epsilon/(T+\epsilon)$, suppressing it when $T \gg \epsilon$. The ablation therefore changes both the learned component and field conservativity, preventing an isolated test of whether curl benefits accuracy. This is our test of a candidate mechanism, not an attribution by the original authors. Table 2 summarizes the properties preserved by each control; the machine-readable supplement retains all discovery endpoints.

For PMLF, let $z = (S, T)$ and $w_i = e^{\beta z_i} / \sum_j e^{\beta z_j}$. With stop-gradient, the executed coordinate field is $g_i = w_i = \partial \Phi_\beta / \partial z_i$. The chain rule then gives $\nabla_\theta \Phi_\beta = w_S \nabla S + w_T \nabla T$ on each differentiable

Table 5: Train-only coefficient-regime audit at registered final DLinear checkpoints. The conservative seasonal coefficient c_S^{Ψ} is evaluated counterfactually for both trained objectives; the negative fraction is calculated over fixed 32-window batches.

Source	trained objective	seeds	negative c_S^{Ψ} (%)	median c_S^{Ψ}	median c_T	median T/τ
Electricity	DBLoss	5	98.44	-0.295	0.131	0.204
Electricity	Ψ_τ	5	92.19	-0.134	0.327	0.281
Traffic	DBLoss	5	100.00	-0.335	0.390	0.188
Traffic	Ψ_τ	5	98.44	-0.121	0.607	0.289
Exchange	DBLoss	5	0.00	+0.476	0.009	0.954
Exchange	Ψ_τ	5	0.00	+0.469	0.010	0.940
Ili	DBLoss	5	40.00	+0.069	0.135	0.426
Ili	Ψ_τ	5	38.89	+0.061	0.142	0.418
Solar	DBLoss	5	98.12	-0.306	0.068	0.200
Solar	Ψ_τ	5	78.75	-0.085	0.053	0.311

stratum, proving Proposition 2. Thus detachment alone does not imply nonconservativity. Ordinary differentiation of the displayed weighted average $\sum_i w_i z_i$ instead gives $w_i + \beta w_i(z_i - \sum_j w_j z_j)$, which generally differs from the executed field. Finite population risk additionally requires integrability of the composed potential.

In the pinned official supplement, instantiating the released class raises a `NameError` because its `super` call references an undefined class. We retain this source finding and use a formula-faithful independent implementation only in the separately frozen ancillary experiment. PMLF's stated potential $k\alpha^{-2}(1 + \alpha e - \log(1 + \alpha e))$ has derivative $ke/(1 + \alpha e)$. Removing its linear and constant terms leaves a *negative* signed logarithm. We therefore analyze the implemented positive symmetric $\log(1 + |e|)$ as a robust penalty, rather than the literal remainder of that potential.

A.7 Post-hoc scalar-direction experiment

The five-source DLinear product experiment is post-hoc. Its code, dependencies, protocol, and analysis were frozen after inspecting the original confirmation results but before execution. An adapter imports the unchanged runner and modifies only the objective, output destinations, and cell list. Logging S, T and gradient scale consumes no random numbers; the original results and selections remain unchanged.

We use EMA $\alpha = 0.2$, $\beta = 0.5$, $\epsilon = 10^{-8}$, batch size 32, and $L = 336$, $H = 96$ (ILI: $L = 104$, $H = 24$). Adam uses zero weight decay, no clipping or schedule, and at most 256 updates per epoch; checkpoints are selected by validation MSE. The train-only persistence anchor τ follows the original rule using at most 2048 windows. Seed 27600 screens $10^{-5}, 3\times10^{-5}, 10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}$ for three epochs. The two best learning rates are restarted with a maximum of ten epochs and patience two. A boundary winner triggers exactly one adjacent full-budget trial at 3×10^{-6} or 10^{-2}. The minimum full-budget validation MSE determines the frozen learning rate. Upper extensions occur on Exchange and ILI. All five learning rates are fixed before any new test evaluation. Final seeds 27600–27604 reuse the original initialization and shuffle rules. Because this protocol uses separate learning-rate selection, the endpoint contrast is not a field intervention at a single fixed step size.

All 42 tuning fits and 25 final fits succeed. Summed fitting time is about 38 minutes across runs on both GPUs. The artifact retains clean, Gaussian-noise, block-zero, and channel-mask evaluations; standardized and raw-unit MSE/MAE; histories; checkpoints; and component traces. The main text reports the prespecified clean MSE. All sources and three contrasts are retained, each with a descriptive exact sign-test $p = 1$.

Table 6 reports paired-seed and source-mean uncertainty. Table 7 shows that the positive rescaling varies substantially within and across training runs. Neither the rescaling range nor a favorable endpoint identifies a causal effect of curl on predictive accuracy.

Table 6: Post-hoc product-control uncertainty. Seed intervals resample five paired seeds within each source. Source-mean intervals enumerate resamples of the five source effects. Neither has an exact coverage guarantee at these sample sizes. Negative favors Q_τ.

Source	Comparator	Effect (%)	Paired-seed 95% interval
Electricity	DBLoss	-0.050	[-0.203, +0.096]
Traffic	DBLoss	+0.321	[+0.118, +0.480]
Exchange	DBLoss	-1.211	[-6.692, +4.643]
ILI	DBLoss	-10.743	[-17.187, -4.211]
Solar	DBLoss	+0.177	[+0.045, +0.337]
Electricity	Fixed	-0.286	[-0.687, +0.025]
Traffic	Fixed	+0.180	[-0.016, +0.333]
Exchange	Fixed	+0.877	[-6.234, +7.233]
ILI	Fixed	-5.184	[-11.729, +1.406]
Solar	Fixed	-0.724	[-1.361, -0.182]
Electricity	MSE	-0.552	[-0.999, -0.133]
Traffic	MSE	+1.231	[+1.122, +1.317]
Exchange	MSE	-1.130	[-6.290, +5.681]
ILI	MSE	+2.179	[-0.056, +5.269]
Solar	MSE	+1.332	[+0.963, +1.702]
Source mean	Comparator	Mean (%)	Source-resampling 95% interval
Five sources	DBLoss	-2.301	[-6.624, +0.189]
Five sources	Fixed	-1.027	[-3.219, +0.412]
Five sources	MSE	+0.612	[-0.522, +1.651]

Table 7: Frozen validation selections and observed Q_τ gradient scale $a = (T + \epsilon)/(\tau + \epsilon)$. The reported scale median is the median of per-seed medians; the range covers every optimization batch. It is not a confidence interval.

Source	Selected LR	Train-only τ	Median scale	All-step scale range
Electricity	0.001	0.8285	0.218	[0.185, 0.619]
Exchange	0.003	0.1990	1.167	[0.779, 2.660]
ILI	0.003	0.6175	0.533	[0.342, 1.132]
Solar	0.001	1.0088	0.209	[0.128, 0.900]
Traffic	0.001	0.8339	0.199	[0.174, 0.598]

B RI: EXPECTED-OBJECTIVE BOUNDARY AND EXECUTED LEARNING

We analyze the full objective averaged over fresh surrogate noise at fixed targets. Its global boundary provides the reference for tests of model attainability, fitted-state derivatives, and random-number consumption during training.

B.1 FINITE CONDITIONAL EXPECTATION AND DISPERSION

Proof of Proposition 3. For iid uniform surrogate rows independent of the fixed residual kernel K, the expected Gaussian kernel has unit diagonal and constant off-diagonal a_H, giving $\mathbb{E}[L] = (1 - a_H)I + a_H\mathbf{1}\mathbf{1}^\top$. Since $C_D\mathbf{1} = 0$, centering removes the rank-one term. Substitution into the released biased statistic $\langle L, C_DKC_D\rangle_F/(D-1)^2$ proves the claim. Linearity extends the result to the released sum over batch examples.

Residual-channel dispersion. Since $C_D^2 = C_D$ and $K_{ii} = 1$, $\mathrm{tr}(C_DKC_D) = D - D^{-1}\mathbf{1}^\top K\mathbf{1} = D-1-D^{-1}\sum_{i\neq j} K_{ij}$, which gives the main-text expression for $\mathbb{E}Z$. Let $c_H = (1-a_H)/(D-1)^2$. At the released unit kernel scale, $K_{ij} = \exp(-\|r_i - r_j\|^2)$. The replacement auxiliary $A = 10\exp(-\mathbb{E}[Z \mid K])$ uses the batch-summed expectation and has gradient, within each example,

$$\nabla_{r_i} A = -\frac{4c_H A}{D}\sum_j K_{ij}(r_i - r_j).$$

Table 8: Post-hoc scalar-direction control. Q_τ uses ordinary differentiation and the same validation-only selection rule. Negative relative MSE effects favor Q_τ. Five paired seeds per source; all sources and contrasts are shown.

Source	Q_τ MSE	vs DBLoss (%)	vs Fixed (%)	vs MSE (%)
Electricity	0.1419	-0.050	-0.286	-0.552
Traffic	0.4173	+0.321	+0.180	+1.231
Exchange	0.0880	-1.211	+0.877	-1.130
ILI	2.3830	-10.743	-5.184	+2.179
Solar	0.2281	+0.177	-0.724	+1.332
Source median	–	-0.050	-0.286	+1.231
Wilcoxon p	–	0.6250	0.6250	0.3125

Its negative gradient sums pairwise separating directions weighted by kernel similarity; these contributions can cancel. This formula describes the conditional-mean replacement auxiliary. A float64 synthetic check gives errors of 5.56×10^{-17} for the rewrite and 1.64×10^{-16} for the gradient. Appendix B.2 establishes the global boundary for the full expected objective and checks local stochastic responses.

B.2 RI GLOBAL OPTIMALITY, CURVATURE, AND EXECUTABLE VERIFICATION

Objective and notation. Let $r \in \mathbb{R}^{B\times D\times H}$ be the residual for fixed targets, with positive integers B, H, integer $D > 1$, and $N = BDH$. Set $C_D = I - \mathbf{1}\mathbf{1}^\top/D$ and $(Cr)_b = C_D r_b$; C is the orthogonal projection onto channel-centered residuals. The surrogate U has iid uniform entries on $[0, 1]$, independent of r. Within example b, define $K_b(r)_{ij} = \exp(-\|r_{bi} - r_{bj}\|^2)$ and $L_{b,ij} = \exp(-\|U_{bi} - U_{bj}\|^2)$. For $\lambda \geq 0$, the helper returns $R_U(r) = \|r\|_F^2/N + \lambda e^{-Z(r,U)}$. Its statistic and full expected objective are

$$Z(r, U) = \frac{1}{(D-1)^2} \sum_b \operatorname{tr}(K_b(r) C_D L_b C_D), \qquad R(r) = \frac{\|r\|_F^2}{N} + \lambda \mathbb{E}_U e^{-Z(r,U)}.$$

For $i \neq j$, let $a_H = \mathbb{E}[L_{b,ij}] \in (0, 1)$ and $c_H = (1 - a_H)/(D-1)^2$. Then $\lambda_{\text{crit}} = 1/(2Nc_H) > 0$.

Proof of Proposition 4. *Global optimality, including the critical weight.* Proposition 3 and $1 - e^{-x} \leq x$ for $x \geq 0$ give

$$\mathbb{E}Z = \frac{c_H}{D} \sum_{b,i,j} \left(1 - e^{-\|r_{bi} - r_{bj}\|^2}\right) \leq \frac{c_H}{D} \sum_{b,i,j} \|r_{bi} - r_{bj}\|^2 = 2c_H \|Cr\|_F^2.$$

If $Cr \neq 0$, at least one pairwise distance is positive, making the inequality strict. Since $R(0) = \lambda$ and $e^{-z} \geq 1 - z$,

$$\begin{aligned} R(r) - R(0) &\geq \frac{\|r\|_F^2}{N} - \lambda \mathbb{E}Z \\ &\geq \frac{\|(I - C)r\|_F^2}{N} + \left(\frac{1}{N} - 2\lambda c_H\right) \|Cr\|_F^2. \end{aligned}$$

For $0 \leq \lambda < \lambda_{\text{crit}}$, the final expression is positive whenever $r \neq 0$. At $\lambda = \lambda_{\text{crit}} > 0$, the second inequality is strict if $Cr \neq 0$, so $R(r) > R(0)$ still holds. If $Cr = 0$ and $r \neq 0$, every $K_b = \mathbf{1}\mathbf{1}^\top$, giving $Z = 0$ for every draw and $R(r) - R(0) = \|r\|_F^2/N > 0$. Thus zero residual is the unique global minimizer throughout the closed interval.

Local curvature and sharpness. For fixed v, $K_b(tv)_{ij} = 1 - t^2\|v_{bi} - v_{bj}\|^2 + O(t^4)$. The matrix $M_b = C_D L_b C_D$ is symmetric, positive semidefinite, and has zero row sums. The constant kernel therefore contributes zero, and

$$-\sum_{ij} (M_b)_{ij} \|v_{bi} - v_{bj}\|^2 = 2 \operatorname{tr}(v_b^\top M_b v_b).$$

Hence

$$Z(tv, U) = \frac{2t^2}{(D-1)^2} \sum_b \operatorname{tr}(v_b^\top C_D L_b C_D v_b) + O(t^4).$$

The smooth kernel and compact surrogate support make the remainder uniform in U near zero, allowing it to be averaged. Using $e^{-Z} = 1 - Z + O(t^4)$ and $\mathbb{E}L_b = (1 - a_H)I + a_H \mathbf{1}\mathbf{1}^\top$ gives

$$R(tv) - R(0) = t^2 \left[\frac{\|v\|_F^2}{N} - \frac{2\lambda(1 - a_H)}{(D-1)^2} \sum_b \|C_D v_b\|_F^2 \right] + O(t^4).$$

The Hessian at agreement, $H_R = \nabla_r^2 R(0)$, is

$$H_R = \frac{2}{N} I - 4\lambda c_H C.$$

Its eigenvalues are $2/N$ on channel-shared directions and $2/N - 4\lambda c_H$ on channel-centered directions. For $\lambda > \lambda_{\text{crit}}$, both signs occur, making zero residual a saddle in residual space. The global bound and local expansion therefore yield the same sharp threshold. At equality, centered quadratic curvature vanishes, but the strict kernel inequality preserves global uniqueness.

The stochastic loss and the conditional-mean replacement. For every fixed surrogate draw, $Z(0, U) = 0$ and $\nabla_r Z(0, U) = 0$. Thus R_U is stationary at zero, with Hessian-vector action

$$H_{R_U} v = \frac{2v}{N} - \frac{4\lambda}{(D-1)^2} \big(C_D L_b C_D v_b\big)_{b=1}^B.$$

Averaging recovers H_R. The replacement $\|r\|^2/N + \lambda e^{-\mathbb{E}Z}$ shares this Hessian at zero, but its gradient generally differs from that of R elsewhere. The sharp global boundary applies to R; fixed-draw curvature still depends on U. Since agreement is stationary for every draw, the trajectory tests below start from a specified nonzero perturbation.

Kernel scale and batch reduction. Let the residual and surrogate kernels have fixed positive denominators σ and η, respectively, with residual kernel $\exp(-\|r_i - r_j\|^2/\sigma)$. If the statistic divides the batch sum by a fixed $q > 0$ ($q = 1$ for sum and $q = B$ for mean), then

$$a_H(\eta) = \left[2 \int_0^1 (1 - u) e^{-u^2/\eta} \, du \right]^H, \qquad \lambda_{\text{crit}}(\sigma, \eta, q) = \frac{q\sigma(D-1)^2}{2BDH[1 - a_H(\eta)]}.$$

The global classification is unchanged: $\mathbb{E}Z \le 2c_H(\eta)\|Cr\|_F^2/(q\sigma)$, with $c_H(\eta) = [1 - a_H(\eta)]/(D - 1)^2$, and the inequality remains strict when $Cr \neq 0$. The auxiliary curvature coefficient is also scaled by $1/(q\sigma)$. Here σ, η are kernel denominators, not their square roots, and MSE remains divided by BDH. Auxiliary-weight comparisons therefore depend on channel count, batch reduction, and kernel scale.

Attainable parameter directions. If zero residual is attainable, every exact-fit parameter globally minimizes the model's expected objective for $0 \le \lambda \le \lambda_{\text{crit}}$; the parameter minimizer need not be unique. For a twice-differentiable model with residual Jacobian J, stationarity in residual space gives parameter Hessian $J^\top H_R J$ at agreement. At unit kernel scale and batch sum, a parameter direction w has negative curvature exactly when

$$4\lambda c_H \|CJw\|^2 > \frac{2}{N} \|Jw\|^2.$$

In particular, any nonzero attainable channel-centered residual direction gives negative parameter curvature above the threshold. These statements concern fixed-target agreement; a noisy conditional-risk minimizer is a different statistical object.

Numerical and model checks. We extract the unmodified published kernel and HSIC functions. At zero residual, fixed-surrogate Hessian-vector checks along unit channel-centered directions use $(B, D, H) = (4, 2, 8), (32, 7, 96), (32, 21, 96)$ and $\lambda = 10$. Maximum absolute errors are 8.88×10^{-16} in float64 and 4.77×10^{-7} in float32. Thirty-six further checks cross these shapes, residual-kernel denominators $0.5, 1, 2$, sum/mean reductions, and $\lambda/\lambda_{\text{crit}} = 0.5, 2$, with surrogate denominator

Table 9: Neural-parameter directions at constructed agreement. Each cell contains three saved models, four validation input batches and two fixed Gaussian directions per batch. Ranges summarize these 24 directions. The directional threshold is $\lambda_{\mathrm{crit}}/\rho_d$; the last two columns count curvature signs computed from the published source.

Source / model	ρ_d: median [range]	Max. threshold	Negative at 10	Positive control
ETTh2 / DL	0.788 [0.584, 0.938]	1.43×10^{-3}	24/24	24/24
ETTh2 / PT	0.531 [0.455, 0.596]	1.84×10^{-3}	24/24	24/24
Weather / DL	0.828 [0.722, 0.892]	4.29×10^{-3}	24/24	24/24
Weather / PT	0.812 [0.669, 0.866]	4.63×10^{-3}	24/24	24/24

one. They evaluate the conditional-mean replacement's Hessian at zero, which the proof identifies with the full expected Hessian. For $(4, 2, 8)$ at weight ten, a 20,000-draw check along a unit channel-centered direction gives quadratic coefficient -13.91490 (Monte Carlo standard error 0.01202), versus the analytic -13.91444.

With $B = 4, H = 8$, a realizable model predicts the same channel bias $b \in \mathbb{R}^2$ at every example and horizon position, against zero targets. It starts at $b = (10^{-3}, -10^{-3})/\sqrt{2}$. We use SGD with learning rate 0.05, 250 updates, five surrogate seeds, and $\lambda/\lambda_{\mathrm{crit}} = 0.5, 2$. Each update draws fresh uniform surrogates and uses the stochastic helper. At ratio 0.5, final/initial MSE ranges from 2.87×10^{-6} to 3.63×10^{-6}; at ratio 2, it ranges from 1.187×10^{4} to 1.261×10^{4}. Figure 3 includes all ten trajectories.

B.2.1 Full neural-parameter attainability at constructed agreement

The model-space condition can be checked without searching for a favorable direction. We use all twelve previously saved RI models across ETTh2 and Weather, DLinear and PatchTST, and seeds 27600–27602. For each model, we take the first four batches of the existing validation diagnostic (indices 0, 5, 11, 17). Each batch contains 32 examples, with lookback 336 and horizon 96. Scaling uses training rows only, and computation uses float64 in evaluation mode with dropout off.

For each fixed input X, let $Y_* = f_{\theta_0}(X)$ be the frozen prediction at the saved parameters. This constructs exact agreement with Y_*; observed future values are not used as targets in this agreement experiment. For each batch, we normalize two Gaussian parameter directions to unit norm without optimization or selection. Reverse-over-reverse differentiation computes $v = Jd$, and all parameters are perturbed together.

For each direction, we fix one uniform surrogate U and differentiate the full published stochastic rule through the model. We compare its second derivative with

$$d^\top H_{\theta,U} d = \frac{2\|v\|^2}{BDH} - \frac{4\lambda}{(D-1)^2} \sum_b \mathrm{tr}(v_b^\top C_D L_b C_D v_b).$$

The largest relative discrepancy across 192 curvature comparisons is 4.83×10^{-16}. At θ_0, all parameter gradients are exactly zero. The centered energy fraction ρ_d ranges from 0.455 to 0.938. All 96 directions have negative curvature at $\lambda = 10$ and positive curvature at the prespecified control $\lambda = \lambda_{\mathrm{crit}}/2$ (Table 9).

We also evaluate $\theta_0 \pm 10^{-3} d$ with the same target and surrogate, restoring the saved parameters after every evaluation. All 192 signed perturbations lower the full loss at weight ten; all 192 raise it under the control. Exponential differences are evaluated with a stable $\mathrm{expm1}$ expression. Central finite differences agree with Jd to maximum relative error 5.09×10^{-7}. These results verify attainable directions and finite responses in the actual architectures at constructed agreement. The real-target residual and training measurements in Appendices B.3 and B.4.1 instead address the observed fitted states.

B.3 RI diagnostics at fitted states

The post-hoc protocol uses twelve existing RI checkpoints: ETTh2 and Weather, DLinear and PatchTST, and three original seeds. Each checkpoint supplies 16 evenly spaced complete validation

Table 10: Published RI helper evaluated on models from the explicitly activated controlled runner, with 3 seeds and 16 fixed validation batches per cell. The auxiliary gradient includes the original factor $10\exp(-Z)$. We first take checkpoint medians over batch–draw measurements, then report the median and range across three checkpoints. Ranges are not confidence intervals.

Source	Backbone	Auxiliary/MSE gradient norm	Max. $\lvert\overline{Z} - \mathbb{E}Z\rvert$
ETTh2	DLinear	0.019 [0.018, 0.020]	8.62×10^{-7}
ETTh2	PatchTST	0.015 [0.013, 0.016]	1.01×10^{-6}
Weather	DLinear	0.890 [0.848, 0.942]	2.68×10^{-7}
Weather	PatchTST	1.366 [1.299, 1.952]	2.87×10^{-7}

Table 11: Observed surrogate variation of the released float32 statistic at $H = 96$, unit bandwidth. Each batch has 64 draws; maxima range over all 48 batches in a cell (three checkpoints). s_Z is sample standard deviation and CV is $s_Z/|\overline{Z}|$. These are measurements at finite numerical precision, not variance bounds.

Source / model	Maximum s_Z	Maximum CV
ETTh2 / DLinear	2.10×10^{-6}	3.94×10^{-7}
ETTh2 / PatchTST	2.07×10^{-6}	3.92×10^{-7}
Weather / DLinear	8.50×10^{-7}	5.70×10^{-7}
Weather / PatchTST	6.64×10^{-7}	4.37×10^{-7}

batches of size 32, with lookback 336, horizon 96, and the original train-only scaler. Evaluation mode disables dropout, leaving only surrogate randomness. Each batch uses 64 draws for the statistic and full objective, and eight for parameter-gradient measurements. The ratio $\|\nabla_\theta(10e^{-Z})\|_2/\|\nabla_\theta \mathrm{MSE}\|_2$ includes all trainable parameters. We take the median of 128 batch–draw ratios per checkpoint, then the median and range across the three checkpoints in each source–backbone cell (Table 10).

The vectorized statistic matches the released source on the first draw of all 192 batches to a maximum absolute error of 9.54×10^{-7}. No new model is fitted and no test data are accessed. Checkpoint selection and validation diagnostics are post-hoc; the 192 batches are repeated measurements, not independent tasks.

Sampling variation and gradient direction. Table 11 reports sample standard deviations recomputed from 64 saved draws per batch, measuring variation in the released float32 computation at unit kernel scale and $H = 96$. For $g_m = \nabla_\theta \mathrm{MSE}$ and $g_a = \nabla_\theta(10e^{-Z})$, define

$$\rho = \frac{\langle g_a, g_m\rangle}{\|g_m\|^2}, \qquad \frac{d}{d\eta}\mathrm{MSE}(\theta - \eta(g_m + g_a))\Big|_{\eta=0} = -(1+\rho)\|g_m\|^2.$$

All 1,536 measured gradient pairs are nonzero. Gradients use the released float32 computation; dot products and norms accumulate in float64 over all trainable parameters. Table 12 reports every cell. On Weather, $\rho < -1$ for 104/384 DLinear and 160/384 PatchTST batch–draw pairs; ETTh2 has none. These correspond to 13/48 and 20/48 Weather batches, respectively, with all eight draws agreeing within each batch. The training study below separately measures actual optimizer steps (Figure 6).

We also compare the gradient of $10\exp(-\mathbb{E}Z)$ with the mean of eight stochastic auxiliary gradients. Relative discrepancies use the mean stochastic gradient's norm as denominator. Across the 48 measured batches per cell, they are at most 1.14×10^{-4}, 1.60×10^{-3}, 1.22×10^{-5}, and 8.03×10^{-6} in the table's order. These values quantify the empirical approximation at the measured states. Appendix B.4.1 compares training with matched randomness.

B.4 Source activation and RI training with matched randomness

Distinguishing the helper from the training entry. The pinned RI repository defines a helper that permutes residuals to $[B, D, H]$, draws a same-shaped surrogate, and adds $10e^{-Z}$ to MSE. The supplied DLinear training class selects MSE; its training-loop helper call is commented out. We execute the unmodified training-method syntax tree with a synthetic loader and linear model, tracing

Table 12: RI auxiliary direction. Cosine and projection ρ are summarized first by checkpoint medians over 16 batches $\times$ eight draws, then by median [range] across three checkpoints. The last column counts draws satisfying $1+\rho<0$, which indicates a local MSE-ascent direction for unpreconditioned combined descent. Draws reuse batches and are not independent tasks; ranges are not confidence intervals.

Source / model	Cosine	Projection ρ	Ascent draws
ETTh2 / DLinear	-0.352 [-0.376, -0.325]	-0.003 [-0.004, -0.003]	0/384
ETTh2 / PatchTST	-0.285 [-0.360, -0.231]	-0.004 [-0.004, -0.003]	0/384
Weather / DLinear	-0.622 [-0.631, -0.597]	-0.442 [-0.463, -0.393]	104/384
Weather / PatchTST	-0.333 [-0.481, -0.183]	-0.735 [-0.815, -0.169]	160/384

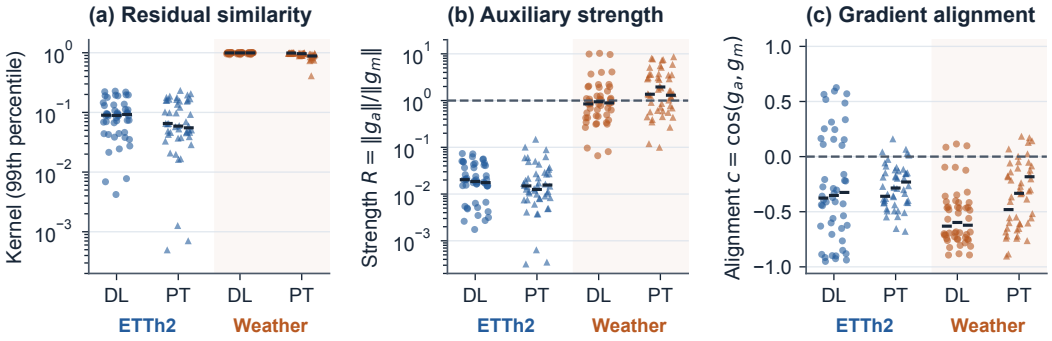


Figure 5: **Similar residual pairs can coexist with a strong opposing auxiliary.** Each small marker represents one of 192 validation batches from twelve saved models; (b–c) use medians over eight draws. Black ticks mark checkpoint medians over 16 kernels or 128 gradient pairs. Color distinguishes datasets; DL/PT denote DLinear/PatchTST. Weather is lightly shaded. Here $g_a = \nabla(10e^{-Z})$ and $g_m = \nabla\text{MSE}$. Each checkpoint contributes repeated batch and draw measurements.

calls through the first actual Adam step. The trace records one MSE call, zero helper calls, and one optimizer step. The maximum absolute prediction-gradient error against analytical MSE is 1.49×10^{-8}. Our controlled runner separately records an HSIC call and a loss above MSE. Thus the theory concerns the published helper, and our RI training results concern its explicit activation in the controlled runner. The default-entry trace does not reconstruct the original paper's training histories.

Earlier source certificates detected function names without establishing training activation; this trace corrects that label, and the earlier certificates remain available. The audit records source reachability separately: InfoGeo has formula-only evidence, and PMLF's pinned snapshot has a constructor failure. For AliO, the executable official example is distinguished from the failing package entry.

B.4.1 Five-arm training design

We use four ETTh2/Weather $\times$ DLinear/PatchTST anchors, three seeds, lookback 336, horizon 96, and batch size 32. Within each anchor and seed, all arms share initialization, loader order, and the original RI-selected learning rate, with exactly 256 Adam updates. The rate is 10^{-4} for Weather–DLinear and 3×10^{-4} elsewhere. There is no early stopping or new test evaluation. The five arms are:

1. Original: the helper draws surrogates from the model's global generator.
2. Expected with draws: replace Z by its conditional expectation, while generating and discarding the original same-shaped random tensor.
3. Isolated: draw surrogates from a separate generator without consuming the model's random stream.
4. Expected: use the conditional-mean statistic without surrogate draws.
5. MSE: use MSE with the same learning rate and update budget.

The matched groups are (1,2) and (3,4,5). All arms receive identical batches at each update. Within each group, we verify the model-forward random state and each dropout call's tensor shape, probability, and incoming random state, ensuring identical masks for matching calls.

At updates 1,32,128,256, we measure $g_m = \nabla_\theta \mathrm{MSE}$ and $g_a = \nabla_\theta(L_{\mathrm{arm}} - \mathrm{MSE})$, where L_{arm} is the arm's complete loss, together with their norm ratio and projection. For the MSE arm, $g_a = 0$. We record the actual Adam change Δ, $g_m^\top \Delta$, and the same-batch MSE change using the original dropout states. Measurements include all parameters. Buffers and generators are restored afterward, leaving subsequent training unchanged. Full validation MSE is evaluated after update 256.

Table 13: All sixty 256-update training runs. Validation MSE is averaged across three seeds. Counts summarize twelve measured steps per cell (four per seed). They separately report raw-gradient MSE ascent, actual Adam first-order ascent and positive same-mask finite-step MSE change.

Task	Arm	Validation MSE	Raw ascent	Adam ascent	Finite ascent
ETTh2 / DL	Original	0.227304	0/12	0/12	0/12
ETTh2 / DL	Expected + draws	0.227304	0/12	0/12	0/12
ETTh2 / DL	Isolated	0.227304	0/12	0/12	0/12
ETTh2 / DL	Expected	0.227304	0/12	0/12	0/12
ETTh2 / DL	MSE	0.225293	0/12	0/12	0/12
ETTh2 / PT	Original	0.216271	0/12	0/12	0/12
ETTh2 / PT	Expected + draws	0.216269	0/12	0/12	0/12
ETTh2 / PT	Isolated	0.215328	0/12	0/12	0/12
ETTh2 / PT	Expected	0.215328	0/12	0/12	0/12
ETTh2 / PT	MSE	0.214741	0/12	0/12	0/12
Weather / DL	Original	0.462420	2/12	1/12	1/12
Weather / DL	Expected + draws	0.462420	2/12	1/12	1/12
Weather / DL	Isolated	0.462420	2/12	1/12	1/12
Weather / DL	Expected	0.462420	2/12	1/12	1/12
Weather / DL	MSE	0.445157	0/12	0/12	0/12
Weather / PT	Original	0.418739	0/12	1/12	1/12
Weather / PT	Expected + draws	0.418741	0/12	1/12	1/12
Weather / PT	Isolated	0.418552	0/12	1/12	1/12
Weather / PT	Expected	0.418550	0/12	1/12	1/12
Weather / PT	MSE	0.408715	0/12	0/12	1/12

Table 14: Relative validation-MSE effects after 256 updates, computed from the three-seed means. The first contrast changes model randomness; the other two preserve it.

Task	Expected / original (%)	Expected + draws / original	Expected / isolated
ETTh2 / DL	-0.000002	-0.000002	-0.000002
ETTh2 / PT	-0.435704	-0.000670	-0.000012
Weather / DL	-0.000004	-0.000004	-0.000008
Weather / PT	-0.045185	+0.000513	-0.000389

Interpreting the longer historical factorial. The historical comparison of Expected and Original changes both the statistic and model randomness. Its DLinear effects on clean test MSE are below 0.005% in magnitude; PatchTST effects are $+1.473\%$ on ETTh2 and -0.415% on Weather. The isolated-surrogate arm separates these changes. On ETTh2/Weather with PatchTST, isolated versus original gives $+1.538\%/-0.158\%$, and expected versus isolated gives $-0.064\%/-0.258\%$. All contrasts use the same selected learning rates and five paired seeds per anchor. The latter comparison holds the model's random stream fixed; call-by-call dropout matching is verified in the short runs. The longer-run endpoints also reflect accumulated optimization under the original validation selection and budgets. Complete factorial results remain in the supplement.

C Experimental protocol and evidence scope

C.1 Method selection and evidence scope

The ten cases are QDF (Wang et al., 2026a), DistDF (Wang et al., 2026b), FreDF (Wang et al., 2025a), Time-o1 (Wang et al., 2025b), PS Loss (Kudrat et al., 2025), InfoGeo (Fang et al., 2026), OLMA (Shi et al., 2025), DBLoss (Qiu et al., 2025), RI-Loss (Wang et al., 2026c), and PMLF (Chen et al.,

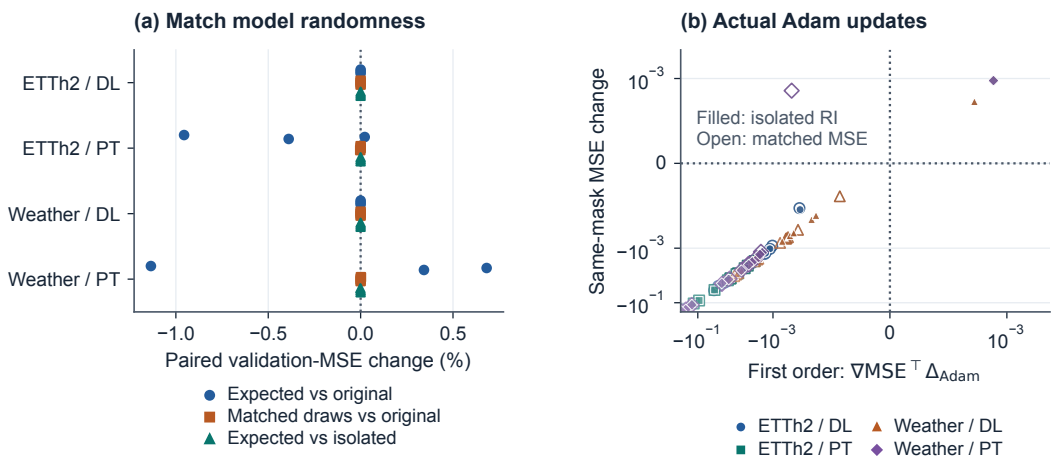


Figure 6: **Matched random streams isolate statistic replacement; measured steps show optimizer responses.** Four source–backbone anchors, three seeds, and 256 Adam updates. (a) All paired validation-MSE changes; both matched contrasts cluster near zero. (b) All 96 measurements at updates 1,32,128,256 for stream-isolated RI (filled) and matched MSE (open). Both axes in (b) use symmetric logarithmic scales, linear within 10^{-5} of zero. Points above zero indicate increases in same-mask batch MSE.

2025). The discovery sample covers contrasting objectives for multi-horizon point forecasting whose contributions centre on a checkable loss formula or implementation. It was assembled from retained papers and proceedings and searches of OpenReview, arXiv, and official repositories, rather than a prospectively enumerated population. Subsequent literature updates and probes of AliO, AMRC, and CvLoss broaden context and mechanism coverage without changing the original selection.

Training experiments cover a subset of the formula and source analyses. An ETTh2 development pilot identified mechanism signals that motivated selecting OLMA and DBLoss; RI was added for its distinct residual–noise construction. These choices preceded standard-split test evaluation. PMLF remains a separate ancillary comparison, and InfoGeo is analyzed only at the formula level. For PMLF, source inspection is distinguished from the ancillary runner because the pinned constructor fails. The supplement retains the selection history and every completed or failed run. Table 27 summarizes the evidence depth.

C.2 Discovery forecasting comparisons

The discovery matrix uses ETTh1, ETTh2, ETTm1, ETTm2, and Weather, with chronological splits, train-only scaling, lookback 336, horizon 96, batch 32, and Adam. We compare MSE, MAE, OLMA, DBLoss, Fixed, and RI using channel-shared DLinear and compact PatchTST. The latter uses three layers, four heads, width 64, patch length 16, stride 8, and RevIN. At seed 27600, validation MSE selects the checkpoint and learning rate from $\{10^{-4}, 3 \times 10^{-4}, 10^{-3}\}$ before the five-seed test runs. The supplement retains all 252 validation and 420 primary final runs in the discovery records, including runs for the four ETTh2/Weather mechanism anchors.

PMLF is a late-discovery ancillary comparison with 42 validation and 70 final runs. Its post-training serialization failure and subsequent repair remain documented. The ten task–backbone cells share three source families, so their historical tests are descriptive. The source-level confirmation below provides the primary forecasting comparison; complete discovery endpoints remain machine-readable.

C.3 Independent-source coverage and completion

The public Solar Google Drive endpoint repeatedly failed through the execution proxy. We therefore used a mirror verified against two independently hosted copies. The supplementary material documents data sources and split indices. Preprocessing reads public raw data; test-metric evaluation is deferred until model selection is complete.

Table 15: Independent-source data and frozen coverage. L/H gives primary lookback/horizon. Every source has DLinear and all five primary objectives (MSE, MAE, Fixed, DBLoss and Ψ_τ), with five final seeds per cell. Transfers use three seeds per cell; PT denotes compact PatchTST and iT denotes the iTransformer-style implementation described in the protocol. A dash means no registered transfer.

Source	Rows	Channels	L/H	Transfers	Provenance
Electricity	26,304	321	336/96	PT, iT	THUML
Traffic	17,544	862	336/96	—	THUML
Exchange	7,588	8	336/96	PT, iT	THUML
ILI	966	7	104/24	—	THUML
Solar	52,560	137	336/96	PT, iT	Matched mirrors

Registered stress horizons are 192/336/720 (ILI: 36/48/60). Solar follows the PULSE data chain; the mirror was checked against two independently hosted copies. Transfer backbones reuse sources and do not add independent comparison units.

The third backbone provides an iTransformer-style transfer check rather than an official-score reproduction. It preserves temporal normalization, variate-token embedding, full cross-variate attention, horizon projection, and de-normalization, while replacing repository-specific encoder wrappers with compact PyTorch encoder blocks and omitting unused calendar tokens. A frozen certificate verifies output shape, finite gradients, exact seeded replay, and channel-permutation equivariance to 7.2×10^{-7}. The artifact documents the implementation and its deviations.

Completed cells and failures. All 57 confirmation cells completed their planned final seeds: 25 primary DLinear cells with five seeds each (125 fits), 30 transfer cells with three each (90 fits), and two limited uncentered-potential cells with three each (six fits). All 476 validation-screening or full-budget selection fits completed. Each of the 133 stress cells has three final seeds (399 fits). No failed final seed is dropped or imputed, and every reported contrast uses the complete common set of planned seeds. The machine-readable supplement lists planned and completed counts per cell. No optimization failures were recorded in these final matrices. Historical infrastructure incidents and deterministic source-edge failures are reported separately and are not counted as forecasting outcomes. The paired means, uncertainty summaries, and source comparisons therefore use the full planned seed sets rather than a selected subset of successful runs.

C.4 Objective-side complexity

Let B, H, D denote batch size, horizon, and channels, and let $K = 25$ for PMLF. Excluding the common backbone, pointwise MSE/MAE cost $O(BHD)$. OLMA costs $O(BHD(\log H + \log D))$ for temporal and channel FFTs plus a linear Haar transform. Released DBLoss and our equivalent recursive decomposition cost $O(BHD)$ via a weighted cumulative sum or recurrence. Released RI-Loss costs $O(BD^2H)$ under the published helper's $[B, D, H]$ permutation, and PMLF costs $O(BHDK)$ for fixed-width pooling. For MSE/MAE, OLMA, DBLoss, RI-Loss, and PMLF, the dominant additional storage is respectively $O(BHD)$, $O(BHD)$, $O(BHD)$, $O(BHD+BD^2)$, and $O(BHD)$. At $L = 336, H = 96$, DLinear has 64,704 trainable parameters (ILI's $L = 104, H = 24$ model has 5,040). Legacy PatchTST has 362,353–362,381 and independent-transfer PatchTST 362,355–362,981, depending on channel count. The compact iTransformer-style backbone has 94,880. None of the audited objectives adds trainable parameters in our runners. These orders describe the audited implementations rather than an avoidable dense-matrix formulation. Shared-device wall time is not used as an intrinsic efficiency ranking.

D Forecasting results, training budgets, and sensitivity

D.1 Independent-source confirmation and stress tests

D.1.1 Independent confirmation metrics

The main paper reports the primary source-level contrasts. Table 16 gives all original contrasts, and Table 17 reports standardized clean-test MSE for every objective and source. The artifact includes raw-

Table 16: All seven original source-level contrasts with matching point estimates and intervals. Exact tests are descriptive; five sources yield minimum two-sided $p = 0.0625$.

Contrast	Mean (%)	Mean 95% interval	Median (%)	Better	Sign p	Rank p
DBLoss vs Fixed	+1.413	[-0.483, +3.980]	-0.140	3/5	1.0000	0.8125
DBLoss vs Ψ_τ	-24.608	[-44.605, -4.611]	-26.535	4/5	0.3750	0.1875
Ψ_τ vs Fixed	+47.189	[+9.592, +84.785]	+34.895	0/5	0.0625	0.0625
DBLoss vs MSE	+3.223	[-0.035, +8.933]	+0.907	1/5	0.3750	0.1875
Fixed vs MSE	+1.726	[-0.774, +5.020]	+1.048	2/5	1.0000	0.4375
Ψ_τ vs MSE	+49.217	[+12.549, +85.981]	+37.690	0/5	0.0625	0.0625
MAE vs MSE	+2.301	[-0.324, +5.518]	+1.616	2/5	1.0000	0.3125

Table 17: Independent-source DLinear clean standardized MSE (mean $\pm$ standard deviation; five paired seeds in every cell; all completed). No entry is highlighted as the best result.

Source	MSE	MAE	Fixed	DBLoss	Ψ_τ
Electricity	0.1427 ± 0.0007	0.1416 ± 0.0002	0.1423 ± 0.0007	0.1419 ± 0.0002	0.3013 ± 0.0005
Traffic	0.4123 ± 0.0006	0.4189 ± 0.0001	0.4166 ± 0.0004	0.4160 ± 0.0004	0.7613 ± 0.0005
Exchange	0.0890 ± 0.0049	0.0882 ± 0.0086	0.0872 ± 0.0085	0.0891 ± 0.0039	0.0898 ± 0.0043
Ili	2.3322 ± 0.1899	2.5286 ± 0.0461	2.5133 ± 0.0425	2.6698 ± 0.0424	2.6032 ± 0.0525
Solar	0.2251 ± 0.0010	0.2321 ± 0.0014	0.2298 ± 0.0006	0.2277 ± 0.0011	0.3100 ± 0.0005

unit MSE/MAE, paired seed differences, architecture transfers, horizon profiles, corruptions, learning curves, failures, and exact machine-readable summaries. The input corruptions are standardized Gaussian noise with standard deviation 0.1, zeroing the last 12 history steps, and a seeded 10% channel mask. All use the same clean-trained checkpoint. The allocation of results to tables was fixed before test evaluation and did not depend on their ranks.

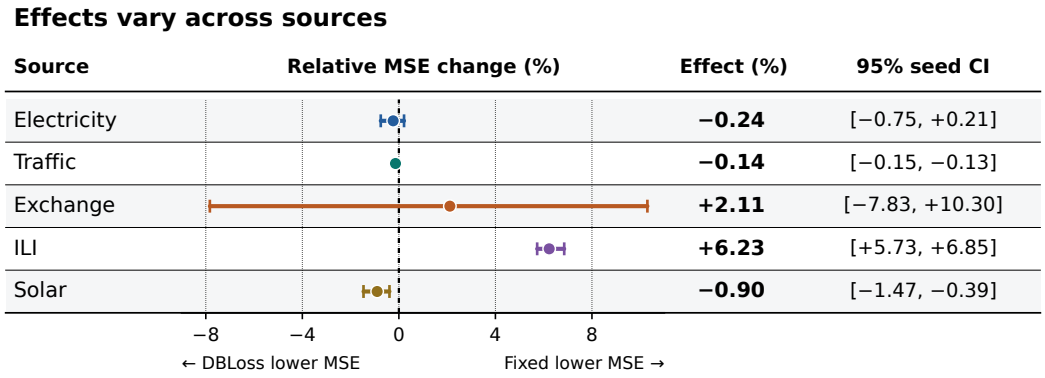


Figure 7: **DBLoss versus Fixed across five sources with DLinear.** Effects are computed from mean test MSE over five paired seeds; negative values favor DBLoss. Intervals show seed variability within each source. Table 1 reports the across-source mean effect and its source-bootstrap interval.

The fixed-multiset experiment uses only development rows and was completed before test evaluation. Each condition fixes 64 windows and a seeded DLinear initialization, then compares the example-count-weighted gradient across smaller partitions with the single-batch gradient. Table 18 pools all nonreference rows within each objective, using the same aggregation over sources, seeds, and orders. The maximum relative fixed-control discrepancy is 7.64×10^{-5}, so numerical partition invariance is approximate. Dataset, seed, order, and batch conditions provide repeated local checks rather than independent samples. For this stateless DLinear model, the weighted partition gradient also equals the gradient from synchronous data-parallel averaging at a common parameter state, with statistics computed locally. Stateful batch normalization and asynchronous drift are outside this check's scope. The same frozen stress protocol also evaluates forecasting endpoints across additional horizons, lookbacks, batch sizes, and both epoch and fixed-step budgets. These results are reported separately from the local gradient checks and do not alter the registered conditions.

Table 18: Fixed-multiset repartition certificate with pooling within each objective. All nonreference rows ($B < 64$) across two sources, three seeds and two orders are pooled within objective. These local measurements are not independent replications.

Objective	Rows	Median relative Δ	Maximum relative Δ	Minimum cosine
Fixed	36	1.398e-07	7.640e-05	1.000000
DBLoss	36	3.268e-03	4.032e-02	0.999190
Ψ_τ	36	2.006e-02	1.509e-01	0.993804

Table 19: Stress-condition source-level summaries. Horizon/lookback cells are first averaged within each independent source; negative favors the first method.

Contrast	median source effect (%)	sources better	Wilcoxon p
DBLoss vs Fixed	+0.008	2/5	0.6250
DBLoss vs Ψ_τ	-15.562	4/5	0.1875
Ψ_τ vs Fixed	+17.691	0/5	0.0625

Partition sensitivity and integrability. Across two sources, three seeds, two orders, and $B < 64$, the median relative gradient discrepancy is 1.40×10^{-7} for fixed decomposition, 3.27×10^{-3} for DBLoss, and 2.01×10^{-2} for the scalar potential Ψ_τ. Partition sensitivity therefore also occurs for a scalar potential and does not identify nonconservativity. Table 18 additionally reports maximum discrepancies and minimum gradient cosines for these local measurements.

Table 20 retains all architecture-transfer effects. For the three input corruptions listed above, the DBLoss–Fixed source-median effects are -0.142%, -0.120%, and -0.104%, respectively, each favoring DBLoss on three of five sources. Figure 8 shows all 33 planned DBLoss–Fixed contrasts across horizon, lookback, and batch-budget settings. The supplement retains the full numerical matrices, including the anchored-potential comparisons.

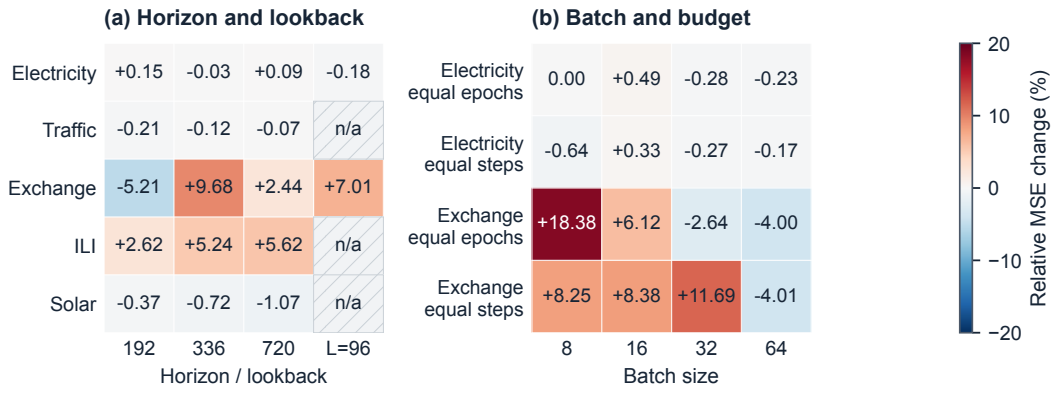


Figure 8: All 33 predefined DBLoss–Fixed stress contrasts, with three paired seeds per cell and a shared linear color scale. Entries give relative MSE change in percent; negative values favor DBLoss. ILI uses horizons 36/48/60. Hatched cells were not included in the design. Conditions within each source are repeated measurements.

D.2 Symmetric training-budget extension

Under the original budget, validation selected the ten-epoch cap in 9/27 PatchTST, 7/27 iTransformer-style, and 3/75 DLinear runs across MSE, Fixed, and DBLoss. Electricity–PatchTST was preselected for follow-up: all three DBLoss checkpoints and two each for MSE and Fixed were selected at epoch ten. All three objectives receive the same extended maximum of thirty epochs, patience five, and the original 256-step epoch cap. The selected learning rates remain 0.003 for MSE/DBLoss and 0.001 for Fixed. The three seeds are 27600–27602; every model is trained afresh.

Table 20: Architecture-transfer clean-MSE effects (%). Each cell averages three complete paired seeds. Sources are reused, so these are sensitivity analyses rather than independent replications; negative values favor the first method.

Source	backbone	DBLoss–Fixed	DBLoss–Ψ_τ	Ψ_τ–Fixed	DBLoss–MSE	MAE–MSE
Electricity	patchtst	+0.331	-66.391	+198.521	+0.111	+0.153
Electricity	itransformer	+0.019	-73.185	+273.001	-0.282	-1.212
Exchange	patchtst	+35.253	+1.834	+32.817	+27.920	-5.845
Exchange	itransformer	+0.609	+1.025	-0.411	-2.899	-3.658
Solar	patchtst	-2.122	-83.531	+494.302	-2.852	-1.439
Solar	itransformer	+1.391	-47.395	+92.740	-0.996	-2.147

Each new trajectory retains two validation-selected checkpoints: the best within its first ten epochs and the best over its full stopping period. The budget comparison is therefore paired within each trajectory. All nine fits finish before any new test evaluation. The original results remain a separate comparison under their original stopping rule; the new first-ten prefix does not apply the original patience-two rule.

Table 21: Symmetric Electricity–PatchTST budget extension. Test entries are mean $\pm$ sample standard deviation over three seeds. Best epochs are selected by validation MSE; checkpoints are paired within each trajectory. Run length gives the evaluated prefix or full stopping time.

Objective	Budget	Validation MSE	Test MSE	Best epochs	Run length
MSE	First ten	0.114636	0.133323 ± 0.000273	10,9,10	10,10,10
MSE	Extended	0.113201	0.132814 ± 0.000756	18,9,19	23,14,24
Fixed	First ten	0.114588	0.133030 ± 0.000044	10,10,8	10,10,10
Fixed	Extended	0.110486	0.130471 ± 0.000864	26,30,27	30,30,30
DBLoss	First ten	0.114574	0.133470 ± 0.000408	10,10,10	10,10,10
DBLoss	Extended	0.113879	0.132907 ± 0.000198	16,10,16	21,15,21

Mean test MSE decreases for all three objectives between the paired checkpoint sets: by 0.382% for MSE, 1.924% for Fixed, and 0.422% for DBLoss. The DBLoss–Fixed relative MSE difference increases from $+0.331\%$ to $+1.867\%$. Fixed selects epochs 26, 30, and 27, while DBLoss selects 16, 10, and 16. All three Fixed runs reach the extended cap, with one selecting that cap as its best epoch. These trajectories show that the comparison depends on the training budget in this source–backbone configuration.

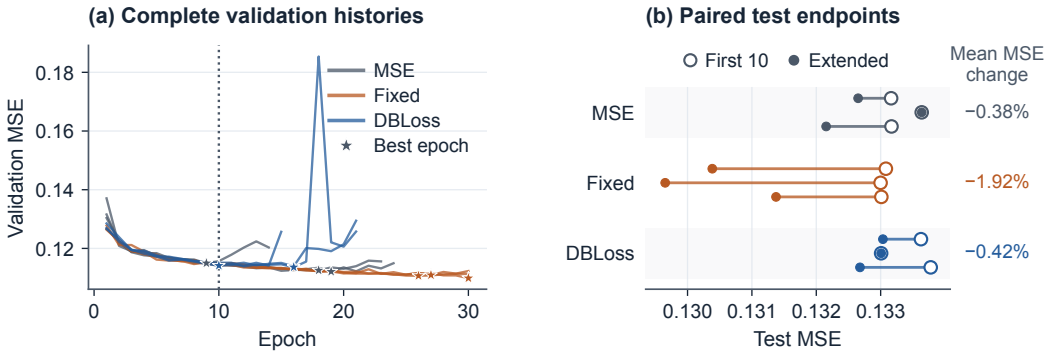


Figure 9: Complete results of the symmetric budget follow-up. All nine validation trajectories and their first-ten/extended test endpoints are retained. Lines in (b) pair checkpoints from the same training trajectory, selected solely by validation MSE.

D.3 Discovery-stage endpoint interpretation

Against MSE, DBLoss and OLMA have discovery-stage median effects of -1.754% and -1.791% over ten dependent cells. These results do not establish source-level confirmation. The DBLoss–Fixed

median effect is -0.344% (7/10 favorable; descriptive $p = 0.160$), with $p = 0.25$ when evaluated over the three source families. Against MAE, DBLoss has a median effect of -0.629% (7/10 favorable; $p = 0.131$), while Fixed has a median effect of -0.095% ($p = 0.846$). The supplement retains all paired effects; these comparisons do not establish detach as the mechanism behind the gains.

E TIME-O1 AND PS: REPRESENTATION AND BATCH-STATISTIC CONTROLS

This appendix extends the audit to Time-o1's representations and PS Loss's batch statistics using published functions. The paired gradient checks hold predictions and targets fixed while changing how the learning signal is computed. Related AMRC and CvLoss checks use, respectively, the published embedding penalty with an identity encoder and a reproduction of the paper formula (Appendices F.13 and F.14).

Data and selection. The real-data follow-ups use ETTh1 and ETTh2 with lookback and horizon 96, seven channels, training observations 0–8639, and validation observations 8640–11519. Channel scaling uses training data only, and data readers stop at the validation boundary without accessing test observations. Hyperparameters, seeds, and interventions are fixed before measurement; all specified encodings, gradient comparisons, and continuations are retained. The records include 64 real-data Time-o1 encodings, 720 PS gradient comparisons, and eighteen PS continuations, alongside 144 AMRC configurations and ten CvLoss formula probes. The complete-objective Time-o1 extension adds 4,608 paired comparisons: 1,536 independent-sign interventions and 3,072 preserving controls. No PS cells failed. These mechanism follow-ups are separate from the five-source forecasting comparison.

E.1 TIME-O1: AVERAGING FULL-RANK BASES CAN ERASE RESIDUAL SENSITIVITY

Executed objective. Time-o1's published temporal PCA path (Wang et al., 2025b) standardizes horizon coordinates and fits one PCA basis per channel. It supports shared-basis and channelwise projection. We examine shared-basis mode with all 96 components and no explained-variance weighting. This mode averages channel bases before applying the transform. Let r_d be the standardized residual in channel d, U_d the matrix of PCA component rows, and $\bar{U} = D^{-1} \sum_d U_d$. The auxiliary is the mean absolute value of $\bar{U} r_d$. Flipping any component row's sign preserves the validity of the PCA basis and its explained variance.

Proposition 7 (Full rank need not survive basis averaging)**.** *Let two channels share a positive-definite temporal covariance with distinct eigenvalues, and retain a full orthogonal PCA basis U. The encodings (U, U) and $(U, -U)$ are equivalent PCA solutions up to row signs. For the shared-basis transformed MAE auxiliary, the second encoding gives zero loss and zero output gradient for every residual. Channelwise transformed MAE has the same value and mapped-back gradient under either encoding wherever the derivative exists.*

Proof. The shared transforms are $(U + U)/2 = U$ and $(U - U)/2 = 0$. The zero transform removes all dependence on the residual. In channelwise mode, $\|(-U)r_d\|_1 = \|Ur_d\|_1$, and the signs cancel in $(-U)^\top \operatorname{sign}(-Ur_d)$. This invariance also holds with the zero subgradient used by the published absolute-value implementation. Distinct eigenvalues remove rotational ambiguity but leave each row's sign free. □

A construction using the published PCA fit. We use 192 training rows and two identical channels with 96 horizon coordinates. Their covariance comprises 48 blocks $\left(\begin{smallmatrix} 1 & \rho_k \\ \rho_k & 1 \end{smallmatrix}\right)$, with evenly spaced $\rho_k \in [0.1, 0.8]$. Each coordinate has unit variance; the 96 eigenvalues $1 \pm \rho_k$ are positive and distinct, with minimum gap 0.01489. The published PCA helper fits two rank-96 bases. Post-fit sign flips produce the equivalent encodings evaluated by the published transformation in Table 22. This construction tests representations of a fitted basis; a fixed solver can still select a deterministic sign convention. An MSE anchor remains sensitive to the residual when the auxiliary vanishes.

Table 22: Equivalent encodings of the published PCA fit. Both channel bases have rank 96. The same targets and predictions are used in every row; values concern the transformed auxiliary.

Computation	Encoding	Auxiliary value	Output-gradient norm
Shared average	(U, U)	0.793021	0.025516
Channelwise	(U, U)	0.793021	0.025516
Shared average	$(U, -U)$	0.000000	0.000000
Channelwise	$(U, -U)$	0.793021	0.025516

Fixed predictions on real windows. Each dataset provides 8,449 training target windows for PCA fitting. We use 32 equidistant validation windows, with the last observed value repeated as the fixed prediction. Seed 923002 generates 32 independent sets of row signs across all seven channel bases, each evaluated by the published transformation. Two controls reuse these draws: applying one common row-sign matrix to every channel basis, or projecting each channel with its own basis before taking absolute values. Each control is compared with its original encoding in the same projection mode. Both preserve values and gradients across all 64 sign settings. Table 23 reports the independent-sign results. Relative gradient change is $\|g_{\text{new}} - g_{\text{ref}}\|/\|g_{\text{ref}}\|$ and can exceed one. Because predictions remain fixed, the measured differences concern auxiliary supervision rather than forecast error.

Table 23: All 32 independent sign encodings per dataset. For independent-sign interventions, ratios compare the same fixed predictions with their original shared-basis encoding. IQR is the interquartile range. Each invariant control is compared with the original encoding in its own projection mode; both have value ratio and cosine one, and gradient change zero.

Data	Quantity	Median	IQR	Full range
ETTh1	Auxiliary value ratio	0.689	[0.639, 0.751]	[0.579, 0.859]
ETTh1	Gradient cosine	0.432	[0.421, 0.443]	[0.388, 0.490]
ETTh1	Gradient rel. change (%)	103.975	[102.738, 104.965]	[98.102, 108.672]
ETTh2	Auxiliary value ratio	0.748	[0.704, 0.807]	[0.662, 0.917]
ETTh2	Gradient cosine	0.444	[0.431, 0.456]	[0.398, 0.471]
ETTh2	Gradient rel. change (%)	102.859	[101.746, 104.121]	[99.771, 106.973]

E.1.1 Complete-objective gradients and matched optimizer steps

The persistence experiment isolates the auxiliary. We extend it to learned DLinear predictions and the complete objective $L = w\,\text{MSE} + (1-w)L_{\text{PCA}}$. The six saved MSE-128 states in Appendix E.2 are copied to float32 to match the source PCA cache. Each receives sixteen training-batch updates with the original basis, separately for $w = 0.2$ and 0.8, yielding twelve model and optimizer states. Adam uses learning rate 10^{-3}, betas $(0.9, 0.999)$, epsilon 10^{-8}, and no weight decay. Training windows are fixed by seed $924100 + s$ for model seed $s \in \{0, 1, 2\}$.

The published iTransformer script uses $w = 0.2$, all temporal components, restandardization, and unweighted transformed MAE on Electricity and Traffic at horizon 96. We transfer this complete-loss configuration to ETT/DLinear; $w = 0.8$ is a prespecified stronger-MSE control. Each weight has its own adapted state. Encoding comparisons hold the model state and weight fixed within each pair.

Each state uses four batches of 32 validation windows, selected with seed $924200 + s$; all targets lie within the validation interval. The earlier experiment's 32 sign encodings independently change the channel bases. Every encoding uses the published transformation and complete MSE-plus-auxiliary loss. We measure output and full parameter gradients, then take one step with SGD, fresh-state Adam, or Adam restored to its saved sixteen-step history. Parameters, optimizer history, and inputs are identical within each pair. The shared-sign and channelwise-projection controls are each compared with their own original encoding at the same model state.

Across the independent-sign interventions, median weighted auxiliary/MSE parameter-gradient norm ratios at $w = 0.2$ are $0.825/1.114$ on ETTh1/ETTh2; at $w = 0.8$, they are $0.049/0.039$. All shared-sign and channelwise controls preserve values, gradients, and measured steps exactly in this float32 execution. The explicit linear-projection derivative matches the source output gradient exactly;

Table 24: Time-o1 complete-objective sign intervention. Each row contains 384 paired comparisons (three states, four batches, 32 encodings). Entries report median relative changes in percent. Interquartile ranges are given for the complete parameter gradient and the state-matched Adam update. Each weight has its own sixteen-update adaptation. All shared-sign and channelwise controls have zero change in the measured gradients and steps.

Data	w	Full gradient	Matched Adam	Fresh Adam	Prediction step[†]
ETTh1	0.2	52.48 [43.84, 59.04]	31.13 [28.96, 34.31]	83.49	34.98
ETTh1	0.8	5.10 [3.70, 6.51]	3.63 [3.35, 4.15]	26.19	3.76
ETTh2	0.2	58.35 [53.61, 62.75]	27.10 [25.27, 30.51]	90.67	54.93
ETTh2	0.8	4.05 [3.29, 4.67]	2.70 [2.55, 2.88]	26.01	3.67

[†]Relative difference between the forecast changes induced by the two matched Adam steps, evaluated on the same inputs; this is not a forecast-error comparison.

the full parameter gradient agrees with the weighted component gradients to relative error below 4.04×10^{-7}. All 4,608 comparisons complete. A prescribed replay uses the first seed, first batch, and first four signs on both datasets at both weights, reproducing all 48 comparison records exactly. No test values or validation-based configuration choices enter this follow-up.

E.2 PS Loss: the scope of batch statistics

Complete objective and model. We extract the published PS Loss methods for patch creation, Fourier-based patch selection, structural components, gradient-based dynamic weighting, and the complete PS auxiliary (Kudrat et al., 2025). The reference adds the auxiliary to MSE with weight three, as specified by the published parser; the patch threshold is 24. The published shared DLinear model has neither batch normalization nor dropout. We train it in float64 for 128 MSE updates using batch size 32, Adam learning rate 10^{-3}, and seeds 0, 1, 2 on each dataset. The resulting six states are the references for subsequent interventions.

Let k_B be the patch length selected from batch B and $w_{j,B}$ the detached weight of structural component j. At a fixed model state, the rule is

$$L_B = \mathrm{MSE}_B + 3 \sum_j w_{j,B}\, C_{j,k_B}(B).$$

With k_B and $w_{j,B}$ fixed, component reductions separate over examples. A size-weighted average of microbatch gradients therefore recovers the whole-batch field. Recomputing k or w locally can change the components or coefficients before averaging.

Five controlled comparisons. We partition eight fixed training batches per saved state into groups of 16, 8, or 4. On the same predictions, we recompute (i) both patch length and weights, (ii) only weights, or (iii) only patch length within each microbatch. Control (iv) restores both whole-batch quantities in every microbatch; (v) split MSE provides a separable reference. Gradients include every trainable parameter. The whole-batch helper agrees with the explicit restored formula to relative error below 6.5×10^{-17}. Table 25 reports all settings; MSE partition errors are zero throughout. Local-weight and local-patch effects are not generally additive. At microbatch size eight, all ETTh1 batches retain the whole-batch patch length; 21/24 ETTh2 batches change it in at least one group.

Matched training continuations. From each saved state, three arms receive the same 128 training batches and fresh Adam state: whole-batch PS, local PS with microbatch size eight, and restored whole-batch statistics with microbatch size eight. All eighteen runs complete. Every 32 updates, we evaluate MSE on all complete validation-target windows. Table 26 reports final MSE for every seed. All six restored runs match their whole-batch counterparts in the measured validation trajectories and final parameters to numerical precision. The maximum relative final-parameter discrepancy is below 2.4×10^{-16}. Local recomputation yields both higher and lower validation MSE. Mean paired relative changes against whole-batch PS are $+0.156\%$ on ETTh1 and $+0.013\%$ on ETTh2. These controls identify which batch statistics define the update and are needed to reproduce it. The runs partition the loss computation on full-batch model outputs; parallel-training memory and communication costs are not measured.

Table 25: Complete PS partition contrasts, consolidated across statistic controls. The three change columns report median relative parameter-gradient change in percent, with interquartile ranges (24 paired batches per cell). Cosine is the median similarity to the whole-batch gradient when both patch length and weights are recomputed locally. Restoring both global statistics gives relative error below 6.5×10^{-17} in every setting; split MSE has zero error. Individual interventions are not additive.

Data	Microbatch	Local patch + weights	Local weights only	Local patch only	Cosine
ETTh1	16	8.12 [6.81, 12.82]	8.12 [6.81, 12.82]	0.00 [0.00, 0.00]	0.997
ETTh1	8	13.97 [10.84, 15.81]	13.97 [10.84, 15.81]	0.00 [0.00, 0.00]	0.992
ETTh1	4	16.51 [15.16, 18.80]	15.88 [14.42, 17.84]	0.00 [0.00, 0.00]	0.988
ETTh2	16	22.75 [10.31, 27.25]	9.84 [6.71, 13.10]	18.96 [0.00, 23.57]	0.975
ETTh2	8	27.03 [21.35, 31.31]	16.79 [15.38, 21.18]	21.49 [11.74, 24.20]	0.966
ETTh2	4	32.69 [28.85, 37.52]	24.90 [19.31, 28.23]	22.04 [17.46, 27.27]	0.954

Table 26: All eighteen PS continuations after 128 updates: six whole-batch, six ordinary local and six restored runs. Lower validation MSE is better. The last column compares ordinary local microbatch training with whole-batch training within the same seed. Restoring global statistics reproduces the whole-batch trajectory in all six matched pairs.

Data	Seed	Whole batch	Local microbatch	Restored	Local change
ETTh1	0	0.682598	0.687588	0.682598	+0.731%
ETTh1	1	0.684150	0.689344	0.684150	+0.759%
ETTh1	2	0.711314	0.704053	0.711314	-1.021%
ETTh2	0	0.221946	0.220055	0.221946	-0.852%
ETTh2	1	0.217174	0.217773	0.217174	+0.276%
ETTh2	2	0.221753	0.223114	0.221753	+0.614%

F Supporting audit tools and additional cases

F.1 Shared probes and complementary properties

The shared probe contract specifies tensor axes, reductions, admissible transformations, randomness, and the implemented backward rule. It checks gradients at agreement, gradients mapped back through transformations, responses to companion interventions, and symmetry of the recomputed field Jacobian. Controls then restore a property or isolate the computation that changes it. Bayes acts are derived analytically; mean elicitation by Bregman losses is established theory (Banerjee et al., 2005). These checks characterize individual computational properties rather than assign an overall correctness score to a method. Full-rank bases can cancel when averaged, and identical effective batches can use different statistics. A scalar descent direction can retain biased stationary forecasts; stationarity alone does not establish a minimum.

F.2 Evidence map and source certificates

The source certificates below complement the complete-update controls in Appendices E.1–E.2. InfoGeo is assessed at the formula level. PMLF combines a failure of the published constructor with a formula-faithful ancillary runner.

Across ten dataset–backbone cells, ancillary PMLF has a median relative clean test-MSE change of $+0.660\%$ versus MSE, using five-seed means within cells. It nearly matches its four fixed-weight anchors. The supplement retains complete robustness, seed-dependence, timing, random-factor, and learning-rate records. These comparisons are descriptive. Infrastructure failures are recorded separately; no failed final seed is omitted from a reported contrast.

F.3 From original explanation to bounded audit conclusion

DBLoss motivates its adaptive ratio through scale alignment in Section 3.3, Eqs. (2)–(3), whereas Appendix C.4, Eq. (12), analyzes fixed weights (Qiu et al., 2025). Our curl–accuracy hypothesis is an additional test, not a claim attributed to its authors. RI-Loss motivates residual–noise dependence

Table 27: **Ten audit cases, with evidence depth stated explicitly.** Layers: target (T), representation (R), sample coupling (S), optimization (O). Evidence: formula analysis (F), source inspection/probes (C), complete-loss model gradients or steps (U), controlled training comparisons (L). The cases cover contrasting mechanisms; endpoint training concerns a subset.

Objective	Layers	Finding tested by the audit	Evidence
QDF	T/O	Under the stated fixed-metric assumptions, the conditional mean remains optimal; a learned scale can reduce the score without changing the predictions.	F,C
DistDF	T/S	Joint matching can shift the preferred prediction, while empirical gradients depend on how samples are grouped.	F,C
FreDF	T/R	Invariance of squared transform energy does not imply invariance of the modulus score.	F,C
Time-o1	R/O	Valid bases can cancel when averaged; independent admissible sign changes can alter full gradients and Adam steps.	F,C,U
PS Loss	S/O	Local batch statistics change updates; reusing whole-batch statistics restores the reference training trajectory.	C,U,L
InfoGeo	T	Local temporal smoothing can shift the preferred prediction even when a deterministic target is repeated.	F
OLMA	T/R	Nonquadratic scores and channel order affect which predictions the objective prefers.	F,C,L
DBLoss	O/S	A direction-equivalent potential preserves the boundary on which nonzero bias is stationary.	F,C,U,L
RI-Loss	S/O	Finite kernel dispersion yields a sharp expected-objective exact-fit optimality boundary.	F,C,U,L
PMLF	O/S	Detached softmax has a smooth-maximum potential.	F,C†,L†

†PMLF: static source plus a formula-faithful ancillary runner; the pinned constructor fails. RI: the published helper is explicitly activated in our controlled runner. Selection and source scope are detailed in the appendix.

and presents a U-statistic in Definition 2 (Wang et al., 2026c). Our analysis concerns the published centered-kernel helper, which treats channels as samples, and the full expected MSE-plus-auxiliary objective. Source activation, matched randomness, and optimizer state specify the scope of the implementation-based conclusions in the preceding appendices.

F.4 Quadratic and transformed objectives

Proposition 8 (Fixed conditional metric)**.** *Suppose $\mathbb{E}\|Y\|_2^2 < \infty$. Let $A(X)$ be symmetric, measurable with respect to X, positive definite almost surely, and such that the risk below is finite at the conditional mean. The unique unrestricted minimizer, up to almost-sure equality, of*

$$\mathbb{E}\big[(Y - f(X))^\top A(X)(Y - f(X))\big]$$

is $f^*(X) = \mathbb{E}[Y \mid X]$.

Proof. Condition on $X = x$, write $\mu = \mathbb{E}[Y \mid x]$, and abbreviate $A = A(x)$. The centered cross term vanishes, giving

$$\mathbb{E}[(Y - a)^\top A(Y - a) \mid x] = \operatorname{tr}(A \operatorname{Cov}(Y \mid x)) + (a - \mu)^\top A(a - \mu).$$

Positive definiteness makes μ the unique conditional minimizer. Integrating over X proves the result. □

Target- or parameter-dependent metrics fall outside these assumptions. Singular positive-semidefinite metrics leave their null directions unidentified. Different metrics can yield different projections within a restricted model class. We examine QDF's unconstrained covariance scale below.

Proposition 9 (Transform audit)**.** *Let $Y \in L^2$ and let predictions have finite transformed risk. For a fixed orthogonal matrix Q, $\|Q(\hat{y} - y)\|_2^2 = \|\hat{y} - y\|_2^2$. For a fixed invertible linear map A, squared transformed error still uniquely elicits the conditional mean. These conclusions do not extend in general to nonquadratic scores or rank-deficient transforms.*

Table 28: Source-level audit certificates and their scope.

Method	Source operation and intervention	Decisive output and scope
QDF	Learned-covariance quadratic: fix residuals, scale the covariance factor, and decompose the query gradient.	Metric loss falls ($0.2150 \rightarrow 0.00084$) at fixed MSE 2.0063; the first-order indirect query gradient is zero. Pinned class/scripts, without likelihood correction.
DistDF	Batchwise empirical Gaussian transport: solve for the Gaussian act and repartition a fixed sample multiset.	Source act 1.374 is near joint act 1.366, rather than the conditional act. Repartitioning changes its gradient by 0.617; the MSE gradient changes only at roundoff. Constructed Gaussian source-level certificate.
FreDF	Fourier auxiliary path: distinguish Parseval squared energy from the released modulus MAE.	Parseval holds for squared energy; the released score prefers median 0 to mean 3.333 for the skewed population. Score interpretation, not an assessment of Fourier optimization.
Time-o1	PCA aggregation: insert a truncation-null-space vector and apply an admissible post-fit basis sign flip.	Rank-24 null-space loss is numerically zero; the sign flip changes ETTh1 source MAE by 2.15%. Pinned implementation averaging bases.
PS Loss	Four published patch/loss routines: change only a companion's period and test spectral endpoints.	Focal-gradient relative change is 1269; constant and Nyquist inputs reproduce source failures. Batch and edge-case certificates; prevalence on typical data is untested.
InfoGeo	Paper formula/public snapshot without executable code: apply the local Gaussian construction to $y_t = t^2$ and repeat the trajectory 1,000 times.	Radius-three local mean remains $y_t + 4$ after repetition. Formula-level construction, not source-code execution.
OLMA	Fourier/Haar score: compare skewed-population optimal acts and permute channels.	Released score prefers act 0 to mean 1; permuting seven channels changes the mapped-back gradient by 18.82%. Exact pinned implementation.
DBLoss	EMA/adaptive decomposition: vary companions and batch partitions, examine curl, and compare matched fixed-weight/scalar-objective controls.	Nonzero curl coexists with $\nabla Q = aG$, $a > 0$; nonzero constant residuals can be stationary. Gradient directions averaged over 64 batches differ by 0.23°–3.97°. Local updates; no causal accuracy claim.
RI-Loss	HSIC with the released sample axis: derive the finite-sample expectation, replicate batches, and test $D = 1$.	At $H = 96$, $D = 7$, analytic value 0.166666565 agrees with 5,000 source-draw Monte Carlo; replication scales the sum; $D = 1$ is nonfinite. Published DLinear helper. Traced original entry defaults to MSE; controlled runner explicitly enables RI.
PMLF	Detached component weighting in the official supplement: compare its field with the log-sum-exp gradient, repartition samples, and instantiate unmodified source.	Potential-gradient relative error is 2.1×10^{-16}; unmodified constructor raises NameError. Official supplement snapshot.

Proof. Orthogonality gives $Q^\top Q = I$. For invertible A, the transformed quadratic has positive-definite metric $A^\top A$, so Proposition 8 applies. Rank deficiency leaves $\ker A$ unidentified; non-quadratic scores need not satisfy either quadratic conclusion. □

FreDF's Fourier-modulus auxiliary, Time-o1's PCA-transformed absolute error, and OLMA's Fourier/Haar absolute errors are nonquadratic. Their full objectives therefore require analysis beyond squared-transform identities. OLMA's channel Fourier term also imposes cyclic geometry on the channel order.

For a nonsingular Gaussian distribution, diagonalizing the covariance by an orthogonal rotation can reduce the sum of marginal Gaussian entropies. The rotation preserves joint entropy and total conditional-mean MMSE, so this entropy reduction does not reduce irreducible squared error. Orthogonal transforms can still affect optimization or approximation within restricted model classes.

F.5 Conditional versus joint distribution alignment

Proposition 10 (Deterministic conditional transport)**.** *Suppose $P_{Y|X=x}$ has a finite second moment. For squared Euclidean ground cost and a finite deterministic forecast a,*

$$W_2^2(P_{Y|X=x}, \delta_a) = \mathbb{E}[\|Y - a\|_2^2 \mid X = x].$$

Its Bayes act is therefore the conditional mean.

Proof. The unique coupling to δ_a pairs every Y with a. Its cost is the displayed conditional squared error, equal to conditional variance plus $\|a - \mathbb{E}[Y \mid X = x]\|^2$. □

DistDF matches empirical joint history–future moments, allowing transport across histories. The Gaussian counterexample below distinguishes this relaxation from conditional transport. Regrouping fixed examples also changes its empirical update.

F.6 Batch rules and nonconservative updates

PS Loss chooses a shared patch scale from the batch-average target spectrum before scoring local structure. Focal-example interventions show that changing only a companion can substantially alter the focal gradient. Constant and Nyquist-only inputs expose source-domain singularities. These results are edge-input certificates for the pinned implementation.

Sample coupling and field integrability are independent, even in two parameters.

Proposition 11 (Coupling–integrability independence)**.** *Let $\theta = (\theta_1, \theta_2)$ and let a batch contain $m \geq 2$ scalars $a_1, \ldots, a_m$, with $\bar{a} = m^{-1} \sum_i a_i$. All four combinations of example separability or batch coupling and conservative or nonconservative fields exist. Here, separability means representability, across batches, as an average of batch-independent per-example fields.*

Proof. The field $(\theta_1 - \bar{a}, \theta_2)$ is separable and conservative; $\nabla_\theta\{(\theta_1 - \bar{a}^2)^2/2 + \theta_2^2/2\}$ is coupled and conservative; $(0, \theta_1)$ is separable and nonconservative; and $(0, \theta_1 \bar{a}^2)$ is coupled and nonconservative when $\bar{a} \neq 0$. The latter two can arise from explicit custom backward rules. The first field averages gradients of $\{(\theta_1 - a_i)^2 + \theta_2^2\}/2$. The second is a scalar gradient, but its dependence on $\bar{a}^2$ has nonzero mixed derivatives in distinct batch elements, precluding separability. The third is batch-independent with asymmetric Jacobian $\left(\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right)$. The fourth scales that Jacobian by $\bar{a}^2$ and adds nonseparable batch dependence. □

F.7 Quadratic scale and transform boundaries

Suppose a learned covariance factor is constrained only by $\Sigma = LL^\top$, and the residual objective contains $r^\top \Sigma^{-1} r$ without a $\log\det\Sigma$ term or scale normalization. Along $L = cI$, the quadratic equals $\|r\|_2^2/c^2$ and tends to zero as $c \to \infty$ for every fixed nonzero residual. Positive definiteness alone therefore does not prevent the score from falling through covariance inflation. A log-determinant term or explicit scale constraint changes this optimization problem.

For an unnormalized length-H DFT F, Parseval gives

$$H^{-1} \sum_k |(Fr)_k|^2 = \sum_t r_t^2.$$

For real inputs and even H, recovering full energy from the real FFT requires weight two on interior frequencies and weight one on DC and Nyquist. An unweighted half-spectrum changes their relative contributions. Replacing squared magnitude by $\sum_k |(Fr)_k|$ changes the score; at $H = 1$, it reduces to absolute error and elicits the possibly set-valued median.

Appendix E.1 examines independent PCA sign choices and channelwise preserving controls. Basis aggregation, truncation, and score choice are distinct checks.

F.8 Joint transport can select a different deterministic act

Let $X \sim \mathcal{N}(0,1)$, $Y = aX + \epsilon$ with independent $\epsilon \sim \mathcal{N}(0, \sigma^2)$, and forecast $F = bX$. The centered joint covariance matrices of (X, Y) and (X, F) are

$$C_Y = \begin{pmatrix} 1 & a \\ a & a^2 + \sigma^2 \end{pmatrix}, \qquad C_b = \begin{pmatrix} 1 & b \\ b & b^2 \end{pmatrix}.$$

The squared Gaussian joint Wasserstein distance simplifies to

$$D_{\text{joint}}(b) = 2 + a^2 + \sigma^2 + b^2 - 2\sqrt{1 + 2ab + (a^2 + \sigma^2)b^2}.$$

Conditional squared transport instead gives $D_{\text{cond}}(b) = (a - b)^2 + \sigma^2$, minimized at $b = a$. When $a = 0$ and $\sigma > 1$, the joint objective has nonzero minimizers $b = \pm\sqrt{\sigma^2 - \sigma^{-2}}$. For $\sigma = 2$, their magnitude is approximately 1.9365, although $\mathbb{E}[Y \mid X] = 0$. The joint objective favors dependence on history to mimic conditional response variance that a deterministic forecast cannot represent.

More generally, let $p \geq 1$ and suppose the joint laws have the same history marginal and finite pth moments. A coupling that preserves histories is feasible for ordinary joint transport, so an additive product cost gives

$$W_p^p(P_{X,Y}, P_{X,F}) \leq \int W_p^p(P_{Y|X=x}, P_{F|X=x})\, dP_X(x).$$

Unrestricted joint transport is therefore a relaxation. At attainable zero, equality of joint laws still implies equality of conditionals P_X-almost surely. The counterexample concerns the closest joint law when zero is unattainable within the deterministic model class.

F.9 Local temporal targets

InfoGeo constructs local Gaussian targets from neighboring values on a realized path. For a symmetric radius-w local mean and the deterministic path $y_t = t^2$,

$$\frac{1}{2w+1}\sum_{j=-w}^{w} y_{t+j} - y_t = \frac{1}{2w+1}\sum_{j=-w}^{w} j^2 = \frac{w(w+1)}{3}.$$

At radius three, the shift is four and persists when the trajectory is repeated. InfoGeo's mean-coordinate term thus targets a smoothed path in this construction. Without additional stationarity or sampling assumptions, neighboring values are not repeated draws from $Y_t \mid X$. The pinned snapshot lacks training source, so this is a formula certificate.

F.10 Applying the shared probe contract to AliO

With thresholds and twelve seeds fixed in advance, we test agreement, channel permutations, companion changes, and symmetry of the recomputed field Jacobian. Finite-difference steps 10^{-5} and 5×10^{-6} check its stability.

AliO (Park et al., 2025) aligns overlapping forecasts while detaching the prediction closer to the target. We use its official DLinear training example; the pinned package entry fails during forward execution. Inputs have shape $[2, 2, 8, 3]$ (batch, overlapping predictions, horizon, channels), with lag one. Companion interventions replace another batch group, preserving the intended coupling between overlapping forecasts within each group. Targets are overlapping windows from the same random length-nine path; predictions add Gaussian noise with standard deviation 0.3. Both component weights are one, and both discrepancies use squared error. The full time-plus-frequency rule and its time component receive the same twelve inputs; MSE and symmetric alignment serve as controls.

Table 29: Shared probes on twelve inputs per rule. The first three columns report maximum absolute gradient discrepancies; the final column reports the range of $\|J - J^\top\|_F / \|J\|_F$. All finite-difference step discrepancies are below 4.2×10^{-12}.

Rule	Prediction–target agreement	Channel	Companion	Jacobian asymmetry
AliO full	0.0e+00	3.5e-18	0.0e+00	[0.929, 0.95]
AliO time	0.0e+00	0.0e+00	0.0e+00	[1, 1]
MSE	0.0e+00	0.0e+00	0.0e+00	[0, 0]
symmetric alignment	0.0e+00	0.0e+00	0.0e+00	[2.72e-12, 4.63e-12]

The full rule's relative Jacobian asymmetry ranges from 0.9295 to 0.9502, yet its negative is an output-MSE descent direction on all twelve inputs. For the time-only pair, write $e_x = x - t$, $e_y = y - t$, with $|e_x| > |e_y|$. The source field is $(2(x-y), 0)$, so

$$\langle \nabla[(e_x^2 + e_y^2)/2], g\rangle = 2e_x(e_x - e_y) \geq 2|e_x|(|e_x| - |e_y|) > 0.$$

The negative field strictly decreases MSE to first order despite unequal mixed partials. At $(x, y, t) = (2, 1, 0)$, the source gives $g = (2, 0)$ and directional derivative -4.

F.11 Alignment-family probe design and limits

This post-hoc mechanism illustration uses MSE, soft-DTW (Cuturi & Blondel, 2017), its self-cost-corrected divergence (Blondel et al., 2021), and DILATE (Le Guen & Thome, 2019). Twelve fixed seeds generate independent standard-normal predictions and targets shared by all four formulas, with $B = 4$, $H = 8$, $D = 2$. Exact agreement provides a second input condition. Soft-DTW uses squared Euclidean local cost and $\gamma = 1$; DILATE uses $\alpha = 1/2$ and timing cost $(i-j)^2/H^2$. We average the per-example formulas. The vector-valued DILATE extension is explicit; released-source comparisons use univariate inputs.

Common translation by $(0.4, -0.7)$ and a common orthogonal rotation of angle 0.7 test representation invariance and mapped-back gradients. Companion predictions are shifted by $+2$ with their targets unchanged; a separate check splits the fixed batch into equal halves. We measure Hessian antisymmetry on the first example of each seed. The largest relative discrepancy across these checks is below 10^{-8} in float64. An independent enumeration of all monotone paths for length-three sequences also checks the soft-DTW dynamic program. These tests cover the specified local properties on all twelve inputs. The nonzero soft-DTW agreement gradient and its divergence correction recover the known self-bias result cited above.

The released DILATE routine downloaded from the authors' repository on 22 September 2026 uses batch means in its forward soft-DTW and path operations. Its custom backward routines omit the corresponding $1/B$ factors. The supplementary code includes the pinned snapshot and its license. On all twelve fixed univariate inputs, the $B = 1$ value and gradient agree with our differentiable formula. At $B = 4$, values still agree, and dividing the source gradient by four restores agreement. Both the absolute value tolerance and relative gradient tolerance are 5×10^{-4}, allowing for float32 intermediates. At fixed B, the field therefore retains the potential BL_{mean} and the same descent direction. Batch-dependent scale is part of the executed field and must be included when comparing objective normalizations.

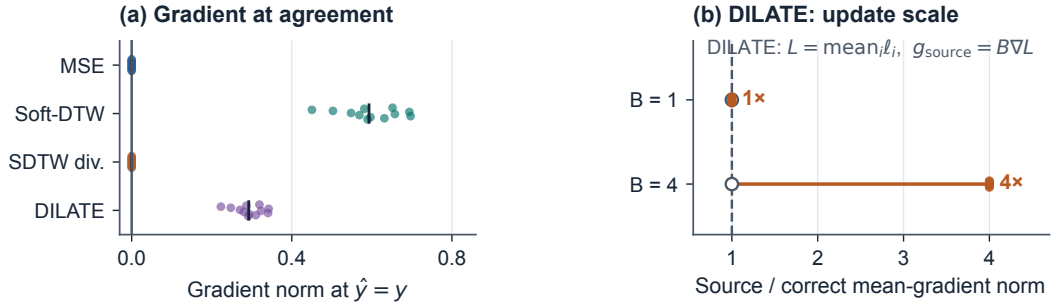


Figure 10: **Shared probes recover known behavior and identify a source-level scale mismatch.** (a) Agreement gradients over twelve seeds preserve true zeros and recover soft-DTW self-bias and its divergence correction. (b) The pinned univariate DILATE source reports the correct mean value, but source-to-formula gradient-norm ratios are one and four at the two tested batch sizes. Open markers indicate the correct scale. All four per-example formulas pass rotation, companion, and Jacobian checks at 10^{-8} tolerance. Appendix F.11 specifies inputs and source configuration.

F.12 Scope of comparisons and neighboring objectives

Table 27 distinguishes formula and source coverage from training evidence. Q_τ and AliO are post-hoc extensions. Evaluation follows the declared target, including timing, constraints, or distributional accuracy where appropriate. Related objectives address transformations, overlap, and component structure (Lee et al., 2022; Park et al., 2025; Jhin & Park, 2026), or peak, cross-variable, directional, and gradient relations (Hari et al., 2026; Ding et al., 2026; Lee et al., 2026; Park et al., 2026). Elicitation/calibration (Fissler & Ziegel, 2016; 2021; Gneiting & Resin, 2023), representation geometry (Kristiadi et al., 2023), dispersion regularization (Zbontar et al., 2021; Bardes et al., 2022), and stop-gradient analysis (Chen & He, 2021; Ponce et al., 2026) provide complementary context. Our deductions concern the specified composite losses and controlled responses.

F.13 AMRC: two-example normalization suppresses a relational penalty

AMRC's representation consistency includes an embedding similarity penalty (ESP) (Liang et al., 2025). The published iTransformer helper computes pairwise mean squared distances for embeddings and labels. It divides each distance matrix by its own batch mean plus $\epsilon = 10^{-8}$, then averages the absolute difference between the normalized matrices. Equations (11)–(13) and Algorithm 1 of the paper instead describe feature-dimension normalization, which we include as a separate formula

control. For two examples with positive pair distances a, b, the source penalty is

$$\mathrm{ESP}_2 = \frac{1}{2}\left|\frac{a}{a/2+\epsilon} - \frac{b}{b/2+\epsilon}\right| = \frac{2\epsilon|a-b|}{(a+2\epsilon)(b+2\epsilon)}.$$

At $\epsilon = 0$, both normalized distance matrices become $\left(\begin{smallmatrix} 0 & 2 \\ 2 & 0 \end{smallmatrix}\right)$, regardless of their original scales. For fixed $a, b > 0$ independent of ϵ, the penalty and its smooth-region derivatives with respect to the distances are $O(\epsilon)$ as $\epsilon \to 0$. This scaling need not hold if a pair distance approaches zero with ϵ. Accumulating two-example microbatch losses does not recover the cross-microbatch relations excluded by this computation.

We call the published penalty with an identity encoder to probe its embedding input. Ten fixed synthetic cases use embeddings of shape $[32, 3, 8]$ and labels $[32, 12, 3]$, with seeds 923500–923509. Two real cases use the 32 equidistant validation windows per dataset from Appendix E.1, treating input histories as embeddings. Every case compares full batch 32 with microbatches 16, 8, 4, 2, and 1, under both the source and paper-formula rules. In float32, all ten synthetic two-example source penalties and embedding gradients are exactly zero; the formula control remains nonzero. At microbatch size two, ETTh1's source value and gradient are also zero. ETTh2 retains a value of 2.61×10^{-8} and an embedding-gradient norm 3.82×10^{-7} times its full-batch source norm (Table 30). The finite-ϵ expression explains why exact zero depends on numerical scale and precision.

Table 30: AMRC ESP at embedding inputs. Synthetic rows summarize all ten cases; real rows use untrained identity encoders. Effective batch is 32 throughout. The last column is the source-rule embedding-gradient norm relative to its full-batch norm. Source and formula values have different normalizations and are not accuracy metrics.

Input	Microbatch	Source ESP	Formula ESP	Gradient ratio
Synthetic (median)	32	0.304751	0.586778	1
Synthetic (median)	16	0.293340	0.579153	1.09
Synthetic (median)	8	0.288654	0.538749	1.3
Synthetic (median)	4	0.256030	0.469947	1.61
Synthetic (median)	2	0.000000	0.311122	0
Synthetic (median)	1	0.000000	0.000000	0
ETTh1 (identity)	32	0.375622	1.044606	1.000000
ETTh1 (identity)	2	0	0.481353	0
ETTh2 (identity)	32	0.291454	0.440143	1.000000
ETTh2 (identity)	2	2.61×10^{-8}	0.084433	3.82×10^{-7}

This certificate concerns the ESP function at its embedding input. The encoder is untrained; adaptive masking and full AMRC optimization are outside the experiment. The reusable check counts the examples participating in each relational computation. Recovering the full-batch relation requires cross-microbatch pairs as well as the normalization constants.

F.14 CvLoss: a structural penalty with a corrective anchor

CvLoss (Ding et al., 2026) provides a formula-level positive contrast. We reproduce Equations (10)–(13) of its preprint, without testing an official implementation. For a fixed graph, let A map residual patches linearly to edgewise differences:

$$L_\alpha(r) = (1-\alpha)\frac{\|r\|^2}{N} + \alpha\frac{\|Ar\|_1}{M}, \qquad 0 \le \alpha < 1,$$

where $N, M > 0$ and M counts scalar edge differences. The positive quadratic term makes $r = 0$ the unique global minimizer for a fixed target. For every $g \in \partial L_\alpha(r)$,

$$r^\top g = 2(1-\alpha)\frac{\|r\|^2}{N} + \alpha\frac{\|Ar\|_1}{M} > 0 \qquad (r \neq 0).$$

Indeed, each $s \in \partial\|\cdot\|_1(Ar) \subset \mathbb{R}^M$ satisfies $s^\top Ar = \|Ar\|_1$, and $g = 2(1-\alpha)r/N + \alpha A^\top s/M$. Taking the inner product with r gives the identity. Output-space negative subgradient flow therefore decreases squared residual norm away from agreement. Jointly relabeling channels and graph

preserves the objective. A graph defined within each example also permits exact sample partitioning with matching reductions.

Ten fixed probes use $[B, H, D] = [8, 24, 4]$, patch length six, and $\alpha = 1/2$. The graph includes every pair of patch nodes from different channels, including pairs at different patch positions. Channel-and-graph relabeling gives relative gradient error below 7.66×10^{-17}; partition-gradient error is zero. The radial identity has absolute error below 2.23×10^{-16}. For a nonzero constant residual, the graph term vanishes while the MSE anchor retains correction. This fixed-target analysis identifies a property preserved by the structural design. Population elicitation and finite optimizer trajectories require separate checks.